ELECTRICAL ENGINEERING AND ELECTRONICS

A Series of Reference Books and Textbooks

1. Rational Fault Analysis, *edited by Richard Saeks and S. R. Liberty*
2. Nonparametric Methods in Communications, *edited by P. Papantoni-Kazakos and Dimitri Kazakos*
3. Interactive Pattern Recognition, *Yi-tzuu Chien*
4. Solid-State Electronics, *Lawrence E. Murr*
5. Electronic, Magnetic, and Thermal Properties of Solid Materials, *Klaus Schröder*
6. Magnetic-Bubble Memory Technology, *Hsu Chang*
7. Transformer and Inductor Design Handbook, *Colonel Wm. T. McLyman*
8. Electromagnetics: Classical and Modern Theory and Applications, *Samuel Seely and Alexander D. Poularikas*
9. One-Dimensional Digital Signal Processing, *Chi-Tsong Chen*
10. Interconnected Dynamical Systems, *Raymond A. DeCarlo and Richard Saeks*

11. Modern Digital Control Systems, *Raymond G. Jacquot*
12. Hybrid Circuit Design and Manufacture, *Roydn D. Jones*
13. Magnetic Core Selection for Transformers and Inductors: A User's Guide to Practice and Specification, *Colonel Wm. T. McLyman*
14. Static and Rotating Electromagnetic Devices, *Richard H. Engelmann*
15. Energy-Efficient Electric Motors: Selection and Application, *John C. Andreas*
16. Electromagnetic Compossibility, *Heinz M. Schlicke*
17. Electronics: Models, Analysis, and Systems, *James G. Gottling*
18. Digital Filter Design Handbook, *Fred J. Taylor*
19. Multivariable Control: An Introduction, *P. K. Sinha*
20. Flexible Circuits: Design and Applications, *Steve Gurley, with contributions by Carl A. Edstrom, Jr., Ray D. Greenway, and William P. Kelly*
21. Circuit Interruption: Theory and Techniques, *Thomas E. Browne, Jr.*
22. Switch Mode Power Conversion: Basic Theory and Design, *K. Kit Sum*
23. Pattern Recognition: Applications to Large Data-Set Problems, *Sing-Tze Bow*
24. Custom-Specific Integrated Circuits: Design and Fabrication, *Stanley L. Hurst*
25. Digital Circuits: Logic and Design, *Ronald C. Emery*
26. Large-Scale Control Systems: Theories and Techniques, *Magdi S. Mahmoud, Mohamed F. Hassan, and Mohamed G. Darwish*
27. Microprocessor Software Project Management, *Eli T. Fathi and Cedric V. W. Armstrong (Sponsored by Ontario Centre for Microelectronics)*
28. Low Frequency Electromagnetic Design, *Michael P. Perry*
29. Multidimensional Systems: Techniques and Applications, *edited by Spyros G. Tzafestas*
30. AC Motors for High-Performance Applications: Analysis and Control, *Sakae Yamamura*
31. Ceramic Motors for Electronics: Processing, Properties, and Applications, *edited by Relva C. Buchanan*
32. Microcomputer Bus Structures and Bus Interface Design, *Arthur L. Dexter*
33. End User's Guide to Innovative Flexible Circuit Packaging, *Jay J. Miniet*
34. Reliability Engineering for Electronic Design, *Norman B. Fuqua*
35. Design Fundamentals for Low-Voltage Distribution and Control, *Frank W. Kussy and Jack L. Warren*
36. Encapsulation of Electronic Devices and Components, *Edward R. Salmon*
37. Protective Relaying: Principles and Applications, *J. Lewis Blackburn*
38. Testing Active and Passive Electronic Components, *Richard F. Powell*
39. Adaptive Control Systems: Techniques and Applications, *V. V. Chalam*

40. Computer-Aided Analysis of Power Electronic Systems, *Venkatachari Rajagopalan*
41. Integrated Circuit Quality and Reliability, *Eugene R. Hnatek*
42. Systolic Signal Processing Systems, *edited by Earl E. Swartzlander, Jr.*
43. Adaptive Digital Filters and Signal Analysis, *Maurice G. Bellanger*
44. Electronic Ceramics: Properties, Configuration, and Applications, *edited by Lionel M. Levinson*
45. Computer Systems Engineering Management, *Robert S. Alford*
46. Systems Modeling and Computer Simulation, *edited by Naim A. Kheir*
47. Rigid-Flex Printed Wiring Design for Production Readiness, *Walter S. Rigling*
48. Analog Methods for Computer-Aided Circuit Analysis and Diagnosis, *edited by Takao Ozawa*
49. Transformer and Inductor Design Handbook: Second Edition, Revised and Expanded, *Colonel Wm. T. McLyman*
50. Power System Grounding and Transients: An Introduction, *A. P. Sakis Meliopoulos*
51. Signal Processing Handbook, *edited by C. H. Chen*
52. Electronic Product Design for Automated Manufacturing, *H. Richard Stillwell*
53. Dynamic Models and Discrete Event Simulation, *William Delaney and Erminia Vaccari*
54. FET Technology and Application: An Introduction, *Edwin S. Oxner*
55. Digital Speech Processing, Synthesis, and Recognition, *Sadaoki Furui*
56. VLSI RISC Architecture and Organization, *Stephen B. Furber*
57. Surface Mount and Related Technologies, *Gerald Ginsberg*
58. Uninterruptible Power Supplies: Power Conditioners for Critical Equipment, *David C. Griffith*
59. Polyphase Induction Motors: Analysis, Design, and Application, *Paul L. Cochran*
60. Battery Technology Handbook, *edited by H. A. Kiehne*
61. Network Modeling, Simulation, and Analysis, *edited by Ricardo F. Garzia and Mario R. Garzia*
62. Linear Circuits, Systems, and Signal Processing: Advanced Theory and Applications, *edited by Nobuo Nagai*
63. High-Voltage Engineering: Theory and Practice, *edited by M. Khalifa*
64. Large-Scale Systems Control and Decision Making, *edited by Hiroyuki Tamura and Tsuneo Yoshikawa*
65. Industrial Power Distribution and Illuminating Systems, *Kao Chen*
66. Distributed Computer Control for Industrial Automation, *Dobrivoje Popovic and Vijay P. Bhatkar*
67. Computer-Aided Analysis of Active Circuits, *Adrian Ioinovici*
68. Designing with Analog Switches, *Steve Moore*
69. Contamination Effects on Electronic Products, *Carl J. Tautscher*
70. Computer-Operated Systems Control, *Magdi S. Mahmoud*
71. Integrated Microwave Circuits, *edited by Yoshihiro Konishi*

72. Ceramic Materials for Electronics: Processing, Properties, and Applications, Second Edition, Revised and Expanded, *edited by Relva C. Buchanan*
73. Electromagnetic Compatibility: Principles and Applications, *David A. Weston*
74. Intelligent Robotic Systems, *edited by Spyros G. Tzafestas*
75. Switching Phenomena in High-Voltage Circuit Breakers, *edited by Kunio Nakanishi*
76. Advances in Speech Signal Processing, *edited by Sadaoki Furui and M. Mohan Sondhi*
77. Pattern Recognition and Image Preprocessing, *Sing-Tze Bow*
78. Energy-Efficient Electric Motors: Selection and Application, Second Edition, *John C. Andreas*
79. Stochastic Large-Scale Engineering Systems, *edited by Spyros G. Tzafestas and Keigo Watanabe*
80. Two-Dimensional Digital Filters, *Wu-Sheng Lu and Andreas Antoniou*
81. Computer-Aided Analysis and Design of Switch-Mode Power Supplies, *Yim-Shu Lee*
82. Placement and Routing of Electronic Modules, *edited by Michael Pecht*
83. Applied Control: Current Trends and Modern Methodologies, *edited by Spyros G. Tzafestas*
84. Algorithms for Computer-Aided Design of Multivariable Control Systems, *Stanoje Bingulac and Hugh F. VanLandingham*
85. Symmetrical Components for Power Systems Engineering, *J. Lewis Blackburn*
86. Advanced Digital Signal Processing: Theory and Applications, *Glenn Zelniker and Fred J. Taylor*
87. Neural Networks and Simulation Methods, *Jian-Kang Wu*
88. Power Distribution Engineering: Fundamentals and Applications, *James J. Burke*
89. Modern Digital Control Systems: Second Edition, *Raymond G. Jacquot*
90. Adaptive IIR Filtering in Signal Processing and Control, *Phillip A. Regalia*
91. Integrated Circuit Quality and Reliability: Second Edition, Revised and Expanded, *Eugene R. Hnatek*
92. Handbook of Electric Motors, *edited by Richard H. Engelmann and William H. Middendorf*
93. Power-Switching Converters, *Simon S. Ang*
94. Systems Modeling and Computer Simulation: Second Edition, *Naim A. Kheir*
95. EMI Filter Design, *Richard Lee Ozenbaugh*
96. Power Hybrid Circuit Design and Manufacture, *Haim Taraseiskey*
97. Robust Control System Design: Advanced State Space Techniques, *Chia-Chi Tsui*
98. Spatial Electric Load Forecasting, *H. Lee Willis*

99. Permanent Magnet Motor Technology: Design and Applications, *Jacek F. Gieras and Mitchell Wing*
100. High Voltage Circuit Breakers: Design and Applications, *Ruben D. Garzon*
101. Integrating Electrical Heating Elements in Appliance Design, *Thor Hegbom*
102. Magnetic Core Selection for Transformers and Inductors: A User's Guide to Practice and Specification, Second Edition, *Colonel Wm. T. McLyman*
103. Statistical Methods in Control and Signal Processing, *edited by Tohru Katayama and Sueo Sugimoto*

Additional Volumes in Preparation

Statistical Methods in Control and Signal Processing

edited by

Tohru Katayama
Kyoto University
Kyoto, Japan

Sueo Sugimoto
Ritsumeikan University
Shiga, Japan

MARCEL DEKKER, INC. NEW YORK • BASEL • HONG KONG

Library of Congress Cataloging-in-Publication Data

Statistical methods in control and signal processing / edited by Tohru Katayama, Sueo Sugimoto.
p. cm. — (Electrical engineering and electronics ; 103)
Includes bibliographical references and index.
ISBN 0-8247-9948-8 (hardcover : alk. paper)
1. Real-time control — Statistical methods. 2. Signal processing — Statistical methods. I. Katayama, Tohru. II. Sugimoto, Sueo. III. Series.
TJ217.7.S73 1997
621.382'2'01176—dc21

97-22476
CIP

This book is printed on acid-free paper.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016
http://www.dekker.com

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

Preface

Engineers and applied scientists are interested in stochastic methods because real physical systems are often subjected to random disturbances and unpredictable environments. With the rapid development of VLSI and computer technology, stochastic and statistical models can now be obtained from observed data and begin to be widely available for real-time control, communication, signal processing, and other related fields. Therefore, stochastic and statistical methods for the analysis and design of various technological systems are being increasingly utilized and becoming more relevant in many areas in engineering and applied sciences.

It is well known that statistical and stochastic concepts play a fundamental role in the area of control and signal processing. In fact, the Markovian (or state-space) model has been a key idea in linear quadratic Gaussian control and in Kalman filtering; since the 1960s, most system identification techniques have been based on statistical estimation theory, and adaptive control is best studied and understood in a stochastic framework. Also autoregressive (AR) and autoregressive moving average (ARMA) models have been a major tool for spectral analysis and signal processing as well as for economic time-series analysis, especially after the Akaike information criteria (AIC) revolutionized the order selection procedure of parametric models.

Probability and statistics have a long history, as their beginning can be said to go as far back as the 17th century. In the early 1960s, these disciplines began to spread into the areas of control engineering and signal processing. For more than 30 years, methods of probability, statistics, and stochastic processes have been employed in these fields and have been a source of interest and fascination for many researchers. Many important results have been obtained and are currently being implemented and used in industrial and technological areas.

During the past decades, however, new techniques of fuzzy logic, fractal, wavelets and neural networks modeling, genetic algorithms, etc., have been introduced, which are also playing an increasingly important role in control and signal processing. Some of the papers in this volume are partially related to these topics. There has also been rapid development in robust design and estimation techniques based on deterministic or minimax concepts. We often encounter many students and young engineers who have good knowledge of these new concepts and techniques but are not familiar with probability, statistics, and stochastic processes.

We still believe, however, that stochastic methods will continue to be one of the key concepts in control and signal processing, and one of the aims of this volume is to emphasize their importance and usefulness.

This book intends to give the state of the art of certain areas of statistical and stochastic methods in control and signal processing. The 20 chapters provide a good sample of recent trends in the area of stochastic modeling, identification, and signal processing together with the relative open problems, application areas, and most representative results. Overlap of topics has been kept to a minimum and there are substantial contributions that represent Japanese activity in these areas.

Part I covers some of latest developments in modeling, identification, and estimation, including stochastic realization, state-space modeling, identification and parameter estimation, filtering, fuzzy modeling, statistical models of economic behavior, etc. Part II deals with topics in signal processing, including time-series analysis, blind deconvolution, array signal processing, detection, image processing, etc.

This volume is suitable for senior undergraduate and graduate students as well as for applied scientists and professional engineers. Most contributors present tutorial or review articles requiring only elementary knowledge of statistics and stochastic processes. Reference to original research papers is made whenever appropriate. Hence, each chapter is self-contained and easy to read. The range of topics is so varied that a reader with special interests should be able to find chapters relevant to his or her particular problem.

Chapter 1, by G. Picci, discusses recent advances in modeling and identification of stationary processes based on the geometric theory of stochastic realization. In particular, the problem of identifying models with exogenous inputs is formulated and a procedure for constructing minimal state-space models is developed based on a generalization of a stochastic realization technique for time-series. It is also shown that this approach naturally leads to subspace identification methods.

Chapter 2, by W. Gersch and G. Kitagawa, integrates and communicates the authors' works on general state-space modeling for time-series. The importance of AIC for evaluating the parametric models fitted to data is emphasized, and the numerical approximation, Gaussian sum approximation, and Monte Carlo methods for general state-space filter/smoother implementation are discussed. Various applications are also considered.

Chapter 3, by W. E. Larimore, addresses the identification of linear, time-invariant dynamical systems with noise disturbances and possible feedback. The canonical variate analysis (CVA) method is applied to the past and future of a process to obtain the optimal determination of the states and state order of a dynamical system. Several applications are also described, including chemical process control, spectral analysis of a vibrating system, process monitoring and fault detection, adaptive control of aircraft wing flutter, etc.

Chapter 4, by G. P. Rao and A. V. B. Subrahmanyam, considers a continuous-time multivariable system identification by using generalized MA models (GMAMs) in which the model output is expressed as a linear combination of certain MA components of the input. It is shown that good low-order approximations of complex systems are obtained by embedding prior knowledge of system dynamics in the parameterizations of the GMAMs.

Chapter 5, by Z.-J. Yang, S. Sagara, and T. Tsuji, develops a technique of identifying a continuous-time impulse response function with local high frequency components from sampled input–output data. The impulse response is approximated by a multiresolution neural network composed of scaling and wavelet functions and is then successively estimated from a coarse resolution level to a fine resolution level via a least-squares method with the help of the AIC.

The classic approach to model order estimation is based on the AIC or its improved versions. Another approach to model order estimation is based on rank test methods. Chapter 6, by J. Sorelius, T. Söderström, P. Stoica, and M. Cedervall, presents a number of rank test methods relevant to ARMA order estimation that are closely related to subspace methods estimation. The chapter provides a numerical evaluation of these methods based on extensive simulations studies.

In Chapter 7, S. Aihara and A. Bagchi consider the maximum a posteriori (MAP) state estimation problem for a nonlinear system in which the observation noise is modeled by a finitely additive white noise to avoid the classic Brownian motion modeling for which the observation is provided in integral form and is nowhere differentiable. The basic equation for the MAP state estimate is derived and a numerical procedure to solve the basic equation is also developed.

Chapter 8, by K. Takaba and T. Katayama, is concerned with the stochastic performance of an H_∞ filter in the case where the underlying noise is zero mean white noise. The relation between the bound γ and the performance of the H_∞ filter is examined based on the monotonicity of the solution of the associated H_∞ Riccati difference equation. Numerical examples are included in order to analyze the sensitivity of the H_∞ filter in the presence of process and observation noises.

Chapter 9, by T. Nakamizo, develops a method for deriving a reduced-order filter that yields an estimate of a linear function of the state for a linear stochastic system based on a system coordinate transform. The relationship between the reduced-order and the full-order filters is discussed and the design method is illustrated by examples.

In Chapter 10, by M. Aoki, the interaction processes of a large number of agents are modeled as Markov processes on the set of exchangeable random partitions of a set of integers. A diffusion equation approximation describing the time evolution of the population compositions of agents is derived. It is shown that the distribution of the order statistics of the fractions of agents is described as the average of multinomial distributions conditioned on the vector random fractions governed by Dirichlet distributions.

T. Fukuda, in Chapter 11, explores fuzzy random vectors (FRVs) as a theoretical basis for studying multidimensional fuzzy stochastic processes. The FRVs are introduced from the point of view of vague perception of nonfuzzy random vectors. The first- and second-order moments of FRVs are defined by applying the multivalued logic, and some properties of FRVs are derived.

Chapter 12, by H. Sakai and S. Ohno, considers the spectral theory of discrete-time cyclostationary processes with applications to the design of an optimal filter bank matched to input signal statistics. A parametric formula is derived for the cyclic spectral density matrix of a periodic AR process. The result is then applied to the optimal design of a low pass filter in the analysis part under perfect reconstruction conditions.

Chapter 13, by J. K. Tugnait, deals with multivariable identification by using the error-in-variables models. Attention is focused on frequency-domain approaches in which the integrated polyspectrum of the input and the integrated cross-polyspectrum of the input–output process are employed to derive two new classes of parametric identification methods. Identifiability and consistency of these methods are analyzed under mild conditions. Performance analysis of the estimators is also presented.

In Chapter 14, by Y. Inouye, the design of an equalizer to perform blind deconvolution for nonminimum-phase systems is formulated as the problem of maximizing the fourth-order cummulant of the equalizer output. A necessary and sufficient condition is derived in order for multichannel blind deconvolution to be solvable. Numerical results are included to show the effectiveness of proposed criteria for blind deconvolution.

Chapter 15, by A. Lee Swindlehurst and M. Viberg, reviews historical development of sensor array signal processing techniques and then considers robust Bayesian approaches to improve the direction of arrival and beam-forming performance by utilizing the a priori information on the array model errors. The MAP estimator for the problem is formulated, and a computationally feasible algorithm is developed by using a subspace technique based on the eigendecomposition of the array covariance matrix.

Chapter 16, by T. Sen Lee, reviews the use of stochastic methods and signal processing techniques in tracking radar, imaging radar, and Doppler weather radar systems. Topics included are state and parameter estimation by extended Kalman filters, resolution enhancement based on bandwidth expansion techniques, use of a priori information, and adaptation.

Chapter 17, by K. Kumamaru, J. Hu, K. Inoue, and T. Söderström, presents two robust change detection algorithms for dynamic systems with model uncertainty based on the Kullback information criterion. In a soft bound approach, unmodeled dynamics are described by a random quantity with soft bound. Moreover, in a nonbound approach, the unmodeled dynamics are defined by the difference between a complex model and its reduced-order submodel. Simulation examples are included.

In Chapter 18, A. Burrell, A. Kogiantis, and P. Papantoni-Kazakos present a sequential algorithm that detects changes from an acting stochastic model to any one of a number of alternatives. Based on a discrete approximation of the stochastic models, they employ stochastic binary neural networks pretrained to produce the statistical measures associated with these models. The overall system performance is discussed and numerical results are presented.

K. Kamejima presents in Chapter 19 a stochastic computation scheme to recognize complicated patterns generated via self-similarity processes. Based on the conditional distribution of the observed images, an invariant subset in the feature pattern is obtained. The existence of an invariant feature implies that the generation process of self-similarity patterns can be detected in a preassigned dictionary in finite steps. The detection scheme is verified by computer simulation.

In the final chapter, M. Suwa and S. Sugimoto discuss a Gibbs random field model and its samplng algorithm for image processing. A relaxation-type algorithm based on the Metropolis algorithm is proposed in the sampling process. Experimental results for generating texture images from the Gibbs random field are given in order to examine the phase transition phenomena.

We would like to express our sincere thanks to all the contributors for their efforts in providing excellent chapters and their patience. This volume will become an important reference work in the field, useful to students and scientists interested in algorithms for design and analysis of signal and control systems in the presence of uncertainty.

We finally acknowledge with pleasure the fruitful collaboration with Russell Dekker and Joseph Stubenrauch of Marcel Dekker, Inc.

Tohru Katayama
Sueo Sugimoto

Contents

Preface iii
Contributors xi

I. Modeling, Identification, and Estimation

1. Stochastic Realization and System Identification 1
Giorgio Picci

2. General State Space Modeling 37
Will Gersch and Genshiro Kitagawa

3. Canonical Variate Analysis in Control and Signal Processing 83
Wallace E. Larimore

4. Models in Generalized MA Form for Identification of Continuous-Time Systems 121
Ganti Prasada Rao and A. V. B. Subrahmanyam

5. Multiresolution Approach to Identification of System Impulse Response 149
Zi-Jiang Yang, Setsuo Sagara, and Teruo Tsuji

6. Comparative Study of Rank Test Methods for ARMA Order Estimation 179
Joakim Sorelius, Torsten Söderström, Petre Stoica, and Mats Cedervall

7. A MAP Recursive Nonlinear Filtering 217
Shin Ichi Aihara and Arunabha Bagchi

8. Stochastic Properties of the H_∞ Filter 239
Kiyotsugu Takaba and Tohru Katayama

9. Reduced Order Functional Estimator for Linear Stochastic Systems 257
Takayoshi Nakamizo

10. Shares in Emergent Markets: Dynamics and Statistical Properties of Equilibrium Classification of Agents in Evolutionary Models 273
Masanao Aoki

11. Fuzzy Random Data Obtained as Vague Perceptions of Random Phenomena 299
Tokuo Fukuda

II. Signal Processing

12. Theory of Cyclostationary Processes and Its Applications 327
Hideaki Sakai and Shuichi Ohno

13. Stochastic System Identification Using Polyspectra 355
Jitendra K. Tugnait

14. Blind Deconvolution of Multichannel Linear Time-Invariant Systems of Nonminimum Phase 375
Yujiro Inouye

15. Bayesian Approaches for Robust Array Signal Processing 399
A. Lee Swindlehurst and Mats Viberg

16. Selected Stochastic Methods and Signal Processing Used in Radar Systems 431
T. Sen Lee

17. Statistical Methods for Robust Change Detection in Dynamical Systems with Model Uncertainty 453
Kousuke Kumamaru, Jinglu Hu, Katsuhiro Inoue, and Torsten Söderström

18. Detecting Changes in Acting Stochastic Models and Model Implementation via Stochastic Binary Neural Networks 481
Anthony Burrell, Achilles G. Kogiantis, and P. Papantoni-Kazakos

19. Invariant Features Associated with a Conditional Distribution Induced by Self-Similar Patterns 499
Kohji Kamejima

20. Gibbs Random Image Models and Sampling 525
Masaki Suwa and Sueo Sugimoto

Index 545

Contributors

Shin Ichi Aihara Department of Management and Systems Science, Science University of Tokyo, Suwa College, Nagano, Japan

Masanao Aoki Center for Computational Economics, University of California, Los Angeles, Los Angeles, California

Arunabha Bagchi Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Anthony Burrell Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, Alabama

Mats Cedervall Systems and Control Group, Uppsala University, Uppsala, Sweden

Tokuo Fukuda Faculty of Management, Otemon Gakuin University, Osaka, Japan

Will Gersch Department of Information and Computer Sciences, University of Hawaii, Honolulu, Hawaii

Jinglu Hu Department of Control Engineering and Science, Kyushu Institute of Technology, Iizuka, Japan

Katsuhiro Inoue Department of Control Engineering and Science, Kyushu Institute of Technology, Iizuka, Japan

Yujiro Inouye Department of Systems Engineering, Osaka University, Osaka, Japan

Kohji Kamejima Industrial Management, Osaka Institute of Technology, Osaka, Japan

Tohru Katayama Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan

Genshiro Kitagawa Department of Prediction and Control, The Institute of Statistical Mathematics, Tokyo, Japan

Achilles G. Kogiantis Department of Electrical and Computer Engineering, University of Southwestern Louisiana, Lafayette, Louisiana

Kousuke Kumamaru Department of Control Engineering and Science, Kyushu Institute of Technology, Iizuka, Japan

Wallace E. Larimore Adaptics, Inc., McLean, Virginia

T. Sen Lee Ground Segment, National Space Program Office, Hsin-Chu, Taiwan

Takayoshi Nakamizo Department of Mechanical Engineering, Ashikaga Institute of Technology, Tochigi, Japan

Shuichi Ohno Department of Mathematics and Computer Science, Shimane University, Shimane, Japan

P. Papantoni-Kazakos Department of Electrical and Computer Engineering, University of Alabama, Tuscaloosa, Alabama

Giorgio Picci Department of Electronics and Informatics, University of Padova, Padova, Italy

Ganti Prasada Rao* Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India

Setsuo Sagara Department of Communication and Computer Engineering, Fukuoka Institute of Technology, Fukuoka, Japan

Hideaki Sakai Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan

Torsten Söderström Systems and Control Group, Uppsala University, Uppsala, Sweden

Joakim Sorelius Systems and Control Group, Uppsala University, Uppsala, Sweden

*Presently on deputation at the International Foundation for Water Science and Technology, PO Box 25862, Abu Dhabi, U.A.E.

Petre Stoica Systems and Control Group, Uppsala University, Uppsala, Sweden

A. V. B. Subrahmanyam Computervision R&D (I) Pvt. Ltd., Pune, India

Sueo Sugimoto Department of Electrical and Electronic Engineering, Ritsumeikan University, Shiga, Japan

Masaki Suwa* Department of Electrical and Electronic Engineering, Ritsumeikan University, Shiga, Japan

A. Lee Swindlehurst Department of Electrical and Computer Engineering, Brigham Young University, Provo, Utah

Kiyotsugu Takaba Department of Applied Mathematics and Physics, Kyoto University, Kyoto, Japan

Teruo Tsuji Department of Electrical Engineering, Kyushu Institute of Technology, Kitakyushu, Japan

Jitendra K. Tugnait Department of Electrical Engineering, Auburn University, Auburn, Alabama

Mats Viberg Department of Applied Electronics, Chalmers University of Technology, Göteborg, Sweden

Zi-Jiang Yang Department of Computer Science and Electronics, Kyushu Institute of Technology, Fukuoka, Japan

**Current affiliation:* OMRON Corporation, Kyoto, Japan.

This Page Intentionally Left Blank

1

Stochastic Realization and System Identification

Giorgio Picci

University of Padova, Padova, Italy

I. INTRODUCTION

Stochastic realization theory deals with modeling of random processes. Given a vector (say m-dimensional) process $y = \{y(t)\}$, construct models of y representing it in terms of simpler and more basic random processes, such as white noise, Markov processes, etc. In particular it deals with procedures for constructing models of (wide-sense) stationary processes, of the form

$$\begin{aligned} x(t+1) &= Ax(t) + Bw(t) \\ y(t) &= Cx(t) + Dw(t) \end{aligned} \tag{1}$$

where $\{w(t)\}$ is some vector normalized white noise, i.e.

$$E\{w(t)w(s)'\} = I\delta(t-s) \qquad E\{w(t)\} = 0$$

where δ is the Kronecker delta function. This representation is called a *state space realization* of the process y. It involves auxiliary variables (i.e., random quantities which are not given as a part of the original data) such as the *state process* x (a stationary Markov process) and the *generating white noise* w, whose peculiar properties lead to representations of y by models having the desired structure. Constructing these auxiliary processes is part of the realization problem.

Wide-sense linear models of this kind are extremely important in

Supported in part by MURST.

applications for a variety of reasons, including the relative simplicity of the probabilistic treatment and the fact that most of the time in practice only second-order statistics are available to describe random phenomena. They are the starting point for popular estimation and control algorithms like Kalman filtering, LQG control, etc.

A. Classical Stochastic Realization

"Classical" stochastic realization theory [5,6,15,18] was developed in the late 1960s. It deals with the problem of computing the parameters A, B, C, D of a state space realization starting from a suitable parametrization of the spectrum or covariance function of the process. This problem is closely related to spectral factorization.

The $m \times m$ *spectral density* matrix of a purely nondeterministic (p.n.d. hereafter) zero-mean process y is the matrix function

$$\Phi(z) = \sum_{t=-\infty}^{\infty} \Lambda(t) z^{-t}$$

where

$$\Lambda(t) := E\{y(t+k)y(k)'\} = E\{y(t)y(0)'\}$$

It is well known that the spectral density matrix of a process admitting a state space realization is a rational function of z. This fact follows easily by the classical Kintchine and Wiener formula for the spectrum of a filtered stationary process [29,65]. The explicit computation of the spectrum is due to Kalman and Anderson [25,5].

Proposition 1.1 *The transfer function $W(z) = C(zI - A)^{-1}B + D$ of any state space representation* (1.1) *of the stationary process y is a* spectral factor *of Φ, i.e.,*

$$W(z)W(1/z)' = \Phi(z) \tag{2}$$

Indeed it can be checked directly that, writing

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)' \tag{3}$$

where $\Phi_+(z) = \frac{1}{2}\Lambda(0) + \Lambda(1)z^{-1} + \Lambda(2)z^{-2} + \cdots$ is the "causal" (i.e., analytic outside of the unit circle) component of $\Phi(z)$, one has

$$\Phi_+(z) = C(zI - A)^{-1}\bar{C}' + \tfrac{1}{2}\Lambda(0) \tag{4}$$

where

$$\Lambda(0) = CPC' + DD' \qquad \bar{C}' = APC' + BD' \tag{5}$$

$P = P'$ being a solution of the Lyapunov equation $P = APA' + BB'$. In other words, the spectrum of a process y described by a state-space model (1) is a rational function expressible in parametric form directly in terms of the parameters of the realization.

Classical stochastic realization is the inverse problem of computing the parameters (A, B, C, D) which define (minimal) spectral factors $W(z) = C(zI - A)^{-1}B + D$, of a spectral density matrix. Of course one assumes here that the process is p.n.d. and has a spectral density which is a rational function of $z = \mathrm{e}^{j\omega}$.

This inverse problem for rational spectral density functions was essentially solved in the 1960s by Kalman, Yakubovich and Popov [26,67,52] in the context of stability theory and was cast into its present spectral factorization form by Anderson [5,6]. The minimal degree spectral factors of $\Phi(z)$ are in one-to-one correspondence with the symmetric $n \times n$ matrices P solving the *Linear Matrix Inequality* (LMI)

$$M(P) := \begin{bmatrix} P - APA' & \bar{C}' - APC' \\ \bar{C} - CPA' & \Lambda(0) - CPC' \end{bmatrix} \geq 0 \tag{6}$$

in the following sense:

Corresponding to each solution $P = P'$ of (6), consider the full column rank matrix factor $\left[\begin{smallmatrix} B \\ D \end{smallmatrix}\right]$ of $M(P)$,

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} [B' \, D'] \tag{7}$$

(this factor is unique modulo right multiplication by orthogonal matrices) and form the rational matrix

$$W(z) := C(zI - A)^{-1}B + D \tag{8}$$

Then (8) is a minimal realization of a minimal analytic spectral factor of $\Phi(z)$. All minimal factors can be obtained in this way.

Under some mild regularity conditions (6) is equivalent to an *Algebraic Riccati Inequality* (ARI) [6,18].

These inequalities have been much studied both from a theoretical and a numerical viewpoint. They play an important role in many areas of system theory such as stability theory, dissipative systems and are central in H^∞ control and estimation theory. It seems to be much less appreciated in the scientific community that they play a very basic role in modeling of stationary random signals as well. Certain solutions of the LMI (or of the ARI) have special probabilistic properties and are related to Kalman filter or "innovations-type" realizations. We shall refer the reader to the literature [66,18,37,41] for a full discussion of these aspects of the problem.

B. Geometric Stochastic Realization

The classical "wide-sense" realization theory is purely distributional as it says nothing about representation of random quantities in a truly probabilistic sense (e.g., how to generate the random variables or the sample paths of a given process, not just its covariance function). In the last two decades a *geometric* or *coordinate-free* approach to stochastic modeling has been put forward in a series of papers by Lindquist, Picci, and Ruckebusch *et al.* [35,36,42,54,55,56] which aims at the representation of random processes in this more specific sense. This idea is also present in the early papers of Akaike [1,2]. It is by now very well understood how to form state spaces and build realizations in a "constructive" manner by a series of geometric operations on a background Hilbert space of linear statistics (random variables) available to the modeler. These statistics are typically just linear functions of the variables of the process y to be modeled but in some situations other random data may be available to construct the model.

Since in most control applications there are also exogenous "input" signals whose influence is to be taken into account in the modeling process, we shall discuss in this paper also some generalizations of the basic ideas of geometric realization theory from time series to random processes in the presence of inputs.

C. Realization and Identification

It is one of the main goals of this paper to persuade the reader that the geometric procedures of realization theory form the conceptual basis of "subspace" identification algorithms. Some of them can actually be numerically implemented very simply after a proper identification of the data Hilbert space has been made.

The statistical problem of identification of a linear dynamic model, in particular a state space model of the type (1), from an observed time series can be approached from (at least) two conceptually different viewpoints.

1. The Traditional "Optimization" Approach

This is based on the principle of minimizing a suitable scalar measure of the discrepancy between the observed data and the data described by the probability law underlying a certain chosen model class. Well-known examples of distance functions are the likelihood function, or the average squared prediction error of the observed data corresponding to a particular model [43]. Except for trivial model classes, these distance functions depend nonlinearly on the model parameters and the minimization can only be done numerically. Hence the optimization approach leads to iterative algorithms in the space of the parameters, say in the space of minimal (A, B, C, D) matrix quadruples which parametrize the chosen model class. In spite of the fact that this has been almost the only accepted paradigm in system identification in the past three decades [43,57], this approach has several well-known drawbacks, among which are the fact that the cost function generally has complicated local minima which are very difficult to detect, for moderate or large dimension of the model there is often structural insensitivity of the cost to variations in the parameters and corresponding ill-posedness of the estimation problem, and there are difficulties in taking consistently into account the (unknown) initial conditions, so that the methods only work "asymptotically", etc.

These limitations, it seems to us, are a consequence of the intrinsically "blind" philosophy which underlies setting the problem as a parameter optimization problem. For, almost all classical problems in control and estimation could be trivially formulated as parametric optimization, and pushing this philosophy to the extreme, one could well say that, in principle, we wouldn't need the maximum principle, Kalman filtering nor H^∞ theory, since we could just formulate and solve everything as a nonlinear programming problem in a suitable space of controller or estimator parameters. It is a fact, however, that no real progress has occurred in the field of control and estimation by following this type of paradigm.

2. Subspace Identification

This is the approach more or less explicitly suggested by [3,8,13,31,58] and also discussed in [39] and in this paper. The basic idea is to introduce a preliminary step in the state space model identification problem consisting in constructing the state space of the model. This is done by geometric operations of projection of certain subspaces generated by the data. For this reason these procedures can be regarded as stochastic realization. The mathematical structure of the problem and its inherent nonlinearity are well known and well understood. The nonlinearity, in particular, has to do with the quadratic nature of the stochastic modeling problem and the diverse manifestations of this fact (Riccati equations or rational spectral factorization equations), depending on the class of parametric models one uses.

In more concrete terms, the paradigm advocated here and in [39,47,48] for identification of state space models from data is to transform the geometric procedures of stochastic realization theory into algorithms of numerical linear algebra. The translation of the abstract geometric operations in the Hilbert space generated by the data will be discussed in Sections II and VI. These operations will have an immediate reformulation as statistical procedures in a "concrete" vector space generated by the shifted tail sequences of the observed data. In practice, when the data are finite, the (approximate) inner product in this Hilbert space will be just the Euclidean inner product

$$\langle \xi, \eta \rangle \cong \frac{1}{N+1} \sum_{t=0}^{N} \xi_t \eta_t$$

and this makes geometric realization procedures particularly simple to translate into vector space computations.

It should be said that the idea of formulating state space identification as a stochastic realization problem is not entirely new and has been present in the literature for some time. In particular Faurre [14,16,17] seems to be the first to systematically attempt to formulate identification as stochastic realization. His context is however still heavily coordinate-dependent. It is actually the geometric viewpoint and the vector space (actually the Hilbert space) characterization of the state space as a subspace of a certain data space that allows systematic introduction of numeric linear algebra and efficient computational tools to solve the problem.

II. THE HILBERT SPACE OF THE OBSERVED DATA

Naturally, in practice, instead of random variables one has a collection of input–output data,

$$\{u_0, u_1, \ldots, u_t, \ldots, u_N\} \qquad \{y_0, y_1, \ldots, y_t, \ldots, y_N\} \tag{9}$$

with $u_t \in \mathbb{R}^p$, $y_t \in \mathbb{R}^m$, measured during some experiment.*

For reasons of clarity and mathematical simplicity, we shall initially consider an idealized situation in which the time series (9) are *infinitely long* and originate in the remote past at $-\infty$. The finite data length situation is discussed in Section VI. The geometric approach of this article is based on the following basic "statistical" assumption on the data.

Assumption 2.1 For $N \rightarrow \infty$ and for any $\tau \geq 0$, the time averages

$$\frac{1}{N+1} \sum_{t=t_0}^{N+t_0} \begin{bmatrix} u_{t+\tau} \\ y_{t+\tau} \end{bmatrix} \begin{bmatrix} u_t \\ y_t \end{bmatrix}' \qquad \tau \geq 0 \tag{10}$$

converge and the limits are independent of the initial time t_0.

*The time averages of all time series considered in this paper will be assumed to be zero. This is clearly no loss of generality.

This assumption can be read as some kind of "statistical regularity" of the (future) data. It is of course unverifiable in practice as it says something about data which have not yet been observed. Some assumption of this sort about the mechanism generating future data seems however to be necessary to even formulate the identification problem.

In a continuous-time setting, functions admitting an "ergodic" limit of the sample correlation function (10) have been studied in depth by Wiener in his famous work on generalized harmonic analysis [63,64]. Although a systematic translation of the continuous-time results of Wiener into discrete-time seems not to be available in the literature, it is quite evident that a totally analogous set of results must hold for discrete-time signals. In particular the reader may show rather easily by adapting Wiener's proof for continuous time that the limits of the time averages (10)

$$\lim_{N\to\infty}\frac{1}{N+1}\sum_{t=t_0}^{N+t_0}\begin{bmatrix}u_{t+\tau}\\ y_{t+\tau}\end{bmatrix}\begin{bmatrix}u_t\\ y_t\end{bmatrix}' := \Lambda(\tau)=\begin{bmatrix}\Lambda_{uu}(\tau) & \Lambda_{uy}(\tau)\\ \Lambda_{yu}(\tau) & \Lambda_{yy}(\tau)\end{bmatrix}\qquad \tau\geq 0 \quad (11)$$

form a matrix function Λ of *positive type*, in other words a *bona fide stationary covariance matrix* sequence. We shall call Λ the *true covariance* of the time series $\{y_t, u_t\}$.

Now, for each $t\in\mathbb{Z}$ define the $p\times\infty$ and $m\times\infty$ matrices

$$u(t) := [u_t, u_{t+1}, u_{t+2}, \ldots] \qquad (12a)$$

$$y(t) := [y_t, y_{t+1}, y_{t+2}, \ldots] \qquad (12b)$$

and consider the sequences $u := \{u(t)\,|\,t\in\mathbb{Z}\}$ and $y := \{y(t)\,|\,t\in\mathbb{Z}\}$. We shall show that these sequences play a very similar role to two jointly wide-sense stationary processes u and y, as those referred to in the previous section.

Define the vector spaces $\mathcal{U}$ and $\mathcal{Y}$ of scalar semi-infinite real sequences obtained as finite linear combinations of the components of u and y,

$$\mathcal{U} := \{\Sigma a_k' u(t_k) \quad a_k\in\mathbb{R}^p, t_k\in\mathbb{Z}\} \qquad (13)$$

$$\mathcal{Y} := \{\Sigma a_k' y(t_k) \quad a_k\in\mathbb{R}^m, t_k\in\mathbb{Z}\} \qquad (14)$$

These vector spaces can be seen as the row spaces of two infinite matrices U and Y having as block rows the semi-infinite entries $u(t)$ and $y(t)$ of (12) for t running on $\mathbb{Z}$.

At this point we need to set notation. In what follows the symbols $\vee$, $+$, and $\oplus$ will denote vector sum, *direct* vector sum, and *orthogonal* vector sum of subspaces; the symbol $\mathbf{X}^{\perp}$ will denote the orthogonal complement of a (closed) subspace $\mathbf{X}$ of a Hilbert space with respect to some predefined ambient space. The orthogonal projection onto the subspace $\mathbf{X}$ will be denoted by the symbol $E(\cdot|\mathbf{X})$ or by the shorthand $E^{\mathbf{X}}$. The notation

$E(z|\mathbf{X})$ will be used also when z is vector valued. The symbol will then just denote the vector with components $E(z_k|\mathbf{X})$, $k = 1, \ldots$. For vector quantities, $|v|$ will denote Euclidean length (or absolute value in the scalar case).

The vector sum $\mathcal{U}\vee\mathcal{Y}$, which can be seen as the row space of the compound infinite matrix $\left[\begin{smallmatrix}U\\Y\end{smallmatrix}\right]$, will originate our basic ambient space. It can be naturally made into an inner product space in the following way. First, define the bilinear form $\langle\cdot,\cdot\rangle$ on the generators of the space by letting

$$\langle a'\begin{bmatrix}u(k)\\y(k)\end{bmatrix}, b'\begin{bmatrix}u(j)\\y(j)\end{bmatrix}\rangle := \lim_{N\to\infty}\frac{1}{N+1}\sum_{t=0}^{N} a'\begin{bmatrix}u_{t+k}\\y_{t+k}\end{bmatrix}\begin{bmatrix}u_{t+j}\\y_{t+j}\end{bmatrix}' b = a'\Lambda(k-j)b \tag{15}$$

for $a, b \in \mathbb{R}^{p+m}$ (the prime denotes transpose). Then extend it by linearity to all elements of $\mathcal{U}\vee\mathcal{Y}$.

Indeed, let $\mathbf{a} := \{a_k, k\in\mathbb{Z}\}$ be a sequence of vectors $a_k \in \mathbb{R}^{p+m}$, with compact support in $\mathbb{Z}$, and let $\mathbf{a}' := \{a_k'\}$. A generic element ξ of the vector space $\mathcal{U}\vee\mathcal{Y}$ can be represented as

$$\xi = \mathbf{a}'\begin{bmatrix}U\\Y\end{bmatrix} := \sum_k a_k'\begin{bmatrix}u(k)\\y(k)\end{bmatrix}$$

Introducing the infinite block-symmetric positive semidefinite Toeplitz matrix

$$T = \begin{bmatrix}\Lambda(0) & \Lambda(1) & \ldots & \Lambda(k) & \ldots\\ \Lambda(1)' & \Lambda(0) & \Lambda(1) & \ldots & \ldots\\ \vdots & & \ddots & & \vdots\\ \Lambda(k)' & & & \Lambda(0) & \\ \ldots & & & & \end{bmatrix} \tag{16}$$

constructed from the "true" covariance sequence $\{\Lambda(0), \Lambda(1), \ldots, \Lambda(k), \ldots\}$ of the data, the bilinear form (15) on $\mathcal{U}\vee\mathcal{Y}$ can be represented by the quadratic form

$$\langle\xi,\eta\rangle = \langle\mathbf{a}'\begin{bmatrix}u\\y\end{bmatrix}, \mathbf{b}'\begin{bmatrix}u\\y\end{bmatrix}\rangle = \sum_{kj} a_k'\Lambda(k-j)b_j = \mathbf{a}'T\mathbf{b}$$

We shall identify elements whose difference has zero norm (this means $\langle\xi,\xi\rangle = 0 \Leftrightarrow \xi = 0$). From this expression it may be seen that the bilinear form is nondegenerate (unless $\Lambda = 0$ identically) and defines a bona fide inner product. In the following we shall assume that for every k, each square block-Toeplitz matrix T_k in the upper left-hand corner of T is positive definite. By closing the vector space $\mathcal{U} \vee \mathcal{Y}$ with respect to convergence in the norm induced by the inner product (15), one obtains a real Hilbert space $\mathbf{H} := \mathbf{U} \vee \mathbf{Y}$ (the wedge now means closed vector sum). This is the basic data space on which hereafter the models will be defined.

Note that if in the limits of the sum (15) $t = 0$ is replaced by an arbitrary initial instant t_0 the limit does not change, so that

$$\langle a' \begin{bmatrix} u(k) \\ y(k) \end{bmatrix}, b' \begin{bmatrix} u(j) \\ y(j) \end{bmatrix} \rangle = \langle a' \begin{bmatrix} u(t_0 + k) \\ y(t_0 + k) \end{bmatrix}, b' \begin{bmatrix} u(t_0 + j) \\ y(t_0 + j) \end{bmatrix} \rangle$$

for all t_0 (wide-sense stationarity). We shall define a *shift operator* σ on the family of semi-infinite matrices (12), by setting

$$\sigma a' u(t) = a' u(t+1) \quad t \in \mathbb{Z}, \quad a \in \mathbb{R}^p$$
$$\sigma a' y(t) = a' y(t+1) \quad t \in \mathbb{Z}, \quad a \in \mathbb{R}^m,$$

defining a linear map which is isometric with respect to the inner product (15) and extendable by linearity to all of $\mathbf{H}$.

This Hilbert space framework was first introduced in [39] for time series. It is shown in this reference that it is isomorphic to the standard stochastic Hilbert space setup widely used in the L^2-theory of second-order random processes [30,53]. By virtue of this isomorphism one can formally think of the observed time series (9) as an ergodic sample path of some Gaussian stationary stochastic process $(\mathbf{u}, \mathbf{y})$ defined on a true probability space and having joint covariance matrices equal to the limit (11) of the sum (10) as $N \to \infty$.

Linear functions and operators on the tail sequence u and y defined in (12) correspond to the same linear functions and operators on the processes $\mathbf{u}$ and $\mathbf{y}$. In particular the second-order moments of the two random processes can equivalently be calculated in terms of the tail sequences u and y provided expectations are substituted by ergodic limits of the type (15). Since second-order properties are all that matters in this paper, one may even regard the tail sequences u and y of (12) as being the *same* object as the two underlying stochastic processes $\mathbf{u}$ and $\mathbf{y}$. This will be done in the rest of this paper. The probabilistic language can be adopted in the present setting provided one identifies real random variables as semi-infinite strings of numbers having the "ergodic property" described at the beginning of this section. The inner product of two semi-infinite strings ξ and η in $\mathbf{H}$ will be called *expectation* and denoted $E\{\xi\eta\}$,

$$\langle \xi, \eta \rangle = E\{\xi\eta\} \tag{17}$$

In the following we shall allow $E\{\cdot\}$ to operate on matrices, taking inner products row by row.

This unification of language permits us to carry over in its entirety the geometric theory of stochastic realization derived in the abstract L^2 setting of [34,36,37] to the present framework. One may just reinterpret everything in the current setting, starting from the definition of $\mathbf{U}_t^-$, $\mathbf{Y}_t^-$, $\mathbf{U}_t^+$, $\mathbf{Y}_t^+$, the *past* and *future* subspaces of the "processes" u and y at time t. These are defined as the closure of the linear vector spaces spanned by the relative past or future "random variables" $u(t)$ and $y(t)$, in the metric of the Hilbert space $\mathbf{H}$. We shall use the notation

$$\begin{aligned}\mathbf{U}_t^- &:= \overline{\text{span}}\{u(s)\,|\,s<t\}\\ \mathbf{Y}_t^- &:= \overline{\text{span}}\{y(s)\,|\,s<t\}\\ \mathbf{U}_t^+ &:= \overline{\text{span}}\{u(s)\,|\,s\geq t\}\\ \mathbf{Y}_t^+ &:= \overline{\text{span}}\{y(s)\,|\,s\geq t\}\end{aligned}$$

Note that, according to a widely accepted convention, the present is included in the future only and not in the past. The only difference to keep in mind here is the different interpretation that representation formulas like (1) have in the new context. The equalities involved in the representation

$$\begin{aligned}x(t+1) &= Ax(t) + Bw(t)\\ y(t) &= Cx(t) + Dw(t)\end{aligned} \tag{18}$$

are now to be understood in the sense of equalities of elements of $\mathbf{H}$, i.e. as asymptotic equality of sequences in the sense of Cesàro limits. In particular the equality signs in the model (18) imply nothing about the possibility that the same relations would be holding for the *sample values* y_t, x_t, w_t at a particular instant of time t. This is in a certain sense similar to the "with probability one" interpretation of the equality sign given to the model (18) in case the variables are bona fide random variables in a probability space.

Modeling and estimation of stationary processes on infinite or semi-infinite time intervals naturally involve various linear operations on the variables of the process which are *time-invariant*, i.e. independent of the particular instant of time chosen as the "present". In this setting it is possible (and convenient) to fix the present instant of time to an arbitrary value say $t = 0$ and work as if time was "frozen" at $t = 0$. At the occurrence one then "shifts" the operations in time by the action of the unitary operator σ^t on the data. In particular, the future and past subspaces of the processes y and u will often be considered referred to time $t = 0$ and denoted $\mathbf{Y}^+$ and $\mathbf{Y}^-$. For an arbitrary present instant t we have

$$\mathbf{Y}_t^+ = \sigma^t\mathbf{Y}^+, \qquad \mathbf{Y}_t^- = \sigma^t\mathbf{Y}^-$$

Last but not least, it should be appreciated that in this setting the projection operator has an immediate *statistical* meaning.

Consider the orthogonal projection $E[\xi|\mathbf{X}]$ of a (row) random variable ξ onto a subspace $\mathbf{X}$ of the space $\mathbf{H}$. This has the well-known interpretation of conditional expectation given the random variables in $\mathbf{X}$, in the probabilistic Gaussian setting. Whenever $\mathbf{X}$ is given as the rowspace of some matrix of generators X, the projection $E[\xi|\mathbf{X}]$ has exactly the familiar aspect of the least squares formula expressing the best approximation of the vector ξ as a linear combination of the rows of X. For, writing $E[\xi|X]$ to denote the projection expressed (perhaps nonuniquely) in terms of the rows of X, the classical linear "conditional expectation" formula leads to

$$E[\xi|X] = \xi X'[XX']^{\#} X \tag{19}$$

In the case of linearly independent rows we can substitute the pseudoinverse $\#$ with a true inverse.

III. INPUT–OUTPUT MODELS

In this section we shall review briefly some of the conceptual steps which lead to the construction of state space models of a jointly stationary "input–output" pair of random processes $[y, u]$.

Very often, especially in control applications, there are "input" or "exogenous" variables that must be taken into account to describe accurately the dynamics of the output signal y. One is however not interested in modeling the dynamics of the input signal *per se* and would like to estimate a (causal) dynamical model relating u and y which is independent of the particular input signal.

A typical route which is commonly taken is to impose on the data a causal (rational) transfer function model* $\hat{y} = W(z)\hat{u}$ which relates the input to a "nominal" undisturbed output. The transfer function is estimated by some variant of linear regression as if u was a known deterministic sequence. Sometimes in the literature it is explicitly "assumed" that u is a "deterministic" signal. The residual regression error appears as an additive stochastic uncertainty term modeling all unpredictable causes *independent of* u, which make y deviate from the nominal trajectory caused by the action of the input signal.

Although this is often a reasonable scheme to follow, one should be aware that it may lead to serious errors whenever there are "stochastic

*Here we use the Z-transform formalism.

components" in u, correlated with the past history of y. This is so because the input variable $u(t)$ could then itself be described by a nontrivial causal dynamical model involving the "output" process y playing in this case the role of an exogenous variable (which should be called the "input" in this context) to determine u. This "feedback structure" is extensively discussed in the literature [10,20,7,21] where it is shown that it is *nonunique* and that there are in general infinitely many causal feedback schemes which could represent a given joint process. Even if, as in many other instances of stochastic modeling, this nonuniqueness could be resolved by selecting one particular representative, say an *innovation* feedback scheme, there is in general no hope of recovering a fixed open loop subsystem independent of the feedback link generating u.

Identification of a causal input–output relation in the presence of feedback is hence an ill-posed problem (this of course in the absence of any other specific information on the structure of the feedback link) and the problem is better formulated as the identification of the *joint process* (y, u) on the basis of the joint corresponding observed time series. This in turn falls into the general setup of time series identification. In the rest of this section we shall discuss specifically the case when there is absence of feedback (from y to u) in the observed data.

A. Feedback-Free Processes

The appropriate setup for discussing feedback-free models is the theory of *feedback* and *causality* between stationary processes à la Granger [23]. See also [10,7,20,21,49]. We shall rephrase it in the language of conditionally orthogonal subspaces. The notation $\mathbf{A} \perp \mathbf{B} \mid \mathbf{X}$ means that the two subspaces $\mathbf{A}, \mathbf{B} \subset \mathbf{H}$ are *conditionally orthogonal* given a third subspace $\mathbf{X}$, i.e.,

$$\langle \alpha - E^{\mathbf{X}}\alpha, \beta - E^{\mathbf{X}}\beta \rangle = 0 \quad \text{for } \alpha \in \mathbf{A}, \beta \in \mathbf{B} \tag{20}$$

When $\mathbf{X} = 0$, this reduces to the usual orthogonality $\mathbf{A} \perp \mathbf{B}$. Conditional orthogonality is orthogonality after subtracting the orthogonal projections onto $\mathbf{X}$. This concept is discussed in depth in [36,37].

One says that there is *absence of feedback* from y to u in the sense of Granger if the future of u is conditionally uncorrelated (which is the same as independent in the Gaussian case) from the past of y given the past of u itself. In our Hilbert space setup this is written as,

$$\mathbf{U}_t^+ \perp \mathbf{Y}_t^- \mid \mathbf{U}_t^- \tag{21}$$

where $\mathbf{U}_t^-, \mathbf{Y}_t^-, \mathbf{U}_t^+, \mathbf{Y}_t^+$ are the past and future subspaces of the processes u and y at time t.

This conditional orthogonality condition will be another basic assumption throughout this paper. It is quite easy to see that, in conditions of absence of feedback, the "causal estimation error" process

$$y_s(t) := y(t) - E[y(t) \mid \mathbf{U}^-_{t+1}] \tag{22}$$

coincides with $y(t) - E[y(t) \mid \mathbf{U}]$ and hence is uncorrelated with the whole history of the input process u,

$$y_s(t) \perp \mathbf{U} \quad \text{for all } t$$

see [47]. Hence the process y_s may be called the *stochastic component* of y.

It also follows that the stochastic process y_u defined by the complementary projection

$$y_u(t) := E[y(t) \mid \mathbf{U}] \quad t \in \mathbb{Z}$$

is uncorrelated with y_s. It will be named the *deterministic component* of y.

In the present feedback-free setting there is no non-uniqueness, arbitrariness, or "user choice" in modeling and we have a natural unique "input–output" linear model

$$y(t) = y_u(t) + y_s(t) = E[y(t) \mid u(s); s \leq t] + y_s(t) \tag{23}$$

where $E[y(t) \mid u(s); s \leq t]$ is the best (in the sense of minimum variance of the error) estimate of the output $y(t)$ based on the past of u up to time t. Under some regularity conditions on the input process to be made precise later on, this estimate is described by a causal and stable linear convolution operator.

Identifying the model (23) means identifying both the input–output "deterministic" part (described by a transfer function $W(z)$) and the additive "noise" process y_s. This last component is always present and may well be the most important for a realistic description of the output.

It is obvious that state space descriptions for the process y can be obtained by combining two separate state space models for y_s and y_u. For example, a (forward) innovation representation of y is obtained by combining together the (forward) innovation representation of y_s

$$x_s(t+1) = A_s x_s(t) + B_s e_s(t) \tag{24a}$$

$$y_s(t) = C_s x_s(t) + e_s(t) \tag{24b}$$

where $e_s(t)$ is the one-step prediction error of the process y_s based on its own past, i.e., the (forward) innovation process of y_s, and the "deterministic" state space model for y_u

$$x_u(t+1) = A_u x_u(t) + B_u u(t) \tag{25a}$$
$$y_u(t) = C_u x_u(t) + D_u u(t) \tag{25b}$$

The process e_s has then the meaning of conditional innovation of y [47].

By combining together (24) and (25), the state space innovation model of the process y "with inputs" has the following structure,

$$\begin{aligned} \begin{bmatrix} x_s(t+1) \\ x_u(t+1) \end{bmatrix} &= \begin{bmatrix} A_s & 0 \\ 0 & A_u \end{bmatrix} \begin{bmatrix} x_s(t) \\ x_u(t) \end{bmatrix} + \begin{bmatrix} 0 \\ B_u \end{bmatrix} u(t) + \begin{bmatrix} B_s \\ 0 \end{bmatrix} e_s(t) \\ y(t) &= [C_s \; C_u] \begin{bmatrix} x_s(t) \\ x_u(t) \end{bmatrix} + D_u u(t) + e_s(t) \end{aligned} \tag{26}$$

Models of this kind are naturally interpreted as state space realizations of the familiar ARMAX-type "input–output" relations $\hat{y} = W(z)\hat{u} + G(z)\hat{e}$ (here we have $W(z) = D_u + C_u(zI - A_u)^{-1}B_u$ and $G(z) = D_s + C_s(zI - A_s)^{-1}B_s$) often used in the identification literature.

It may happen that even if the realizations of the two subsystems (stochastic and deterministic) are minimal, (26) may give redundant descriptions of the signals in certain particular cases, as there may be loss of observability when the transfer functions $W(z)$ and $G(z)$ have common poles and common corresponding eigenspaces.

These cases are highly nongeneric and in practice one need not worry about this unlikely evenience in black-box identification. However, in certain structured problems one may have a priori some knowledge about the way the input or the noise enters in the system and there may be noise effects which one specifically wants to model as being subject to the same dynamics as the input. In these cases there is actually a need to use models which allow for common dynamics.

IV. CONSTRUCTING THE STATE SPACE OF THE "STOCHASTIC" COMPONENT

In this section we shall discuss geometric realization theory for the stochastic component of y. There will be no input processes in this section and for notational simplicity we shall drop the subscript "s".

The geometric theory centers on the idea of *Markovian splitting subspaces* for the process y. This concept is the probabilistic analog of the deterministic notion of state space of a dynamical system and captures at an abstract level the property of "dynamic memory" that the state variables have in deterministic system theory. Once a stochastic state space is given, the procedures for the construction of the auxiliary random quantities which enter in the model and in particular the state process are fairly obvious. Since the state variables $x(t)$ of a particular realization can be regarded just as a particular basis for the state space, once a (minimal) state space is constructed, finding state equations is just a matter of choosing a basis and computing coordinates.

Let y be a stationary vector process and $\mathbf{Y}$ the relative Hilbert space of linear functionals. Let $\mathbf{X}$ be a subspace of some large stationary Hilbert space $\mathbf{H}$ of wide-sense random variables containing $\mathbf{Y}$. Define

$$\mathbf{X}_{\mathrm{t}} := \sigma^t\mathbf{X}, \qquad \mathbf{X}_t^- := \vee_{s\leq t}\mathbf{X}_s, \qquad \mathbf{X}_t^+ := \vee_{s\geq t}\mathbf{X}_s$$

Definition 4.1 *A Markovian splitting subspace* $\mathbf{X}$ for the process y is a subspace of $\mathbf{H}$ making the vector sums $\mathbf{Y}^-\vee\mathbf{X}^-$ and $\mathbf{Y}^+\vee\mathbf{X}^+$ conditionally orthogonal (i.e., uncorrelated) given $\mathbf{X}$, denoted

$$\mathbf{Y}^-\vee\mathbf{X}^-\perp\mathbf{Y}^+\vee\mathbf{X}^+\,|\,\mathbf{X}. \tag{27}$$

The conditional orthogonality condition (27) can be equivalently written as

$$E[\mathbf{Y}^+\vee\mathbf{X}^+\,|\,\mathbf{Y}^-\vee\mathbf{X}^-] = E[\mathbf{Y}^+\vee\mathbf{X}^+\,|\,\mathbf{X}] \tag{28}$$

which gives the intuitive meaning of the splitting subspace $\mathbf{X}$ as a dynamic memory of the past for the purpose of predicting the joint future.

The subspace $\mathbf{X}$ is called *proper*, or *purely nondeterministic* if

$$\cap_t\mathbf{Y}_t^-\vee\mathbf{X}_t^- = \{0\} \quad \text{and} \quad \cap_t\mathbf{Y}_t^+\vee\mathbf{X}_t^+ = \{0\}$$

Obviously for the existence of proper splitting subspaces y must also be purely nondeterministic [53]. Properness is, by the Wold decomposition theorem, equivalent to the existence of two vector white noise processes w and $\bar{w}$ such that

$$\mathbf{Y}^-\vee\mathbf{X}^- = \mathbf{H}^-(w), \qquad \mathbf{Y}^+\vee\mathbf{X}^+ = \mathbf{H}^+(\bar{w})$$

Here the symbols $\mathbf{H}^-(w)$, $\mathbf{H}^+(w)$, etc., denote the Hilbert subspaces linearly generated by the past and future of the process w. The spaces

$$\mathbf{S} := \mathbf{Y}^-\vee\mathbf{X}^- \quad \text{and} \quad \bar{\mathbf{S}} := \mathbf{Y}^+\vee\mathbf{X}^+ \tag{29}$$

associated to a Markovian splitting subspace $\mathbf{X}$, play an important role in the geometric theory of stochastic systems. They are called the *scattering pair* of $\mathbf{X}$ as they can be seen to form an incoming–outgoing pair in the sense of Lax–Phillips scattering theory [33].

Definition 4.2 Given a stationary Hilbert space $(\mathbf{H},\sigma)$ containing $\mathbf{Y}$, a *scattering pair* for the process y is a pair of subspaces $(\mathbf{S},\bar{\mathbf{S}})$ satisfying the following conditions:

1. $\sigma^*\mathbf{S}\subset\mathbf{S}$ and $\sigma\bar{\mathbf{S}}\subset\bar{\mathbf{S}}$, i.e., $\mathbf{S}$ and $\bar{\mathbf{S}}$ are invariant for the left and right shift semigroups (this means that $\mathbf{S}_t$ is increasing and $\bar{\mathbf{S}}_t$ is decreasing with time).
2. $\mathbf{S}\vee\bar{\mathbf{S}}=\mathbf{H}$.
3. $\mathbf{S}\supset\mathbf{Y}^-$ and $\bar{\mathbf{S}}\supset\mathbf{Y}^+$.
4. $\mathbf{S}^\perp\subset\bar{\mathbf{S}}$ or, equivalently, $\bar{\mathbf{S}}^\perp\subset\mathbf{S}$.

The following representation theorem provides a link between Markovian splitting subspaces and scattering pairs.

Theorem 4.3 *The intersection*

$$\mathbf{X}=\mathbf{S}\cap\bar{\mathbf{S}} \tag{30}$$

of any scattering pair of subspaces of $\mathbf{H}$ *is a Markovian splitting subspace. Conversely every Markovian splitting subspace can be represented as the intersection of a scattering pair. The correspondence* $\mathbf{X}\leftrightarrow(\mathbf{S},\cap\bar{\mathbf{S}})$ *is one-to-one, the scattering pair corresponding to* $\mathbf{X}$ *being given by*

$$\mathbf{S}=\mathbf{Y}^-\vee\mathbf{X}^-,\qquad \bar{\mathbf{S}}=\mathbf{Y}^+\vee\mathbf{X}^+ \tag{31}$$

The process of forming a scattering pair associated to $\mathbf{X}$ should be thought of as an "extension" of the past and future spaces of y. The rationale for this extension is that scattering pairs have an extremely simple splitting geometry due to the fact that

$$\mathbf{S}\perp\bar{\mathbf{S}}\,|\,\mathbf{S}\cap\bar{\mathbf{S}} \tag{32}$$

which is called *perpendicular intersection*. It is easy to show that Property 4 in the definition of a scattering pair is actually equivalent to perpendicular intersection. This property of conditional orthogonality given the intersection can also be seen as a natural generalization of the Markov property.* Note that $\mathbf{A}\perp\mathbf{B}\,|\,\mathbf{X}\Rightarrow\mathbf{A}\cap\mathbf{B}\subset\mathbf{X}$ but the inclusion of the intersection in the splitting subspace $\mathbf{X}$ is only *proper* in general. For perpendicularly intersecting subspaces, the intersection is actually the *unique minimal subspace* making them conditionally orthogonal.

Any basis vector $x(0):=[x_1(0),\,x_2(0),\,\ldots,\,x_n(0)]'$ in a (finite-dimensional) Markovian splitting subspace $\mathbf{X}$ generates a stationary Markov process $x(t):=\sigma^t x(0)$, $t\in\mathbb{Z}$ which serves as a *state* of the process y. If $\mathbf{X}$ is proper, the Markov process is purely nondeterministic.

*In which case $\mathbf{S}=\mathbf{X}^-$, $\bar{\mathbf{S}}=\mathbf{X}^+$, and $\mathbf{X}=\mathbf{X}^-\cap\mathbf{X}^+$.

Denote by $\mathbf{W}_t$, $\bar{\mathbf{W}}_t$ the spaces spanned by the components, at time t, of the generating noises $w(t)$ and $\bar{w}(t)$, of the scattering pair of $\mathbf{X}$. Since

$$\mathbf{S}_{t+1} = \mathbf{S}_t \oplus \mathbf{W}_t \tag{33}$$

we can write

$$\mathbf{X}_{t+1} \subset \mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_t = (\mathbf{S}_t \cap \bar{\mathbf{S}}_t) \oplus (\mathbf{W}_t \cap \bar{\mathbf{S}}_t) \tag{34}$$

since $\bar{\mathbf{S}}_t$ is decreasing in time, we have $\mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_t \supset \mathbf{X}_{t+1}$ and by projecting the shifted basis $\sigma x(t) := x(t+1)$ onto the last orthogonal direct sum above, the time evolution of any basis vector $x(t) := [x_1(t), x_2(t), \ldots, x_n(t)]'$ in $\mathbf{X}_t$ can be represented by a linear equation of the type $x(t+1) = Ax(t) + Bw(t)$. It is also easy to see that by the p.n.d. property, A must have all its eigenvalues strictly inside of the unit circle. Naturally, by decomposing instead $\bar{\mathbf{S}}_{t-1} = \bar{\mathbf{S}}_t \oplus \bar{\mathbf{W}}_t$ we could have obtained a *backward difference equation* model for the Markov process x, driven by the backward generator $\bar{w}$.

Note also that by definition of the past space, we have $y(t) \in (\mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_t)$. Inserting the decomposition (33) and projecting $y(t)$ leads to a state output equation of the form $y(t) = Cx(t) + Dw(t)$. Here again we could have equivalently obtained an equation driven by the backward noise $\bar{w}$ instead.

As we have seen, any basis in a Markovian splitting subspace produces a stochastic realization of y. It is easy to reverse the implication. We have in fact the following fundamental characterization.

Theorem 4.4 [37,42] *The state space* $\mathbf{X} = \operatorname{span}\{x_1(0), x_2(0), \ldots, x_n(0)\}$ *of any stochastic realization* (1) *is a Markovian Splitting Subspace for the process* y.

Conversely, given a finite-dimensional Markovian splitting subspace $\mathbf{X}$, *to any choice of basis* $x(0) = [x_1(0), x_2(0), \ldots, x_n(0)]'$ *in* $\mathbf{X}$ *there corresponds a stochastic realization of* y *of the type* (1).

Once a basis in $\mathbf{X}$ is available, there are obvious formulas expressing the coefficient matrices A, C and $\bar{C}$ in terms of the processes x and y:

$$A = Ex(t+1)x(t)'P^{-1} \tag{35}$$

$$C = Ey(t)x(t)'P^{-1} \tag{36}$$

$$\bar{C} = Ey(t-1)x(t)' \tag{37}$$

where P is the Gramian matrix of the basis (equal to the state covariance matrix). The matrices B and D however are related to the (unobservable) generating white noise w and require the solution of the LMI.

Stochastic realizations are called *internal* when $\mathbf{H} = \mathbf{Y}$, i.e., the state space is built from the Hilbert space made just of the linear statistics of the process y. For identification the only realizations of interest are the internal ones.

A central problem of geometric realization theory is to construct and to classify the *minimal* state spaces, i.e., the minimal Markovian splitting subspaces for the process y.

The obvious ordering of subspaces of $\mathbf{H}$ by inclusion induces an ordering on the family of Markovian splitting subspaces. The notion of minimality is most naturally defined with respect to this ordering. Note that this definition is independent of assumptions of finite dimensionality and applies also to infinite dimensional Markovian splitting subspaces, i.e., to situations where comparing dimension would not make much sense.

Definition 4.5 A Markovian splitting subspace is *minimal* if it doesn't contain (properly) other Markovian splitting subspaces.

Contrary to the deterministic situation minimal Markovian splitting subspaces are *nonunique*. Two very important examples are the *forward* and *backward predictor spaces* (at time zero):

$$\mathbf{X}_- := E^{\mathbf{H}^-}\mathbf{H}^+, \qquad \mathbf{X}_+ := E^{\mathbf{H}^+}\mathbf{H}^- \tag{38}$$

for which we have the following characterization [36].

Proposition 4.6 *The subspaces* $\mathbf{X}_-$ *and* $\mathbf{X}_+$ *are the unique minimal splitting subspaces contained in the past* $\mathbf{H}^+$, *and, respectively, in the future* $\mathbf{H}^+$, *of the process* y.

The study of minimality forms an elegant chapter of geometric system theory. There are several known geometric and algebraic characterizations of minimality of splitting subspaces and of the corresponding stochastic state space realizations. Since however the discussion of this topic would take us too far from the main theme of the paper we shall refer the reader to the literature [36,37].

V. STATIONARY REALIZATION OF THE DETERMINISTIC COMPONENT

State space modeling of the deterministic component y_u may be based on the same principle of constructing splitting subspaces for the future of the process y_u and the past of the input process u as discussed in the previous sections. However this approach leads to state space models which are driven by white noise and "include" also the dynamics of the process u, which is not interesting for identification and we do not want to appear explicitly in the model. For example, it was shown in [46] and in [1] that one may choose as state space for y_u the predictor space $E[\hat{\mathbf{Y}}_t^+ | \mathbf{U}_t^-]$, but that this choice leads to an innovation model for y_u where the state process is driven by the forward innovation process of u. This model in general includes as a cascade subsystem a state space innovation representation for u and therefore does involve the particular dynamics of the input process.

As we shall see, to construct state space descriptions driven by a nonwhite process u, it is necessary to generalize the geometric theory of stochastic realization of the previous section.

In order to streamline notation, we shall drop the subscript u throughout this section and whenever possible fix $t = 0$. Assume that y and u are two jointly stationary p.n.d. processes of dimensions m and p. We shall call a model of the type

$$x(t+1) = Ax(t) + Bu(t) \tag{39a}$$

$$y(t) = Cx(t) + Du(t) \tag{39b}$$

a *deterministic realization of y* with input process u. Models of this kind reduce to the standard (Markovian) stochastic models when, of course, u is white noise. As usual a realization is called *minimal* if the dimension of the state vector is as small as possible. For minimal realizations it must necessarily hold that (A, B, C) is a minimal triplet. If A has all eigenvalues inside the unit circle ($|\lambda(A)| < 1$), both $x(t)$ and $y(t)$ can be expressed as functionals of the infinite past of u, i.e. their components belong to $\mathbf{U}_t^-$. Realizations with this property will be called *causal.*

We shall now discuss the geometric construction of the state space of a deterministic realization of y. To this purpose we shall introduce a technical assumption of "sufficient richness" of the input process.

Assumption 5.1 For each t the input space $\mathbf{U}$ admits the direct sum decomposition

$$\mathbf{U} = \mathbf{U}_t^- + \mathbf{U}_t^+ \tag{40}$$

An analogous condition (namely $\mathbf{U}_t^- \cap \mathbf{U}_t^+ = 0$) is discussed in [39] where it is shown that it is equivalent to strict positivity of the spectral density matrix of u on the unit circle, i.e. $\Phi_u(e^{j\omega}) > cI$, $c > 0$, or to all canonical angles between the past and future subspaces of u being strictly positive (or, in turn, to all canonical correlation coefficients between past and future of the input process being strictly less than one). A slightly stronger version of this condition is found in [53], Chapter II, Sect. 7.

We assume all through this section that $y(t) \in \mathbf{U}_{t+1}^-$ (this is actually the feedback-free property). Because of this, and in virtue of Assumption 5.1, $y(t)$ has a unique representation as a causal functional

$$y(t) = \sum_{-\infty}^{t} W_{t-k}u(k) \tag{41}$$

where $\hat{W}(z) = \Sigma_0^{+\infty} W_k z^{-k}$ is analytic in $\{|z| > 1\}$. Indeed, $\hat{W}(z)$ is just the transfer function of the Wiener filter $y(t) = E[y(t)|\mathbf{U}_{t+1}^-]$ and can be expressed as

$$\hat{W}(z) = [\Phi_{yu}(z)G(1/z)^{-T}]_+ G(z)^{-1}$$

where $G(z)$ is the outer (or minimum-phase) spectral of Φ_u and the symbol $[\cdot]_+$ means "analytic part", see, e.g., [53], Chapter II. It is evident that $\hat{W}(z)$ is analytic and, because of the nonsingularity of Φ_u on the unit circle, unique almost everywhere.

Since the input–output map relating u and y must be causal, it follows that in our case the only realizations of interest are the causal ones.

The *oblique* projection of a random variable $\eta \in \mathbf{U}$ onto $\mathbf{U}_t^-$ along $\mathbf{U}_t^+$ will be denoted by $E_{\|\mathbf{U}_t^+}[\eta|\mathbf{U}_t^-]$. Clearly, if u is a white noise process, this is the ordinary orthogonal projection onto $\mathbf{U}_t^-$.

Definition 5.2 We shall call a subspace $\mathbf{X} \subset \mathbf{U}^-$ *a oblique (causal) splitting subspace* for the pair $(\mathbf{Y}^+, \mathbf{U}^-)$ if

$$E_{\|\mathbf{U}^+}[\mathbf{Y}^+ \vee \mathbf{X}^+ | \mathbf{U}^-] = E_{\|\mathbf{U}^+}[\mathbf{Y}^+ \vee \mathbf{X}^+ | \mathbf{X}] \tag{42}$$

Note that this condition is a generalization of the conditional orthgonality condition (28) of the Markovian case. For the reasons explained a moment ago, we shall only consider causal splitting subspaces in this paper so the "incoming" subspace $\mathbf{S}$ is now fixed equal to $\mathbf{U}^-$. The *oblique predictor space* $\mathbf{X}^{+/-} := E_{\|\mathbf{U}^+}[\mathbf{Y}^+|\mathbf{U}^-]$ is obviously contained in $\mathbf{U}^-$ and is oblique splitting. It is in fact the *minimal* causal oblique splitting subspace.

Write

$$y(t) = (H_W u)(t) + (W^+ u)(t) \tag{43}$$

where,

$$(H_W u)(t) := \sum_{-\infty}^{-1} W_{t-k}u(k), \qquad (W^+ u)(t) := \sum_{0}^{t} W_{t-k}u(k) \tag{44}$$

Evidently $(H_W u)(t) \in \mathbf{U}^-$ and $(W^+ u)(t) \in \mathbf{U}^+$ for $t \geq 0$. This obvious fact is formally recorded in the followimg lemma.

Lemma 5.3 *For $t \geq 0$ the random variable $(H_W u)(t)$ is the oblique projection of $y(t)$ onto $\mathbf{U}^-$ and belongs to $\mathbf{X}^{+/-}$. Consequently it belongs to all causal oblique splitting subspaces $\mathbf{X}$.*

Lemma 5.4 *Let $\mathbf{X}$ be an oblique splitting subspace and define*

$$\bar{\mathbf{S}} := \mathbf{Y}^+ \vee \mathbf{X}^+$$

Then $\mathbf{X} = E_{\|\mathbf{U}^+}[\bar{\mathbf{S}} \mid \mathbf{U}^-]$ (hence $\mathbf{X}$ is the minimal oblique splitting subspace for $\bar{\mathbf{S}}$ and $\mathbf{U}^-$ contained in $\mathbf{U}^-$).

Proof. Every element $\bar{s}$ of $\bar{\mathbf{S}}$ has the form $\bar{s} = y + x$, $y \in \mathbf{Y}^+$, $x \in \mathbf{X}^+$ so that $E_{\|\mathbf{U}^+}[y \mid \mathbf{U}^-] \in \mathbf{X}^{+/-} \subset \mathbf{X} \subset \mathbf{U}^-$. On the other hand, by definition of oblique splitting we have

$$E_{\|\mathbf{U}^+}[\mathbf{X}^+ \mid \mathbf{U}^-] = E_{\|\mathbf{U}^+}[\mathbf{X}^+ \mid \mathbf{X}] \qquad x \in \mathbf{X}^+$$

therefore

$$\overline{\text{span}}\{E_{\|\mathbf{U}^+}[\bar{s} \mid \mathbf{U}^-] \mid \bar{s} \in \bar{\mathbf{S}}\} = \mathbf{X}$$

This implies that $\mathbf{X}$ is the oblique predictor space of $\bar{\mathbf{S}}$ with respect to $\mathbf{U}^-$ and hence it is minimal splitting. □

The following is the "deterministic" analog of perpendicular intersection.

Lemma 5.5 *Let the symbols have the same meaning as in Lemma 5.2. Then*

$$\bar{\mathbf{S}} \cap \mathbf{U}^- = \mathbf{X}$$

Proof. First note that $\mathbf{X}$ contains the intersection $\bar{\mathbf{S}} \cap \mathbf{U}^-$. For if $\eta \in \bar{\mathbf{S}} \cap \mathbf{U}^-$ then clearly it belongs to $E_{\|\mathbf{U}^+}[\bar{\mathbf{S}} \mid \mathbf{U}^-]$ which is equal to $\mathbf{X}$ in view of the previous lemma.

Then just observe that, conversely, the intersection contains $\mathbf{X}$, since $\bar{\mathbf{S}} \supset \mathbf{X}$ and $\mathbf{U}^- \supset \mathbf{X}$. This proves the lemma. □

The result in particular applies to the extended future space $\bar{\mathbf{Y}}^+ = \mathbf{Y}^+ \vee (\mathbf{X}^{+/-})^+$ (this is in a sense the "minimal" $\bar{\mathbf{S}}$).

Theorem 5.6 *The oblique predictor space can be computed as the intersection*

$$\mathbf{X}^{+/-} = (\mathbf{Y}^+ \vee \mathbf{U}^+) \cap \mathbf{U}^- \tag{45}$$

The proof of this result requires minimality theory and will not be given here. It will be found in a forthcoming publication [50].

Note that it is in general *not true* that $\mathbf{X}^{+/-} = \mathbf{Y}^+ \cap \mathbf{U}^-$ as

$$E_{\|\mathbf{U}^+}[\mathbf{Y}^+ | \mathbf{U}^-] \supset \mathbf{Y}^+ \cap \mathbf{U}^-$$

properly, unless some special conditions are satisfied. Some partial results pointing in the direction of Theorem 5.6 can be found in [44], Thm. 3.

Lemma 5.7 *Let the symbols have the same meaning as in Lemma 5.2. Then*

$$\bar{\mathbf{S}} = (\bar{\mathbf{S}} \cap \mathbf{U}^-) + (\bar{\mathbf{S}} \cap \mathbf{U}^+) \tag{46}$$

Proof. That

$$\bar{\mathbf{S}} \supset (\bar{\mathbf{S}} \cap \mathbf{U}^-) + (\bar{\mathbf{S}} \cap \mathbf{U}^+)$$

is obvious since both terms in the right-hand side are subspaces of $\bar{\mathbf{S}}$. We shall show that the opposite inclusion also holds.

We shall first show that $\mathbf{Y}^+ \subset (\bar{\mathbf{S}} \cap \mathbf{U}^-) + (\bar{\mathbf{S}} \cap \mathbf{U}^+)$. In effect, decomposing $y(t)$ for $t \geq 0$ as in (43), i.e., $y(t) = (H_W u)(t) + (W^+ u)(t)$, from Lemma 5.3 above we have $(H_W u)(t) \in \mathbf{X} \subset \bar{\mathbf{S}}$, so that for $t \geq 0$ necessarily $(W^+ u)(t) = y(t) - (H_W u)(t) \in \bar{\mathbf{S}}$ as well. In fact $(H_W u)(t) \in \bar{\mathbf{S}} \cap \mathbf{U}^-$ and $(W^+ u)(t) \in \bar{\mathbf{S}} \cap \mathbf{U}^+$, given the explicit dependence on the past and future of u. Taking finite linear combinations of the form $\Sigma a_k' y(t_k)$, $a_k \in \mathbb{R}^m$, $t_k \geq 0$, and then closing in the Hilbert space norm of second-order random variables gives immediately the inclusion we want. Second, by projecting obliquely $x^+ \in \mathbf{X}^+$ onto the direct sum (40), we obtain

$$x^+ = E_{\|\mathbf{U}^+}[x^+ | \mathbf{U}^-] + E_{\|}\mathbf{U}^-[x^+ | \mathbf{U}^+].$$

The first term belongs to $\mathbf{X} = (\bar{\mathbf{S}} \cap \mathbf{U}^-)$ in view of the splitting property (42), so since $x^+ \in \bar{\mathbf{S}}$ by definition, the second term in the sum must belong to the same subspace. Evidently, then $E_{\|}\mathbf{U}^-[x^+ | \mathbf{U}^+] \in (\bar{\mathbf{S}} \cap \mathbf{U}^+)$. Hence $\mathbf{X}^+$ satisfies the same subspace inclusion as $\mathbf{Y}^+$. This concludes the proof. □

This intersection representation extends the formula $\bar{\mathbf{S}} = (\bar{\mathbf{S}} \cap \mathbf{S}) \oplus (\bar{\mathbf{S}} \cap \mathbf{S}^\perp)$, known for "orthogonal" splitting subspaces [36,37].

The following argument shows how state space realizations can be constructed by a procedure based on the geometry of oblique splitting subspaces.

Denote by $\mathbf{U}_t$ the p-dimensional subspace of $\mathbf{U}_t^+$ spanned by the components of $u(t)$. By Assumption 5.1

$$\mathbf{U}_{t+1}^- = \mathbf{U}_t^- + \mathbf{U}_t$$

and since $\bar{\mathbf{S}}_{t+1} \subset \bar{\mathbf{S}}_t$, we can then write

$$\bar{\mathbf{S}}_{t+1} \cap \mathbf{U}^-_{t+1} \subset (\bar{\mathbf{S}}_t \cap \mathbf{U}^-_t) + (\bar{\mathbf{S}}_t \cap \mathbf{U}_t) \tag{47}$$

Now pick a basis vector $x(t)$, say of dimension* n in $\mathbf{X}_t$ and let $x(t+1)$ be the corresponding vector shifted by one unit of time. The n scalar components of $x(t+1)$ span $\bar{\mathbf{S}}_{t+1} \cap \mathbf{U}^-_{t+1}$ so, by projecting $x(t+1)$ onto the two components of the direct sum decomposition (47) we obtain a unique representation of the type

$$x(t+1) = Ax(t) + Bu(t)$$

Similarly, since $y(t) \in \cap \mathbf{U}^-_{t+1}$, we have

$$y(t) \in \bar{\mathbf{S}}_t \cap \mathbf{U}^-_{t+1} = (\bar{\mathbf{S}}_t \cap \mathbf{U}^-_t) + (\bar{\mathbf{S}}_t \cap \mathbf{U}_t)$$

and by projecting $y(t)$ onto the two components of the direct sum above we immediately obtain the state-output equation

$$y(t) = Cx(t) + Du(t)$$

This leads to the following theorem.

Theorem 5.8 *Assume the joint spectral density of y and u is rational and that the input process satisfies Assumption 5.1. Then the oblique predictor subspace $\mathbf{X}^{+/-}$ is finite dimensional. To any choice of basis vector $x(t)$ in a finite-dimensional oblique splitting subspace $\mathbf{X}_t$, there correspond unique matrices (A, B, C, D) such that*

$$x(t+1) = Ax(t) + Bu(t) \tag{48a}$$
$$y(t) = Cx(t) + Du(t) \tag{48b}$$

and the realization (48) is causal, i.e. $|\lambda(A)| < 1$.

Conversely, the state space of any other causal realization of y is an oblique splitting subspace.

Proof. We shall take for granted the statement that rationality implies finite dimensionality of $\mathbf{X}^{+/-}$.

Now, it is evident that the state process $x(t)$ of the representation obtained by the geometric argument illustrated before the theorem statement is stationary by construction. Hence, since $x(t)$ is a functional

*Here for the sake of illustration we assume that $\mathbf{X}_t$ is finite dimensional.

of the past history $\mathbf{U}_t$, there must be an $n \times p$ matrix function $F(z)$ analytic in $\{|z|>1\}$, with rows in the space $L_p^2[\Phi_u \,\mathrm{d}\omega/2\pi]$ of functions square integrable on the unit circle with respect to the matrix measure $\Phi_u(\mathrm{e}^{\mathrm{j}\omega})\,\mathrm{d}\omega/2\pi$, such that

$$x(t) = \int_{-\pi}^{+\pi} \mathrm{e}^{\mathrm{j}\omega t} F(\mathrm{e}^{\mathrm{j}\omega})\,\mathrm{d}\hat{u}$$

where $\hat{u}$ denotes the Fourier transform (random orthogonal measure) of the process u [53]. By substituting this into the state equation for x derived above, we see, by uniqueness of the spectral representation, that $F(z) = (zI - A)^{-1}B$. Note that $F(z)$ is rational and actually analytic also on the unit circle, since poles of modulus 1 would prevent integrability of the spectrum of x, $F(z)\Phi_u(z)F(1/z)'$, on the unit circle. [Recall that $\Phi_u(z)$ has no zeros on the unti circle, so there cannot be cancellations with the zeros of $\Phi_u(z)$]. One easily deduces from the analyticity of $F(z)$ that the eigenvalues of the reachable subsystem of (A, B) must lie inside the unit disk. On the other hand there cannot be eigenvalues of A with $|\lambda(A)| \geq 1$, since eigenvalues with absolute value larger than one would contradict stationarity and eigenvalues on the unit circle (necessarily unreachable) would imply that x has a purely deterministic component. This is impossible since $\mathbf{X}_t^- \subset \mathbf{U}_t^-$ and u is purely nondeterministic by Assumption 5.1.

The proof of the last statement is a simple verification and will be omitted. □

VI. FINITE-INTERVAL REALIZATION WITH INPUTS

The analysis in the previous Sections III, IV, and V is based on the idealized assumption that we have access to a doubly infinite sequence of data. In reality we will have only a *finite* string of observed data (9) where, however, N may be quite large. More specifically, we shall assume that N is sufficiently large that replacing the ergodic limits (11) by the finite time averages of $N+1$ elements yields good approximations of the true covariances

$$\{\Lambda(0), \Lambda(1), \Lambda(2) \ldots, \Lambda(T)\} \tag{49}$$

for some lag $T \ll N$. This is equivalent to saying that N is sufficiently large for the time averages (10) to be essentially the same as the inner products

$$E\{\begin{bmatrix} u(\tau) \\ y(\tau) \end{bmatrix}\begin{bmatrix} u(0) \\ y(0) \end{bmatrix}'\} = \Lambda(\tau)$$

for $\tau = 0, \ldots, T$. Hence we may in our analysis proceed as if we had two finite sequences of random vectors

$$\{u(0), u(1), u(2), \ldots, u(T)\}, \qquad \{y(0), y(1), y(2), \ldots, y(T)\} \tag{50}$$

where each $u(t)$ and $y(t)$ is for our purposes the same as a semi-infinite string (12) of data. It is of course to be understood that the inner products are replaced by averages of finite sums when it comes to practical implementation. This approximation is unavoidable and its effects will not be discussed further here.

Now, the state space construction of Sections IV and V was done in a stationary setting, where the state space model has to represent the output process on an *infinite* time horizon. In a situation where only the finite segment of data (50) is available it is necessary to understand the relation between data and models which realize them on a *finite* interval of time.

A finite-interval realization describes the process on a finite interval $[0, T]$ without bringing in the history of the process outside of $[0, T]$. Unfortunately, even if the process is stationary, a finite-interval realization turns out to be in general a time-varying (nonstationary) system. The notion of Markovian splitting subspaces applies without difficulty to finite-interval realizations. For example, it is easy to see that the finite-interval predictor spaces for the (stochastic component of the) process y

$$\hat{\mathbf{X}}_t^{+/-} := E^{\mathbf{Y}_{[0,t]}}\mathbf{Y}_{[t,T]}, \qquad \hat{\mathbf{X}}_t^{-/+} := E^{\mathbf{Y}_{[t,T]}}\mathbf{Y}_{[0,t]} \tag{51}$$

are minimal Markovian splitting subspaces for the process y on the finite interval $[0, T]$. These state spaces lead essentially to transient Kalman filter (innovation) representations of the process. These state models are initialized at $x(0) = 0$ and to $x(T) = 0$, respectively.

Even if the ultimate goal of modeling and identification is the construction of a *stationary model* describing the (stationary) data which are being observed, it is important to view model building in practice (i.e., when only a finite segment of data is available) as the construction of finite-interval realizations, since this viewpoint only gives the correct way of dealing with the (unknown) initial conditions of a stationary model. Using state space models here facilitates things greatly as the transient Kalman filter realizations are known to have exactly the same constant (A, C) and $\bar{C}$ parameters as the stationary model. This point lies at the foundations of the successful "subspace" approach of Van Overschee and De Moor [58] to time series identification.

Now while subspace methods identification of "purely stochastic" systems (i.e., of signals or time series) seem to be reasonably well understood, see, e.g., the book [8], the influential paper [58], and the subsequent discussions in [39], for systems which are driven by "inputs" or, *exogenous* variables, the picture still looks a bit unsatisfactory. Of the various algorithms given in the literature [45,62,59] some require a rather complicated analysis to motivate [59]. The assumptions on the input signal are different (sometimes assumed to be a white noise process, a "deterministic" known signal, etc.) and in particular the finite-interval modeling issue seems to have been largely overlooked.

Regarding this last point, it has been shown in [47] that when data are finite the two (stochastic and deterministic) realization problems cannot be decoupled and solved separately as could be done for the stationary infinite-interval situation. We shall try to explain below that this is an important point to keep in mind, especially for what concerns the order estimation step of the identification algorithm (this may perhaps sound surprising to practitioners used to neglecting initial conditions and only "thinking asymptotically").

The finite history subspaces of **H** generated by the finite stochastic data (50) will be denoted

$$\mathbf{U}_{[0,T]} := \text{span}\{u(t) \,|\, 0 \le t \le T\}$$

$$\mathbf{Y}_{[0,T]} := \text{span}\{y(t) \,|\, 0 \le t \le T\}.$$

The orthogonal complement of $\mathbf{U}_{[0,T]}$ in $\mathbf{U}_{[0,T]} \vee \mathbf{Y}_{[0,T]}$ will be denoted by $\mathbf{U}^{\perp}_{[0,T]}$, so that $\mathbf{U}_{[0,T]} \oplus \mathbf{U}^{\perp}_{[0,T]} = \mathbf{U}_{[0,T]} \vee \mathbf{Y}_{[0,T]}$. The practical computation of $\mathbf{U}^{\perp}_{[0,T]}$ can be done by an LQ factorization of the data matrix generating $\mathbf{U}_{[0,T]} \vee \mathbf{Y}_{[0,T]}$.

The following lemma is straightforward. It is taken from [47].

Lemma 6.1 *Let y_u be described by the deterministic realization* (25). *Then*

$$E[y(t) \,|\, \mathbf{U}_{[0,T]}] = E[y_u(t) \,|\, \mathbf{U}_{[0,T]}] := \hat{y}_u(t)$$

where $\hat{y}_u(t)$ is described by the same state space model but started at a different initial state, namely

$$\hat{x}_u(t+1) = A_u \hat{x}_u(t) + B_u u(t) \tag{52a}$$

$$\hat{y}_u(t) = C_u \hat{x}_u(t) + D_u u(t) \tag{52b}$$

$$\hat{x}_u(0) = E[x_u(0) \,|\, \mathbf{U}_{[0,T]}] \tag{52c}$$

It follows from this lemma that the deterministic part of (26), namely the system matrices (A_u, B_u, C_u, D_u) (in a suitable basis) can be identified by using any "deterministic" identification procedure, based on the data

$\{\hat{y}_u(t), u(t) \mid t = 1, 2, \ldots, T\}$. Once the system matrices are computed the estimate of the initial state $\hat{x}_u(0)$ can also be reconstructed.

The identification of the stochastic subsystem can be based on the projections of the output data onto the complementary subspace $\mathbf{U}_{[0,T]}^{\perp}$. To this end, we introduce the random vectors

$$\hat{y}_s(t) := y(t) - E[y(t) \mid \mathbf{U}_{[0,T]}], \quad 0 \leq t \leq T$$

These can be computed from the available data and actually we have

$$\mathbf{U}_{[0,T]}^{\perp} := \operatorname{span}\{\hat{y}_s(t) \mid 0 \leq t \leq T\}$$

The following proposition, taken from [47], shows how the finite-time estimate $\hat{y}_s(t)$ relates to the stochastic component y_s.

Proposition 6.2 *Let $\tilde{y}_u(t) := y_u(t) - \hat{y}_u(t)$ (the "smoothing error" of $y_u(t)$). Then*

$$\hat{y}_s(t) = y_s(t) + \tilde{y}_u(t), \quad 0 \leq t \leq T \tag{53}$$

Hence for finite data length, the projection $\hat{y}_s(t)$ of the output on the complementary subspace $\mathbf{U}_{[0,T]}^{\perp}$ does not coincide with the stochastic component $y_s(t)$, as would have happened for an *infinite* data length. The "ideal" projection $y_s(t)$ is affected by an additional "smoothing error" term $\tilde{y}_u(t)$ which depends on the error on the estimate of the initial state of the deterministic component, $\tilde{x}_u(0) := x_u(0) - \hat{x}_u(0)$. In fact,

$$\tilde{y}_u(t) = C_u A_u^t \tilde{x}_u(0), \quad 0 \leq t \leq T$$

This additional term is a source of difficulty in identification of the stochastic part since, if its different dynamics is not properly subtracted off, it tends to produce a stochastic model of y_s of a much higher dimension than the true order $\tilde{n}$. In fact, the estimated model will tend to include also the dynamics of the deterministic subsystem. Therefore a preliminary step is necessary for the identification of the stochastic component, i.e., to filter out $\tilde{y}_u(t)$ somehow.

Note that (53) has the explicit form

$$\begin{bmatrix} \hat{y}_s(0) \\ \hat{y}_s(1) \\ \cdot \\ \cdot \\ \cdot \\ \hat{y}_s(T) \end{bmatrix} = \begin{bmatrix} C_u \\ C_u A_u \\ \cdot \\ \cdot \\ \cdot \\ C_u A_u^T \end{bmatrix} \tilde{x}_u(0) + \begin{bmatrix} y_s(0) \\ y_s(1) \\ \cdot \\ \cdot \\ \cdot \\ y_s(T) \end{bmatrix} \tag{54}$$

which can be rewritten more compactly in vector form as

$$\hat{y}_s = \Omega_u \tilde{x}_u(0) + y_s$$

having denoted the (deterministic) observability matrix by Ω_u. Note that the two terms on the right-hand side are uncorrelated. Once the parameters (A_u, C_u) of the deterministic subsystem are identified, the matrix Ω_u can be assumed to be known.

It seems that one may easily filter out $\tilde{y}_u$ from $\hat{y}_s$ by just premultiplying the vector equation above by $\Omega_u \Omega_u^{\perp}$ where $\Omega_u^{\perp}$ is a matrix with rows spanning the left nullspace of Ω_u. With a procedure of this kind however a distortion is introduced on the time series $\{y_s(t),\ t = 0, \ldots, T\}$ which seems to be very hard to remove. To keep control on the reconstruction errors, one needs to filter out the unwanted term $\tilde{y}_u$ by a *sequential* algorithm. Below we shall describe an algorithm [48] which in principle only distorts a small finite initial segment of the time series $\{y_s(t),\ t = 0, \ldots, T\}$.

Algorithm 6.3 *Assume A_u and A_s have no common eigenvalues. The following algorithm recovers the time series $\{y_s(t),\ t = 0, \ldots, T\}$ asymptotically.*

1. *Compute a left-coprime factorization of the rational matrix $C_u(I - z^{-1}A_u)^{-1}$. Let the $m \times m$ and $m \times n$ polynomial matrices in the unit backward shift z^{-1}, $D(z^{-1}) = \Sigma_0^{\nu} D_k z^{-k}$, and $N(z^{-1}) = \Sigma_0^{\nu-1} N_k z^{-k}$ be such a left-coprime factorization, i.e., let*
$$D(z^{-1})^{-1} N(z^{-1}) = C_u(I - z^{-1}A_u)^{-1} \tag{55}$$
2. *Compute*
$$\check{y}(t) := D(z^{-1}) \hat{y}_s(t) \tag{56}$$
so that $\check{y}(t) = N(z^{-1})\bar{x}_u(0) + D(z^{-1})y_s(t)$. Since the first term has finite support $t = 0, \ldots, \nu - 1$,
$$\check{y}(t) = D(z^{-1}) y_s(t), \quad t \geq \nu$$
3. *Solve the following vector difference equation in the unknown variable $z(t)$*
$$D(z^{-1}) z(t) = \check{y}(t), \quad t \geq \nu \tag{57}$$
started with initial conditions $z(0) = z(1) = \ldots\ z(\nu - 1) = 0$. Then $z(t) = y_s(t)$, for t large enough.

Proof. In symbolic notation we have
$$\tilde{y}_u(t) = C_u(I - z^{-1}A_u)^{-1} \bar{x}_u(0)$$
since $\tilde{y}_u(t)$ is the zero-input response of the system (52) to the initial state

$\tilde{x}_u(0)$. Hence, by applying the difference operator $D(z^{-1})$ to the left-hand side we see that

$$D(z^{-1})\tilde{y}_u(t) = \sum_{0}^{\nu-1} N_k z^{-k}\tilde{x}_u(0),$$

in particular, $D(z^{-1})\tilde{y}_u(t) = 0$ for $t \geq \nu$. Then the difference equation (57) reduces to $D(z^{-1})z(t) = D(z^{-1})y_s(t)$, $t \geq \nu$, subject to zero initial conditions $z(0) = z(1) = \ldots = z(\nu - 1) = 0$. Now since the linear system represented (of course nonminimally) by the left MFD $D(z^{-1})^{-1}D(z^{-1})$ has all transmission zeros different from the poles of the spectrum of $y_s(t)$, it is obvious that this equation has the steady-state solution $z(t) = y_s(t)$. □

The left-coprime MFD (65) can be computed by reducing the polynomial matrix $[(I - z^{-1}A_u)'C_u']'$ to upper triangular form by left-multiplication by a unimodular polynomial matrix, namely

$$\begin{bmatrix} U_1(z^{-1}) & V_1(z^{-1}) \\ U_2(z^{-1}) & -V_2(z^{-1}) \end{bmatrix} \begin{bmatrix} (I - z^{-1}A_u) \\ C_u \end{bmatrix} = \begin{bmatrix} R(z^{-1}) \\ 0 \end{bmatrix}$$

where $R(z^{-1})$ is upper triangular. From this one gets

$$V_2(z^{-1})^{-1}U_2(z^{-1}) = C_u(I - z^{-1}A_u)^{-1}$$

VII. POSITIVITY

A warning is in order concerning the implementation of the "subspace" identification methods in that some nontrivial mathematical questions related to positivity of the estimated spectrum have been completely overlooked in the discussion. This issue is thoroughly discussed in [39] and here we shall just present a short summary. The problem occurs only in identification of the stochastic component, which we shall here name y for short.

It is shown in [39,40] that subspace identification is equivalent to the following three-step procedure which is essentially the same as that discussed in [8].

1. The first step is estimation of a finite sequence of covariance matrices

   $$\{\Lambda_0, \Lambda_1, \ldots, \Lambda_T\} \tag{58}$$

   from the observed data.
2. The second step is identification of a rational model for the covariance sequence (58). This is a minimal partial realization (also called "rational extension") problem. Given a finite set of "experimental"

covariance data one is asked to find a minimal value of n and a minimal* triplet of matrices $(A, C, \bar{C})$, of dimensions $n \times n$, $m \times n$, and $m \times n$, respectively, such that

$$\Lambda_k = CA^{k-1}\bar{C}', \quad k = 1, \ldots, T \tag{59}$$

The solution of the partial realization problem leads to "estimates" of the parameters $(A, C, \bar{C})$ of a minimal realization of a rational spectral density matrix of the process.

3. The third step is to compute a stationary state space model (typically the forward innovation model) by solving the linear matrix inequality (6), or an appropriate equivalent Riccati equation, relative to the rational estimated spectrum computed in step 2.

The estimation of $(A, C, \bar{C})$ is done by simply *matching second-order moments*, i.e., by solving the equations (59).

This is an instance of *estimation by the method of moments* described in statistical textbooks, e.g., [11] p. 497, which is a very old idea used extensively by Pearson in the beginning of the century. The underlying estimation principle is that the parameter estimates should match exactly the sample second-order moments and is close in spirit to the wide-sense setting that one is normally working in. It does not involve optimality or minimal distance criteria between the "true" and the model distributions.

Once a minimal triplet $(A, C, \bar{C})$ interpolating the partial sequence (58) has been found, so that $CA^{k-1}\bar{C}' = \Lambda_k$, $k = 1, 2, \ldots, T$, we also completely determine the infinite sequence

$$\{\Lambda_0, \Lambda_1, \Lambda_2, \Lambda_3, \ldots\} \tag{60}$$

by setting $\Lambda_k = CA^{k-1}\bar{C}'$ for $k = T+1, T+2, \ldots$. This sequence is called a *minimal rational extension* of the finite sequence (58). The attribute "rational" is due to the fact that

$$Z(z) := \tfrac{1}{2}\Lambda_0 + \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \cdots = \tfrac{1}{2}\Lambda_0 + C(zI - A)^{-1}\bar{C}' \tag{61}$$

is a rational function. In order for (60) to be a bona fide covariance sequence, however, it is necessary, but *not* sufficient, that the Toeplitz

*Recall that $(A, C, \bar{C})$ is minimal if (A, C) is completely observable and $(A, \bar{C}')$ is completely reachable.

matrix

$$T = \begin{bmatrix} \Lambda_0 & \Lambda_1 & \Lambda_2 & \dots & \Lambda_\nu \\ \Lambda_1' & \Lambda_0 & \Lambda_1 & \dots & \Lambda_{\nu-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_\nu' & \Lambda_{\nu-1}' & \Lambda_{\nu-2} & \dots & \Lambda_0 \end{bmatrix} \tag{62}$$

be nonnegative definite. In fact, it is required that the function (the spectral density corresponding to (60))

$$\Phi(z) = \Lambda_0 + \sum_{k=1}^{\infty} \Lambda_k(z^k + z^{-k}) = Z(z) + Z(z^{-1})' \tag{63}$$

be nonnegative on the unit circle. This is equivalent to the function $Z(z)$ being *positive real*. Consequently, the partial realization needs to be done subject to the extra constraint of positivity.

The constraint of positivity is a rather tricky one and in all identification methods which are directly or indirectly, as in the subspace methods described in the literature, based on the interpolation conditon (59) it is normally disregarded. For this reason these methods may fail to provide a positive extension and hence may lead to data $(A, C, \bar{C})$ for which there are no solutions of the LMI and hence to totally inconsistent results.

It is important to appreciate the fact that the problem of positivity of the extension has little to do with the "noise" or "sample variability" superimposed on the covariance data and is present equally well for (finite) data extracted from a true rational covariance sequence. For there is no guarantee, even in this idealized situation, that the order of a minimal rational extension (60) of the first T covariance matrices of the sequence would be sufficiently high to equal the order of the infinite sequence and hence to generate a positive extension. A minimal partial realization may well fail to be positive because its order is too low to guarantee positivity.

Neglecting the positivity constraint amounts to tacitly assuming that

Assumption 7.1 The covariance data (58) can be generated exactly by some (unknown) stochastic system whose dimension is equal to the rank of the block Hankel matrix

$$H_\mu = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \Lambda_3 & \dots & \Lambda_\mu \\ \Lambda_2 & \Lambda_3 & \Lambda_4 & \dots & \Lambda_{\mu+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_\mu & \Lambda_{\mu+1} & \Lambda_{\mu+2} & \dots & \Lambda_{2\mu-1} \end{bmatrix}, \tag{64}$$

where $\mu = [T/2]$.

This assumption is not "generically satisfied" and it can be shown [39] that there are relatively "large" sets of data (58) for which it does not hold. It is not even enough to assume that the data are generated from a "true" finite-dimensional stochastic system: the rank condition is also necessary. Otherwise, for a minimal triplet $(A, C, \bar{C})$ which satisfies the interpolation condition (59), the positivity condition will not be automatically fulfilled, and the matrix A may even fail to be stable.

A. Statistical Properties of Subspace Methods

There are also important questions concerning the statistical significance (what are the uncertainty bounds on the parameters and on the estimated transfer functions, etc.) of the subspace/realization approach which have not been touched on in the previous sections. Some do not seem to have been completely answered in the literature as yet.

Havenner and Aoki [24,9] have given an instrumental variable interpretation of the formulas for the partial realization estimates of $A, C, \bar{C}$ described in [8]. In particular, it is pointed out in [9] that using the state variable as an instrument gives the most efficient estimates. It is not clear however when these particular estimates are asymptotically of minimal variance (statistically efficient). The statistical properties of the estimates of the B and D matrices are also difficult to analyze. The paper [12] also relates to the statistical significance of subspace methods but discusses a particular class of estimation procedures which appear to be different from the method of [58]. In any case the question of asymptotic relative efficiency of subspace methods compared with ML (or PEM) methods (a "central statistical question" according to [12]) is left open in the above references.

Some argue that estimation by the method of moments is in general "nonefficient" and it is generally claimed in the literature that one should expect better results (in the sense of smaller asymptotic variance of the estimates) by optimization methods. Actually if the covariance estimates are maximum likelihood estimates and the partial realization problem has a unique solution (modulo similarity), then, choosing $(A, C, \bar{C})$ in a suitable canonical form, there is a map

$$\{\Lambda_0, \Lambda_1, \ldots, \Lambda_T\} \rightarrow (A, C, \bar{C})$$

which is locally smooth and one-to-one. It follows by a well-known theorem of Zehna [68] that the canonical parameter estimates are also maximum likelihood estimates. Hence in this case we have efficient estimates.

This argument is elaborated in [51] where it is pointed out that the subspace estimates of $A, C, \bar{C}$ (and also those of B, D) can in principle be as efficient as ML. In fact, they are ML estimates if so are the sample covariance estimates which are implicitly associated to the subspace identification method.

The model classes described above are wide-sense. In case the signal y is believed to be *Gaussian* they can equivalently be interpreted as defining the spectrum or the covariance function of a family of Gaussian probability laws for the underlying stochastic process. These probability laws are uniquely determined by a corresponding model and are then also parametrized by the parameters $(A, C, \bar{C})$.

REFERENCES

1. H Akaike. IEEE Trans Automat Contr AC-19:667–674, 1974.
2. H Akaike. SIAM J Control 13:162–173, 1975.
3. H Akaike. In: R Mehra and D Lainiotis, eds. System Identification: Advances and Case Studies. New York: Academic Press, 1976, pp 27–96.
4. BDO Anderson. SIAM J Control 5:171–182, 1967.
5. BDO Anderson. IEEE Trans Autom Control AC-12:410–414, 1967.
6. BDO Anderson. J Stat Phys 1:133–147, 1969.
7. BDO Anderson, MR Gevers. Automatica 18:195–213, 1982.
8. M Aoki. State Space Modeling of Time Series. 2nd ed. Springer, 1990.
9. M Aoki. Europ J Operat Res 73:265–278, 1994.
10. PE Caines, CW Chan. In: R Mehra and D Lainiotis, eds. System Identification: Advances and Case Studies. New York: Academic Press, 1976, pp 349–405.
11. H Cramer. Mathematical Methods of Statistics. Princeton, 1949.
12. M Deistler, K Peternell, W Scherrer. Automatica 31:1865–1875, 1995.
13. UB Desai, D Pal, RD Kirkpatrick. Int J Control 42:821–838, 1985.
14. P Faurre. Identification par minimisation d'une representation Markovienne de processus aleatoires, Symposium on Optimization, Nice 1969, Springer Lect. Notes in Math., 132.
15. P Faurre. Realisations Markovienne de processus stationnaires, Report de Recherche n. 13, INRIA, 1973.
16. P Faurre, P Chataigner. Identification en temp reel et en temp differee par factorisation de matrices de Hankel French-Swedish colloquium on process control, IRIA Roquencourt, 1971.
17. P Faurre, JP Marmorat. CR Acad Sc Paris, A 268:978–981, 1969.
18. P Faurre, M Clerget, F Germain. Opérateurs Rationnels Positifs. Dunod, 1979.
19. PA Fuhrmann. Linear Operators and Systems in Hilbert Space. McGraw Hill, 1981.
20. MR Gevers, BDO Anderson. Intern J Control 33:777–809, 1981.

21. MR Gevers, BDO Anderson. IEEE Trans Autom Control AC-27:431–436, 1982.
22. GH Golub, CR Van Loan. Matrix Computations. 2nd ed. The Johns Hopkins Univ. Press, 1989.
23. CWJ Granger. Informat Control 6:28–48, 1963.
24. A Havenner, M Aoki. J Econ Dynam Control 12:333–336, 1988.
25. RE Kalman. On a new Criterion of Linear Passive Systems. Proc of the First Allerton Conference Univ of Illinois, pp 456–470, Nov 1963.
26. RE Kalman. Proc Natnl Acad of Sciences (USA) 49:201–205, 1963.
27. RE Kalman. Realization of covariance sequences. Proc Toeplitz Memorial Conference, Tel Aviv, Israel, 1981.
28. RE Kalman, PL Falb, MA Arbib. Topics in Mathematical Systems Theory. McGraw-Hill, 1969.
29. A Kintchine. Math Annalen 109:604–615, 1934.
30. AN Kolmogorov. Bull Math Moscow State University 2:40, 1941.
31. WE Larimore, System identification, reduced-order filtering and modeling via canonical variate analysis. Proc American Control Conference, 1983, pp 445–451.
32. WE Larimore. Canonical variate analysis in identification, filtering, and adaptive control. Proc 29th IEEE Conference on Decision and Control, 1990, pp 596–604.
33. PD Lax, RS Phillips. Scattering Theory. New York: Academic Press, 1967.
34. A Lindquist, M Pavon. IEEE Tr Autom Contr AC-29:418–432, 1984.
35. A Lindquist, G Picci. SIAM J Contr Optimiz 17:365–389, 1979.
36. A Lindquist, G Picci. SIAM J Contr Optimiz 23:809–857, 1985.
37. A Lindquist, G Picci. J Math Syst Estim Cont 1:241–333, 1991.
38. A Lindquist, G Picci. On "subspace-methods" identification and stochastic model reduction. Proc 10th IFAC Symposium on System Identification, Copenhagen, DK, 2, 1994, pp 397–403.
39. A Lindquist, G Picci. Automatica 32:709–733, 1996.
40. A Lindquist, G Picci. In: Identification, Adaptation, Learning (Lectures given at the NATO-ASI School, From Identification to Learning held in Como, Italy, 1994). Springer Verlag, 1996.
41. A Lindquist, G Michaletzky, G Picci. SIAM J Control Optimiz 33:365–401, 1995.
42. A Lindquist, G Picci, G Ruckebusch. Math System Theory 12:271–279, 1979.
43. L Ljung. System Identification—Theory for the User. Prentice-Hall, 1987.
44. M Moonen, B De Moor, L Vanderberghe, J Vandewalle. Int J Control 49:219–232, 1989.
45. M Moonen, J Vandewalle. Int J Control 51:1133–1146, 1990.
46. G Picci. Proc of the IEEE 64:112–122, 1976.
47. G Picci, T Katayama. Signal Processing 52:145–160, 1996.
48. G Picci, T Katayama. A simple "subspace" identification algorithm with exogenous inputs, Proceedings of the 1996 triennial IFAC Congress, San

Francisco, Ca., paper n.0916, session 3a-06-5.
49. G Picci, S Pinzoni. IMA J Math Control Inform 3:185–210, 1986.
50. G Picci. Application of Stochastic Realization to System Identification (submitted for publication).
51. G Picci. Statistical properties of certain subspace identification methods. Proc IFAC Symposium SYSID 97, Fukuoka, Japan (to appear).
52. VM Popov. Revue Rumaine des Sciences Techniques, Ser Elechtrotechnique et Energetique 9:629–690, 1964.
53. Y Rozanov. Stationary Random Processes. Holden-Day, 1963.
54. G Ruckebusch. CR Acad Sc Paris Series A 282:649–651, 1976.
55. G Ruckebusch. A state space approach to the stochastic realization problem. Proc 1978 IEEE Intern Symp Circuits and Systems, 1978, pp. 972–977.
56. G Ruckebusch. Factorisations minimales de densités spectrales et répresentations markoviennes. Proc 1re Colloque AFCET-SMF, Palaiseau, France, 1978.
57. T Söderström, P Stoica. System Identification. Prentice-Hall, 1989.
58. P Van Overschee, B De Moor. Automatica 29:649–660, 1993.
59. P Van Overschee, B De Moor. Automatica 30:75–93, 1994.
60. P Van Overschee, B De Moor. A unifying theorem for subspace system identification algorithms and its interpretation. Proc 10th IFAC Symposium on System Identification 2, 1994, pp 145–156.
61. JH van Schuppen. Acta Applicandae Methematicae 35:193–212, 1994.
62. M Verhaegen. Automatica 30:61–74, 1994.
63. N Wiener. Acta Mathematica 55:117–258, 1930.
64. N Wiener. The Fourier Integral and Certain of its Applications. Cambridge, 1933.
65. N Wiener. Extrapolation, Interpolation and Smoothing of Stationary Time Series. Cambridge, Mass: The M.I.T. Press, 1949.
66. JC Willems. IEEE Trans Autom Control AC-16:621–634, 1971.
67. VA Yakubovich. Avtomatika i Telemekhanika 24:293–303, 717–731, 1963.
68. PW Zehna. Annals of Math Statist 37:744, 1966.
69. HP Zeiger, AJ McEwen. IEEE Trans Autom Control AC-19:153, 1974.

This Page Intentionally Left Blank

2
General State Space Modeling

Will Gersch

University of Hawaii, Honolulu, Hawaii

Genshiro Kitagawa

The Institute of Statistical Mathematics, Tokyo, Japan

I. INTRODUCTION

General state space modeling, initiated in Kitagawa [1], is a Bayesian modeling of not-necessarily linear not-necessarily Gaussian time series. It is an extension of our earlier work on "smoothness priors", a Bayesian linear Gaussian modeling of time series. That work in smoothness priors (reviewed in [2]) was motivated by Akaike [3], a penalized likelihood constrained least squares computational approach. Kalman filter-type state space approaches to that modeling were introduced in [4] and [5]. The term "smoothness priors", adopted from Shiller [6], was used in [7–9]. Subsequently we attempted to redo conventional time series analysis from the linear Gaussian smoothness priors approach as well as extend that approach to the analysis of previously unaddressed time series modeling problems including the modeling of multivariate nonstationary covariance time series.

Reference [1], a recursive computational algorithm for the realization of an algorithm for not-necessarily linear not-necessarily Gaussian state space modeling required numerical approximations of the densities or Gaussian-sum approximations. Work on general state space methods continued with emphasis on the development of computational methods and on new applications. The initial computations of conditional distributions by numerical integration methods was adequate for systems with a small number of states. For the analysis of seasonal time series with as many as 13 states, Kitagawa [10] used a Gaussian sum approximation to state estimation. A two-filter formula for Gaussian sum smoothing in

Kitagawa [11], completed the Gaussian sum approximation modeling. A more recent development (Kitagawa [12,13]) is a Monte Carlo approach to state estimation. This chapter attempts to coherently integrate and communicate our work and present the original smoothness priors state space and more recent general state space approach to time series analysis. It is important to note that the AIC, ([14,15]), Akaike's information theoretic criterion for evaluating the parametric models fitted to data has a crucial role in our approach to time series modeling.

Following this introduction, in Section II, background material including Akaike's AIC, smoothness priors and linear-Gaussian smoothness priors state space modeling and some of its applications are briefly reviewed. In the applications we present trend estimation, seasonal time series estimation, and the estimation of a time-varying autoregressive model for the estimation of gradually changing spectra. The general state space model is treated in Section III. Included are brief treatments of the numerical approximation, Gaussian sum approximation and Monte Carlo methods of general state space filter/smoother implementations. Finally, the emphasis of the chapter, on illustrating applications of general state space modeling is in Section IV. In the applications of mostly real data analysis, we present abrupt trend estimation, abrupt seasonal estimation, and estimation of abruptly changing variance and abruptly changing spectrum estimation. Also included are the modeling of an inhomogeneous discrete process, of a quasi-periodic process, the filtering, and the smoothing of a nonlinear process. The examples are realized by the different methods of general state space implementation. A more complete treatment of our work, including many more examples, is in Kitagawa and Gersch [16]. A list of references concludes the chapter.

II. BACKGROUND

A. The AIC

Akaike's AIC [14,15] is an information theoretic criterion for the selection of the best of alternative parametric models based on observed data. The AIC has proven to be extensively applicable in statistical data analysis and engineering modeling (see [17] and [18] for example). It is an approximately unbiased estimate of the expected log-likelihood, the essential part of the Kullback–Leibler information. The Kullback–Leibler information or the K–L number [19] is an information theoretic measure of the dissimilarity between two distributions. The larger the measure, the greater the difference between the two distributions.

The maximized log-likelihood is a biased estimator of the average expected log-likelihood. The bias is approximately equal to the number of parameters estimated in the model [15]. An approximate correction of the bias is reflected in the definition of the AIC given below.

$$\begin{aligned}\mathrm{AIC}(m) &= -2(\text{maximized log-likelihood of the fitted model}) \\ &\quad +2(\text{number of estimated parameters in the model}) \qquad (1) \\ &= -2\sum_{n=1}^{N} \log f_m(y_n | \hat{\theta}_m) + 2|\hat{\theta}_m|\end{aligned}$$

In (1), $f_m(y_n | \hat{\theta}_m)$ denotes the likelihood and $|\hat{\theta}_m|$ denotes the dimension of the vector $\hat{\theta}_m$. A derivation of the AIC is in Chapter 2 of [16].

B. Smoothness Priors

"Smoothness priors" originated as a normal distribution theory-linear model-stochastic regression treatment of stationary and nonstationary time series [6,3]. A conceptual predecessor of smoothness priors can be seen in a smoothing problem posed in Whittaker [20]. In [2], a review of some of our smoothness priors approach to the time series analyses of what are essentially signal-in-noise estimation problems was presented. The method is Bayesian. The Bayesianness provides a framework for doing statistical inference. A prior distribution on the model parameters is expressed in the form of a stochastic difference equation and parametrized by hyperparameters which in turn have a crucial role in the analysis. The maximization of the likelihood of a small number of hyperparameters permits the robust modeling of a time series with relatively complex structure and a very large number of implicitly inferred parameters. The critical statistical ideas in smoothness priors are the likelihood of the Bayesian model and the use of likelihood as a measure of the goodness of fit of the model. Reference [1], an extension of smoothness priors to not-necessarily Gaussian not-necessarily linear model smoothness priors, initiated work in a general state space modeling methodology.

Our own work was motivated by Akaike [3]. We applied Akaike's linear regression model method to a variety of time series modeling problems not considered by Akaike and also we embedded that method into a state space model framework and applied the state space method to additional problems.

For illustrative purposes, consider the simple case of a time series with nonstationarity in the mean. That can be expressed as

$$y_n = t_n + \varepsilon_n \tag{2}$$

where for simplicity ε_n is a stationary Gaussian white noise and t_n is a mean value function (or trend component). The problem is to estimate the unknown t_n, $n = 1, \ldots, N$, given the observations y_n, $n = 1, \ldots, N$. Following the approach in [20] let the solution balance a tradeoff of goodness of fit to the data and goodness of fit to a smoothness criterion. This idea can be expressed by minimizing

$$\sum_{n=1}^{N} (y_n - t_n)^2 + \mu^2 \sum_{n=k+1}^{N} (\nabla^k t_n)^2 \tag{3}$$

for an appropriately chosen smoothness tradeoff parameter μ^2. In (3) $\nabla^k t_n$ expresses a kth-order difference constraint on the solution t_n, with $\nabla t_n = t_n - t_{n-1}$, $\nabla^2 t_n = \nabla(\nabla t_n)$, etc.

The properties of the solution to the problem in (2)–(3) are apparent. If $\mu^2 = 0$, $t_n = y_n$, and the solution is a replica of the observations. As μ^2 becomes increasingly large, the smoothness constraint dominates the solution and the solution satisfies the kth-order constraint. For large μ^2 and $k = 1$, the solution is a constant, for $k = 2$, it is a straight line, etc. Whittaker [20] left the choice of μ^2 to the investigator.

In Bayesian analysis [21], with the trend vector given by $t = (t_1, t_2, \ldots, t_n)^T$: the conditional data distribution of the observation vector, $y = (y_1, \ y_2, \ \ldots, \ y_n)^T$ is in the form $p(y|t, \tau^2, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\{-(y-t)'(y-t)/2\sigma^2\}$. Then, $\pi(t|y, \tau^2, \sigma^2)$, the posterior distribution of the trend parameters, is proportional to the product of the prior distribution of the trend, $\pi(t|\tau^2)$, and the conditional data distribution and may be seen to be

$$\pi(t|y, \tau^2, \sigma^2) \propto p(y|t, \sigma^2)\pi(t|\tau^2) \tag{4}$$

As seen below in (6), the prior distribution on the trend parameters is a consequence of the difference equation constraints. The integration of the right-hand side of (4) yields $L(\tau^2, \sigma^2)$, the likelihood for the unknown parameters τ^2 and σ^2,

$$L(\tau^2, \sigma^2) = \int_{-\infty}^{\infty} p(y|t, \sigma^2)\pi(t|\tau^2)\,\mathrm{d}t \tag{5}$$

I.J. Good [22] referred to the maximization of (5) for the unknown parameters τ^2 and σ^2 as a Type II maximum likelihood method.

Ignoring initial values, the smoothness priors trend modeling problem corresponds to the maximization of

$$\exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - t_n)^2 \right\} \exp\left\{ -\frac{1}{2\tau^2} \sum_{n=1}^{N} (\nabla^k t_n)^2 \right\} \tag{6}$$

To within constant terms, the first term in (6) is the conditional data distribution and the second term corresponds to the prior distribution of the trend.

The role of the hyperparameter τ^2 as a measure of the uncertainty in the belief of the prior distribution is clear from (6). Relatively large τ^2 implies a relatively wiggly trend component. The ratio of τ^2/σ^2 can be interpreted as a signal-to-noise ratio.

More complete treatments of this Bayesian approach to time series analysis appear in [2], [23] and Kitagawa and Gersch [16].

C. Smoothness Priors State Space Modeling

First the standard state space analysis is reviewed. That is followed by some smoothness priors state space analysis applications. The applications include nonstationarity in the mean, which itself includes modeling time series with nonstationary trends and seasonal components, and nonstationary covariance modeling.

1. Review of Standard State Space Analysis

Let a state space model be given by [24]

$$\begin{aligned} x_n &= F_n x_{n-1} + G_n w_n \\ y_n &= H_n x_n + \varepsilon_n \end{aligned} \tag{7}$$

where y_n is an l-dimensional time series, and x_n is an m-dimensional state vector. w_n and ε_n are k- and l-dimensional white noises with $w_n \sim N(0, Q_n)$ and $\varepsilon_n \sim N(0, R_n)$. F_n, G_n, and H_n are $m \times m$, $m \times k$, and $l \times m$ matrices, respectively.

2. State Estimation

The problem of state estimation is to evaluate $p(y_n | Y_m)$, the distribution of y_n given the observations $Y_m \equiv (y_1, \ldots, y_m)$, and the initial values $x_{0|0}$ and $V_{0|0}$. Various problems in time series analysis can be solved by estimating the state vector. In the case of the standard state space model, they are given by the following Kalman filter.

(a) Prediction.

$$\begin{aligned} x_{n|n-1} &= F_n x_{n-1|n-1} \\ V_{n|n-1} &= F_n V_{n-1|n-1} F_n^T + G_n Q_n G_n^T \end{aligned} \tag{8}$$

(b) Filtering.

$$\begin{aligned} K_n &= V_{n|n-1} H_n^T (H_n V_{n|n-1} H_n^T + R_n)^{-1} \\ x_{n|n} &= x_{n|n-1} + K_n (y_n - H_n x_{n|n-1}) \\ V_{n|n} &= (I - K_n H_n) V_{n|n-1} \end{aligned} \tag{9}$$

Using the outputs of the Kalman filter, the smoothed state x_n given the entire observations Y_N is given by the following fixed interval smoother.

(c) Smoothing.

$$\begin{aligned} A_n &= V_{n|n} F_{n+1}^T V_{n+1|n}^{-1} \\ x_{n|N} &= x_{n|n} + A_n(x_{n+1|N} - x_{n+1|n}) \\ V_{n|N} &= V_{n|n} + A_n(V_{n+1|N} - V_{n+1|n})A_n^T \end{aligned} \tag{10}$$

(d) Identification of the Model. The state space model and the Kalman filter yield a very efficient method for the computation of the likelihood of the time series models. The likelihood can be expressed by using the conditional distributions as follows:

$$L(\theta) = f(y_1, \ldots, y_N|\theta) = \prod_{n=1}^{N} f(y_n|Y_{n-1}) \tag{11}$$

Each individual term in the last expresssion is given by

$$f(y_n|Y_{n-1}) = \frac{1}{\sqrt{2\pi r_n}} \exp\left\{ -\frac{(y_n - H_n x_{n|n-1})^2}{2r_n} \right\} \tag{12}$$

with $r_n = H_n V_{n|n-1} H_n^T + R_n$. Therefore, the log-likelihood of the model is given by

$$l(\theta) = \log L(\theta) = -\frac{1}{2}\left\{ N \log 2\pi + \sum_{n=1}^{N} \log r_n + \sum_{n=1}^{N} \frac{(y_n - H_n x_{n|n-1})^2}{r_n} \right\} \tag{13}$$

$\hat{\theta}$, the maximum likelihood estimate of the parameter θ, is obtained by maximizing (13) with respect to θ. Then, computing the AIC as defined by (1) for each of the candidate fitted models yields a unified procedure for fitting and selecting the best of alternative time series models. Examples of this procedure of fitting alternative models and selecting as best the one with the smallest AIC are illustrated throughout the chapter.

3. Nonstationarity in the Mean

As before, a time series with nonstationarity in the mean can be expressed as

$$y_n = t_n + \varepsilon_n \tag{14}$$

with ε_n a stationary Gaussian white noise, and t_n a mean value function (or trend component) with

$$\nabla^k t_n = w_n \tag{15}$$

w_n is assumed to be a Gaussian white noise with mean zero and variance

τ^2. The state space models for difference constraint orders $k = 1$ and $k = 2$ are given by

$$\text{for } k = 1: \quad F^{(1)} = G^{(1)} = H^{(1)} = 1, \qquad x_n = t_n \tag{16}$$

$$\text{for } k = 2: \quad F^{(2)} = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \qquad G^{(2)} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$H^{(2)} = [1 \;\; 0], \qquad x_n = \begin{bmatrix} t_n \\ t_{n-1} \end{bmatrix} \tag{17}$$

(a) A Simple Trend Model Example. The state space model in (17) is used here for the estimation of the trend of nonstationary time series embedded in white Gaussian noise. We consider an $N = 200$ simulated data example in which the smooth trend has the functional form of an asymmetrically truncated Gaussian function. The problem is: Given the noisy observations $\{y_n, \; n = 1, \ldots, N\}$, estimate the unknown smooth function that is corrupted by noise. That is, specify $\hat{t}_{n|N}$, $n = 1, \ldots, N$. The "true" smooth function t_n, $n = 1, \ldots, N$, and the smooth function with the superposition of t_n and the additive noise for the fixed trend model order $k = 2$ for different values of the hyperparameter are shown in Fig. 1. The critical role of the hyperparameter is clearly exhibited in this example. The estimated trend is the most irregular for the smallest value of τ^2 and becomes increasingly smooth with increasing values of τ^2.

One generalization of the simple trend plus uncorrelated observation noise model is to a trend plus autoregressive (AR) component plus observation noise model. In addition in [16], we considered a further generalization for multiple time series in which there is a common trend and an individual AR process model for each individual time series component. An observation model for the set of C simultaneous time series data with common trend and individual AR components takes the form

$$y_{i,n} = t_n + v_{i,n} + \varepsilon_{i,n}, \quad i = 1, \ldots, C \tag{18}$$

where t_n, $v_{i,n}$, $\varepsilon_{i,n}$ are, respectively, the common trend, the individual AR processes and the unobserved added noise process for each time series component. The generalizations of the simple trend model to a trend model with an AR component and also the multiple time series with a common trend component and individual AR components are treated in [16] with real data examples.

(b) Seasonal Component Time Series Modeling. Time series with seasonal components arise for example in meteorological, oceanographic, and econometric data modeling. Here we consider an example of econometric data modeling. A model for the seasonal adjustment of

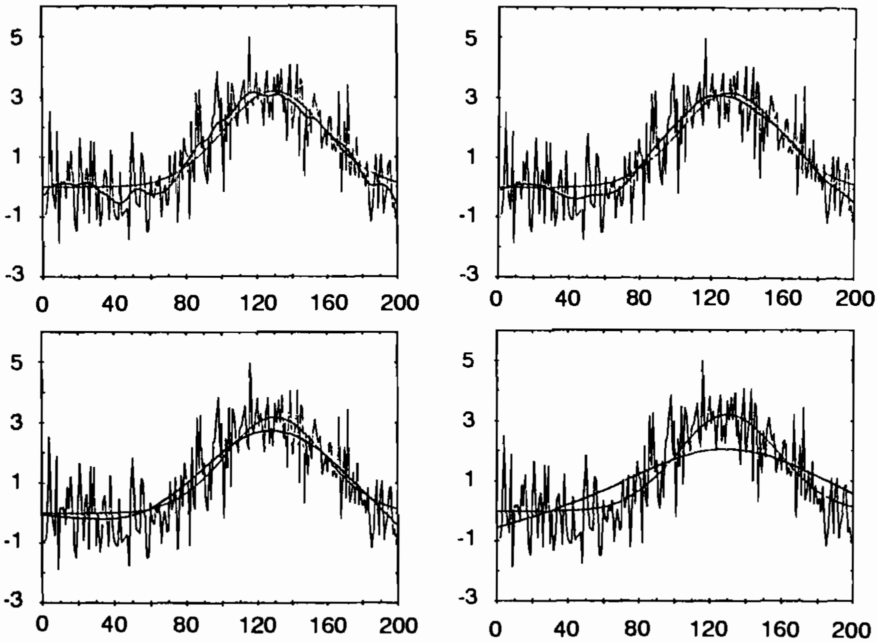


Figure 1 Trend model order 2, for hyperparameter values $\tau^2 = 8.0 \times 10^{-8}$, 8.0×10^{-7}, 8.0×10^{-5}, 8.0×10^{-3}.

econometric data [7,25] is

$$y_n = t_n + s_n + v_n + \varepsilon_n \tag{19}$$

where t_n is a trend component, s_n is a seasonal component, v_n is a globally stationary AR component, and ε_n is an unobserved observation error component.

As before, let the trend component t_n satisfy a kth-order stochastically perturbed difference equation

$$\nabla^k t_n = w_{1,n} \tag{20}$$

where $w_{1,n}$ is an i.i.d. sequence with $w_{1,n} \sim N(0, \tau_1^2)$.

The stationary AR component v_n is assumed to satisfy an AR model of order p. That is given by

$$v_n = a_1 v_{n-1} + \cdots + a_p v_{n-p} + w_{2,n} \tag{21}$$

In (21) $w_{2,n}$ is an i.i.d. sequence with $w_{2,n} \sim N(0, \tau_2^2)$.

The seasonal component of the L period difference equation is

$$s_n = -s_{n-1} - s_{n-2} - \cdots - s_{n-L+1} + w_{3,n} \tag{22}$$

In (22), $w_{3,n}$ is an i.i.d. sequence with $w_{3,n} \sim N(0, \tau_3^2)$.

These components can be incorporated and expressed in the generic state space model form (7). x_n is the state vector at time n and y_n is the observation at time n. For any particular model of the time series, the matrices F, G, and H_n are known and the observations are generated recursively starting from an initial state that is assumed to be normally distributed with mean x_0 and covariance matrix V_0.

The state space model that includes the local polynomial trend, stationary AR component, seasonal, and observation error components can be written in the orthogonal decomposition form

$$\begin{aligned} x_n &= \begin{bmatrix} F_1 & 0 & 0 \\ 0 & F_2 & 0 \\ 0 & 0 & F_3 \end{bmatrix} x_{n-1} + \begin{bmatrix} G_1 & 0 & 0 \\ 0 & G_2 & 0 \\ 0 & 0 & G_3 \end{bmatrix} w_n, \\ y_n &= [\, H_1 \quad H_2 \quad H_3 \,] x_n + \varepsilon_n \end{aligned} \tag{23}$$

The component models (F_j, G_j, H_j) in order $(j = 1, \ldots, 3)$ represent the trend, stationary AR, and seasonal components, respectively. The particular trend, AR, and seasonal component difference equation constraints that we have employed and that have representations in the (F_j, G_j, H_j) matrices in (23) are shown in (20), (21), and (22).

The state or system noise vector w_n and observation noise ε_n are assumed to be independent normally distributed random variables with zero mean and diagonal covariance matrix

$$\begin{bmatrix} w_n \\ \varepsilon_n \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_1^2 & 0 & 0 & 0 \\ 0 & \tau_2^2 & 0 & 0 \\ 0 & 0 & \tau_3^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} \right) \tag{24}$$

An example of a state space model that incorporates each of the components with trend order 2, AR model order 2, and seasonal component with period $L = 4$ is

$$\left[\begin{array}{c} t_n \\ t_{n-1} \\ \hline v_n \\ v_{n-1} \\ \hline s_n \\ s_{n-1} \\ s_{n-2} \end{array}\right] = \left[\begin{array}{cc|cc|ccc} 2 & -1 & & & & & \\ 1 & 0 & & & & & \\ \hline & & a_1 & a_2 & & & \\ & & 1 & 0 & & & \\ \hline & & & & -1 & -1 & -1 \\ & & & & 1 & 0 & 0 \\ & & & & 0 & 1 & 0 \end{array}\right] \left[\begin{array}{c} t_{n-1} \\ t_{n-2} \\ \hline v_{n-1} \\ v_{n-2} \\ \hline s_{n-1} \\ s_{n-2} \\ s_{n-3} \end{array}\right] + \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \begin{bmatrix} w_{1,n} \\ w_{2,n} \\ w_{3,n} \end{bmatrix}$$

$$y_n = [1 \;\; 0 \,|\, 1 \;\; 0 \,|\, 1 \;\; 0 \;\; 0] x_n + \varepsilon_n$$

Again ignoring initial values, the smoothness priors problem that includes all of the components in the decomposition identified above corresponds to the maximization of

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - t_n - s_n)^2\right\}\exp\left\{-\frac{1}{2\tau_1^2}\sum_{n=1}^{N}(\nabla^k t_n)^2\right\}$$
$$\times\exp\left\{-\frac{1}{2\tau_2^2}\sum_{n=1}^{N}\left(v_n - \sum_{i=1}^{p} a_i v_{n-i}\right)^2\right\}$$
$$\times\exp\left\{-\frac{1}{2\tau_3^2}\sum_{n=1}^{N}\left(\sum_{i=0}^{L-1} s_{n-i}\right)^2\right\} \tag{25}$$

The first term in (25) corresponds to the conditional data distribution. The remaining terms in (25) in order correspond to the priors on the trend, the globally stochastic component and the seasonal component.

The role of the hyperparameters τ_1^2 and τ_3^2 as measures of the uncertainty in the belief of the priors is clear from (25). Relatively large τ_1^2 (τ_3^2) imply relatively wiggly trend (seasonal) components. Relatively small τ_1^2 (τ_3^2) imply relatively smooth trend (seasonal) components. The ratios of τ_j^2/σ^2, $j = 1$ or 3, can be interpreted as signal-to-noise ratios. Here, too, the value of σ^2 in (25) is estimated essentially free of computational cost in the Kalman filter algorithm.

(c) A Seasonal Adjustment Example. Here we examine the modeling and prediction performance of seasonal models with and without AR components in the 27-year monthly ($N = 324$), log-transformed sale of household appliance data (US Bureau of the Census). The modeling and prediction performance of globally stochastic AR components with AR components of different order were fitted and the AIC best component model was determined.

The computational and graphical results shown here were obtained using the DECOMP program which appeared in [26]. Table 1 shows the AR model order and corresponding values of the log-likelihood, estimated innovations variance and AICs.

The AIC best model is the one with the AR model order $m = 2$. The AIC is computed from the usual formula and includes the log-Jacobian of the transformation. The number of parameters estimated is the sum of the number of hyperparameters, plus the number of AR components plus the state dimension. As expected, the largest absolute and relative changes in the log-likelihood, estimated innovations variance $\hat{\sigma}^2$ and AIC in Table 1 occur in going from the AR(0) model to the AR(1) model. Subsequent to identifying the AR(2) model as the single best model, we

Table 1 AR Model Order and AICs

AR model order	Log likelihood	Variance	AIC
0	962.410	0.369×10^{-4}	3347
1	1004.602	0.277×10^{-4}	3266
2	1008.097	0.226×10^{-4}	3261
3	1008.333	0.110×10^{-4}	3263
4	1008.331	0.145×10^{-4}	3265

withheld the 24 observations, $n = 301, \ldots, 324$, recomputed the AR(0) and AR(2) models, and predicted the withheld data.

Graphical results from these computations are in Fig. 2. The left-hand column results, in that illustration, correspond to the AR(0) model; the right-hand column refer to the AR(2) model. It is noteworthy that the trend of the AR component model is smoother than the trend of the without-AR component model. Also, the trend plus AR component is almost indistinguishable from the trend in the without-AR component model. The seasonal component and the noise component of the AR(2) model were very similar to those of the AR(0) model and are not shown. The last row of graphs in Fig. 2 shows the true data $n = 276, 277, \ldots, 324$, the (out-of-sample) forecast values for $n = 301, 302, \ldots, 324$, and the corresponding $\pm 1, 2, 3\sigma$ forecast bands around the forecast values. Those graphs illustrate an important property of the AIC best trend plus seasonal plus AR component model, as compared to the trend plus seasonal model. That is, the forecast bands of the AIC best, trend plus AR plus seasonal model are much tighter than those for the model without an AR component. Therefore we can be more confident about the reliability of the forecast using the trend plus AR plus seasonal model than from the trend plus seasonal model. We have seen many time series in which the trend plus AR plus seasonal model is the better AIC criterion model than the trend plus seasonal model.

4. Smoothly Changing Spectrum Estimation

The problem of modeling time series with changing spectra arises in numerous applications areas, including meteorology, oceanography, speech processing, and automatic analysis of biomedical signals such as electroencephalograms and electrocardiograms. Of the approaches that have been proposed for the modeling of changing spectra included are, partitioning the time series into stationary segments and modeling each segment as an AR or ARMA model (for example [27,28]), state

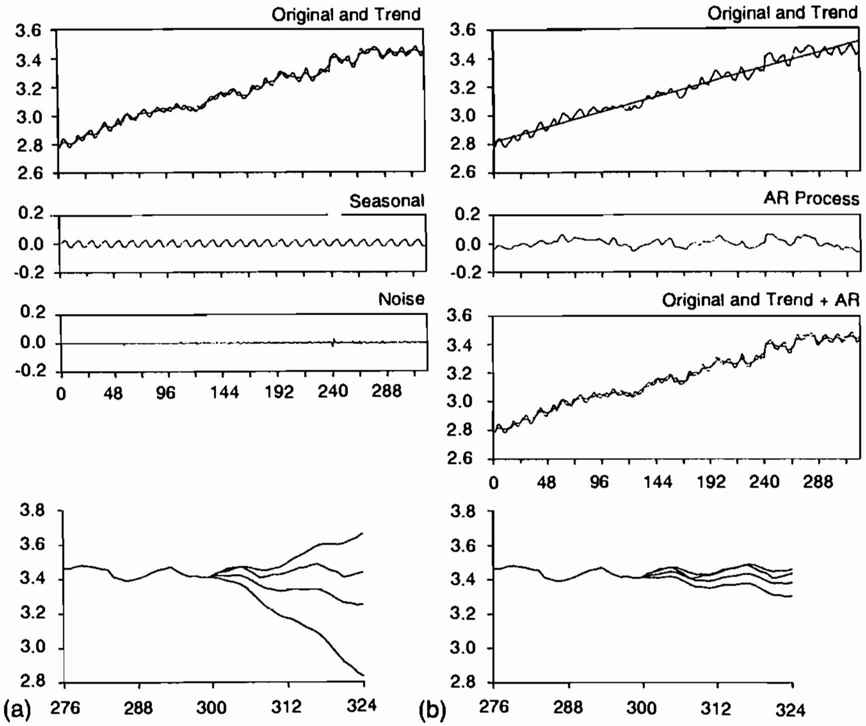


Figure 2 Linear Gaussian model analysis of household appliances data. (a) Results from models without an AR component, original and trend, seasonal and noise components. True data, $n = 276, \ldots, 324$, out of sample forecast, $n = 301, \ldots, 324$, and 3σ band. (b) Results from models with an AR component, original and trend, AR component and original and trend plus AR components. True data, $n = 276, \ldots, 324$, out of sample forecast, $n = 301, \ldots, 324$, and 3σ band.

space modeling with time-dependent coefficients [29], and orthogonal polynomial expansion of the AR coefficients of a time varying AR (TVAR) coefficient model [30]. Other notable research in time varying AR coefficient modeling is known.

Kitagawa [31] introduced a smoothness priors state space approach for TVAR coefficient modeling. Subsequently variations on smoothness priors state space TVAR modeling appeared in [7,32] and Kitagawa's program TVCAR in [26].

The problem in modeling nonstationary covariance time series is to achieve an efficient parametrization to capture the local and global statistical relationships in the time series. That objective is achieved here via the TVAR coefficient model by imposing smoothness priors constraints in the form of stochastically perturbed difference equations directly on the evolution of the individual AR coefficients. The variances of the white noise stochastic perturbations are the hyperparameters of the AR coefficient distributions. An alternative approach, imposing smoothness priors constraints in the form of stochastically perturbed difference equations on the evolution of the PARCOR coefficient in a Levinson algorithm type lattice structure AR modeling is presented in Section IV.D where we treat the modeling of time series with abruptly changing spectra. The TVAR model is used in the computation of an "instantaneous spectral density". A formal definition of the concept of evolutionary spectra in the context of nonlinear wavelet thresholding appears in [33]. (That approach to changing spectrum estimation is too remote from our interests to be treated here.)

(a) A Scalar Time-Varying AR Coefficient Model. The generic scalar time-varying AR coefficient model of the observed data $y = y_1, \ldots, y_N$ is given by

$$y_n = \sum_{i=1}^{m} a_{i,n} y_{n-i} + w_n, \quad w_n \sim \text{dist}(0, \sigma_n^2) \tag{26}$$

In general, in (26), the innovations w_n, $n = 1, \ldots, N$, are constrained to be independent but not necessarily Gaussian distributed or necessarily with constant variance. Initially we consider the case in which they are Gaussian.

Here, we assume that we are modeling a nonstationary covariance time series whose covariance structure changes slowly in time. In that case, the coefficients $a_{i,n}$ in (26) are assumed to change "gradually" with time and w_n is assumed to be a normally distributed white noise sequence with perhaps an instantaneous variance σ_n^2. Since there are $m \times N$ AR coefficients in the model in (26), an attempt to fit the parameters by least squares or any other ordinary means to the N observations $y_1, \ldots, y_N$, will yield poor parameter estimates. We consider the unknown AR coefficients to be random variables and impose Gaussian distributed stochastic constraints on those coefficients. Those constraints define a Gaussian smoothness prior distribution on the time history of the AR coefficients.

A simple and useful constraint for a time-varying AR coefficient model is obtained by imposing the stochastically perturbed difference equation constraint model on the evolution of the AR model parameters,

$$\nabla^k a_{i,n} = \delta_{i,n}, \quad i = 1, \ldots, m \tag{27}$$

For convenience, in (27) $\delta_{i,n}$ is assumed to be a zero-mean Gaussian white noise sequence with variance τ_i^2 independent of i and n. That is, $\tau_i^2 = \tau^2$, $i = 1, \ldots, m$. In addition, the frequency domain constraints shown in [8] are also incorporated in this model.

The smoothness priors constraints on the AR coefficients mitigate the problem of overparametrization by permitting the AR coefficients to be expressed as the solution of the constrained least squares problem

$$\sum_{n=1}^{N}\left[y_n - \sum_{i=1}^{m} a_{i,n}y_{n-i}\right]^2 + \zeta^2\sum_{n=1}^{N}\sum_{i=1}^{m}[\nabla^k a_{i,n}]^2 + \nu^2\sum_{n=1}^{N}\sum_{i=1}^{m} a_{i,n}^2 + \lambda^2\sum_{n=1}^{N}\sum_{i=1}^{m} i^4 a_{i,n}^2 \qquad (28)$$

In (28), m and k are assumed known and the third and fourth terms in (28) relate to the additional frequency domain smoothness priors constraint terms ([8,16]). ζ^2, ν^2, and λ^2, the tradeoff parameters which balance the infidelity of the model to the data, and the infidelity of the model to the smoothness constraints are not known.

Equation (28) yields a Bayesian interpretation of the least squares problem. Multiply (28) by $-1/2\sigma^2$ and exponentiate. Then, to within a constant term,

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[y_n - \sum_{i=1}^{m} a_{i,n}y_{n-i}\right]^2\right\}\exp\left\{-\frac{\zeta^2}{2\sigma^2}\sum_{n=1}^{N}\sum_{i=1}^{m}[\nabla^k a_{i,n}]^2\right\} \times \exp\left\{-\frac{\nu^2}{2\sigma^2}\sum_{n=1}^{N}\sum_{i=1}^{m} a_{i,n}^2\right\}\exp\left\{-\frac{\lambda^2}{2\sigma^2}\sum_{n=1}^{N}\sum_{i=1}^{m} i^4 a_{i,n}^2\right\} \qquad (29)$$

expresses the product of the conditional data distribution, the prior distribution of the smoothness of the AR parameters and the prior distribution of the smoothness of the spectrum. As before, integration of the posterior distribution of the AR parameters yields the likelihood for the smoothness tradeoff parameters.

The TVCAR program [26] uses this approach to the modeling of time varying AR models. That program incorporates a provision for human intervention which permits the modeling of nonstationary covariance time series with abruptly changing covariance structure as occurs for example in seismic events time series. At the specific instants of abruptly changing covariance structure, the observation variance is made very large. In effect, that artifice is equivalent to starting the program over again at those specified time points. In Section 4, we show a non-Gaussian method of time varying AR coefficient modeling which achieves such an effect automatically without human intervention.

A complete state space model which incorporates the smoothness priors constraints in (28) is

$$
\begin{gathered}
x_n = \begin{bmatrix} a_{1,n} \\ \vdots \\ a_{m,n} \end{bmatrix} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{bmatrix} \begin{bmatrix} a_{1,n-1} \\ \vdots \\ a_{m,n-1} \end{bmatrix} + \begin{bmatrix} w_{1,n} \\ \vdots \\ w_{m,n} \end{bmatrix} \\
\begin{bmatrix} y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} y_n & y_{n-1} & \cdots & y_{n-m} \\ 1 & & & \\ & \ddots & & \\ & & 1 & 0 \end{bmatrix} x_n + \begin{bmatrix} \varepsilon_n \\ u_{1,n} \\ \vdots \\ u_{m,n} \end{bmatrix} \\
\begin{bmatrix} w_{1,n} \\ \vdots \\ w_{m,n} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \zeta^{-2} & & \\ & \ddots & \\ & & \zeta^{-2} \end{bmatrix} \right) \\
\begin{bmatrix} \varepsilon_n \\ u_{1,n} \\ \vdots \\ u_{m,n} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & (\nu^2+\lambda^2)^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & (\nu^2+m^4\lambda^2)^{-1} \end{bmatrix} \right)
\end{gathered}
\tag{30}
$$

For a fixed difference order k, the best fit of the state space smoothness priors constraints-time varying AR coefficient model to the data $y_1, \ldots, y_N$ is the one for which the likelihood of the hyperparameters is maximized. The likelihood is computed using the recursive formulas shown in Section II.C.2.

(b) Instantaneous Spectrum. The applications we are concerned with here, in the fitting of time varying autoregressive models to data, involve the estimation of an instantaneous power spectrum. That term is due to Page in a pioneering paper [34]. Other contributions to the topic include, for example, [35] and [36].

The instantaneous power spectral density is the distribution of energy in the time series as a function of frequency at an instant in time. The more conventional analysis of stationary time series by parametric or Fourier transform methods extended to an evolutionary spectrum [35], do not have the time-frequency resolution properties necessary to capture the transitory characteristics of relatively rapidly changing nonstationary covariance data.

Our own approach is to define an instantaneous spectrum operationally. In a natural way we extend the definition of the power spectral density of a stationary process to that of an instantaneous power spectrum in terms of the time varying autoregressive model. That approach, while not mathematically formal, appears to yield meaningful computational results. Examples were shown for the scalar case in [8] and [31] in the modeling of scalar nonstationary covariance time series. An adaptive spectrum estimation variation [37] combines a smoothness priors long AR modeling [9] and state space modeling.

Therefore motivated by earlier work on spectrum estimation and the considerations discussed above, we define the instantaneous spectrum of a time-varying coefficient AR process by

$$S_n(f) = \frac{\sigma_n^2}{|1 - \Sigma_{k=1}^{m} a_{k,n} \exp(-2\pi ikf)|^2}; \quad -\tfrac{1}{2} \leq f \leq \tfrac{1}{2} \tag{31}$$

The value of the instantaneous spectrum is obtained by substituting the smoothed estimates of the TVAR coefficients and a smoothed estimate of the innovations variance σ_n^2 (Section IV.C.3) into (31). The method of estimation of instantaneous frequency introduced here has superior time-frequency resolution properties than any of the methods reported in survey articles on the subject [38].

(c) An Example. The MYE2F Hokkaido seismic data, $n = 2600$, was modeled using the TVCAR program [26]. Two different runs of the program were made with the AR model order $m = 8$. One was a linear Gaussian modeling, the other was a "human intervention" run in which the P-wave and S-wave arrival times estimated by using the locally stationary AR model were inputs to the program. The likelihoods for the Gaussian and human intervention runs were respectively −2792.35 and −2723.97. Correspondingly the AICs were 5592.70 and 5455.95. The human intervention model was the better AIC model. Figure 3a respectively shows the original earthquake data, the envelope function and the variance normalized data for the human intervention run. Figure 3b illustrates the evolution of the PARCORs. The abrupt changes in the appearance of the envelope function and PARCORs is a direct consequence of the intervention. In Fig. 3c the evolution of the instantaneous spectra computed from both the linear Gaussian model and the human intervention model are shown. The intervention model instantaneous spectrum captures the abrupt changes in the spectrum. Initially we observe the spectrum of the background or ongoing vibrations of the earth. The dramatic changes in the spectrum coincident with the arrival of the P-wave and the S-wave are clearly discernible. After the S-wave, the spectrum tends to return to its original steady-state form. The appearance of the evolutionary spectrum in the linear Gaussian nonintervention modeling

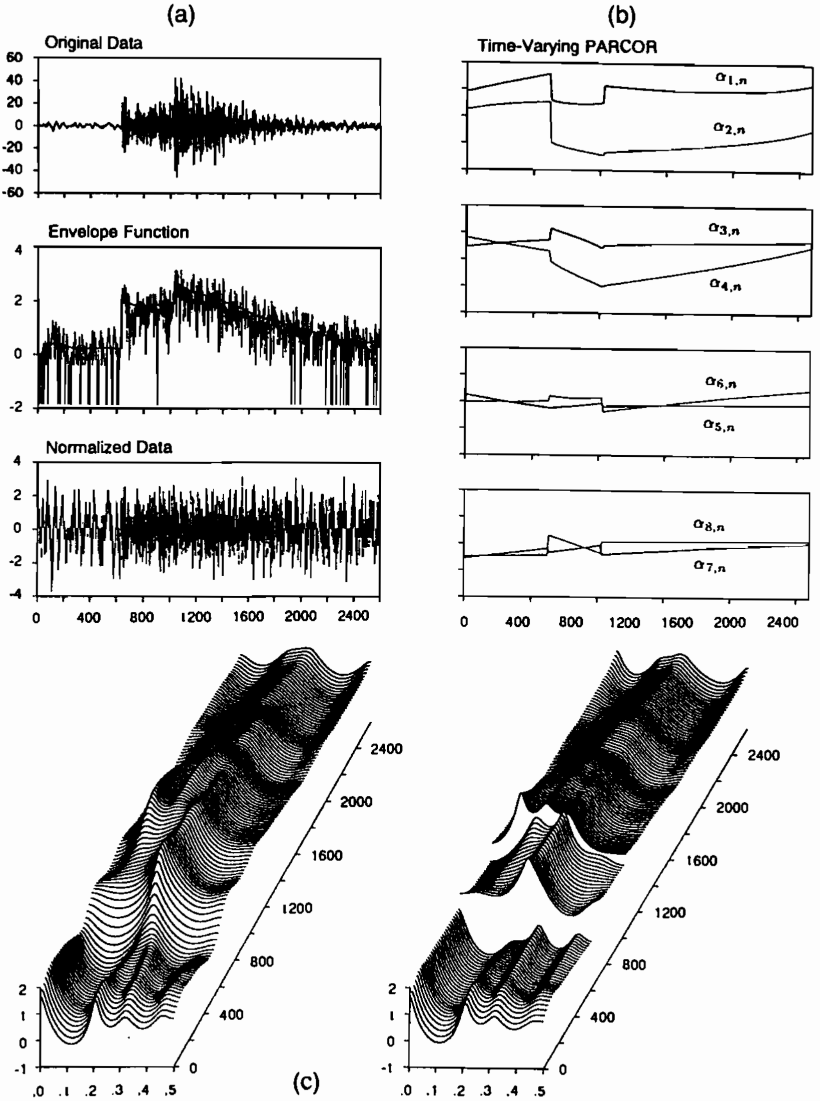


Figure 3 (a) Original earthquake data, envelope function and variance normalized data. (b) Evolution of the PARCORs. (c) The evolution of the instantaneous spectra computed from the linear Gaussian model and the human intervention model.

is seen to only change gradually and the transitions from background to P-wave to S-wave spectra are blurred.

III. THE GENERAL STATE SPACE MODEL

Consider a system described by a general state space model

$$\begin{aligned} &X_n \sim q(x_n | x_{n-1}) \\ &Y_n \sim r(y_n | x_n) \end{aligned} \tag{32}$$

where y_n is the time series and x_n is the unknown state vector. q and r are conditional distributions of x_n given x_{n-1} and of y_n given x_n, respectively. The initial state vector x_0 is distributed according to the distribution $p(x_0 | Y_0)$. This general state space model includes various important time series models including a linear state space model with non-Gaussian white noises, nonlinear state space modeling, and modeling of inhomogeneous discrete processes.

(a) General Filtering and Smoothing. For the state estimation of the general state space model, we need to evaluate $p(x_n | Y_m)$, the conditional distribution of x_n given observations Y_m. It can be shown that for the general state space model, the recursive formulas for obtaining the one step ahead predictor, the filter, and the smoother are given as follows ([1,16,39]).

(b) One Step Ahead Prediction.

$$p(x_n | Y_{n-1}) = \int_{-\infty}^{\infty} q(x_n | x_{n-1}) p(x_{n-1} | Y_{n-1})\, \mathrm{d}x_{n-1} \tag{33}$$

(c) Filtering.

$$p(x_n | Y_n) = \frac{r(y_n | x_n) p(x_n | Y_{n-1})}{p(y_n | Y_{n-1})} \tag{34}$$

where $p(y_n | Y_{n-1})$ is obtained by $\int r(y_n | x_n) p(x_n | Y_{n-1})\, \mathrm{d}x_n$.

(d) Smoothing.

$$p(x_n | Y_N) = p(x_n | Y_n) \int_{-\infty}^{\infty} \frac{p(x_{n+1} | Y_N) q(x_{n+1} | x_n)}{p(x_{n+1} | Y_n)}\, \mathrm{d}x_{n+1}. \tag{35}$$

The formulas (33), (34), and (35) show recursive relations between state distributions. However, the conditional distribution of the state $p(x_n | Y_m)$ is in general non-Gaussian. In the next subsection, a numerical method for the realization of the formulas is shown.

For the general state space model the log likelihood is obtained by

$$l(\theta) = \sum_{n=1}^{N} \log p(y_n | Y_{n-1}) \tag{36}$$

It should be noted here that $p(y_n | Y_{n-1})$ is the denominator of (34). Therefore the log-likelihood is obtained as the by-product of the non-Gaussian filter.

A. Implementations of the General Filter and Smoother

In this subsection, we briefly show methods for implementing the general filtering and smoothing formulas for the case when the distribution has a density function.

1. Numerical Approximation

The general filtering and smoothing formulas can be realized by using numerical approximation to the densities. In this approach, each density function is approximated by a step function, continuous piecewise linear function, or spline function. In the step-function approximation, each function is expressed by the number of segments, k, location of nodes, x_i, $(i = 0, \ldots, k)$, and the value of the density at each segment, p_i, $(i = 0, \ldots, k)$. Specifically, we use the following notation: $p(x_n | Y_{n-1}) \sim \{k, x_i, p_{ni}\}$, $p(x_n | Y_n) \sim \{k, x_i, f_{ni}\}$, $p(x_n | Y_N) \sim \{k, x_i, s_{ni}\}$, $q(x) \sim \{kq, xq, q_i\}$.

In a typical situation, the filtering and smoothing formulas can be implemented by using the nonlinear transformation of state, the convolution of two densities, Bayes formula, and normalization. They can be realized as follows:

- Convolution: $p_{ni} = \Delta x \, \Sigma_{j=1}^{k} q_{i-j} f_{n-1,j}$
- Normalization: $C = \Delta x \, \Sigma_{i=1}^{k} p_{ni} r_{yi}$
- Bayes formula: $f_{ni} = C^{-1} p_{ni} r_{yi}$.

(a) Remark. In the above implementation, most of the computing time is spent for the convolution. This computation can be significantly reduced by using an FFT algorithm [39].

2. Gaussian Sum Approximation

In the case of a state space model with densities, another way of implementing the non-Gaussian filter is to use Gaussian sum (mixture) approximations to the densities [40]. In this method each density is approximated by a Gaussian sum:

$$p(x_n|x_{n-1}) = \sum_{i=1}^{m_q} \alpha_i \varphi_i(x_n|x_{n-1}), \qquad p(y_n|x_n) = \sum_{j=1}^{m_r} \beta_j \varphi_j(y_n|x_n)$$

$$p(x_n|Y_{n-1}) = \sum_{k=1}^{m_{p_n}} \gamma_{kn} \varphi_k(x_n|Y_{n-1}), \qquad p(x_n|Y_n) = \sum_{l=1}^{m_{f_n}} \delta_{ln} \varphi_l(x_n|Y_n)$$

where each φ_i is a Gaussian density with appropriate mean and covariance matrix. Using this approximation, the formulas for prediction and filtering [11] are:

(a) Prediction.

$$p(x_n|Y_{n-1}) = \sum_{i=1}^{m_q} \sum_{l=1}^{m_{f_{n-1}}} \alpha_i \delta_{l,n-1} \varphi_{il}(x_n|Y_{n-1}) \equiv \sum_{k=1}^{m_{p_n}} \gamma_{kn} \varphi_k(x_n|Y_{n-1}) \tag{37}$$

(b) Filtering.

$$p(x_n|Y_n) \propto \sum_{j=1}^{m_r} \sum_{k=1}^{m_{p_n}} \delta_{jk,n} \varphi_l(x_n|Y_n) = \sum_{l=1}^{m_{f_n}} \delta_{ln} \varphi_l(x_n|Y_n) \tag{38}$$

Here $\equiv$ means reordering, $\gamma_{kn} = \alpha_i \delta_{l,n-1}$ (for some k), $\delta_{jk,n} = \beta_j \gamma_k \varphi_{jk}(y_n|Y_{n-1})$ (for some l) and φ_{il} and φ_{jk} are obtained by the Kalman filter.

The Gaussian-sum version of the smoothing can be derived by using the two-filter formula for smoothing [12]. Let $Y^n \equiv \{y_n, \ldots, y_N\}$ denote the set of present and future observations. Then

$$p(x_n|Y_N) = p(x_n|Y_{n-1}, Y^n) = \frac{p(x_n|Y_{n-1})p(Y^n|x_n)}{p(Y^n|Y_{n-1})} \tag{39}$$

and $p(Y^n|x_n)$ can be evaluated by the following backward filtering

$$p(Y^{n+1}|x_n) = \int_{-\infty}^{\infty} p(Y^{n+1}|x_{n+1}, x_n) q(x_{n+1}|x_n)\, dx_{n+1} \tag{40}$$

$$p(Y^n|x_n) = p(Y^{n+1}|x_n) r(y_n|x_n) \tag{41}$$

Therefore, by using this two-filter formula for smoothing, the Guassian-sum version of the smoother can be obtained.

3. A Monte Carlo Filtering and Smoothing Method

The use of Monte Carlo-Gibbs sampling methods for statistical data analysis received a considerable impetus from [41,42] and many subsequent papers. More recently Monte Carlo methods have been applied to time series analysis, and to non-Gaussian and nonlinear modeling as well [43,44]; however they have primarily been used to achieve numerical integration.

Here, we show a direct Monte Carlo method for state space prediction, filtering and smoothing. In this method, each conditional distribution is expressed by many of its realizations, and the movement of each "particle" (sample from a distribution) is simulated by using the assumed model. It can be applied to a wide class of nonlinear non-Gaussian higher dimensional state space models if the dimensions of the system noise and the observational noise are low. Examples of the application of the Monte Carlo method include the modeling of time series with abrupt trend discontinuities, time series with abrupt trend and seasonal discontinuities, and nonlinear time series. Additional examples and many technical details that are not shown here appear in [12,13]. An interesting variation of the resampling scheme in the filtering procedure is in [45].

(a) One Step Ahead Prediction. Let $\{s_{n-1}^{(1)}, \ldots, s_{n-1}^{(m)}\}$ be independent realizations of $p(x_{n-1}|Y_{n-1})$. That is, for $i = 1, \ldots, m$

$$s_{n-1}^{(i)} \sim p(x_{n-1}|Y_{n-1}) \tag{42}$$

Here, m is the number of "particles" used for the approximation. Using these realizations, generate an ith sample at time n, $t_n^{(i)}$ by

$$t_n^{(i)} \sim q(x_n|s_{n-1}^{(i)}) \tag{43}$$

Then, $\{t_n^{(1)}, \ldots, t_n^{(m)}\}$ can be considered as realizations of the one step ahead predictor $p(x_n|Y_{n-1})$.

(b) Filtering. Given the current observation y_n and $t_n^{(j)} \sim p(x_n|Y_{n-1})$, compute

$$\alpha_n^{(j)} = H(y_n|t_n^{(j)}) \tag{44}$$

for $j = 1, \ldots, m$. Then, given y_n, the posterior probability is given by

$$\Pr(x_n = t_n^{(j)}|Y_n) = \frac{\alpha_n^{(j)} \cdot \frac{1}{m}}{\sum_{i=1}^m \alpha_n^{(i)} \cdot \frac{1}{m}} = \frac{\alpha_n^{(j)}}{\sum_{i=1}^m \alpha_n^{(i)}} \tag{45}$$

For the next prediction step, it is necessary to represent this distribution function by an empirical distribution. This can be done by generating m independent realizations $\{s_n^{(1)}, \ldots, s_n^{(m)}\}$ by the resampling of $\{t_n^{(1)}, \ldots, t_n^{(m)}\}$ with probabilities

$$\Pr(s_n^{(i)} = t_n^{(j)}|Y_n) = \frac{\alpha_n^{(j)}}{\alpha_n^{(1)} + \cdots + \alpha_n^{(m)}} \quad \text{for } j = 1, \ldots, m \tag{46}$$

(c) An Algorithm for Filtering and Computation of the Likelihood. The following is a summary of the prediction and filtering algorithm.

1. *Determine m, the number of realizations to be used for the approximation of each distribution.*
2. *Generate a k-dimensional random number $s_0^{(j)} \sim p_0(x)$ for $j = 1, \ldots, m$.*
3. *Repeat the following steps for $n = 1, \ldots, N$.*
    (a) *Generate $t_n^{(j)} \sim q(x|s_{n-1}^{(j)})$ for $j = 1, \ldots, m$.*
    (b) *Compute $\alpha_n^{(j)} = r(g(y_n, t_n^{(j)}))|\partial g/\partial y|$ for $j = 1, \ldots, m$.*
    (c) *Generate $s_n^{(j)} \sim (\Sigma_{i=1}^m \alpha_n^{(i)})^{-1} \Sigma_{i=1}^m \alpha_n^{(i)} I(x, t_n^{(i)})$ for $j = 1, \ldots, m$ by the resampling of $t_n^{(1)}, \ldots, t_n^{(m)}$. ($I$ is the indicator function.)*

(d) Smoothing. In principle, the algorithm for smoothing can be obtained by a simple generalization of the filter algorithm. A careful treatment required to avoid numerical difficulties is in [13]. We consider a smoothing algorithm based on storing the state vector. Assume that $(s_1^{(j)}, \ldots, s_{n-1}^{(j)})$ is the jth realization of $(x_1, \ldots, x_{n-1})$ and $w_n^{(j)} \sim q(w)$, and define $(t_1^{(j)}, \ldots, t_n^{(j)})$ by $t_{i|l}^{(j)} = s_i^{(j)}$ for $i = 1, \ldots, n-1$ and $f(s_{n-1}^{(j)}, w_n^{(j)})$ for $i = n$. Then $(t_{1|n-1}^{(j)}, \ldots, t_{n|n-1}^{(j)})$ can be considered as realizations from $p(t_1^{(j)}, \ldots, t_n^{(j)}|Y_{n-1})$.

Next, given the observation y_n, $p(t_1^{(j)}, \ldots, t_n^{(j)}|Y_{n-1})$ is updated as follows:

$$p(t_1^{(j)}, \ldots, t_n^{(j)}|Y_n) = \frac{p(y_n|t_n^{(j)})p(t_1^{(j)}, \ldots, t_n^{(j)}|Y_{n-1})}{p(y_n|Y_{n-1})} \tag{47}$$

This indicates that realization of the fixed interval smoother, $p(x_1, \ldots, x_n|Y_n)$, can be obtained by storing and resampling m sets of realizations $(t_1^{(j)}, \ldots, t_n^{(j)})$, $j = 1, \ldots, m$, with the same probability as for the filtering case.

In principle, this algorithm realizes the fixed interval smoothing for the nonlinear non-Gaussian state space model. However, in practice, since the number of realizations is finite, the repetition of the resampling will gradually decrease the number of different realizations and will deteriorate the shape of the distribution. Therefore, it is recommended to stop the smoothing algorithm after repeating the resampling fewer times (at the largest, less than 5 per cent of m is recommended).

The outcome is equivalent to applying an L-lag fixed lag smoother rather than the fixed interval smoother. The increase of lag L will improve the accuracy of the $p(x_n|Y_{n+L})$ as an approximation to $p(x_n|Y_N)$, while it will decrease the accuracy of $\{s_n^{(1)}, \ldots, s_n^{(m)}\}$ as representatives of $p(x_n|Y_{n+L})$. Since $p(x_n|Y_{n+L})$ usually converges quickly to $p(x_n|Y_N)$, it is recommended to take L not so large.

Another way of achieving fixed interval smoothing is to apply the two-filter formula used in [11]. Examples of the application of the Monte

Carlo filtering method are in Section IV. Additional examples are shown in Kitagawa [13] and Kitagawa and Gersch [16].

IV. APPLICATIONS OF GENERAL STATE SPACE MODELING

A. Modeling Trends with Abrupt Discontinuities

Here we consider the modeling of a synthesized data set time series with a trend which has abrupt discontinuities. To achieve reliable modeling of such time series it is necessary to use the more general state space model with either non-Gaussian state and/or non-Gaussian observation noise processes. The performance of the Gaussian disturbances smooth trend model, the numerical integration, and the Gaussian sum methods of non-Gaussian state space modeling (described in Section III) are shown.

Consider the data generated from the following model

$$Y_n \sim N(t_n, 1)$$
$$t_n = \begin{cases} 0 & n = 1, \ldots, 100 \\ 1.5 & n = 101, \ldots, 200 \\ -1 & n = 201, \ldots, 300 \\ 0 & n = 301, \ldots, 400 \end{cases} \tag{48}$$

In the top left insert in Fig. 4, the theoretical constant mean trend values in successive intervals, are superimposed on the observed data. The problem is to estimate the abruptly changing mean value function t_n.

For these data we used the model

$$\begin{aligned} \nabla^k t_n &= w_n \\ y_n &= t_n + \varepsilon_n \end{aligned} \tag{49}$$

As before, ∇ is the difference operator defined by $\nabla t_n = t_n - t_{n-1}$ and w_n and ε_n are white noise sequences that are not necessarily normally distributed. For simplicity in the analysis, we assume that the difference order k is one. In this case, the state space model takes on a particularly simple form with the state, transition, input and observation matrices respectively given by $x_n = t_n$, $F = G = H = 1$. As a consequence of earlier experience, the following system and observation noise model classes were considered:

$$\begin{aligned} &\text{Model (a): } \varepsilon_n \sim N(0,1),\ w_n \sim \alpha N(0, \tau^2) + (1-\alpha)N(0, \tau_x^2) \\ &\text{Model (b): } \varepsilon_n \sim N(0,1),\ w_n \sim Q(b, \tau^2). \end{aligned} \tag{50}$$

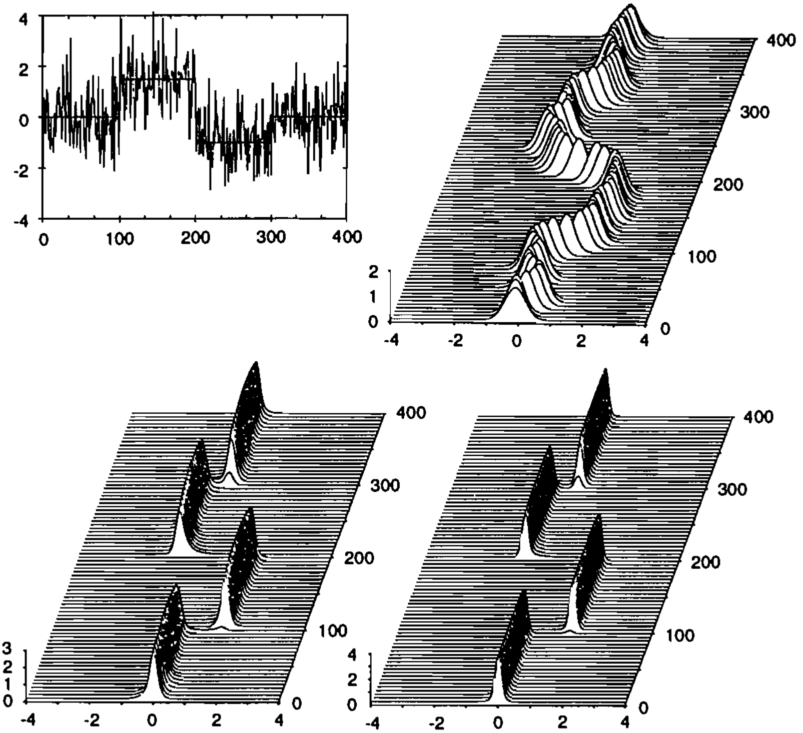


Figure 4 Nonstationary mean data with an abruptly changing theoretical mean value superimposed and marginal posterior densities, for the best Gaussian, Pearson and sum of Gaussian models.

Model (a) denotes a mixture of Gaussian noises. In Model (b), $Q(b, \tau^2)$ denotes the distribution of the Pearson system with density $q(x; b, \tau^2) = C(\tau^2 + x^2)^{-b}$ with $\frac{1}{2} < b \leq \infty$ and $C = \tau^{2b-1}\Gamma(b)/\Gamma(b - \frac{1}{2})\Gamma(\frac{1}{2})$ [46]. This broad family of distributions includes the Cauchy distribution ($b = 1$) and the Gaussian distribution ($b = \infty$). These models were fitted by the numerical integration method.

In the analysis of these data, in Model (a), τ_x^2 was arbitrarily set to 4.0, approximately the sample variance of the simulated data. The maximum likelihood estimate of τ^2 for the Gaussian model, Model (a), with $\alpha = 1.0$ or equivalently Model (b) with $b = \infty$, was $\hat{\tau}^2 = 0.0197$. The AIC of the model was 1189.44. For the mixture of Gaussian system noises model, $\hat{\alpha} = 0.99$, $\hat{\tau}^2 = 0.000004$, and AIC = 1179.59. We tried six Pearson family models: $b = 0.6, 0.75, 1.0, 1.5, 3.0$ and ∞. $b = 0.60$ is the AIC best Pearson family model with $\hat{\tau}^2 = 0.90 \times 10^{-10}$ and AIC = 1180.98. The AIC best model is the mixture of Gaussian noises model. The graphical results in Fig. 4 clearly confirm that appraisal.

B. Non-Gaussian Seasonal Component Time Series

As in Section II.C.3, nonstationary trend estimation is here extended to include nonstationary non-Gaussian seasonal components. Here neither the system noise w_n nor the observational noise ε_n is assumed to be Gaussian. We assume that they are distributed as a mixture of Gaussian distributions

$$p(x) = \alpha\varphi_1(x) + (1 - \alpha)\varphi_2(x)$$

where α is the mixture weight and $\varphi_1(x)$ and $\varphi_2(x)$ are Gaussian densities with appropriate means and variances. Since the state dimension of the seasonal adjustment model is large, we used a Gaussian sum approximation.

The generic state space model that we consider for non-Gaussian seasonal adjustment is the same as that in (7) with the exception that instead of Gaussian process noise and Gaussian observation noise we permit one or the other or both noise processes to be the sum of Gaussian processes. A real data example of non-Gaussian seasonal adjustment modeling realized by both the Gaussian sum-two filter and Monte Carlo methods is shown.

(a) A Real Data Abrupt Trend and Abrupt Seasonal Model Example. A real data example of the analysis of an economic time series, the quarterly series of the increase of the inventories of private companies in Japan 1965–1983 is discussed here. The data were modeled by a Gaussian disturbance model and also modeled under the assumption of Gaussian observation noise and a two-component Gaussian mixture process noise model. A second-order trend model was used, therefore the state dimension was 5. Gaussian and Gaussian mixture modeling results by both the two-filter formula and the Monte Carlo filter are shown in Fig. 5.

The original data are in the top row of Fig. 5. The observation interval includes the time of the 1973 oil price crisis so, as expected, abrupt changes can be seen in the original data. Graphical results of the estimated trend and seasonal components for the Gaussian and Gaussian mixture two-filter formula modeling methods are shown in the left- and right-hand columns respectively. The Gaussian model smooths out the abrupt changes in the trend. Those abrupt changes can be clearly seen in both the Gaussian mixture two-filter formula and Monte Carlo modeling methods' graphical results. Also the seasonal component in the linear Gaussian modeling is quite different from that seen in the non-Gaussian modeling method.

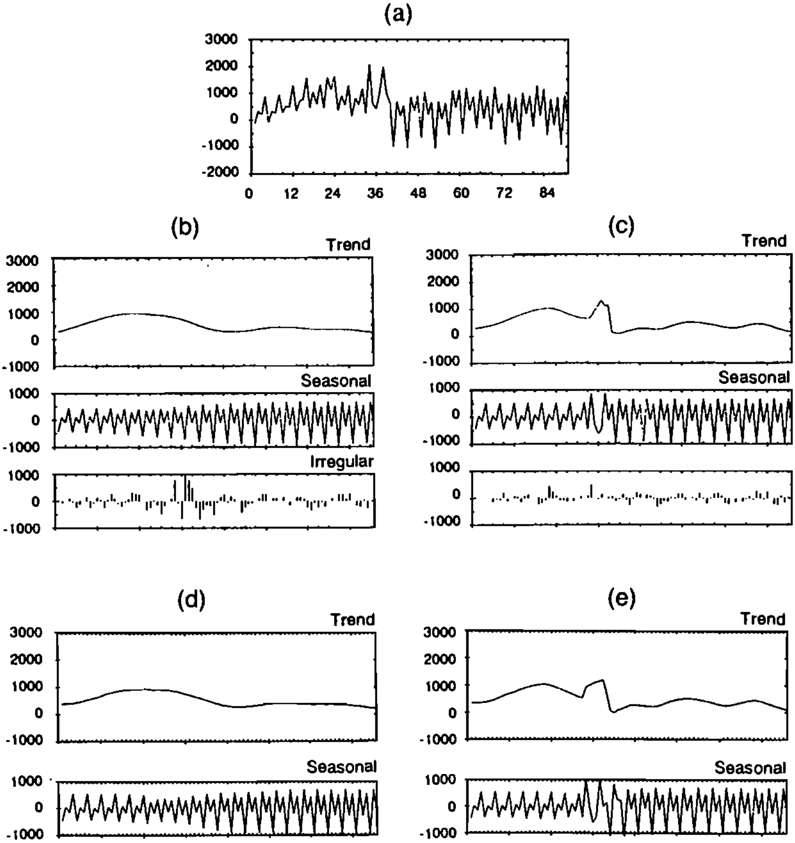


Figure 5 Analysis of quarterly increase of inventories series. (a) Top row, the original data. (b) Linear Gaussian model analysis, posterior means of trend seasonal and noise components. (c) Gaussian mixture two-filter analysis, posterior medians of trend, seasonal and noise components. (d) Linear Gaussian model analysis by a Monte Carlo smoother. (e) Gaussian mixture modeling by a Monte Carlo smoother.

C. Modeling Time-Varying Variance

First we consider a Gaussian state space modeling of changing variance. Consider a realization of white noise s_n, $n = 1, \ldots, N$, where $s_n \sim N(0, \sigma_n^2)$ with unknown time-varying variance σ_n^2. Then, the stochastic process χ_m^2 defined by

$$\chi_m^2 = (s_{2m-1}^2 + s_{2m}^2)/2, \quad m = 1, \ldots, N/2 \tag{51}$$

constitutes an independent sequence of chi-square random variables with two degrees of freedom ($\chi^2_m \sim \chi^2_2$). Then from [47] the transformation

$$t_m = \log \chi^2_m + \gamma \tag{52}$$

where $\gamma = 0.57722$ is the Euler constant, leaves the independent random variable t_m with distribution that is almost normal and with the mean and variance

$$E[t_m] \simeq \log \sigma^2_{2m}, \qquad \mathrm{var}[t_m] \simeq \pi^2/6 \tag{53}$$

That tranformation justifies our use of a smoothness priors procedure for the estimation of t_m, and hence for the estimation of the unknown variance σ^2_{2m}. This approximate normal property of t_m for smoothing the log-periodogram was used in [48].

To obtain a smooth estimate of the variance σ^2_n, consider a kth-order difference equation constraint on the log variance defined by

$$\nabla^k t_m = w_m. \tag{54}$$

In (54), $w_m \sim N(0, \tau^2)$ i.i.d. Then, as before, embed the difference equation constraint model in (54) into a state space form

$$\begin{aligned} x_m &= F x_{m-1} + G w_m \\ t_m &= H x_m + \varepsilon_m \\ \begin{bmatrix} w_m \\ \varepsilon_m \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \right) \end{aligned} \tag{55}$$

For convenience, assume here that $k = 2$. Define the state vector $x_m = (t_m, t_{m-1})^T$. Then the matrices F, G, H associated with (55) are

$$F = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad H^T = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{56}$$

Application of the Kalman filter and smoothing algorithms described in Section II yields the smoothed value of $t_{n|N}$, the logarithm of the smoothed estimate of the changing variance. $\sigma^2_{2m|N} = \sigma^2_{2m-1|N} = \exp\{t_{m|N} - \gamma\}$ will then be our smoothed estimate of the changing variance.

(a) A Seismic Data Example. Here, the changing variance structure of the Urakawa-Oki, Hokkaido, Japan March 21 1982 earthquake data [49] is estimated by both Gaussian and non-Gaussian state space models.

First we consider a linear state space Gaussian modeling of the seismic data. Figure 6a shows an $N = 1000$ seismometer earthquake signal observed in Hokkaido, March 21, Japan 1982 [49]. The original data were

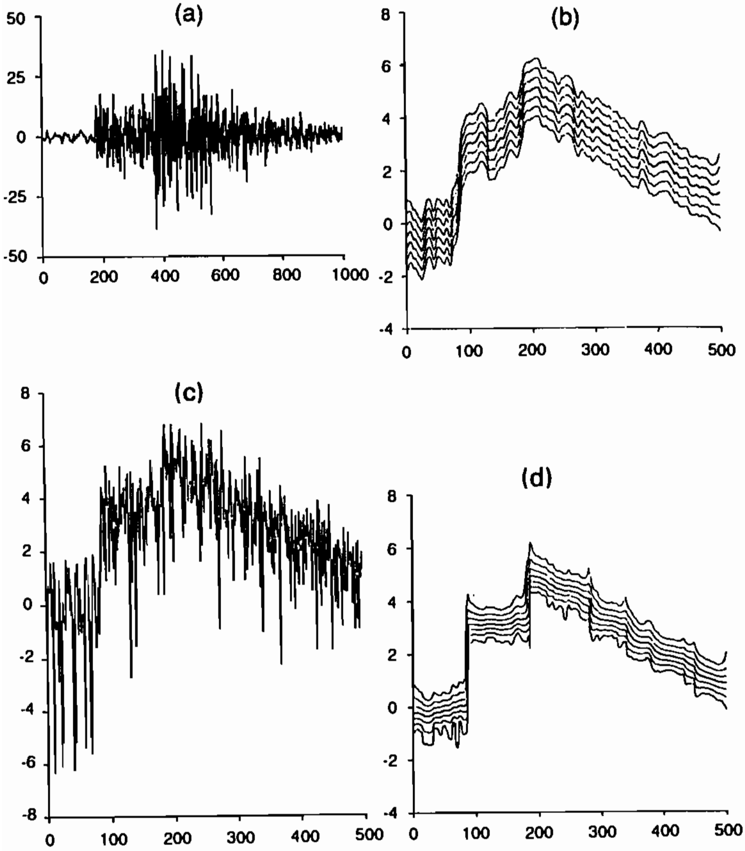


Figure 6 (a) Seismometer recording of an earthquake observed in Hokkaido, Japan 1982. (b) The transformed series χ^2_m. (c) Changing log-variance estimated by the best Gaussian model, mean and $\pm 1, 2, 3\sigma$ intervals. (d) Changing log-variance estimated by a non-Gaussian model, median, and 0.13, 2.27, 15.87, 84.13, 97.73 and 99.87 percentage intervals.

regularly sampled at intervals of about 0.01 seconds. Four successive observations were averaged to yield samples at intervals of 0.04 seconds. The problem here is the estimation of the changing variance (envelope function) of the time series, $y_1, \ldots, y_N$. Figure 6b shows the sequence of transformed observations, (51), χ^2_m, $m = 1, \ldots, N/2$. There are two abrupt changes in that sequence that correspond to the arrival of the P-wave and the S-wave. Since the density of the observational noise is assumed known in this case, the model has only one parameter τ^2, the variance of the system noise. The maximum likelihood estimate of the parameter was $\hat{\tau}^2 = 0.04318$ with AIC = 1861.64.

The estimated mean value by this model is wiggly (Fig. 6c). If we use a smaller value of τ^2, we can get a smoother curve but that curve cannot follow the sudden change of the variance. It is possible to alleviate the difficulty by allowing several outliers in the system noise, but that requires the positioning of outliers and thus is applicable only when we know the position of the outliers beforehand (e.g., the exact arrival times of P-wave and S-wave), or when an objective and effective outlier detection procedure is available.

A second approach to the modeling of these data is via a linear state space non-Gaussian modeling. That is, if we consider modeling the transformed series by a smoothness priors non-Gaussian model, we may define the system and observation noise distributions respectively by

$$\begin{aligned} q(x) &= \tau\{\pi(\tau^2 + x^2)\}^{-1} \\ r(x) &= \exp\{x - e^x\} \end{aligned} \tag{57}$$

Since χ_m^2 is exponentially distributed, the distribution of the transformed data t_m, $r(x)$, is double exponential [46].

These distributions are used in the state space model defined in (55). The numerical integration procedure described in Section III is employed to compute the likelihood of the unknown parameter τ^2. Application of the non-Gaussian model yields the maximum likelihood estimate of τ^2 to be 0.000 112 with AIC = 1718.15. The AICs indicate that in this case the non-Gaussian model is significantly better than the Gaussian model.

Figures 6c and 6d are respectively the Gaussian and non-Gaussian estimates of the log-periodogram. The estimated curves shown in Fig. 6d capture the abrupt changes in the sequence of transformed variables. Elsewhere they are smoother and have less variability than the corresponding Gaussian modeling. This figure clearly shows that the non-Gaussian models with heavy-tailed noise distributions have the ability to estimate both smooth changes and jumps of parameters simultaneously. That was not possible with the simple Gaussian model. The robustness of the non-Gaussian model to large deviations is apparent in this example. Due to the nature of the sampling distribution of χ_m^2, there are frequent large negative deviations. Those deviations act as outliers for the Gaussian model. As a result, the estimated curves are wiggly. For the non-Gaussian model, however, they are no longer outliers and do not severely affect the estimates. In Fig. 6d, we see that only the lowermost 0.13 per cent curve is affected by these large deviations.

It is interesting to note that if we use non-Gaussian modeling, the data can be modeled directly without requiring the use of the transformations in (51)–(53). If as before $s_n \sim N(0, \sigma_n^2)$, we can immediately use $y = \log s^2$ and the distribution of y is the double exponential $y \sim \frac{1}{2\pi} \exp\{\frac{y}{2} - \frac{1}{2}e^y\}$.

D. Estimation of Abruptly Changing Spectrum

Here our objective it to "automatically" achieve the modeling of a time series whose spectrum changes abruptly. Here too an instantaneous spectrum will be estimated by first fitting a scalar time-varying AR coefficient model of the observed data $y_1, \ldots, y_N$. Again the time-varying AR model is given by

$$y_n = \sum_{i=1}^{m} \alpha_{i,n} y_{n-i} + w_n, \qquad w_n \sim \mathrm{dist}(0, \sigma_n^2) \tag{58}$$

In general, in (58), the innovations w_n, $n = 1, \ldots, N$, are constrained to be independent but not necessarily Gaussian distributed or necessarily with constant variance. To achieve the modeling we introduce a PARCOR time-varying AR model.

(a) PARCOR Time-Varying AR Modeling. The coefficients of a stationary AR model can be estimated recursively via the estimation of the partial autocorrelation coefficients (PARCORs) in a lattice AR structure model [50]. This concept was extended to the time-varying AR coefficient model in (58) in Kitagawa [39]. It is used here to model the PARCORs in terms of stochastic trend models. In contrast with the approach in Section II.C.4, in which stochastic constraints are placed on each of m time-varying AR coefficients, in the time-varying PARCOR method, stochastic constraints are only placed on the PARCORs (in successively increasing model orders). In that case, models for smoothing time-varying partial autocorrelation coefficients are

$$\begin{aligned}
f_n^{(m-1)} &= \alpha_{mn}^{(m)} b_{n-m}^{(m-1)} + f_n^{(m)} \\
\nabla \alpha_{mn}^{(m)} &= v_n^{(m)} \\
b_n^{(m-1)} &= \gamma_{mn}^{(\mathrm{m})} f_{n+m}^{(m-1)} + b_n^{(m)} \\
\nabla \gamma_{mn}^{(m)} &= u_n^{(m)}
\end{aligned} \tag{59}$$

with $f_n^{(0)} = b_n^{(0)} = y_n, f_n^{(m)}$, and $b_n^{(m)}$ respectively the forward and backward prediction errors of the autoregressive model of order m. In the stationary case $v_n^{(m)} = u_n^{(m)} = 0$, and $\alpha_{mn}^{(m)} = \gamma_{mn}^{(m)}$ are identical to the partial autocorrelation coefficients. A somewhat brief description of the method is as follows.

The zeroth-order forward and backward innovations are

$$f_n^{(0)} = y_n, \qquad b_n^{(0)} = y_n \tag{60}$$

The relations between the order update innovations and the instantaneous forward and backward PARCORs are in general given by

$$\begin{aligned} f_n^{(j)} &= f_n^{(j-1)} - \alpha_{j,n}^{(j)} b_{n-j}^{(j-1)} \\ \nabla \alpha_{jn}^{(j)} &= v_n^{(j)}, \qquad v_n^{(j)} \sim N(0, \tau_1^2) \\ b_n^{(j)} &= b_n^{(j-1)} - \gamma_{j,n}^{(j)} f_{n+j}^{(j-1)} \\ \nabla \gamma_{jn}^{(j)} &= u_n^{(j)}, \qquad u_n^{(j)} \sim N(0, \tau_2^2) \end{aligned} \tag{61}$$

The instantaneous updated forward and backward AR model parameters for model order $i = 1, \ldots, j-1$, then become

$$\begin{aligned} \alpha_{i,n}^{(j)} &= \alpha_{i,n}^{(j-1)} + \alpha_{j,n}^{(j)} \gamma_{j-i,n}^{(j-1)} \\ \gamma_{i,n}^{(j)} &= \gamma_{i,n}^{(j-1)} + \gamma_{j,n}^{(j)} \alpha^{(j)}_{j-i,n} \end{aligned} \tag{62}$$

In (62) the order updated instantaneous PARCORs, $\alpha_{j,n}^{(j)}$, $\gamma_{j,n}^{(j)}$ are scalar regression coefficients and the updated forward and backward innovations are the residuals of the regressions. The innovation in (62) is the stochastic difference equation representation for the order updated forward and backward instantaneous PARCORs. State space computations for the likelihood of the hyperparameter models corresponding to τ^2 are used (for convenience we let $\tau_1^2 = \tau_2^2 = \tau^2$). Akaike's AIC is used to determine the AIC best order PARCOR model.

The distribution of the noise inputs may be either Gaussian or non-Gaussian. After estimating time-varying AR coefficients by the general smoothing, we can estimate the instantaneous spectrum of the nonstationary process using (31).

(b) An Example. Figure 7 shows the estimated changing spectrum of a seismic data example. The AR coefficients are estimated by assuming that $f_n^{(k)}$ and $b_n^{(k)}$ are Cauchy and w_n is Gaussian. We can see that the arrival of P and S waves are clearly detected by this method.

E. Inhomogeneous Discrete Process

In this section we illustrate the application of the non-Gaussian modeling method to the analysis of inhomogeneous discrete random processes data. A particular example considered is that of Tokyo rainfall data as a nonstationary binary process.

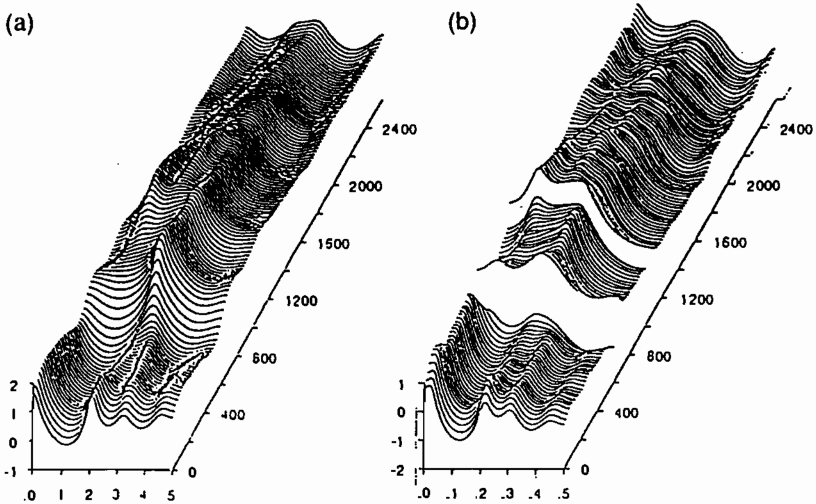


Figure 7 Changing spectrum obtained by the Gaussian and non-Gaussian models.

Consider an inhomogeneous discrete process

$$y_n \sim \mathrm{dist}(\lambda_n)$$

where the parameter of the process λ_n evolves with time. The problem considered here is the estimation of the mean value function λ_n from the observation $y_1, \ldots, y_n$. For this purpose, we consider λ_n as a state and use the smoothness prior model

$$\nabla^k \lambda_n = \omega_n.$$

In the formulation of our state space model, for simplicity we assumed that the random variables have density functions. But actually we only need to compute the convolution involved in the one step-ahead prediction formula for arbitrary distributions and Bayes formula. Therefore we can also handle discrete distributions.

The recursive formulae for obtaining one step ahead prediction and filtering densities for inhomogeneous discrete processes can be obtained from Section III.

(a) Nonstationary Binary Processes. Here we show a problem concerning estimating the time-varying mean of a nonstationary (inhomogeneous) binary process. This problem was considered earlier in [51] under Gaussian assumptions and a quadratic approximation to the likelihood

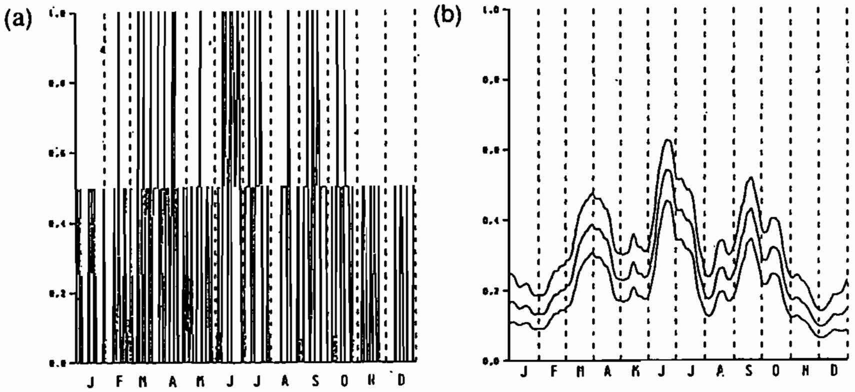


Figure 8 (a) The number of occurrences of rain for each day of the year in Tokyo, 1983–1984. (b) The estimated binomial mean function of rainfall in Tokyo.

function. That assumption is not required in our analysis, and we do not need to approximate the likelihood function.

The general problem in the modeling of such nonhomogeneous binary process data is to estimate the probability p_n of the occurrence of rainfall on a specific calendar day. The probability of rain is believed to be gradually changing with time.

The estimates obtained from only two samples per day (Fig. 8a) are hopelessly irregular. The original data indicating the number of occurrences of rainfall over 1 mm in Tokyo, 1983–1984 appears in [1].

Our model for the probability of occurrence is given by

$$
\begin{aligned}
z(m_n | l_n, p_n) &= \binom{l_n}{m_n} p_n^{m_n}(1-p_n)^{l_n-m_n} \\
q_n &= \log[p_n/(1-p_n)] \\
\nabla^k q_n &= v_n
\end{aligned}
\tag{63}
$$

Here l_n is the number of observation at the nth time point, m_n is the number of occurrences of an event at the nth time point and $z(m_n | l_n, p_n)$ is the probability mass function of the binomial distribution. Also $q_n = \log[p_n/(1-p_n)]$; the logit transformation is applied for compatibility with the analysis in [51]. In their treatment, the transformation from p_n to q_n was applied to guarantee that $0 < p_n < 1$. That is not an essential requirement in our method, since the state space is easily restricted. As indicated above, we apply a trend model computation to the logit of the daily probability of rain and in (63) we assume that $v_n \sim \text{dist}(0, \tau^2)$.

For this system with discrete observations, the analog of the filtering equation is given by

$$p(q_n | Y_n) = C^{-1} z_{l_n}(m_n | l_n, p_n) p(q_n | Y_{n-1}) \tag{64}$$

with $C = \int_{-\infty}^{\infty} z_{l_n}(m_n | l_n, p_n) p(q_n | Y_{n-1}) \, \mathrm{d}q_{n-1}$ and $p_n = \mathrm{e}^{q_n}/(\mathrm{e}^{q_n} + 1)$, where $z_{l_n}(m_n | l_n, p_n)$ is the binomial probability mass function that is now conditioned on the state at time n, q_n.

The results obtained are shown in Fig. 8b. The estimated rainfall probability reveals the known characteristics of weather in Tokyo; dry winter, unsettled spring, clear sky in May, rainy season in late June to mid July, stable hot summer in late July through August, generally fine but with an occasional typhoon in September and October. The estimated 50 per cent curve resembles the one obtained from four years of data (1981–1984) and 10 years of data (1975–1984) although the $\pm\sigma$ intervals differ, depending on the number of observations. Unlike the procedure in [51], here we are free from the Gaussian assumption and the quadratic approximation to the likelihood function.

F. Quasi-Periodic Processes

Many of the time series such as ecological data, climatological data, sunspot data, etc., exhibit the approximate repetition of a pattern but both the period and the amplitude are not very definite and change gradually. Although such series are frequently modeled by AR, ARMA, or AR plus sinusoidal models, none of those models seems quite satisfactory for prediction with more than one lead time. For such a time series with quasi-periodic character, by using a model

$$\begin{aligned} \Delta^k t_n &= v_n \\ \Delta^k c_n &= u_n \\ y_n &= c_n h(t_n) + w_n \end{aligned} \tag{65}$$

we can estimate the phase and amplitude of the model. Here $h(t)$ is a cyclic function and can be expressed, for example, by a Fourier series.

(a) The Quasi-Periodic Model. The observation process of the quasi-periodic model is expressed by

$$y_n = c_n h(\theta_n) + \varepsilon_n, \tag{66}$$

where $h(\theta_n)$ is a cyclic or periodic function, θ_n is a time-varying phase function, and c_n is a time-varying amplitude (modulating) function. Our approach to modeling such quasi-periodic time series is expressed in the form

$$\begin{aligned}
\nabla^{k_1}\theta_n &= v_n \sim N(0,\tau_1^2) \\
\nabla^{k_2}a_n &= u_n \sim N(0,\tau_2^2) \\
c_n &= \exp(a_n) \\
y_n &= c_n\left\{\sum_{j=0}^{m}\alpha_j\cos(2\pi j\theta_n) + \sum_{j=1}^{m}\beta_j\sin(2\pi j\theta_n)\right\} + \varepsilon_n
\end{aligned} \tag{67}$$

That is, we allow the phase θ_n to vary in accordance with a k_1th-order stochastic trend model and assume $v_n \sim N(0,\tau_1^2)$. Also the logarithm of the amplitude fluctuations c_n varies in accordance with a k_2th-order stochastic trend process with $u_n \sim N(0,\tau_2^2)$. From (66) and (67), $\varepsilon_n = y_n - e^{a_n}h(\theta_n)$ is nonlinear, so that despite the fact that we let the distribution of ε_n be Gaussian, i.e., $r(\varepsilon) \sim N(0,\sigma^2)$, this model is in the class of general state sppace models and the computations are realized by numerical integration. Note that the periodic component model in (67) is both cyclically and amplitude modulated to form the output. We also note that since this model is only nonlinear in the observation process the modifications to the filtering formula are rather simple and that no modifications are required for the smoothing computations. This model yields estimates of the posterior distributions of phase and amplitude. A quasi-periodic analysis of the well-known Wolfer sunspot series is shown here. Additional examples of the use of quasi-periodic models are shown in Kitagawa and Gersch [16].

(b) The Wolfer Sunspot Data. Rudolf Wolf introduced a formula for calculating the sunspot number and using historical data reconstructed the annual mean sunspot numbers of the previous 100 years [52]. As a result of considerable work by solar astronomers, records of annual means are available dating back to 1700. The continued interest in the sunspot data reflects the assumption that the data are a good indicator of the evolution of the magnetic oscillation of the sun and there are a variety of physical theories to account for that phenomenon [53].

The earliest linear model to account for these data is probably due to Yule [54]. Some of the many other time series analysis studies of the Wolfer sunspot series data appear for example in [55–58]. Our own approach is to model these data as a quasi-periodic process by the model in (66) and (67). We modeled these data with a first-order trend model, $k_1 = 1$, and a first-order amplitude modulation trend model, $k_2 = 1$. Both the phase noise and the observation noise were assumed to be Gaussian distributed. Since the observation equation is nonlinear, the model is in

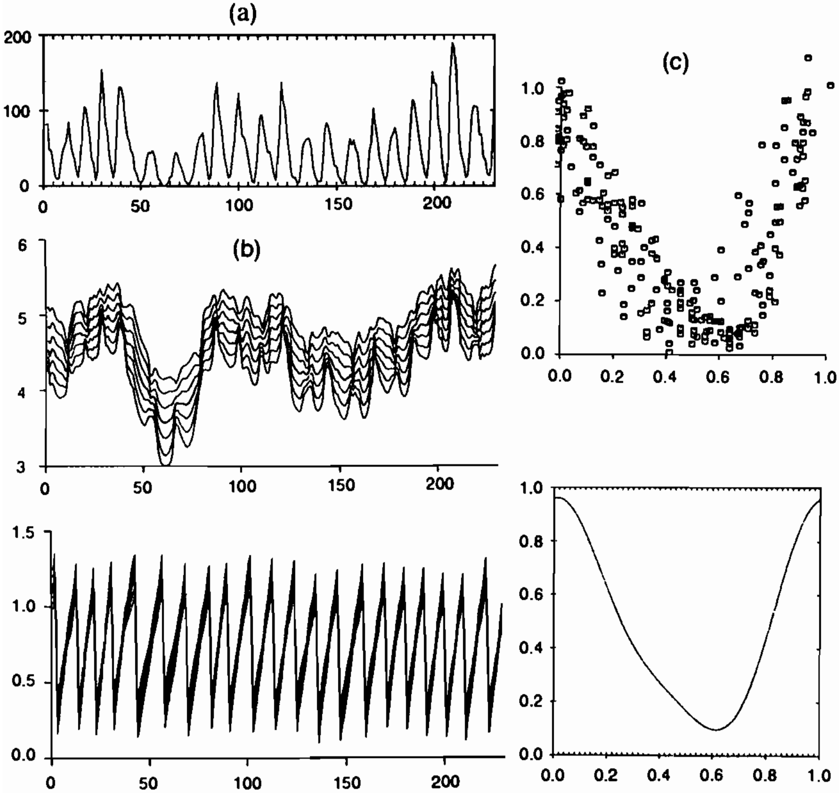


Figure 9 (a) The Wolfer sunspot series data, 1749–1979. (b) The estimated amplitude and phase. (c) The estimated cyclic function and corresponding scatter diagram.

the class of general state space modeling and the computations were realized by numerical integration.

The graphical results obtained are shown in Fig. 9. The annual $N = 231$ data for 1749–1979 are shown in Fig. 9a. Also shown in Fig. 9 are the estimated amplitude and phase and the cyclical or periodic function as well as a scatter diagram which illustrates the observed variability of the cyclical function. The computed results clearly exhibit the amplitude modulated and phase modulated structure of the series.

For completeness, and possible comparison with future modeling of these data, we report the estimated values of the fitted model. The estimated variances of the phase and amplitude noise process are $\tau_1^2 = 0.017$ and $\tau_2^2 = 0.0011$, respectively, and the observation variance is $\hat{\sigma}^2 = 18.5$. The cyclical function h_n was a trigonometric sum of components up to order $m = 2$, with parameters $\alpha_0 = 0.47893$, $\alpha_1 = 0.39369$, $\alpha_2 = 0.08451$, $\beta_1 = 0.11371$, $\beta_2 = -0.021336$. These parameters were estimated by a Householder transformation least squares procedure and the AIC was used to determine the best fitting model order.

G. Nonlinear Smoothing

A nonlinear smoothing methodology for time series analysis shown here is based on the general state space model treated in Section III. The time series is expressed in state space model form where the system model or the observation model contains nonlinear functions of the state vector. The performance of the method is illustrated by the analysis of an example that has been previously considered in the literature. The example requires numerical approximations of the relevant densities and numerical computations for the nonlinear transformations of variables, the convolution of two densities, Bayes formula, and normalization. Results for the example are compared and contrasted with those obtained by the extended Kalman filter method, by a second-moment approximation method, and also by the Monte Carlo filter method. More complete treatments of this material appear in Kitagawa [59] and [16].

The method is particularly useful for time series that cannot be analyzed satisfactorily by the standard linear time series models or by any of a large variety of linear approximation techniques. Some of the well-known approaches to state space modeling of nonlinear systems include the extended Kalman filter [60], the second-order filter, and the Gaussian sum filter [61]. These procedures approximate the non-Gaussian distribution by one or several Gaussian distributions and are known to be satisfactory in various nonlinear problems [24,62]. These methods do however have several drawbacks. The methods based on a single Gaussian density such as the extended Kalman filter may yield disastrous results when the true density is not unimodal. The conventional Gaussian sum filter has several technical difficulties in its implementation. (The Gaussian mixture two-filter method [11] does overcome some of those difficulties.)

The recursive filtering and smoothing formulas for a nonlinear state space model that we show can be easily derived from the formulas for a general state space model (Section III). Each of the probability density functions is approximated by a step function or a continuous piecewise linear function. The necessary operations on the densities are realized by numerical computations. This kind of direct method was attempted in an early stage of the development of nonlinear filters [63,64]. The recent development of fast computing facilities now makes it practical to rely on such direct numerical methods, at least for lower order systems. In return for the intensive numerical computations, our method is free from Gaussian or linearity assumptions, and we also do realize a smoothing algorithm.

Other related works on state space nonlinear systems appear in [65–68] and [43].

1. State Estimation

Consider a system described by a nonlinear state space model

$$\begin{aligned} x_n &= g(x_{n-1}) + w_n \\ y_n &= h(x_n) + \varepsilon_n \end{aligned} \tag{68}$$

where y_n and x_n are l-dimensional observation and m-dimensional state vectors, respectively. w_n and ε_n are m-dimensional and l-dimensional white noise sequences having densities $q(w)$ and $r(\varepsilon)$, respectively, which are independent of the past history of x_n and y_n. The initial state vector x_0 is assumed to be distributed according to the density $p(x_0)$.

The notation used here is as follows. The collections of the states and the observations up to time n are denoted by X_n and Y_n, namely, $X_n \equiv \{x_0, x_1, \ldots, x_n\}$ and $Y_n \equiv \{y_1, \ldots, y_n\}$. The conditional density of x_n given X_i and Y_j is denoted by $p(x_n | X_i, Y_j)$.

Here too the general problem is the evaluation of $p(x_n | Y_j)$, the conditional density function of the state x_n given the observations Y_j. As with other state space modeling, the problems of prediction, filtering, and smoothing are respectively identified with the cases $n > j$, $n = j$, and $n < j$.

The nonlinear system (68) can be expressed in the general evolution and observation equation form as

$$\begin{aligned} x_n &\sim p(x_n | x_{n-1}) \\ y_n &\sim p(y_n | x_n) \end{aligned} \tag{69}$$

for which the conditional densities $p(x_n | x_{n-1})$ and $p(y_n | x_n)$ are

$$\begin{aligned} p(x_n | x_{n-1}) &= q(x_n - g(x_{n-1})) \\ p(y_n | x_n) &= r(y_n - h(x_n)) \end{aligned} \tag{70}$$

The general state space model implies that the conditional distributions satisfy the following Markov properties:

$$\begin{aligned} p(x_n | X_{n-1}, Y_{n-1}) &= p(x_n | x_{n-1}) \\ p(y_n | X_n, Y_{n-1}) &= p(y_n | x_n) \end{aligned} \tag{71}$$

Then, from the general state space treatment in Section IV, we have that the density of x_n conditional on x_{n+1} and the entire set of observations Y_N is

$$p(x_n|x_{n+1}, Y_N) = p(x_n|x_{n+1}, Y_n) \tag{72}$$

Similarly from the general state space model in Section III, we can obtain the recursive formulas for one step ahead prediction, filtering, and smoothing densities for the nonlinear smoothing model.

As observed in Section III, for linear Gaussian systems, the conditional densities $p(x_n|Y_{n-1})$, $p(x_n|Y_n)$, and $p(x_n|Y_N)$ are characterized by the mean vectors and the covariance matrices and hence (33)–(35) are equivalent to the well-known Kalman filter [69] and the fixed interval smoothing algorithms [24]. For nonlinear state space models, however, due to the nonlinear transformation of the state variables, the conditional density $p(x_n|Y_j)$ is non-Gaussian even when both w_n and ε_n are Gaussian and cannot be specified by using the first two moments.

It should be noted that only the structure in (69) and the Markovianness in (71) are the critical model assumptions. Therefore, the additive error structure in (68) is not essential, and the method presented here can be applied to a wider class of nonlinear models than the one formulated in (68).

2. The Andrade Netto Example

In this section the performance of our nonlinear smoother is compared with that of the extended Kalman filter (EKF), a second-moment approximation method as well as the Monte Carlo filter method, by computations on a well-known model. We consider the data artificially generated by the nonlinear model originally considered by Andrade Netto *et al.* [70].

$$\begin{aligned} x_n &= \frac{1}{2}x_{n-1} + \frac{25x_{n-1}}{1+x_{n-1}^2} + 8\cos(1.2n) + w_n \\ y_n &= \frac{x_n^2}{20} + \varepsilon_n \end{aligned} \tag{73}$$

The x_n and y_n shown in Fig. 10a are generated by independent Gaussian random numbers $x_0 \sim N(0,5)$, $w_n \sim N(0,1)$, and $\varepsilon_n \sim N(0,10)$. The problem is to estimate the true signal x_n from the sequence of observations $\{y_n\}$ assuming that the model (73) is known. Our nonlinear filter and smoother were applied to the problem. For comparison, the well-known extended Kalman filter, the second-order filter, and the linearized fixed-interval smoother associated with these filters were also

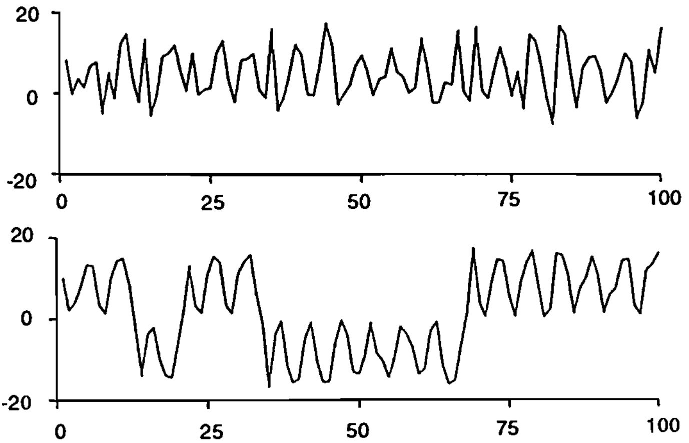


Figure 10a True signal x_n, $n = 1, \ldots, 100$, and the observations y_n, $n = 1, \ldots, 100$.

applied [59]. In filtering and smoothing, the following discretization was (arbitrarily) used: $k = 400$, $z_0 = -30$, $z_k = 30$, and $p(x_0) = N((z_0 + z_k)/2, (z_k - z_0)^2/16)$. Figure 10b shows the posterior densities $p(x_{17}|Y_m)$, $m = 16, \ldots, 20$ and 100. From left to right each column of the figure shows the results obtained by the extended Kalman filter, by the second-order filter, and by our nonlinear filter and smoother, respectively. This figure illustrates a quite typical situation where these algorithms yield substantially different results. Using our nonlinear filter, the one step ahead predictive density $p(x_{17}|Y_{16})$ is very broad and bimodal, and this bimodality extends to the filtered density $p(x_{17}|Y_{17})$ and to the smoothed density $p(x_{17}|Y_{18})$. On the other hand, the extended Kalman filter approximates each density $p(x_{17}|Y_m)$ by a single Gaussian density. Although for $m \geq 19$ the smoothed density obtained by our nonlinear smoother, $p(x_{17}|Y_m)$, also becomes unimodal and resembles a Gaussian density, its location is completely different from the one of the linearized smoother and is actually on the other side of the origin. The second-order filter shown in the middle column also approximates the posterior density by a single Gaussian density. The second-order filter posterior density estimates are very conservative in that they have large variances.

Figure 10c shows the smoothed posterior density $p(x_n|Y_N)$ versus time n obtained by our nonlinear smoother using numerical integration. In that illustration, the bold curve shows the 50% point of the posterior density and two fine curves express the 2.3% and 97.7% points which correspond to the two standard error intervals of the Gaussian densities. + indicates

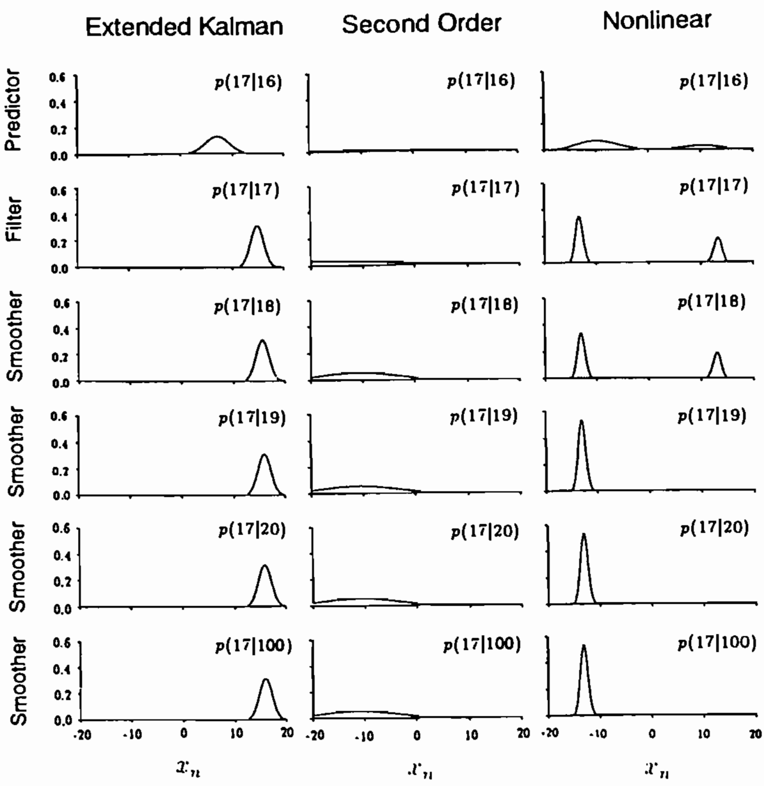


Figure 10b Posterior densities $p(x_{17}|Y_m)$, $m = 16, \ldots, 20$ and 100 obtained by the extended Kalman filter-based smoother (left), the second-order filter-based smoother (middle), and our nonlinear smoother (right).

the true value of x_n. In comparison with the performance of our nonlinear smother, the performance of the extended Kalman filter is very poor.

The smoothed estimates of the state x_n obtained by the Monte Carlo method is shown in Fig. 10d. The results compare quite nicely with those obtained by the "exact" nonlinear smoother in Fig. 10c and are also superior to those obtained by the extended Kalman filter method.

Additional details, including an empirical study on the effect of the selection of the number of nodes, and an evaluation of the extended Kalman filter performance appear in [59] and [16]. That study clearly shows that the EKF performance is much poorer than our nonlinear filter or smoother.

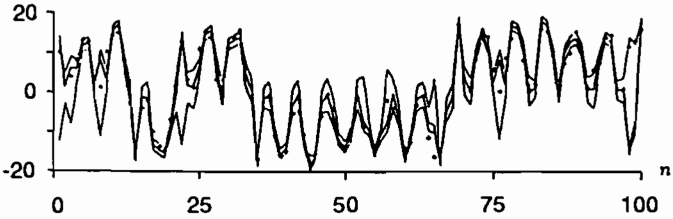


Figure 10c Posterior density $p(x_n|Y_N)$ obtained by our nonlinear smoother. The bold curve shows the median and the fine curves show the two standard error interval.

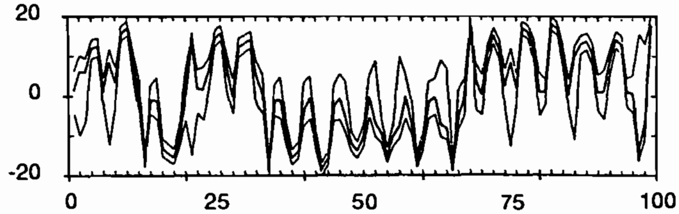


Figure 10d Posterior density $p(x_n|Y_N)$ obtained by the Monte Carlo method. Posterior median plus and minus two standard error intervals.

In summary, this example reveals two important points in the nonlinear filtering problem:

1. The extended Kalman filter and any other filter that approximates the density by a single Gaussian density may produce disastrous results when the true density is not unimodal.
2. The information from future observations is quite important to identify the location of the state. Thus the role of smoothing is essential to get a good estimate of the state.

REFERENCES

1. G Kitagawa. J Amer Stat Assoc 82:1032–1063, 1987.
2. W Gersch, G Kitagawa. In: JC Spall, ed, Bayesian Analysis of Time Series and Dynamic Models. New York: Marcel Dekker, 1988, 431–476.
3. H Akaike. In: JM Bernardo, MH De Groot, DV Lindley, AFM Smith, eds, Bayesian Statistics. Valencia, Spain: University Press, 1980, 143–166.
4. G Kitagawa. J Time Series Analysis 2:103–116, 1981.
5. T Brotherton, W Gersch. A data analytic approach to the smoothing problem

and some of its variations. Proc 20th IEEE Conf on Decision and Control, San Diego, California, 1981, pp 1061–1069.

6. R Shiller. Econometrica 41:775–788, 1973.
7. G Kitagawa, W Gersch. J Amer Stat Assoc 79:378–389, 1984.
8. G Kitagawa, W Gersch. IEEE Trans Automatic Control AC-30:48–56, 1985.
9. G Kitagawa, W Gersch. IEEE Trans Automatic Control AC-30:57–65, 1985.
10. G Kitagawa. Computers & Mathematics with Applications 18:503–514, 1989.
11. G Kitagawa. Ann Inst Statist Math 46:605–623, 1994.
12. G Kitagawa. A Monte-Carlo filtering and smoothing method for non-Gaussian nonlinear state space models. Proceedings of the 2nd US–Japan Joint Seminar on Statistical Time Series Analysis, 1993, pp 110–131.
13. G Kitagawa. J Comp Graph Stat 5:1–25, 1996.
14. H Akaike. In: BN Petrov and F Caski, eds, Second International Symposium on Information Theory. Budapest: Akademiai Kiado, 1973, 267–281.
15. H Akaike. IEEE Trans Automatic Control AC-19:716–723, 1974.
16. G Kitagawa, W Gersch. Smoothness Priors Analysis of Time Series. New York: Springer-Verlag, Lecture Notes in Statistics, 1996.
17. Y Sakamoto, M Ishiguro, G Kitagawa. Akaike Information Criterion Statistics. Tokyo: D Reidel, 1986.
18. L Ljung. System Identification, Theory for the User. New Jersey: Prentice-Hall, 1985.
19. S Kullback. Information Theory and Statistics. New York: Wiley, 1958.
20. ET Whittaker. Proc Edin Math Soc 41:65–89, 1923.
21. JO Berger. Statistical Decision Theory and Bayesian Analysis. 2nd ed. New York: Springer-Verlag, 1985.
22. IJ Good. The Estimation of Probabilities. Cambridge, Mass: MIT Press, 1985.
23. W Gersch. In: D Brillinger, P Caines, J Geweke, E Parzen, M Rosenblatt, MS Taqqu, eds, New Directions in Time Series Part II, IMA Volumes in Mathematics and Its Applications, Vol. 46. Springer-Verlag, 1992, 113–146.
24. BDO Anderson, JB Moore. Optimal FIltering. New Jersey: Prentice-Hall, 1979.
25. AC Harvey. Forecasting, Structural Time Series Models and the Kalman Filter. Victoria, Australia: Cambridge University Press, 1989.
26. H Akaike *et al.* TIMSAC-84, Computer Science Monograph, No. 22, Tokyo: Inst Statist Math, 1985.
27. T Ozaki, H Tong. On the fitting of nonstationary autoregressive models in time series analysis. Proc 8th Hawaii Intl Conf on System Sciences, 1975, pp 224–246.
28. G Kitagawa, H Akaike. Ann Inst Stat Math 30-B:215–363, 1978.
29. T Bohlin. In: R Mehra and DG Lainotis: eds, System Identification Advances

and Case Studies. New York: Academic Press, 1976.

30. F Kozin. Estimation and modeling of nonstationary time series, Proc Symposium Applied Computational Methods in Engineering, University of Southern California, 1977, pp 603–612.
31. G Kitagawa. J Sound and Vibration 89:433–445, 1983.
32. W Gersch, G Kitagawa. Earthquake Eng Structural Dynamics 13:243–254, 1985.
33. R Dahlhaus. Beiträge zur Statistik, 4, Universität Heidelberg, 1993.
34. CH Page. J Appl Phys 23:103–106, 1952.
35. MB Priestley. Spectral Analysis and Time Series (2 Vols). London: Academic Press, 1981.
36. W Martin, P Flandrin. IEEE Trans Acoust Speech Signal Processing ASSP-33:1461–1476, 1985.
37. A Houacine, G Demoment. Fast adaptive spectrum estimation: Bayesian approach and long AR models. Intl Conf Acoust Speech Signal Proc (ICASSP) 4:2085–2088, 1987.
38. B Boahash. Proc IEEE 80:520–538, 540–568, 1992.
39. G Kitagawa. Numerical Approach to Non-Gaussian Smoothing and its Applications. Computing Science and Statistics; Proc of the 20th Symposium on the Interface, EJ Wegman, DT Gantz, JJ Miller, eds, 1988, pp 379–388.
40. HW Sorenson, DL Alspach. Automatica 7:465–479, 1971.
41. AE Gelfand, AFM Smith. J Amer Statist Assoc 85:398–409, 1990.
42. AE Gelfand, SE Hills, A Racine-Poon, AFM Smith. J Amer Statist Assoc 85:972–985, 1990.
43. BP Carlin, NG Polson, DS Stoffer. J Amer Statist Assoc 75:493–500, 1992.
44. CK Carter, R Kohn. Biometrika 81:541–553, 1994.
45. T Higuchi. Res Memo #556, Inst Statist Math, Tokyo, 1995.
46. NL Johnson, S Kotz. Distributions in Statistics, Continuous Univariate Distributions, 1. New York: Wiley, 1970.
47. HT Davis, RH Jones. J Amer Statist Assoc 63:141–149, 1968.
48. G Wahba. J Amer Statist Assoc 75:122–132, 1980.
49. T Takanami. Ann Inst Statist Math 43:605, 1991.
50. S Haykin. Adaptive Filter Theory. 2nd ed. New Jersey: Prentice Hall, 1991.
51. M Ishiguro, Y Sakamoto. Ann Inst Statist Math 35B:115–137, 1983.
52. A Izenman. Mathematical Intelligence 7:27–33, 1985.
53. H Yoshimura. Astrophys J 227:1047–1058, 1979.
54. GU Yule. Philos Trans Roy Soc London A 226:267–298, 1927.
55. PAP Moran. J R Statist Soc B16:112–117, 1954.
56. P Bloomfield. Fourier Analysis of Time Series: An Introduction. New York: Wiley, 1976.
57. H Akaike. J R Statist Soc A140:456, 1977.
58. H Tong. Threshold Models in Non-linear Time Series. New York: Springer-Verlag, 1983.

59. G Kitagawa. Statistica Sinica 1:371–388, 1991.
60. AP Sage, JL Melsa. Estimation Theory with Applications to Communications and Control. New York: McGraw-Hill, 1971.
61. DL Alspach, HW Sorenson. IEEE Trans Automatic Control 17:439–447, 1972.
62. HW Sorenson. ed. Kalman Filtering: Theory and Application. Piscatawney, New Jersey: IEEE Press, 1985.
63. RS Bucy, KD Senne. Automatica 7:287–289, 1971.
64. RJP de Figueiredo, YG Jan. Spline filters, Proc of the 2nd Symposium on Nonlinear Estimation Theory, and Its Applications, San Diego, 1971, pp 127–141.
65. M West, PJ Harrison, HS Mignon. J Amer Statist Assoc 80:73–97, 1985.
66. PJ Harrison, CF Stevens. J R Statist Soc B38:205–247, 1976.
67. M West, PJ Harrison. Bayesian Forecasting and Dynamic Models. Berlin: Springer-Verlag, 1989.
68. AC Harvey, C. Fernandes. J Bus Econ Statist, 1989.
69. RE Kalman. Trans Amer Soc Mech Eng J Basic Engineering 82:35–45, 1960.
70. ML Andrade Netto, L Gimeno, MJ Mendes. IEEE Trans Automatic Control AC-23:1062–1067, 1978.

This Page Intentionally Left Blank

3
Canonical Variate Analysis in Control and Signal Processing

Wallace E. Larimore

Adaptics, Inc., McLean, Virginia

I. OVERVIEW

In recent years, a number of subspace system identification methods have been developed that involve primarily a singular value decomposition computation (see particularly Larimore [27,28], Verhaegen [46], and Van Overschee and De Moor [45]). Such procedures permit the completely automatic and reliable identification of multivariable system. However only the canonical variate analysis (CVA) procedure has been developed on the basis of optimal statistical inference principles, and as a result only it achieves optimal statistical accuracy while the others can be considerably less accurate. In this paper, the major concepts and results involved in CVA are developed, and a number of important applications are discussed.

A major issue in the application of control and signal processing methods has been the availability of suitably accurate models of the system. Once obtained, such models can be used for the design of an appropriate control system or implementation of a signal processing procedure. This has been a particular problem in multivariable systems in the past. With the recent development of CVA system identification, it is now possible to automatically and reliably identify large-scale multivariable stochastic systems including a description of the identified model uncertainty.

As a result, there is a paradigm shift in the way we think of system modeling from system data—it can now be accomplished automatically, reliably, and with optimal accuracy. This has a number of potential

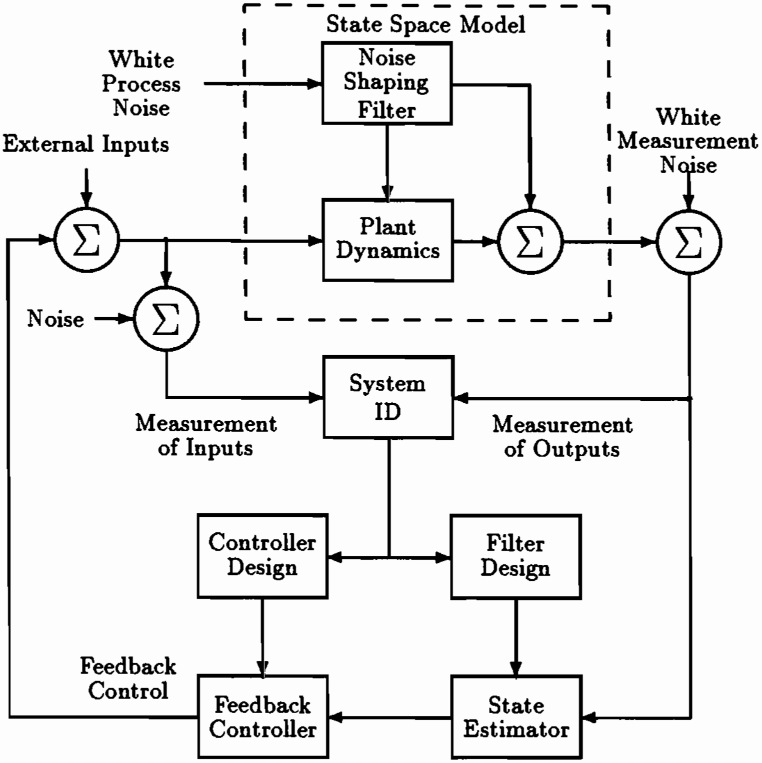


Figure 1 Structure of system identification and adaptive control.

opportunities and implications for computer automated control system design and adaptive signal processing including self-tuning, system monitoring, fault detection, and adaptation that will be explored in this paper. There are numerous industrial systems that have a need for such adaptation and automation based upon automatic system identification.

A particularly impressive example of automated system identification is a wind tunnel test of on-line adaptive control of unstable aircraft wing flutter using CVA system identification and linear quadratic Gaussian (LQG) control design. This example illustrates the use of a single system identification and control design procedure to successfully identify over 100,000 multivariable systems with up to 30 states for a wide range of system dynamics and structural configurations.

A typical situation involving CVA system identification in control is shown in Fig. 1. Input and output data are used for identification of a state space model that includes the input–output transfer function as well as a statistical model of disturbance and measurement noise processes. The identified state space model can then be used for design of a filter for state estimation and/or the design of a feedback controller. The presence of unknown feedback does not degrade the optimal statistical efficiency of the CVA procedure. In the simplest case there may not be any observed system input, so that only a stochastic model of the observed outputs is identified.

In the context of the literature of on-line adaptive control, the approach proposed here is somewhat different. The steps proposed are:

- System identification. Optimal identification of the statistically significant states of the system dynamics and disturbances and identification of the corresponding model parameters. Also determination of the uncertainty of the identified model may be required.
- Control design/model reduction. Control design based upon the identified model and possibly the model uncertainty description if robust control design is used. This may involve a model reduction step to reduce the model order taking into account the control design criteria.
- State estimation. Design and implementation of a state estimator such as a Kalman filter may be required in the controller.

A major difference in the approach proposed here as compared with much of the adaptive control literature (for example see Bitmead, Gevers and Wertz [6]) is that explicit identification of all statistically significant states is first done with the model reduction occurring in the control design step based on the control design philosophy, rather than imposing an *a priori* reduced order model structure in the identification step.

This paper is organized starting with an overview of the CVA method in the rest of this section. In Section II, the CVA method for the static multivariate regression case is developed and shown to lead to an optimal statistical procedure for determining the rank of a relationship between two sets of variables. In Section III, this result is applied to the past and future of a process to obtain the optimal determination of the states and state order of a dynamical process. Once the states are determined, a state space model is estimated simply by multivariate linear regression. To select the statistically optimal state order requires the use of the AIC that is discussed in Section IV. This is followed in Section V by a discussion of the computational aspects of the CVA method including the use of the SVD and fast computational methods. The optimality of the CVA procedure is described in Section VI. The issue of identification in the presence of unknown feedback is developed in Section VII. The CVA method generalizes to very general nonlinear systems as discussed in Section VIII. Finally a number of applications are described in Section IX including chemical process control, spectral analysis of vibrating systems, process monitoring and fault detection, on-line adaptive control of unstable aircraft wing flutter, and robust adaptive control.

A. Heuristic Approaches to Identification

Although the development of objective methods for system identification has long been a goal of system identification, the recent state of the art (Box and Jenkins [7], Ljung and Söderström [38]) is far from being a well-defined objective procedure. A major cause of this situation is the dichotomy between the computationally reliable procedures that are of low to moderate accuracy and the high resolution procedures such as maximum likelihood (ML) estimation that are computationally unreliable. ML has long been used as a benchmark because of its achievement, at least asymptotically for large samples, of the Cramer–Rao lower bound of attainable accuracy. Most often, ML is implemented using iterative parameter maximization that may become ill-conditioned and has no bound in the amount of required computation.

As a result, a toolbox approach has been used where the analyst is supplied a variety of methods for trying to coax an acceptable solution out of the problem, with the particular method being used left largely to the ingenuity of the analyst. This of course precludes any possibility of online or automated use of such procedures unless a computationally reliable procedure of lower accuracy such as least squares (LS) or recursive least squares (RLS) is used. To compound this problem, the analyst is confronted with a host of possible model structures (ARMAX, ARX, MA) with attendant model orders to be specified. The literature on comparison and selection of the best among these has been contradictory, leaving the analyst to make a heuristic choice from a host of measures such as the AIC, BIC, Schwartz, and cross validation criteria to mention only a few presently in vogue.

To add to these problems are difficulties in the ARMAX parametrization that has singularities in the multivariable case as shown by Gevers and Wertz [13]. For any particular ARMAX parametrization, there is a particular set of parameters so that the parametrization is not well defined. Computational algorithms necessarily become ill-conditioned at such parameter values. It is no wonder that general high resolution multivariable identification methods are missing from these toolboxes. A comparison of the various methods for system identification is given in Table 1 including the CVA procedure described below.

Table 1 Comparison of System Identification Procedures

	CVA	RLS	ELS	MLE
Unbiased estimates	yes	no	yes	yes
Statistical accuracy	yes	no	yes	yes
Nonexpert user	yes	yes	yes	no
Reliable computation	yes	yes	no	no
Model order selection	yes	yes	no	no

B. Automated and Objective Identification

To suggest that there is an objective and reliable procedure for system identification would seem to contradict the current body of software and practice of system identification. This possibility stems from the confluence of several very recent developments in computational methods and model order selection:

- The use of an over-parametrized state space model that is globally well defined.
- Optimal determination of the system states for each state order below some maximum using a single singular value decomposition (SVD) computation. The computational requirements are predictable and the computation is stable. This is entailed in the canonical variate analysis (CVA) method (Larimore [27,28]).
- The development of an objective measure of statistical model approximation that measures the predictive ability of a model (Larimore [26], Larimore and Mehra [36]). This holds even for quite small samples, and permits the objective comparison of different model orders and structures using a small sample corrected Akaike information criterion (AIC) (Akaike [1]; Hurvich *et al.* [17–19]).

As a result, for a given model structure or state order, the ML estimate is achieved using a stable, nonrecursive procedure that is bounded in computation. No procedure can improve on the accuracy without additional assumptions or prior information about the parametric structure of the model. In comparing different model structures and orders, an objective and optimal procedure is again used.

Since there are no prior assumptions other than a linear, Gaussian process, the procedure determines the best linear representation of the process and finds all statistically significant dynamics that are present. The error in model identification for the CVA procedure has been demonstrated in Larimore [31] to be very close to the maximum likelihood lower bound, even for quite small sample sizes when the model state order is unknown.

C. Model Assumptions

The CVA method applies to equal-spaced data in time taken from a very general class of multivariable linear time-invariant stochastic systems. The assumptions are that:

- The observations are equal spaced in time.
- The system is finite-dimensional, linear, time-invariant, and possibly multivariable.
- The noise disturbances are finite-dimensional Gaussian processes, i.e., the output of white Gaussian noise exciting a linear time-invariant finite-dimensional system.
- The observations may include the addition of a bias, trend, or a deterministic polynomial function of time.

This class of systems is very general and includes the following: unknown system state order, unstable systems and nonstationary noise, multi-input multi-output systems, arbitrary input signals to the system including the presence of unknown feedback around the system, arbitrary colored or nonstationary state, and measurement disturbance processes. The CVA method explicitly deals with processes involving unit roots, cointegration, feedback, unstable systems, and nonstationary noise without the usual method of differencing that can destroy important dynamical information. For a non-Gaussian process, the CVA procedure is a generalized least squares method. There is no assumption about the system state order which is determined in the model fitting. A state space model is identified in the form

$$x_{t+1} = \Phi x_t + G u_t + w_t \tag{1}$$

$$y_t = H x_t + A u_t + B w_t + v_t \tag{2}$$

where x_t is a k-order Markov state and w_t and v_t are white noise processes that are independent with covariance matrices Q and R, respectively. The parameters to be estimated are the coefficient matrices Φ, G, H, A, and B along with the covariance matrices Q and R. The CVA approach assumes no special parametric structure of these matrices. Imposing special parametric structure on the model requires the use of nonlinear optimization methods in most cases and consequently destroys the outstanding computational and numerical properties of the CVA procedure.

It is useful to contrast the above fully parametrized case with the parametric structure case where these matrices are functions of a vector θ of unknown parameters that imposes some special constraints on the model structure. The use of prior information about special parametric structure of the problem can significantly reduce the number of parameters to be estimated and consequently improve the accuracy of the remaining parameters to be estimated. However if the assumed parametric structure is incorrect, then this can result in much larger errors than not using this prior information and estimating a general parametric model form. Also even if the assumed special structure is correct, in a number of problems it is more important to have a computationally reliable or automatic system identification procedure. If in using the special parametric structure no solution is obtained because of computational problems, then the higher accuracy that is potentially possible is of no consequence.

Once the above state space model is identified, several additional model forms can be calculated directly from it including the state space innovations model form, the state space echelon or overlapping parametrization form, and an ARMAX form. Here the ARMAX form can be chosen so that it is well conditioned. Also if the form (1) and (2) is unstable and/or nonstationary, the corresponding ARMAX form will also be. This provides an elegant solution to the problem of identification of unstable and nonstationary processes without the need for differencing that is problematic.

II. CANONICAL VARIATE ANALYSIS AND OPTIMAL RANK

A number of methods have been developed for modeling and identifying dynamical systems from data that use primarily the singular value decomposition for computation. If there is no noise in the system, these methods can produce very precise results. The major difficulty is that if there is noise in the system, the results can become much less accurate. Various methods have been proposed to deal with such noise, however in most cases the procedures are *ad hoc*. In this section, the problem of determining the rank or dimension of a relationship is discussed from a fundamental statistical approach that results in a statistically optimal procedure. This leads very naturally and directly to canonical variate analysis.

The analysis of canonical correlations and variates is a method of mathematical statistics developed by Hotelling [15] (also see Anderson [5]). Concepts of canonical variables for representing random processes were explored by Gelfand and Yaglom [12], Yaglom [49], and Kailath [21]. The initial application of the canonical correlation analysis method to stochastic realization theory and system identification was done in the pioneering work of Akaike [2–4], primarily involving ARMA structures. The present approach using state space models was first developed in Larimore [27].

Traditionally the development of canonical variate analysis in multivariate statistical analysis involves the successive selection of pairs of variables from "past" and "future" such that at each step of the procedure the pair of variables maximize a correlation measure. The development below directly determines the maximum likelihood estimator for reduced rank regression when the rank of a relationship is constrained. This gives the likelihood ratio test for the rank of the relationship in terms of the canonical correlations. Here we derive the canonical variate analysis for the case of independent and identically distributed vectors which is simpler than the case of correlated time series to be discussed [50].

Consider two vectors of random variables, x and y. We will initially consider the vector x as the set of predictor variables and the vector y as the set of variables to be predicted. We assume that x and y are jointly distributed as normal random variables with mean zero and covariance matrices Σ_{xx}, Σ_{yy}, Σ_{xy}, and that the relationship giving the optimal prediction of y from x is linear. The extension to the case of a nonzero mean is trivial; however assuming a zero mean will simplify the derivation.

We wish to determine an intermediate set of r variables z that may be fewer in number than x such that z contains all of the information in x relevant to predicting y. This problem includes the determination of the rank r of z, i.e., the optimal dimension r of z, and detemination of the linear relationship between x and z as well as that between z and y.

To be precise, consider the model described by the equations

$$y_i = Bz_i + e_i \tag{3}$$

$$z_i = Ax_i \tag{4}$$

where e_i is the error in the linear prediction of y_i from x_i for a particular selection of the prediction relationship given by the matrices A and B, and Σ_{ee} is the covariance matrix of the prediction error e_i. These equations are of course equivalent to predicting y from x as

$$y_i = BAx_i + e_i = Cx_i + e_i \tag{5}$$

where the matrix $C = BA$ has the rank constraint $\text{rank}(C) \leq r$. In the derivation, it will be much easier to deal with A and B with the dimension of z fixed as r rather than dealing with the constraint $\text{rank}(C) \leq r$.

The procedure to be used for determining the matrices A and B will be the maximum likelihood (ML) procedure. The optimality of maximum likelihood procedures will be discussed in detail including the critically important model order selection problem to determine the rank r of z.

To simplify the notation, let the matrices $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ contain the observation vectors for N samples. The joint likelihood of Y and X as a function of A, B, and Σ_{ee} is expressed in terms of the conditional likelihood of Y given X as

$$p(Y, X; A, B, \Sigma_{ee}) = p(Y|X; A, B, \Sigma_{ee})p(X) \tag{6}$$

The density function $p(X)$ of X is not a function of the unknown parameters A, B, and Σ_{ee}, and thus can be ignored. It will be seen below that the solution for the reduced rank problem is the same if we condition X on Y instead of Y on X.

For A fixed, the vectors $z_i = Ax_i$ are given, and the log of the conditional likelihood of Y given X as a function of A, B, and Σ_{ee} is

$$\log p(Y|X; A, B, \Sigma_{ee}) = \frac{N}{2}\log|\Sigma_{ee}| + \frac{1}{2}\operatorname{tr}\{\Sigma_{ee}^{-1}(Y - BAX)(Y - BAX)^T\} \tag{7}$$

From multivariable analysis (Anderson [5]), the ML estimates of B and Σ_{ee} with A fixed are

$$\hat{B} = \hat{\Sigma}_{yx}A^T(A\hat{\Sigma}_{xx}A^T)^{-1}; \qquad \hat{\Sigma}_{ee} = \hat{\Sigma}_{yy} - \hat{\Sigma}_{yx}A^T(A\hat{\Sigma}_{xx}A^T)^{-1}A\hat{\Sigma}_{xy} \tag{8}$$

where

$$\hat{\Sigma}_{xx} = \frac{1}{N}XX^T; \qquad \hat{\Sigma}_{yy} = \frac{1}{N}YY^T; \qquad \hat{\Sigma}_{xy} = \frac{1}{N}XY^T \tag{9}$$

The value of the log likelihood maximized over B and Σ_{ee} with A fixed is

$$\log p(Y|X; A, \hat{B}, \hat{\Sigma}_{ee}) = \frac{N}{2}\log|\hat{\Sigma}_{yy} - \hat{\Sigma}_{yx}A^T(A\hat{\Sigma}_{xx}A^T)^{-1}A\hat{\Sigma}_{xy}| \tag{10}$$

To simply the problem further we use the *canonical variate analysis* (CVA) to transform the x and y to independent identically distributed (iid) random variables that are only pairwise correlated, i.e., with diagonal covariance. The notation I_k is used to denote the $k \times k$ identity matrix.

CVA Theorem. Let $\Sigma_{xx}(m \times m)$ and $\Sigma_{yy}(n \times n)$ be nonnegative definite (satisfied by covariance matrices). Then there exist matrices $J(m \times m)$ and $L(n \times n)$ such that

$$J\Sigma_{xx}J^T = I_{\bar{m}}; \qquad L\Sigma_{yy}L^T = I_{\bar{n}} \tag{11}$$

$$J\Sigma_{xy}L^T = D = \operatorname{diag}(\gamma_1, \ldots, \gamma_r, 0, \ldots, 0) \tag{12}$$

where $\bar{m} = \operatorname{rank}(\Sigma_{xx})$ and $\bar{n} = \operatorname{rank}(\Sigma_{yy})$.

The vector of transformed variables $c = Jx$ has covariance matrix

$$\Sigma_{cc} = J\Sigma_{xx}J^T = I \tag{13}$$

so the CVA chooses the transformation J so that c is a vector of uncorrelated random variables. The covariance matraices of the *canonical variables* $c = Jx$ and $d = Ly$, respectively, are thus identity matrices. The covariance between c and d is diagonal (only pairwise correlated). The CVA is a generalized singular value decomposition (GSVD) where the weightings Σ_{xx} and Σ_{yy} insure that the canonical variables are orthogonal—the random vector equivalent to orthogonality for the usual SVD. Thus the CVA reduces the multivariate relationship between x and y to a set of pairwise univariate relationships between the independent and identically distributed canonical variables.

Substituting the CVA into the log likelihood function maximized over B and Σ_{ee} gives, with the notation $M = AJ^{-1}$,

$$\log p(Y|X; A, \hat{B}, \hat{\Sigma}_{ee}) = \frac{N}{2}\log|\hat{\Sigma}_{yy}|^{-1}|I - DM^T(MM^T)MD| \tag{14}$$

It can be shown that the maximum occurs at $\hat{M} = [I_r 0]$ which implies that

$$\hat{A} = [I_r 0]J; \qquad \text{thus } z = [I_r 0]Jx \tag{15}$$

So $\hat{A}$ is the first r rows of J. Equivalently z, the optimal rank r predictors of y, consists of the first r canonical variables $c_1, \ldots, c_r$. The predictable (with a reduction in error) linear combinations of y are given by the random variables $d = [I_r 0]Ly$, the first r canonical variables $d_1, \ldots, d_r$. Note that the relationship between X and Y is completely symmetric so that if X and Y are interchanged at the beginning, the roles of A and J are interchanged with B and L in the solution.

The log likelihood maximized over A, B, and Σ_{ee}, or equivalently over $C = BA$ with constrained rank r, is given by the first r canonical correlations γ_i between c_i and d_i:

$$\max_{\{\Sigma_{ee}, C:\text{rank}(C)=r\}} \log p(Y|X; C, \Sigma_{ee}) = \frac{N}{2}\sum_{i=1}^{r} \log|\hat{\Sigma}_{yy}|^{-1}(1 - \gamma_i^2) \tag{16}$$

Optimal statistical tests on rank involve likelihood ratios. Thus the optimal rank or order selection depends only on the canonical correlations γ_i. A comparison of potential choices of rank can thus be determined from a single GSVD computation on the covariance structure. The above theory holds exactly for zero-mean Gaussian random vectors with x_i and x_j uncorrelated for $i \neq j$. For time series, this assumption is violated.

However analogous solutions for the time series case is obtained with appropriate modifications.

A number of other statistical rank selection procedures are closely related to CVA. Consider the generalized singular value decomposition (GSVD) problem of finding J and L such that

$$\begin{aligned} &J\Delta J^T = I_{\bar{m}}; \qquad L\Lambda L^T = I_{\bar{n}}; \\ &\qquad J\Sigma_{xy}L^T = D = \mathrm{diag}(\gamma_1, \ldots, \gamma_r, 0, \ldots, 0) \end{aligned} \tag{17}$$

where the weightings Δ and Λ are positive semidefinite symmetric matrices. CVA is given by $\Delta = \Sigma_{xx}$, $\Lambda = \Sigma_{yy}$. Reduced rank regression is $\Delta = \Sigma_{xx}$, $\Lambda = I$. Principal component analysis is $x = y$, $\Lambda = I$. Principal component instrumental variables are $x = y$, $\Lambda = \Sigma_{yy}$. Partial least squares solves for the variables z sequentially. The first step is equivalent to choosing $\Delta = I$, $\Lambda = I$ and selecting z_1 as the first variable. The procedure is repeated using the residuals at the ith step to obtain the next variable z_i. Only CVA is optimal in maximizing the likelihood function under a rank constraint.

Van Overschee and De Moor [45] have shown that CVA is one of the subspace methods that effectively computes a generalized projection that can be described by a generalized singular value decomposition. The difference between the various subspace methods is the weightings Δ and Λ, and the other subspace methods use weightings different from CVA. As a result, the other subspace methods are suboptimal and potentially may have much larger errors.

III. CANONICAL VARIATE ANALYSIS OF TIME SERIES

In this section, the CVA method is applied to the identification of time series data. Although the theory in the previous section applies only to iid multivariate vectors, it is applied to correlated vector time series. Larimore presents the extension of the previous section to the time series case [50].

A. CVA of Past and Future

A fundamental concept in the CVA approach is the *past* and *future* of a process. Suppose that data are given consisting of observed outputs y_t and observed inputs u_t at time points labeled $t = 1, \ldots, N$ that are equally spaced in time. Associated with each time t is a past vector p_t consisting of the past outputs and inputs occurring prior to time t as well as a future vector f_t consisting of outputs at time t or later, specifically

$$p_t = (y_{t-1}^T, y_{t-2}^T, \ldots, u_{t-1}^T, u_{t-2}^T, \ldots)^T, \qquad f_t = (y_t^T, y_{t+1}^T, \ldots)^T \tag{18}$$

For simplicity, consider first purely stochastic processes with no observed deterministic input to the system. A fundamental property of a linear, time invariant, strict sense Markov process of finite state order is the existence of a finite dimensional state x_t which is a linear function of the past p_t

$$x_t = Cp_t \tag{19}$$

The state x_t has the property that the conditional probability of the future f_t conditioned on the past p_t is identical to that of the future f_t conditioned on the finite dimensional state x_t so

$$P(f_t|p_t) = P(f_t|x_t) \tag{20}$$

Thus, only a finite number of linear combinations of the past are relevant to the future evolution of the process.

To extend this concept to processes involving deterministic controls or inputs, the effects of future inputs must first be removed from the future outputs. Let q_t denote the future inputs $q_t^T = (u_t^T, u_{t+1}^T, \ldots,)$ and consider the conditional random variable $f_t|q_t$. Then the process is a *controlled Markov processes* of order k if there exists a k-order state such that the conditional distribution of $f_t|q_t$ given the past p_t is identical to the conditional distribution of $f_t|q_t$ given the state x_t so

$$P((f_t|q_t)|p_t) = P((f_t|q_t)|x_t) \tag{21}$$

This is equivalent of the statement that

$$P(f_t|(q_t, p_t)) = P(f_t|(q_t, x_t)) \tag{22}$$

For Gaussian processes, the probability distributions are characterized by the first two moments and we can replace $P(\cdot)$ by the expectation operation $E\{\cdot\}$ above and obtain a characterization of the Markov property of the state. In this case, the conditional random variable

$$f_t|q_t = f_t - Mq_t \tag{23}$$

is the regression coefficient for predicting f_t from q_t, where $M = \Sigma_{fq}\Sigma_{qq}^{-1}$. In terms of the discussion of CVA in Section II, the random vectors x_i and y_i are replaced respectively by the random vectors p_t and $f_t|q_t$ of this section.

In the computational problem given finite data, the past and future of the process are taken to be finite lags of length l so

$$p_t^T = (y_{t-1}^T, \ldots, y_{t-l}^T, u_{t-1}^T, \ldots, u_{t-l}^T)^T,\ f_t^T = (y_t^T, \ldots, y_{t+l-1}^T)^T \tag{24}$$

and similarly for the future inputs q_t. Akaike [4] proposed choosing the number l of lags by least squares autoregressive modeling using recursive least squares algorithms and choosing the number of lags as that minimizing the AIC criterion discussed below. This insures that a sufficient number of lags are used to capture all of the statistically significant behavior in the data. This procedure is easily generalized to include the case with inputs u_t by using ARX models.

The CVA from (11) and (12) on the past and future give the transformation matrices J and L, respectively, specifying the canonical variables c and d associated with the past p_t and future f_t. For each choice k of state order (the rank r used in Section II), the "memory" of the process is defined in terms of the past as

$$m_t = J_k p_t = [I_k\ 0] J p_t \tag{25}$$

where m_t are the first k canonical variables. The vector m_t is intentionally called "memory" rather than "state". A given selection of memory m_t may not correspond to the state of any well-defined k-order Markov process since truncating states of a Markov process will not generally result in a Markov process for the remaining state variables. In particular, the memory m_t does not usually contain all of the information in the past for prediction of the future values of m_t, i.e., $m_{t+1}, m_{t+2}, \ldots$. For the system identification problem, this is not a problem since many orders k will be considered and the one giving the best prediction will be chosen as the optimal order. This optimal order memory will correspond to the state of a Markov process within the sampling variability of the problem.

B. State Space Model Estimation

In this section, the problem of determining a state space model of a Markov process is considered. The modeling problem is: given the past of the related random processes u_t and y_t, develop a state space model to predict the future of y_t by a k-order state x_t. A k-order linear Markov process has been shown by Lindquist and Pavon [37] to have a representation in the following general state space form

$$x_{t+1} = \Phi x_t + G u_t + w_t \tag{26}$$

$$y_t = H x_t + A u_t + B w_t + v_t \tag{27}$$

where x_t is a k-order Markov state and w_t and v_t are white noise processes that are independent with covariance matrices Q and R, respectively. These state equations are more general than typically used since the noise

$Bw_t + v_t$ in the output equation is correlated with the noise w_t in the state equation. This is a consequence of requiring that the state of the state space equations be a k-order Markov state. Requiring the noises in (26) and (27) to be uncorrelated may result in a state space model where the state is higher dimensional than the Markov order k so that it is not a minimal order realization.

The solution to the optimal reduced rank modeling problem is given above in terms of the canonical variables. For a given choice k of rank, the first k canonical variables are then used as memory m_t in the construction of a k-order state space model. The canonical variables will provide an accurate estimate of the state when k is greater than or equal to the true state order of the system. For k less than the true state order, the truncation to a k-order memory results in a less than optimal selection of the state as measured in terms of maximum likelihood. However for the purpose of determination of the optimal selection of state order, the use of the truncated memory will still lead to an optimal procedure for determining the system state order.

Now consider the estimation of the state space model and then its use in model order selection. Note that if over an interval of time t the state x_t in (26) and (27) was given along with data consisting of inputs u_t and outputs y_t, then the state space matrices Φ, G, H, and A could be estimated easily by simple linear multiple regression methods. For the model state order k equal to or greater than the true state order, the state estimate $m_t = J_k p_t$ provided by the canonical variate analysis is a sufficiently good estimate so that it can be used in place of the true state for estimating these matrices of the state equations. For k less than the true order, a suboptimal solution is obtained which is still suitable for determination of the model state order as discussed in the next seciton.

In particular, consider the state equations (26) and (27) with the state x_t replaced with the memory m_t determined from CVA. The multivariate regression equations are expressed in terms of covariances, denoted by Σ, among various vectors as

$$\begin{pmatrix} \Phi & G \\ H & A \end{pmatrix} = \Sigma\left[\begin{pmatrix} m_{t+1} \\ y_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right]\Sigma^{-1}\left[\begin{pmatrix} m_t \\ u_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right] \tag{28}$$

and the error in prediction has covariance matrix

$$\begin{aligned} S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} &= \Sigma\left[\begin{pmatrix} m_{t+1} \\ y_t \end{pmatrix}, \begin{pmatrix} m_{t+1} \\ y_t \end{pmatrix}\right] \\ &\quad - \begin{pmatrix} \Phi & G \\ H & A \end{pmatrix}\Sigma^T\left[\begin{pmatrix} m_{t+1} \\ y_t \end{pmatrix}, \begin{pmatrix} m_t \\ u_t \end{pmatrix}\right] \end{aligned} \tag{29}$$

and the matrices Q, R, and B are simply expressed in terms of

$$Q = S_{11} \tag{30}$$

$$B = S_{21}S_{11}^{\dagger} \tag{31}$$

$$R = S_{22} - S_{21}S_{11}^{\dagger}S_{12} \tag{32}$$

where † denotes the pseudoinverse. Explicit computation is obtained by the substitution of $m_t = J_k p_t$.

The use of the state space model form has several advantages over the more widely used ARMAX form. The ARMAX parametrization is inherently ill-conditioned in the multivariable case which means that for any global ARMAX parametrization it is possible to find particular values of the parameters that produce arbitrarily bad ill-conditioning (Gevers and Wertz [13]). In his pioneering work on CVA of time series, Akaike [4] developed the fitting of ARMAX forms by the successive selection of a basis for the state space in terms of autoregressive and moving average terms. This procedure requires a succession of SVDs with a decision made at each step which involves much more computation and results in only an approximate procedure.

The state space model form has the major advantage that the model is globally identifiable so that the method is statistically well conditioned. The state space parametrization does not provide a unique parametrization of the problem, however all that is required is one representative from the equivalence class of parameter values specifying equivalent models. In the development, no unique parametrization is required. The CVA guarantees the choice of a well conditioned parametrization from the equivalence class.

The resulting CVA model identification procedure is fundamentally different from most other approaches in that the state is first determined by primarily an SVD computation and the coefficient matrices and covariance matrices of the state space equations are then determined by simple multiple linear regression. Other methods sharing this characteristic are generally known as subspace methods. This is in sharp contrast with extended least squares and maximum likelihood methods that require iterative nonlinear optimization with no upper bound on the required computation and that can be ill-conditioned and fail to converge. Since the CVA method involves primarily an SVD, the computations are numerically stable and accurate with an upper bound on the required computations.

IV. OBJECTIVE MODEL STRUCTURE SELECTION

In this section, the objective comparison and selection of model structure and state order is discussed. The CVA method permits the comparison of very general and diverse model structures such as the presence of an additive deterministic polynomial, the state order of the system dynamics, the presence of an instantaneous or delayed effect on the system output from the inputs, and the feedback and "causality" or influence among a set of variables. The methods discussed below allow for the precise statistical comparison of such diverse hypotheses about the dynamical system.

To decide on the model state order or model structure, recent developments based upon an objective information measure is used. Such a method was originally developed by Akaike [1] and involves the use of the Akaike Information Criterion (AIC) for deciding the appropriate order of a statistical model. Considerations of the fundamental statistical principles of sufficiency and repeated sampling provide a sound justification for the use of an information criterion as an objective measure of model fit (Larimore [26], Larimore and Mehra [36]). In particular, suppose that the true probability density is p_* and an approximating density is p_1, then the measure of approximation of the model p_1 to the truth p_* is given by the Kullback discrimination information [25]

$$I_Y(p_*, p_1) = \int p_*(Y) \log \frac{p_*(Y)}{p_1(Y)} \mathrm{d}Y \tag{33}$$

It can be shown that for large samples the AIC is an optimal estimator of the Kullback information and achieves optimal decisions on model order (Shibata [44]). For small samples, an improved AIC is discussed below that is closer to optimal than the original AIC in minimizing the Kullback information.

The AIC for each order k is defined by

$$\mathrm{AIC}(k) = -2 \log p(Y^N, U^N; \hat{\theta}_k) + 2fM_k \tag{34}$$

where p is the likelihood function based on the observations (Y^N, U^N) at N time points, and $\hat{\theta}_k$ is the maximum likelihood parameter estimate using a k-order model with M_k parameters. The small sample correction factor f is equal to 1 for Akaike's original AIC, and is discussed below for the small sample case. The model order k is chosen corresponding to the minimum value of AIC(k). For the model state order k taken greater than or equal to the true system order, the CVA procedure provides an approximate maximum likelihood solution. For k less than the true order, the CVA estimates of the system are suboptimal so the likelihood function may not be maximized. However this will only accentuate the difference between the calculated AIC of the lower order models and the model of true order so that reliable determination of the optimal state order for approximation in maintained.

The number of parameters in the state space model (26) and (27) is

$$M_k = k(2n + m) + mn + n(n + 1)/2 \tag{35}$$

where k is the number of states, n is the number of outputs, and m is the number of inputs to the system. Candy *et al.* [9] develop this result by considering the size of the equivalence class of state space models having the same input/output and noise characteristics. Thus the number of functionally independent parameters in a state space model is far less than the number of elements in the various state space matrices. The AIC provides an optimal procedure for model order selection in large sample sizes.

A small sample correction to the AIC has been recently developed for model order selection (Hurvich et al. [17–19]). The small sample correction factor f is

$$f = \frac{N}{N - \left(\frac{M_k}{n} + \frac{n+1}{2}\right)} \tag{36}$$

The effective sample size N is the number of time points at which one-step predictions are made using the identified model. For a large sample N, the small sample factor f approaches 1, the value of f originally used by Akaike in defining AIC. The small sample correction has been shown to produce model order selection that is close to the optimal as prescribed by the Kullback information measure of model approximation error.

V. COMPUTATIONAL ASPECTS

The CVA algorithm has been shown by Van Overschee and De Moor [45] to be related to subspace algorithms that effectively compute a weighted singular value decomposition. Such algorithms include the N4SID algorithm of Van Overschee and De Moor [45] and MOESP of Verhaegen [46]. These algorithms have a number of computational advantages over nonlinear optimization methods for computing maximum likelihood estimates. This includes the stable computation of models for ill-conditioned data that is accurate to machine precision. The amount of required computation is predetermined by the problem size and depends on the number of inputs, outputs, the number of past lags used in the CVA computation, and the data length. Larimore and Luk [34] show a considerable speedup in the computation is potentially possible using parallel algorithms on a systolic array of processors where the speedup is proportional to the number of processors.

Table 2 Computational Requirements of Subspace Algorithms

	Flop count	Storage
N4SID, MOESP	$O(MN^2+M^2N+M^3)$	$O(N^2+MN)$
CVA (ADAPT$_X$)	$O(MN+M^3)$	$O(N+M^2)$
FSD	$O(MN+M^2)$	$O(N+M^2)$

Recently, Cho *et al.* [10] have developed fast computational algorithms for subspace methods. The fast subspace decomposition (FSD) algorithm exploits the block shift structure of the various matrices. These methods, however, have potential problems of numerical instability (Cho *et al.* [10]). The ADAPTx implementation of the CVA procedure in Larimore [32] takes advantage of the structure to reduce the computation, but uses only numerically stable SVD computations.

Table 2 gives a comparison of the computation and storage requirements for the various algorithms, where N is the data length or sample size and M is the number of past lags used in the computation. The main difference between the fast algorithms FSD and CVA and the other two is the term N in the fast algorithms whereas the others have term N^2. Thus the N4SID and MOESP algorithms require about a factor of N more computation, which can become very large for large N, and they require a factor of around N more memory. This can make a considerable difference. For example, for a sample size $N=10{,}000$ and the number of past lags $M=100$, N4SID and MOESP require approximately 100 GFLOP and 1 GBYTES, where CVA and FSD require about 10 MFLOP and 100 KBYTES. Usually to apply any of the algorithms requires enough data to obtain accurate estimates of the parameters. In such case it is necessary that $M^2 \leq N$, and the CVA algorithm is then nearly as fast as the FSD algorithm.

The CVA computations are implemented in the ADAPT$_X$ software package (Larimore [32]). It is available running under the Matlab software package on both workstations and IBM/PC compatible computers. A C++ package is also available for UNIX workstations and IBM/PC compatible computers. Future releases are anticipated for the C++ version on high-speed signal processing chips such as the C30 and C40.

VI. OPTIMAL MODEL ACCURACY

In this section, recent results of Larimore [31] demonstrating the optimal accuracy of CVA in small samples for multivariable systems including feedback and colored noise excitation are described. The near optimality of CVA was first noted in Larimore *et al.* [35] for an ARMA process with no inputs. Deistler *et al.* [11] have also noted the optimality of CVA for the case of no inputs.

The expected Kullback information is directly related to the Cramer–Rao bound on the minimum possible parameter estimation error for any unbiased estimation procedure. As derived in Kullback [25], the expansion of the Kullback information in a Taylor series in θ about θ_0 is

$$2I(p(\theta_0);p(\theta)) = (\theta - \theta_0)^T F_\theta(\theta - \theta_0) + o(\theta - \theta_0)^3 \tag{37}$$

where the Fisher information matrix F_θ lower bounds the parameter estimation error covariance Σ_θ of any unbiased parametric estimation procedure $\hat{\theta}$ in that

$$E\{(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)\} - F_\theta^{-1} = \Sigma_\theta - F_\theta^{-1} > 0 \tag{38}$$

that is, the difference is positive semidefinite. For an optimal estimation procedure, the parameter estimation error is asymptotically normally distributed with estimation error covariance F_θ^{-1}. Twice the Kullback information $2I(p(\theta_0);p(\theta))$ evaluated for $\theta = \hat{\theta}$ is a Chi squared random variable with the degrees of freedom equal to the number of estimated parameters. The expected value is the degrees of freedom which is just the number of estimated parameters. In small samples, Hurvich *et al.* [17–19], have shown that

$$\begin{aligned} 2E\{I(p(\theta_0);p(\hat{\theta}))\} &= fM_k \\ &= f \times \text{number of estimated parameters} \end{aligned} \tag{39}$$

is a much better approximation where the small sample correction factor f is given in (36).

An estimate of the expected Kullback information can be obtained directly from simulations using a finite number R of Monte Carlo trials. The estimate is

$$\hat{I}(p(\theta_0);p(\hat{\theta})) = \frac{1}{R}\sum_{i=1}^{R} \log \frac{p(Y_i|\,\theta_0)}{p(Y_i|\,\hat{\theta})} \tag{40}$$

where in each trial $\hat{\theta}$ is estimated using a fit set sample and $\hat{I}$ is calculated using an independent prediction set sample Y_i. The likelihoods $p(Y_i|\,\theta_0)$ and $p(Y_i|\,\hat{\theta})$ are evaluated in Larimore [31] using a Kalman filter.

A 2-input, 2-output, 6-state system was used for Monte Carlo simulation of the input and output data. Two cases were simulated:

- Open-loop. State noise, measurement noise, white noise input excitation, and no feedback.

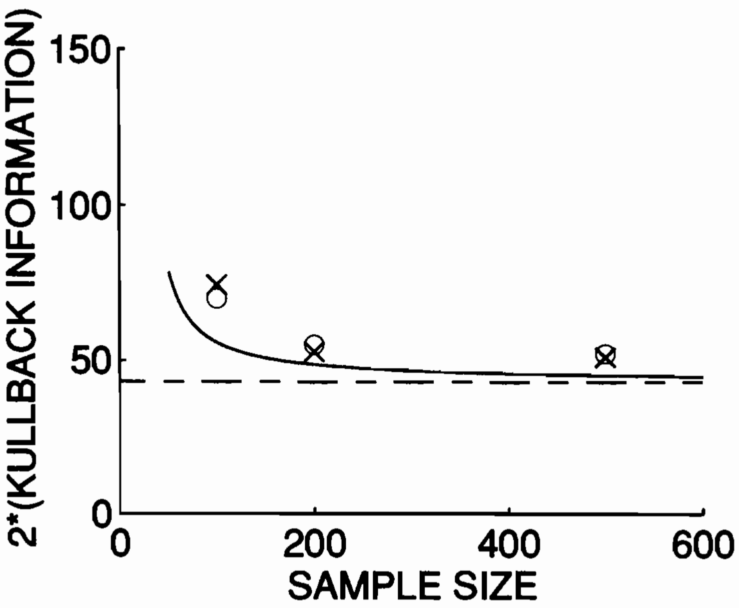


Figure 2 Estimated Kullback information (○ and ×) and lower bound (solid line) for known state order.

- Closed-loop. State noise, measurement noise, colored noise input excitation, and constant gain feedback. The input excitation was somewhat larger in this case than for the open-loop case.

In the system identification, there was no *a priori* information about the presence or absence of feedback. The details of the system dynamics used in the Monte Carlo simulations are in Larimore [31]. The ADAPT$_X$ software described in Larimore [32] for automated identification of multivariable systems was used to perform the system identification and state order estimation including script files for these evaluations of the optimality of identification.

Figure 2 gives the estimate $\hat{I}(p(\theta_0); p(\hat{\theta}))$ of the expected Kullback information when the true state order $k = 6$ is assumed known in the model identification. Asymptotically in large samples for an optimal estimator, this is the number of estimated parameters $M_6 = 43$, indicated by the horizontal dashed line. The solid line shows the small sample approximation (39) of the expected Kullback information. The results of 32 Monte Carlo simulations each for sample sizes N equal to 100, 200, and 500 are shown for both open-loop case (○) and the closed-loop case (×). The

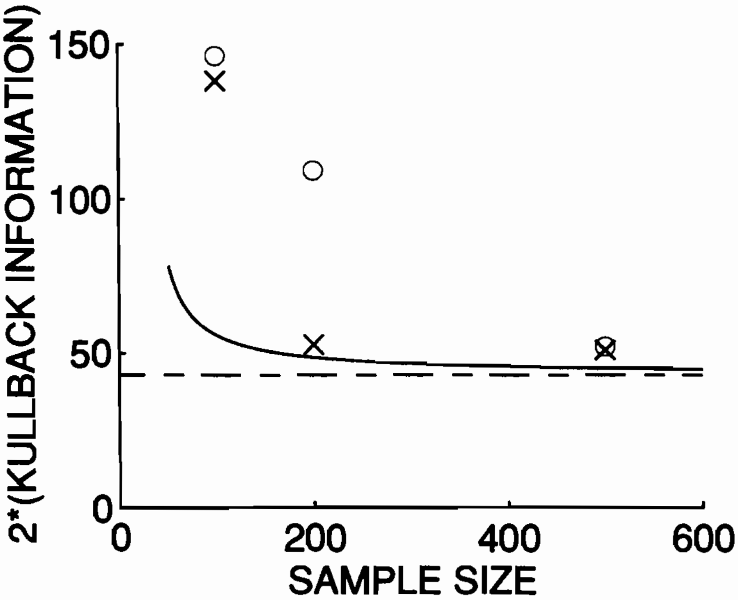


Figure 3 Estimated Kullback information (○ and ×) and lower bound (solid line) for estimated state order.

results are in excellent agreement with the small sample theory. Figure 3 gives the results for the case of unknown state order in the model identification and using the AIC in (34) for model order selection. In three of the cases where the correct state order was selected in all 32 trials, there is no difference in the results. However in the other three cases, the state order was often underestimated leading to much larger modeling errors. The difficulty is that the smaller sample sizes were not sufficient to discriminate the true state order. Such effects depend entirely on the particular system and the amount of input excitation relative to the state and measurement noise. The closed-loop case was more successful because there was more input excitation (apart from the feedback).

VII. IDENTIFICATION UNDER FEEDBACK

In this section, the justification is given for the use of CVA and in particular any maximum likelihood procedure when feedback is present from the plant output to the plant input. As will be seen, very little knowledge is required about the feedback structure, and the CVA or ML identification is done as if there is no feedback. This is in contrast to some other procedures that are not ML where the presence of feedback can cause bias or other anomalies. A procedure is developed that makes possible the comparison and test of hypothesis concerning different feedback structures for the system. A general survey of identification under feedback is given in Gustavsson, Ljung, and Söderström [14].

Consider the identification problem where the observed system inputs u_t and observed outputs y_t are given and it is desired to identify the input–output system and noise model for the system disturbances. The CVA procedure developed above assumes that the input u_t is exogenous, that is, it is uncorrelated with past and future disturbances to the system. We wish to extend the results to apply to the case of feedback from the past outputs to the present input via some sort of dynamical feedback relationship. This would introduce a correlation between the present input and past disturbances of the system.

The following notation will be used in the development, $Y_N = (y_N, \ldots, y_1)$ and similarly for U_N where Y_i for $i<1$ means that no variables are included in the set. Also let p_t denote the vector of inputs and outputs in the strict past of t. The joint likelihood function of the outputs Y_N and the inputs U_N conditional on the initial state expressed by the past p_1 at time $t=1$ and as a function of the unknown parameters θ can be expressed as

$$\begin{aligned} p(Y_N, U_N|p_1;\theta) &= p(y_N, u_N|Y_{N-1}, U_{n-1}, p_1;\theta) \\ &\quad \times p(Y_{N-1}, U_{N-1}|p_1;\theta) \qquad (41) \\ &= p(y_N|u_N, Y_{N-1}, U_{N-1}, p_1;\theta) \\ &\quad \times p(u_N|Y_{N-1}, U_{N-1}, p_1;\theta) \\ &\quad \times p(Y_{N-1}, U_{N-1}|p_1;\theta) \qquad (42) \end{aligned}$$

where first (y_N, u_N) is conditioned on (Y_{N-1}, U_{N-1}) and then y_N is conditioned on u_N. Now successively applying this gives

$$\begin{aligned} p(Y_N, U_N|p_1;\theta) &= \prod_{t=1}^{N} p(y_t|u_t, Y_{t-1}, U_{t-1}, p_1;\theta) \\ &\quad \times \prod_{t=1}^{N} p(u_t|Y_{t-1}, U_{t-1}, p_1;\theta) \qquad (43) \\ &= \prod_{t=1}^{N} p(y_t|u_t, p_t;\theta) \prod_{t=1}^{N} p(u_t|p_t;\theta) \qquad (44) \\ &= \prod_{t=1}^{N} p((y_t, u_t)|p_t;\theta) \qquad (45) \end{aligned}$$

where the last equality is simply the result of successively conditioning the joint process (y_t, u_t) on its past. The probability densities above involve the conditional random variable $y_t|(u_t, p_t)$ that is the usual output innovations process of the plant input–output model. The conditional random variable $u_t|p_t$ is the innovation of the feedback system with a required delay of one time step between y_t and u_t. Also the conditional random variable $(y_t, u_t)|p_t$ is the innovations process for the joint vector process (y_t, u_t). The joint likelihood function of (y_t, u_t) is expressed as the product of two terms that are thus independently distributed. Each of these terms is the product of probabilities of independently distributed innovations processes.

Notice that in the second term of the product there is no instantaneous effect of the output y_t on the feedback u_t so that there is a minimum delay of one time step. The reason for this is that in the joint model process (y_t, u_t) there is in general a correlation between the two components $y_t|p_t$ and $u_t|p_t$ of the innovations process $(y_t, u_t)|p_t$. The conditioning above expresses this correlation as an effect of the input u_t on the output y_t. This imposes a delay in the feedback between y_t and u_t. The two forms are exactly equivalent. The correlation of the joint model has been expressed as an input–output effect in the plant. These are just equivalent ways of expressing a contemporaneous effect in the system. If instead u_t is conditioned on y_t rather than the conditioning above of y_t on u_t, then the delay would have appeared in the plant. The delay can be arbitrarily selected to be in either the plant or feedback part of the system. There is no way of determining which is correct from the data without prior knowledge or some additional inputs in the system. One can write the system of equations either way and obtain an equivalent description of the system as discussed in Lütkepohl [39].

The above factoring of the likelihood function into two terms as in (44) and (45) always holds and is a consequence of simple conditional probability rules. The real usefulness comes, however, when the plant and feedback pieces of the system can be parametrized separately. Suppose that the parameter vector can be written as $\theta = (\theta_p, \theta_f)$ where the two subvectors respectively parametrize the plant and feedback parts of the systems. In this case, the maximum of the likelihood function is the product of the maximums of each of the two pieces. This allows the consideration of the plant–feedback structure for the process as opposed to a joint model with all possible relationships between the process y_t and u_t.

The advantage of this can be seen by considering the number of parameters required to estimate a feedback structure verses a joint model. Suppose that the true process has the plant–feedback structure with 4 states in each of the plant and the feedback parts with y_t and u_t each one dimensional. Then the number of required parameters to estimate the plant is 14 and to estimate tahe feedback part of the system is 13 for a total of 27 parameters. The joint system model requires 35 parameters. On the average it is to be expected that the AIC of the joint model will be larger by 8 than the AIC for the feedback structure model. On the other hand, suppose that the process is really not of the form of the feedback model. For example suppose that the true system is of state order 8 where all parameters are required to adequately describe the system. From the theory, 35 parameters are required to describe the system, and most of the time this would be the optimal description of the system to be identified. Then when each of the separate models for the plant and feedback part of the system is maximized, some common states may appear in each of these models since they are only considered separately. As a result, the total number of parameters will be significantly more than 35. In this way tests of hypotheses can be set up concerning the presence or absence of a feedback structure in the system.

Now notice that if the hypothesis of a feedback structure is correct, then the identification of the two separate pieces is done separately and as if the inputs to the respective pieces were uncorrelated with the past outputs of the system. In doing so, what is implicitly assumed is that

- There is a delay in the feedback from the output y_{t-1} to the input u_t.
- The input innnovations $u_t|p_t$ which are disturbances in the feedback part of the system are uncorrelated with the output innovations $y_t|(u_t,p_t)$ that are disturbances that are effecting the plant.
- The input innovations $u_t|p_t$ are nonzero.

Then ML identification of the system will yield the same result whether the presence of feedback is taken into account or not. The last condition above is necessary to assure identifiability and depends on the particular nature of the input innovation as well as the plant where persistent excitation will guarantee identifiability. The first two conditions guarantee that the various effects are correctly interpreted by the model. The delay can be assumed to be in the feedback and the above statements appropriately modified.

The analysis can be taken one step further to test the lack of feedback. In such a hypothesis, the model fitted for the feedback would be modified as $p(u_t|U_t^-)$ where U_t^- is the past of only u and not y so that there is no feedback from y_t to u_t. We could also test for the absence of any contemporaneous effect by setting the appropriate parameter to zero in the model fitting. Thus there are a number of different structures that can be fitted and compared to determine the most appropriate model for the system.

In some applications such as forecasting, it does not matter which of several equivalent delay structures is used, and the one used is arbitrary. In other applications however it may be necessary to know the delay structure or inject additional signals in the system so that it can be determined. For example if the identified model is to be used for the design of a control system and then modification of the system feedback, it is necessary to know the actual delay structure.

The above results easily generalize to dividing up a multivariate time series into more than two subsets of variables. Following the above procedure, tests for feedback between the various subsets can be constructed as in Larimore and Cavanaugh [33] and Larimore [51].

VIII. EXTENSIONS TO NONLINEAR SYSTEMS

In this survey of CVA methods, there is only space for a brief indication of the nature of the extension to nonlinear systems. This extension is quite surprising in the generality of the results. A detailed discussion of nonlinear CVA is contained in Larimore [29].

The definition of the Markov property of nonlinear processes is the same as in Section III above except that the state $x_t = C(p_t)$ is a nonlinear function of the past. The reduced order prediction problem involves the prediction of nonlinear functions $g(f_t)$ of the future using nonlinear functions $h(p_t)$ of the past.

First consider the problem where g and h are fixed functions and we wish to find the nonlinear function $\hat{g}(h(p))$ such that the relative prediction error

$$\|g(f) - \hat{g}(h(p))\|_{\Sigma_{gg}^{\dagger}} \doteq E\{[g(f) - \hat{g}(h(p))]^T \Sigma_{gg}^{\dagger}[g(f) - \hat{g}(h(p))]\} \tag{46}$$

is minimum, where (†) denotes the pseudoinverse operation. For a given pair of such functions, the optimal predictor is the conditional expectation of $g(f_t)$ given the random vector $h(p_t)$

$$\hat{g} = E\{g(f_t) | h(p_t)\} \tag{47}$$

The conditional expectation operator can be expressed in terms of projection operators in the Hilbert space of nonlinear functions of the past p_t and future f_t.

Now with the optimal prediction $\hat{g}$ given by (47), consider the r-rank nonlinear prediction problem of finding an r-dimensional nonlinear function $h(p)$ and an r-dimensional nonlinear function $g(f)$ of f so as to minimize (46). Specifically consider the following minimization problem.

(a) Rank r Nonlinear Prediction Problem. For a given positive integer r, find r-dimensional vector functions $h(p)$ and $g(f)$ minimizing the relative prediction error

$$\min_{(h,g)} \|g(f) - \hat{g}(h(p))\|_{\Sigma_{gg}^{\dagger}} \tag{48}$$

where $\hat{g} = E(g|h)$.

The optimal solution to this problem reduces to a sequence of univariate function optimization procedures involving maximal correlation. Consider random vectors $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_n)$. The *maximal correlation* of X and Y is defined as

$$\rho^*(X, Y) = \sup_{f,g} \rho(f(X), g(Y)) = \sup_{\substack{f,g \\ \|f\|=1 \\ \|g\|=1}} E[f(X)g(Y)] \tag{49}$$

where f and g run over all Borel measurable functions with zero mean, i.e., $Ef = Eg = 0$, and $\rho(f, g)$ is the correlation coefficient given in this case by $E[f(X)g(Y)]$.

From the theory of continuous operators on Hilbert spaces, under suitable assumptions on the probability distributions there always exist functions g and h for which the maximal correlation is attained. The alternating conditional expectation (ACE) algorithm of Brieman and Friedman [8] is available for computation of the optimal functions g and h by iteratively projecting g on the Hilbert space over p_t and h on the Hilbert space over f_t. ACE has been used extensively in nonlinear regression problems in many variables.

A nonlinear CVA procedure using less optimal methods than ACE was implemented in Larimore [29] and applied to the Lorenz chaotic attractor to obtain a nonlinear model of the process.

IX. APPLICATIONS

A. Chemical Process Control

The CVA approach has been applied to a number of complex multivariable chemical processes using both detailed simulation models as well as real process data. The study of Schaper *et al.* [43] includes a stirred tank reactor, an autothermal reactor, and a pilot scale distillation column. Other studies of chemical processes include an industrial recovery boiler by Kemna *et al.* [23], and the use of model predictive control along with CVA for on-line identification and control of a CSTR process by Kemna *et al.* [22]. The discussion below concerns the results of Schaper *et al.* [43].

A detailed simulation model of a 2-input, 2-output, continuous stirred tank reactor (CSTR) was used for data simulation to test the CVA algorithm. The dynamics were mildly nonlinear for the variations around the set point that were simulated. The step responses produced by the CVA identification were closer to the true response than those of the theoretical linearized model. CVA was shown to handle time delays by augmenting the system state order.

The tubular autothermal reactor exhibits virtually every process nonideality encountered in practice including (1) both stable and unstable steady-state operation, (2) both inverse response (nonminimum phase) and oscillatory dynamic behavior, (3) very nonlinear behavior resulting from exponential dependence of reaction on temperature, (4) distributed process characteristics, (5) highly coupled state variable structure due to internal energy flow paths, (6) inaccessible concentration state variables, (7) practical controllability and observability problems, (8) stiff equations (a need for two time scale techniques), and (9) a relatively high process noise. The detailed simulation of a 2-input, 2-output system involved 36 nonlinear ODEs. CVA identified a 6-state system with the AIC indicating that there was no further significant model structure. In contrast with earlier studies of other identification methods, CVA was able to capture the extreme stiffness of the system dynamics.

B. Modeling Vibrating Systems

In this section, first an example of model identification from shake table data is given, and then a comparison of CVA with the eigensystem realization algorithm (ERA) for a feedback system is discussed.

To illustrate the power of the CVA method, its application to data from a shake table is discussed. There are 2 outputs and 1 input to the system. The CVA procedure was applied to 600 time observations. The canonical variate analysis decomposed the system into 117 states. The AIC is plotted in Fig. 4 with a minimum at 29 states which was chosen as the optimal model order and involved the estimation of 152 parameters. Since there were only 600 time points used, this is a small sample case for system identification.

The frequency response function of the identified system from input 1 to output 1 is given in Fig. 5 along with the probability 0.999 confidence bands. These bands are simultaneous so that the probability is 0.999 that the true transfer function is contained within the bands over all frequencies. A detailed discussion of the confidence band procedure and

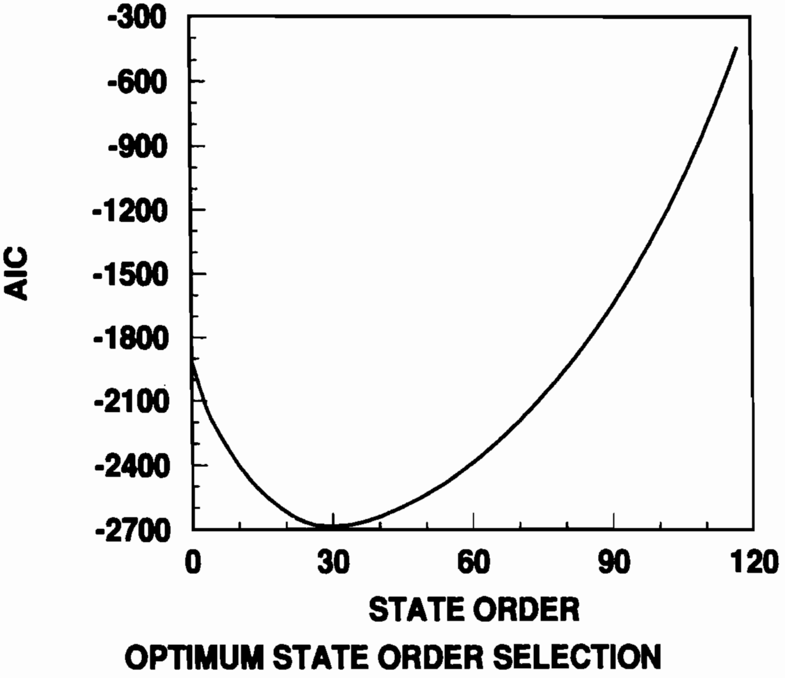


Figure 4 AIC versus model state order.

interpretation is contained in Larimore [30]. The accuracy of the CVA identified model was comparable to a 32,000 sample FFT where FFTs were performed on batches of 512 points using a Hanning window and averaged for batches with 50 percent overlap. The FFT thus required about 50 times as much data and it appeared that the CVA procedure was still more accurate. This is due to the adaptive nature of the spectral estimation. Note how the width of the confidence bands increases where there are very sharp peaks or valleys in the curve while it can be very narrow where it is slowly changing.

In another study, Hunter [16] did an extensive comparison of CVA and the eigensystem realization algorithm (ERA) of Juang and Pappas [20] in simulations of a vibrating structure. The structure was a finite element model of a lithographic stage including actuator and measurement dynamics with a total of 84 states, 3 input actuators, and 3 output sensors. Some of the modes were lightly damped, and four signal-to-noise ratios were considered of infinite (no noise), 6 Db, 0 Db, and −6 Db with the

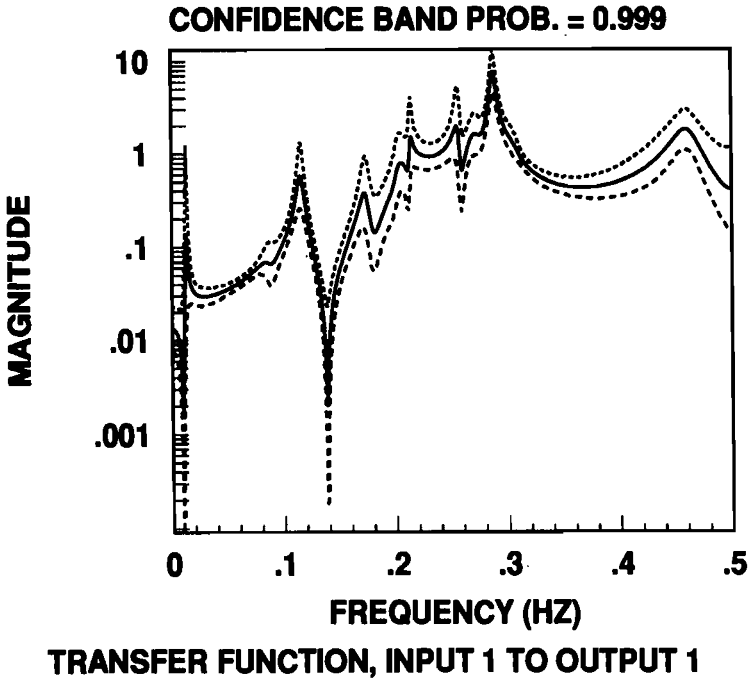


Figure 5 Identified transfer function and confidence bands.

noise exciting the system state. The effects of active control were considered including rate feedback, linear quadratic Gaussian feedback, or no feedback. Two levels of the feedback gains were evaluated for each combination of signal-to-noise level and feedback type. For each case, time series of 988,400 points were simulated and used for doing model identification using CVA and ERA.

In evaluating the identified model accuracy, the error measure used was the RMS value of the difference between the theoretical impulse response and the identified model impulse response expressed as a percent error. The error measure for CVA was between 0.1 and 0.25 of the value for ERA. CVA was typically 7 times more accurate than ERA, and for ERA to achieve the same accuracy by using more data would require 49 times the sample size since RMS error varies with the square root of sample size. ERA has a number of similarities with CVA in that a major part of the computation involves an SVD. However, the first crucial step of ERA involves the estimation of the frequency response function by averaging FFTs of the data. This first step is a nonparametric modeling method that is known to be far less accurate than parametric identification methods.

C. Process Monitoring and Fault Detection

There have recently been a number of studies of process monitoring and fault detection trying to isolate and use a low dimensional subspace where the process is under normal operation. The chemometric methods of partial least squares and principal component analysis have been used to try to identify such a subspace in Kourti and MacGregor [24], Nomikos and MacGregor [41], and MacGregor and Kourti [40]. These methods have been mostly developed and applied for static regression problems and not dynamical systems. The CVA method provides a precise and rigorous approach to identifying the low dimensional state space that contains the process behavior apart from random observation noise.

The approach is to identify the system state space and then look for anomalous behavior either within the state space or behavior that significantly departs from the state space. Let tc be the time t at which a change in the process occurs and denote the future outputs $f_{tc}^T = (y_{tc}^T, y_{tc+1}^T, \ldots, y_{tb}^T)$ and similarly for the future inputs q_{tc}. The future canonical variables d are given by L times the future outputs f_{tc} conditional on the future inputs q_{tc} and have the form

$$d_{tc} = L(f_{tc} - Mq_{tc}) \tag{50}$$

The matrix M has blocks consisting of the system impulse response function of the form $M_{ij} = H\Phi^{i-j}G$ for $i>j$, $M_{ij} = 0$ for $i<j$, and block diagonal $M_{ii} = HA$. Then under the null hypothesis, the future canonical variables d_{tc} lie in a low dimensional subspace with an expected response due to the state x_{tc} at time tc. The dimension of this subspace is equal to the state dimension. Also recall that from the Markov property of the process dynamics, the prediction of the future from the state x_{tc} is identical to the prediction using the past p_{tc} consisting of all past process inputs and outputs. The prediction of the future with zero future inputs is simply $y_{t+i} = H\Phi^i x_t$. The prediction of the future canonical variables d from the past canonical variables that are precisely the states x_{tc} is simply Dx_{tc} where D is the diagonal matrix of the canonical correlations γ_i. From the CVA analysis, the error

$$d_{tc} - Dx_{tc} \tag{51}$$

in predicting the future canonical variables d_i has covariance matrix $I - D^2$. Often at least some of the canonical correlations are close to 1 so that the corresponding prediction error $1 - \gamma_i^2$ for say the future canonical variable d_i is quite small. A disturbance or change in the system can effect the future canonical variables in many different ways. However the CVA takes advantage of the correlation structure in the process to reduce the variance in the future canonical variables from 1 to $1-\gamma_i^2$. The future canonical variables d_i with index i larger than the state order all have canonical correlation $\gamma_i = 0$ and variance 1 and are mutually uncorrelated, and there is no reduction in prediction error using information in the state or the past.

Several tests of the hypothesis of no change in the process can be constructed. From the above discussion, a natural statistic for testing change in the process is the T^2 statistic

$$T^2 = (d_{tc} - Dx_{tc})^T(I - D^2)(d_{tc} - Dx_{tc}) \tag{52}$$

Also we can partition the error vector (52) and the covariance matrix $I - D^2$ to contain the first k and last $l-k$ elements and obtain two additional statistics denoted T_1^2 and T_2^2. The statistic T_1^2 can be used to test that the process satisfies the state space model on the state space subspace, while the statistic T_2^2 can test that the process lies on the state space subspace. The statistic T_1^2 is distributed as a Chi-squared random variable with k degrees of freedom while T_2^2 is Chi-squared with $l-k$ degrees of freedom. Typically l is much larger than k so that $T^2 = T_1^2 + T_2^2$ contains a lot of "noise" in the term T_2^2 in observing any departure in the process dynamics on the state space subspace. This is the advantage of using T_1^2. If changes cause a response off the state space subspace that is large enough compared to the high dimensional noise in T_2^2, then it will be detected in the statistic T_2^2. Thus the CVA decomposition of the past/future can be used to construct tests of hypotheses to provide a transparent and powerful means for checking if the behavior of the process is in control or out of control.

Suppose that such a test has indicated that a possible change has occurred. With more future data, a test can be developed to precisely test the hypothesis that a change in the process has occurred. In particular, the change model consists of two models that are fitted to the process respectively before and after the suspected change. This is compared to fitting a single model over the whole data set to determine if this is a better description of the data than the change model. The primary tools for performing such a test is the CVA model identification that gives ML models and the Akaike information criteria (AIC) for comparing model fit. The approach is quite simple. The data are split into two data intervals D_1 and D_2 that may have a gap between them. CVA is used to fit dynamical models M_1 and M_2 respectively to each of the data set D_1 and D_2. Also the model M_{1+2} is fitted to the union $D_{1+2} = D_1 \cup D_2$ of the data sets. Then the AIC of the composite model consisting of two models $[M_1, M_2]$ is simply the sum of the AICs

$$\mathrm{AIC}[M_1, M_2] = \mathrm{AIC}[M_1] + \mathrm{AIC}[M_2] \tag{53}$$

This is compared with $\mathrm{AIC}[M_{1+2}]$, and the smaller is the preferred model. Keep in mind that the standard error of the sampling variability is of the order of

$$\sqrt{2(D_1 + D_2 - D_{1+2})} \tag{54}$$

where D_i is the number of estimated parameters in model M_i. If the difference is of this order, then there is no clear preference for the change or no change models.

The above test has been applied to detection of the presence of disturbances in simulations of a continuous stirred tank reactor in Wang, Seborg and Larimore [48]. The disturbances included ramps, steps, and sine waves. The CVA procedure was found to be much more sensitive for detection of disturbances than standard Shewhart charts or more recent principal component analysis (PCA) methods. The above monitoring method using CVA was also applied by Wang, Seborg and Larimore [47] to the Tennessee Eastman Challenge Problem and compared with PCA and Shewhart charts. CVA monitoring was able to detect the presence of disturbances missed by the standard monitoring methods.

D. On-Line Adaptive Control of Unstable Wing Flutter

A dramatic example of the use of CVA identification and LQG control design is discussed in Peloubet *et al.* [42]. The problem is to design a fully adaptive control system to stabilize unstable aircraft wing flutter:

- No prior assumptions on the system dynamics or state order.
- Identify a model of the system dynamics strictly from observational data.
- Design a feedback controller based on the identified model.
- Adapt to changes in dynamics due to changes in aircraft speed and altitude.
- Adapt to changes in dynamics due to changes in wing stores (dropping stores or changing fuel in wing tanks).
- When the wing flutter is unstable (above a particular aircraft speed), the controller must stabilize the flutter.

The system was demonstrated in a wind tunnel test at NASA Langley. Previous flutter suppression systems were nonadaptive, or only semiadaptive where some prior knowledge of the system dynamics was assumed.

The adaptive control system consisted of:

- CVA for on-line automated system identification.
- CVA also provides a Kalman filter state estimator for use in LQG control.
- Linear quadratic (LQ) control design implemented on-line using the CVA identified state space model.
- Linear quadratic control (LQG) feedback control using the Kalman filter state estimate and LQ control.

The CVA identification and LQ control design were typically implemented once per second using the last 8 seconds of data sampled at 100 Hz. This was accomplished with a 40 MFLOPS computation rate. The state estimates and feedback control were computed at the 100 Hz rate to provide the feedback.

Wind tunnel tests were conducted at the NASA Langley 16 Foot Transonic Dynamics Tunnel. Extensive tests were conducted over a 3.5 week period for 6–8 hours per day involving three store configurations with unstable flutter. The aircraft model was a $\frac{1}{4}$ scale full span model with a 10 foot wing span. The aircraft system inputs were two flaperon control surfaces, one on each wing, and 6 accelerometer sensor outputs, 3 on each wing. Speeds of up to 30 percent above the critical unstable flutter speed were tested for all three store configurations.

Over 100,000 dynamical system models were identified, controllers designed, and feedback implemented with no failures of the algorithms until the last day, as discussed in the next section. Most of the tests were conducted above the critical flutter speed where the dynamics were unstable. Consequently the identified models typically were unstable and had state orders between 15 and 30 states. The identified model provided a disturbance model for state and measurement noise as well as the input/output dynamics. The disturbance model provided the basis for the Kalman filter state estimator with good disturbance rejection.

Tests were conducted to evaluate the ability of the adaptive control system to adapt to changes in the flutter dynamics. One test was a rapid decleration of the wind tunnel speed which changes flutter dynamics. In another test, the adaptive controller was restarted with no prior knowledge of the system dynamics. In this case, the first identified model used 1 second of data, the second used 2 seconds, etc., until reaching the use of 8 seconds of past data for identification. In some cases it took several seconds of data before the controller achieved stabilizing feedback. Several tests were done where stores were dropped from the wing so that there was an instant change in dynamics as well as an impulse of energy input into the unstable system. Video cameras clearly showed growing vibration of the wingd due to the unstable dynamics and no stabilizing control feedback. It took several seconds of data following the store drop before the CVA identification could determine a sufficiently accurate model so that the corresponding controller would stabilize the system.

E. Robust Adaptive Control

At the end of the wind tunnel tests of the adaptive control of wing flutter described in the previous section, the system went unstable and the wing broke.

In subsequent analysis, the cause of the instability in the controller was found to be due to the uncertainty in the identified model exceeding the controller robustness. This was due to the system identification not providing any measure of the identified model accuracy, and the controller not assessing robustness relative to the model uncertainty. Currently available CVA identification software discussed in Larimore [32] provides confidence bands on the estimated frequency response function developed in Larimore [30] that could be used for at least assessing the robustness of the controller relative to this model uncertainty.

There are several potential approaches to robust adaptive control using the model uncertainty from the identified model:

- Check robustness. Determine if the designed controller has sufficient robustness to contain the identified model uncertainty.
- Robust controller design. Use the identified model uncertainty in the controller design.
- Change input excitation. Determine how the system input excitation spectrum could be changed to reduce the identified model uncertainty with a corresponding increase in controller robustness relative to the model uncertainty.

The information available from CVA identification on model accuracy can thus be used in several ways from simply checking robustness, using it in robust control design, or determining how to change the input excitation.

ACKNOWLEDGMENT

Appreciation is given to Dr Norman Hunter of Los Alamos National Laboratory for the shake table data.

REFERENCES

1. H Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. BN Petrov, F Csaki, eds. 2nd International Symposium on Information Theory, Budapest: Akademiai Kiado, 1973, pp 267–281.
2. H Akaike. IEEE Trans Automatic Control 19(no.6):667–674, 1974.
3. H Akaike. SIAM J Control 13(no.1):162–173, 1975.
4. H Akaike. Canonical Correlation Analysis of Time Series and the Use of an Information Criterion. RK Mehra, DG Lainiotis, eds, System Identification: Advances and Case Studies. New York: Academic Press, 1976, pp 27–96.
5. TW Anderson. An Introduction to Multivariate Statistical Analysis. 2nd ed. New York: Wiley, 1984.
6. RR Bitmead, M Gevers, V Wertz. Adaptive Optimal Control, a Thinking Man's GPC. New York: Prentice Hall, 1990.
7. GEP Box, GM Jenkins. Time Series Analysis Forecasting and Control. San Francisco: Holden-Day, 1976.
8. L Breiman, JH Friedman. J Amer Stat Assoc 80:580–598, 1985.
9. JV Candy, TE Bullock, ME Warren. Automatica 15:493–495, 1979.
10. YM Cho, G Xu, T Kailath. Trans Automatic Control 39:2004–2017, 1994.
11. M Deistler, K Peternell, W Scherrer. Automatica 31:1865–1875, 1995.
12. IM Gelfand, AM Yaglom. Amer Math Soc Trans Series (2) 12:199–236, 1959 (original Usp Mat Nauk 12:3–52, 1956).
13. M Gevers, V Wertz. On the Problem of Structure Selection for the Identification of Stationary Stochastic Processes. Sixth IFAC Symposium on Identification and System Parameter Estimation, G Bekey, G Saridis, eds. Washington, DC: McGregor & Werner, 1982, pp 387–392.
14. I Gustavsson, L Ljung, T Söderström. Automatica 13:59–75, 1977.
15. H Hotelling. Biometrika 28:321–377, 1936.
16. N Hunter. A Comparison of State Model Estimation Using Canonical Variate Analysis and Eigensystem Realization Analysis. Los Alamos Unclassified Report LAUR-95-1275, 1995.
17. CM Hurvich, R Shumway, CL Tsai. Biometrika 77:709–720, 1990.
18. CM Hurvich, CL Tsai. Biometrika 76:297–307, 1989.
19. CM Hurvich, CL Tsai. Biometrika 78:499–510, 1991.
20. J Juang, R Pappas. J Guidance, Control and Dynamics 8:620–627, 1985.
21. T Kailath. IEEE Trans Inform Theory 20:146–181, 1974.
22. AK Kemna, WE Larimore, DE Seborg, DA Mellichamp. On-line Multivariable Identification and Control of Chemical Processes Using Canonical Variate Analysis. Proc 1994 American Control Conference, June 28–July 1, 1994, Baltimore, MD.
23. AK Kemna, DA Mellichamp, DE Seborg, N Sweerus. Dynamic Models of Recovery Boilers from Input-Output Data. 1992 International Chemical Recovery Conference, June 7–11, 1992, Seattle, WA.
24. T Kourti, JF MacGregor. J Quality Tech (in press).
25. S Kullback. Information Theory and Statistics. New York: Dover, 1959.

26. WE Larimore. Biometrika 70:175–181, 1983.
27. WE Larimore. System Identification, Reduced-Order Filtering and Modeling Via Canonical Variate Analysis. Proc 1983 American Control Conference, HS Rao, T Dorato, eds, 1983, pp 445–451.
28. WE Larimore. Canonical Variate Analysis for System Identification, Filtering, and Adaptive Control. Proc 29th IEEE Conference on Decision and Control, Honolulu, Hawaii, December, Vol. 1, 1990, pp 635–639.
29. WE Larimore. In: M Casdagli, S Eubank, eds. Nonlinear Modeling and Forecasting. Reading, MA: Addison-Wesley, 1991.
30. WE Larimore. Accuracy Confidence Bands Including the Bias of Model Underfitting. Proc 1993 American Control Conference, June 2–4, 1993, San Francisco, CA.
31. WE Larimore. The Optimality of Canonical Variate Identification by Example. 10th IFAC Symposium on System Identification, Copenhagen, 4–6 July, 1994.
32. WE Larimore. ADAPT$_X$ Automated System Identification Software Users Manual. Adaptics, Inc, 40 Fairchild Drive, Reading, MA 01867, 1996.
33. WE Larimore, KL Cavanaugh. A Canonical Variate Approach to the Analysis of Causality and Feedback in Multiple Time Series. Annual Meeting of the American Statistical Association, Chicago, IL, August 18–20, 1986.
34. WE Larimore, FT Luk. System Identification and Control Using SVDs on Systolic Arrays. SPIE Symposium on Innovative Science and Technology, Proc of Conference on High Speed Computing, Vol. 880, January, 1988.
35. WE Larimore, S Mahmood, RK Mehra. Multivariable Adaptive Model Algorithmic Control, Proc Conference on Decision and Control. AH Haddad, M Polis, eds, Vol. 2, 1984, pp 675–680.
36. WE Larimore, RK Mehra. Byte 10:167–180, 1985.
37. A Lindquist, M Pavon. Markovian Representation of Discrete-Time Stationary Stochastic Vector Processes. 20th IEEE Conf on Decision and Control, Vol. 3, 1981, pp 1345–1356.
38. L Ljung, T Söderström. Theory and Practice of Recursive Identification. Cambridge: MIT Press, 1983.
39. H Lütkepohl. Introduction to Multiple Time Series Analysis. New York: Springer-Verlag, 1983.
40. JF MacGregor, T Kourti. Control Eng Practice 3:403–414, 1995.
41. P Nomikos, JF MacGregor. Technometrics 37: 1995.
42. RP Peloubet, RL Haller, RM Bolding. On-line Adaptive Control of Unstable Aircraft Wing Flutter. Proc 29th IEEE Conference on Decision and Control, Honolulu, Hawaii, December 1990, Vol. 1, pp 643–651.
43. CD Schaper, WE Larimore, DE Seborg, DA Mellichamp. Comp Chem Eng 18:55–69, 1994.
44. R Shibata. Ann Statistics 9:300–306, 1981.
45. P Van Overschee, B De Moor. A Unifying Theorem for Three Subspace System Identification Algorithms. American Control Conference, June 29–July 1, 1994, Baltimore, MD, pp 1645–1649.

46. M Verhaegen. Automatica 30:61–74, 1994.
47. Y Wang, DE Seborg, WE Larimore. Process Monitoring based on Canonical Variate Analysis: Application to the Tennessee Eastman Problem. Submitted for publication.
48. Y Wang, DE Seborg, WE Larimore. Process Monitoring based on Canonical Variate Analysis. IFAC 1997 Int Symp on Advanced Control of Chemical Processes to be held June 9–11, 1997, Banff, Canada.
49. AM Yaglom. Outline of Some Topics in Linear Extrapolation of Stationary Random Processes. Proc Fifth Berkeley Symp Math Stat and Prob, 1970. Berkeley, California: California Press, pp 259–278.
50. WE Larimore. Optimal Reduced Rank Modeling, Prediction, Monitoring, and Control using Canonical Variate Analysis. IFAC 1997 Int. Symp on Advanced Control of Chemical Processes to be held June 9–11, 1997, Banff, Canada.
51. WE Larimore. System Identification of Feedback and "Causality" Structure using Canonical Variate Analysis. 11th IFAC Symposium on System Identification to be held July 8–11, 1997, Fukuoka, Japan.

This Page Intentionally Left Blank

4

Models in Generalized MA Form for Identification of Continuous-Time Systems

Ganti Prasada Rao*
Indian Institute of Technology, Kharagpur, India

A. V. B. Subrahmanyam
Computervision R&D (I) Pvt. Ltd., Pune, India

I. INTRODUCTION

System identification is concerned with the determination of an appropriate mathematical model which adequately describes the input–output characteristics of a system under consideration. The existing literature on the subject includes a large part which is based on treatment in discrete-time (DT) [1–4] and a considerable part which advocates and describes several approaches based on continuous-time (CT) for identification of physical systems. An elaborate justification for the choice of CT models for physical systems and several methods of identification for a variety of CT model forms may be found in Unbehauen and Rao [5] and Sinha and Rao [6]. This chapter is concerned with identification of dynamic systems in terms of CT models in a particular form which is termed "generalized moving average" for its structural similarity with its discrete-time (DT) counterpart. The term does not reflect the literal sense of moving average in this case.

Consider a CT linear time-invariant asymptotically stable dynamical system $\mathcal{G}$ with input $u(t)$ and output $y(t)$. The input–output description of the system in terms of its unknown transfer function $G^0(p)$, is

$$y(t) = G^0(p)\,u(t) + v(t) \tag{1}$$

*Presently on deputation at the International Foundation for Water Science and Technology, PO Box 25862, Abu Dhabi, U.A.E.

Table 1 Various Choices of the Dynamic Operator δ for CT Modeling

	δ	Description
1	q^{-1}	Backward shift operator
2	$\dfrac{1-q^{-1}}{q^{-1}T_s}$	Forward difference-based approximation
3	$\dfrac{2}{T_s}\dfrac{1+q^{-1}}{1-q^{-1}}$	Trapezoidal approximation
4	$\dfrac{1-q^{-1}}{T_s}$	Backward difference-based approximation

The term $G^0(p)u(t)$ represents the component of response of the system to the input $u(t)$, $v(t)$ represents the stochastic part of $y(t)$, and p is the differential operator. In many practical situations, only discrete samples of the input and output signals are available, requiring an approximate version of the CT model in an appropriate time domain. Consider the set of measurements sampled at equal intervals of length T_S.

$$\mathbf{Y}^N = \{u(k), y(k), k = 1, \ldots, N\} \tag{2}$$

Given $\mathbf{Y}^N$ and some prior knowledge of the dynamics of $\mathcal{G}$, the system identification problem is to obtain $G(\delta, \theta)$, an estimate of the transfer function which best describes the dynamics of $\mathcal{G}$ in some sense by minimizing a chosen norm of the modeling error. δ represents the time domain of approximation of the model. $\theta \in \mathbf{R}^n$ is the parameter vector. Some simple approximations of the CT domain are listed in Table 1 as functions of the backward shift operator q^{-1}.

In terms of $G(\delta, \theta)$, the input–output description becomes

$$y(k) = G(\delta, \theta)u(k) + H(\delta, \theta)e(k) \tag{3}$$

in which the second term accounts for the combined effects of $v(t)$, unmodelled dynamics (due to model simplification), and possibly of unknown initial conditions. This term is generally called the noise model.

The focus of our attention in the present treatment is on the first term. The treatment is also applicable to the second term.

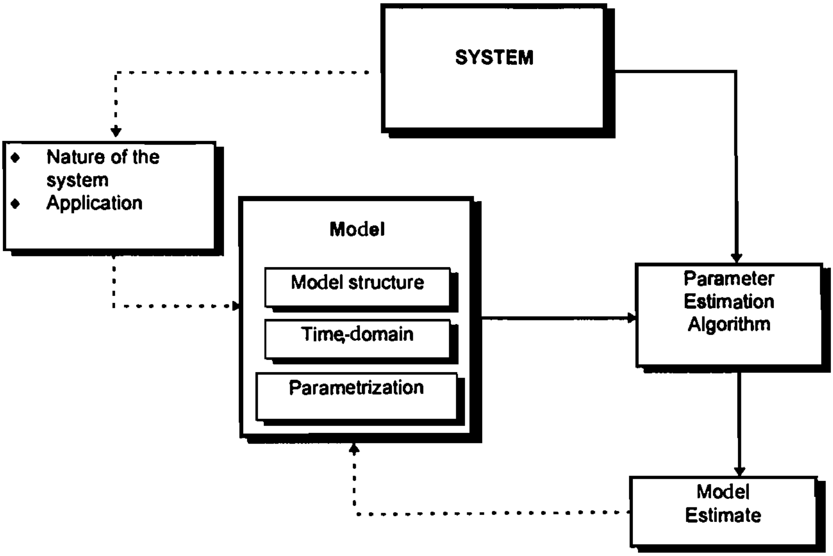


Figure 1 Role of model structure in parametric system identification.

II. IDENTIFICATION OF PARAMETRIC MODELS

In the process of identification by parametric models, it is necessary to select a model with a suitable structural form characterized by an appropriate set of parameters. In the case of CT systems, the handling of the time domain is additionally important. The complexity and quality of the estimation procedure naturally depend on the chosen model structure. It is therefore desirable to make a choice that leads to reasonably good estimates of the system even with very simple parameter estimation approaches. This point is explained in Fig. 1. The measured input–output signals, the specific form of the model, and the related set of parameters happen to be the inputs to a parameter estimation algorithm which gives the estimated model parameters of the model.

The choice of the model (structure and parametrization in a time domain) is governed by the nature of the system under test, as understood from the quanitative or quantitative prior knowledge of the system and the intended application of the estimated model. The quality of the model estimate with a given parameter estimation approach in turn depends on the chosen model (structure and parametrization in a time domain).

III. SYSTEM IDENTIFICATION IN CONTINUOUS-TIME DOMAIN

CT dynamic systems have traditionally been modeled by linear differential equations in the time domain, or rational transfer functions in the complex variable s in the frequency domain. Some models can be processed on digital computers only after discretizing the CT operators and process signals. Consequently, CT systems have been subjected to a process of discretization which gives rise to difference equations in the time domain, or rational transfer functions in the complex variable z in the frequency domain. These models have been in wide use mainly for their simple digital appeal and the availability of analytical tools until the virtue of the CT treatment became evident in the context of identification and control [6–9]. CT approaches avoid the following setbacks which are inherent in the traditional DT approaches:

1. High sensitivity of the model coefficients which is associated with rapid sampling and gives rise to problems in parameter estimation and control,
2. Clustering of poles near (1,0) in the z-plane at high rates of sampling leading to undesirable numerical ill-conditioning,
3. Creation of a non-minimum phase feature which is unnatural and problematic,
4. Loss of uniqueness of models due to their dependence on sampling time,
5. Loss of physical significance of parameters due to discretization, and
6. Loss of available knowledge in the discretization process in partially known CT systems.

Mukhopadhyay *et al.* [10] discuss these in greater detail and cite further literature on this aspect. Investigations in this direction have led to attractive approaches to the DT approximation of CT models. A particular feature of these approaches is that the parameters in the resultant DT approximation retain their strong relationship, which is asymptotically an identify, with those of the original CT version [5,6] and the resulting DT domain is in harmony with its CT counterpart. In these approaches, the DT (sampled or averaged) measurements, and the filtered versions of signals that arise in CT modeling, are computed by way of DT approximations of the CT differentiation operator $p = \mathrm{d}/\mathrm{d}t$, such as items 2–4 of Table 1.

IV. PARAMETRIC FORMS OF CONTINUOUS-TIME MODELS

A general parametric form of description for Eq. (3) is the polynomial black-box model

$$A(\delta,\theta)y(k) = \frac{B(\delta,\theta)}{F(\delta,\theta)}u(k) + \frac{C(\delta,\theta)}{D(\delta,\theta)}e(k) \tag{4}$$

that is the CT counterpart of the DT Box–Jenkins version [11,12]. In this, $u(k)$ and $y(k)$ are the samples of input and output signals, respectively, and $e(k)$ is a sequence of independent and uniformly distributed zero mean random variables. $A(\delta,\theta)$, $B(\delta,\theta)$, $C(\delta,\theta)$, $D(\delta,\theta)$, and $F(\delta,\theta)$ are polynomials in δ whose coefficients are arranged to form the parameter vector θ. Specific cases of these polynomials lead to particular models such as autoregressive (AR), moving average (MA), autoregressive and moving average (ARMA), and so on. In particular, to characterize stationary stochastic processes, the following ARMA model is considered.

$$A(\delta,\theta)y(k) = C(\delta,\theta)e(k) \tag{5}$$

where

$$A(\delta,\theta) = \delta^{nA} + a_1\delta^{nA-1} + \ldots + a_{nA},$$
$$C(\delta,\theta) = c_0\delta^{nC} + c_1\delta^{nC-1} + \ldots + c_{nC}$$

whose coefficients appear in the AR and MA portions, respectively, of the model. In the context of the discussion in this chapter, the terms AR, MA, and ARMA refer to the CT context, where the operator δ denotes an approximation to the CT differentiation operator, and not the usual backward shift operator.

The ARMA model of Eq. (5) is commonly used in spectral estimation and time series analysis. In dynamic system identification, where the goal is to characterize the dynamic input–output relation of the underlying process, the following model is suitable:

$$A(\delta,\theta)y(k) = B(\delta,\theta)u(k) \tag{6}$$

where $A(\delta,\theta) = \delta^n + a_1\delta^{n-1} + \ldots + a_n$, and $B(\delta,\theta) = b_1\delta^{n-1} + b_2\delta^{n-2} + \ldots + b_n$. Here the MA portion is formed from the process input signal which is usually known and measurable. Hence, in the sequel, with a little abuse of notation, this model is also termed as an ARMA model. We refer to

$$G(\delta,\theta) = \frac{B(\delta,\theta)}{A(\delta,\theta)} \tag{7}$$

as "deterministic ARMA" in the same vein. This model attempts to represent the dynamics of $G^0(p)$ through a proper rational transfer function having insightful features such as poles and zeros. However, the nonlinear-in-parameters nature of this model structure is not in favour of simplicity in parameter estimation.

With this model structure, the model output error (OE) in sampled form is

$$\varepsilon_{\mathrm{OE}}(k) = y(k) - \frac{B(\delta,\theta)}{A(\delta,\theta)}u(k) \tag{8}$$

A parameter estimation criterion is to minimize

$$J_{\mathrm{OE}}(k) = \sum_{k=1}^{N} \varepsilon_{\mathrm{OE}}(k)^2 \tag{9}$$

with respect to θ. Since the output error of Eq. (8) is nonlinear-in-parameters, this happens to be a problem of nonlinear optimization. In an attempt to simplify the situation, most of the identification approaches resort to the equation error (EE).

$$\varepsilon_{\mathrm{EE}}(k) = \frac{A(\delta,\theta)}{E(\delta)}y(k) - \frac{B(\delta,\theta)}{E(\delta)}u(k) \tag{10}$$

and a criterion

$$J_{\mathrm{EE}}(k) = \sum_{k=1}^{N} \varepsilon_{\mathrm{EE}}(k)^2 \tag{11}$$

Here $1/E(\delta)$ is a linear-dynamic operator of adequate order for the removal of the need for direct differentiation of process data [12]. These operators also serve the purpose of prefilters used for removing unimportant frequencies from the process data.

Since Eq. (10) is linear-in-parameters, parameter estimation is simplified to linear recursive least squares (RLS) estimation. However, EE minimization has its disadvantages.

- Biased estimation. The parameter estimates will be biased when the EE is not white [1]. Variants of the ordinary least squares (LS) algorithm such as "generalized least-squares" and "instrumental variables" [3] are applied to remove the bias. These and other "bias compensating least-squares" methods [13–15] are computationally demanding [16]. These approaches assume that the measurements are actually generated by an ARMA model, and that the measurement noise is Gaussian distributed. However, the performance of some of these may not be satisfactory when there is significant modeling error, as this component of error may not be Gaussian distributed.

- Reducible models (for MIMO systems). Consider a v_i-input, v_o-output system described by the transfer function matrix (TFM)

$$G(\delta,\theta)=\begin{bmatrix} G_{11}(\delta,\theta) & G_{12}(\delta,\theta) & \dots & G_{1v_i}(\delta,\theta) \\ G_{21}(\delta,\theta) & G_{22}(\delta,\theta) & \dots & G_{2v_i}(\delta,\theta) \\ \vdots & \vdots & \ddots & \vdots \\ G_{v_o1}(\delta,\theta) & G_{v_o2}(\delta,\theta) & \dots & G_{v_ov_i}(\delta,\theta) \end{bmatrix} \tag{12}$$

where,

$$G_{ij}(\delta,\theta)=\frac{B_{ij}(\delta,\theta)}{A_{ij}(\delta,\theta)}$$

EE formulation necessitates a canonical form having a least common denominator (CD) of all the elements of the TFM. The CD considerably inflates the unknown parameter vector. To reduce this inflation partially, the TFM is decomposed into multiple input–single output (MISO) submodels with several CDs limited only to the rows of the TFM. In this way, a two-stage algorithm was proposed by Diekmann and Unbehauen [17] for DT model identification, and its CT version by Mukhopadhyay *et al.* [16]. A Gauss–Seidel type iterative algorithm that does not require a CD was suggested by Rao *et al.* [18]. The approaches to estimation of irreducible CT models are surveyed in [16,19].

- Distribution of estimation errors. Modeling of physical processes is synonymous with approximation. The approximation is tantamount to undermodeling. This coupled with noisy process data results in biased estimates. Though it is possible to eliminate bias due to measurement noise, the bias resulting from undermodeling cannot be eliminated at all. It can be distributed over a range of frequencies by careful design of the identification experiment [20] such that such undermodeling is not harmful in the context of the final application of the resulting model. With ARMA modeling, the problem of experiment design for a prescribed distribution of bias over a range of frequencies is not simple and straightforward.

Using Parseval's theorem, the frequency domain description of the EE criterion (11) in the limit as $T_S \to \infty$, is

$$J_{\mathrm{EE}}(\omega)=\int_0^\infty \left| \frac{A(\mathrm{j}\omega,\theta)}{E(\mathrm{j}\omega)}\, U(\mathrm{j}\omega) \right|^2 \left| G^0(\mathrm{j}\omega)-\frac{B(\mathrm{j}\omega,\theta)}{A(\mathrm{j}\omega,\theta)} \right|^2 \mathrm{d}\omega \tag{13}$$

where $U(j\omega)$ is the Fourier transform of the input signal. The first term on the right-hand side of (13) is a weighting function that manipulates the second term (bias) over a range of frequencies. With the chosen ARMA model structure, it is clear that this weighting function is a function of the yet-unknown $A(\delta, \theta)$ which renders on-line experiment design impossible. Off-line design, however, is shown to be possible [21].

These problems are simplified, as will be clarified in the following section, when one considers linear-in-parameters models. Note that a parametrization $G(\delta, \theta)$ is linear if second and higher order derivatives of $G(\delta, \theta)$ with respect to θ vanish for all θ and linearity of a parametrization is different from the linearity of the model in terms of its input–output behavior. Even nonlinear models can be linearly parametrized.

In recent years, there has been a considerable amount of literature on linear-in-parameters models in the context of identification and approximation of dynamical systems. An attempt to unify various model forms which are linear-in-parameters is made in the following under a "generalized moving-average modeling" framework.

V. MODELS IN GENERALIZED MOVING-AVERAGE FORM

One situation in which the ARMA model of Eq. (6) is linearized (with respect to θ) is when its denominator $A(\delta, \theta)$ is fixed as some appropriate $A(\delta)$ which leads to the description

$$G(\delta, \theta) = \sum_{i=1}^{n-1} \frac{b_i \delta^{n-i}}{A(\delta)} = \theta^T \mathscr{B}(\delta) \tag{14}$$

in which $\theta = [b_1, b_2, \ldots, b_n]^T$ and $\mathscr{B}(\delta) = 1/A(\delta)\ [\delta^{n-1}, \delta^{n-2}, \ldots, \delta^0]$. A linear-in-parameters model is therefore obtained as:

$$y(k) = \frac{B(\delta, \theta)}{A(\delta)} u(k) \tag{15}$$

whereby EE = OE and estimation (minimization) is linear. This leads to an advantageous situation with the following possibilities.

- Robust estimation. In the limit as $N \rightarrow \infty$, the LS estimate $\hat{\theta}$ in the presence of zero-mean disturbances tends to $\hat{\theta}$, where $\hat{\theta}^*$ is the limiting estimate in the absence of disturbances. In particular, if the disturbance term is Gaussian and there is no modeling error, the LS estimate $\hat{\theta}$ is asymptotically normal with mean $\hat{\theta}^*$ and a covariance proportional to the variance of the disturbance. This holds good even for colored disturbances uncorrelated with the input. This implies that the LS estimation is robust to zero-mean disturbances. Note that the estimates will still be "biased" due to the inherent undermodelling.

- Irreducible model estimation. With MIMO TFM models, since the denominators do not include unknown parameters, the CD formulation does not inflate the parameter vector.
- Simplified error distribution problem. The weighting function in Eq. (13) now equals$|[A(j\omega)/E(j\omega)]\,U(j\omega)|^2$. The absence of the unknown θ in this weighting function permits on-line experimental design for a prescribed bias distribution.
- Gray-box modeling. The fixed denominator polynomial $A(\delta)$ in the linear-in-parameters model (14) serves as an additional design variable allowing for effective incorporation of prior knowledge of the process dynamics. By an intelligent choice of this polynomial, even complex systems can be estimated significantly accurately with a smaller number of parameters.

These are the advantages of linear-in-parameters models in system identification. In these models output is expressed as a linear combination of certain MA components of the input. This leads to the "generalized moving average model" (GMAM) formulation as

$$y(k) = \sum_{i=1}^{n} \theta_i \mathcal{F}_i(\delta)\, u(k) \tag{16}$$

In this model, the moving average components of the inputs are formed as the responses of a set of known filters $\{\mathcal{F}_i(\delta)\}$ to $u(k)$. These filters form the basis

$$\mathcal{B}(\delta) = [\mathcal{F}_1(\delta), \mathcal{F}_2(\delta), \ldots, \mathcal{F}_n(\delta)]^T \tag{17}$$

of the GMAM structure

$$G(\delta, \theta) = \sum_{i=1}^{n} \theta_i \mathcal{F}_i(\delta) \tag{18}$$

With such a parametrization, the model output error

$$\varepsilon_{\mathrm{OE}}(k) = y(k) - \sum_{\mathrm{i}=1}^{n} \theta_i \mathcal{F}_i(\delta)\, u(k)$$

is linear in $\{\theta_i\}$, and consequently the minimization problem of the output error criterion (Eq. (9)) is linear.

Models of the above structure evolve very naturally from truncated power series expansions of the rational transfer function. For example, in the DT case, the system transfer function may be written as

$$G^0(q^{-1}) = \sum_{i=1}^{\infty} h_i q^{-i} \tag{19}$$

where $\{h_i\}$ is the impulse response sequence. This suggests the following parametrization

$$G(\delta, \theta) = \theta^T \mathscr{B}(q^{-1}) \tag{20}$$

where $\theta = [\theta_1, \theta_2, \ldots, \theta_n]^T$, and $\mathscr{B}(q^{-1}) = [q^{-1}, q^{-2}, \ldots, q^{-n}]^T$. The quality of this approximation depends on the rate of convergence of the impulse response sequence. The poles of $G^0(z)$ close to the unit circle slow down the rate of convergence. Consequently a high model order is required for a given tolerance. For these reasons, in rapidly sampled CT systems the rate of convergence of the approximation will be very slow, and in the limit as $T_S \rightarrow 0$, the DT poles approach unity and consequently the approximation fails to converge. Furthermore, even in the case of convergent approximations, high model order is required as the memory of the basis (shift operator) is very short (unity). Therefore, model representations having better convergence properties and less sensitivity to sampling rate will be preferable.

In the CT case, the transfer function $G^0(p)$ may be expanded about $p = \infty$ as a complex power series in p^{-1} as

$$G^0(p) = \sum_{i=1}^{\infty} h_i (p^{-1})^i \tag{21}$$

leading to the form

$$G(p, \theta) = \theta^T \mathscr{B}(p^{-1}) \tag{22}$$

where $\theta = [h_1, h_2, \ldots, h_n]^T$ and $\mathscr{B}(p^{-1}) = [p^{-1}, p^{-2}, \ldots, p^{-n}]^T$. It is well known that h_i are the CT Markov parameters of $G^0(p)$ which are defined as

$$h_i = \left. \frac{\mathrm{d}^{i-1}}{\mathrm{d}t^{i-1}} g^0(t) \right|_{t=0} \tag{23}$$

where $g^0(t)$ is the impulse response of $G^0(p)$.

Considering a similar expansion of $G^0(p)$ about $p = 0$, one has models parametrized in terms of normalized time moments of the impulse response $g^0(t)$ of $G^0(p)$, i.e.,

$$G(p, \theta) = \theta^T \mathscr{B}(p) \tag{24}$$

where $\theta = [m_1, m_2, \ldots, m_n]^T$ and $\mathscr{B}(p) = [p, p^2, \ldots, p^n]^T$ and

$$m_i = \frac{(-1)^i}{i!} \int_0^\infty t^i g^0(t)\,\mathrm{d}t \tag{25}$$

are the normalized time moments.

Other basis functions are also possible. Well known among these are Laguerre and Kautz filters. Laguerre filters imply a basis

$$\mathscr{B}_{\mathrm{LAG}}(p) = \left[\frac{1}{p+\lambda}, \frac{1}{p+\lambda}\left(\frac{p-\lambda}{p+\lambda}\right), \ldots, \frac{1}{p+\lambda}\left(\frac{p-\lambda}{p+\lambda}\right)^{n-1}\right]^T \tag{26}$$

with $\lambda > 0$, and Kautz filters imply

$$\mathscr{B}_{\mathrm{KAUTZ}}(p) = [\psi_1(p), \psi_2(p), \ldots, \psi_n(p)]^T \tag{27}$$

where

$$\psi_{2k-1}(p, b, c) = \frac{p}{p^2 + bp + c}\left[\frac{p^2 - pb + c}{p^2 + pb + c}\right]^{k-1}$$

and

$$\psi_{2k}(p, b, c) = \frac{1}{p^2 + bp + c}\left[\frac{p^2 - bp + c}{p^2 + bp + c}\right]^{k-1}$$

with $b > 0$, $c > 0$, and $k = 1, 2, \ldots$. Wahlberg [22] discusses these bases in greater detail. The role played by the basis in continuous and discrete system modeling is discussed by Goodwin *et al.* [23,24].

In the following sections, CT models parametrized in terms of Markov parameters and time moments are presented for CT MIMO system identification. These models are so generalized as to include prior knowledge of system dynamics, and to ensure that even low order models give an adequate representation of the system under consideration.

VI. MARKOV PARAMETER MODELS

The use of Markov parameters for parametrizing models is not new in the field of system identification. There have been references to DT Markov parameters for identification of DT multivariable models. DT models of the form of Eq. (20) were considered for identification by Sinha *et al.* [25]. The problem of direct DT Markov parameter (MP) estimation based on cross-correlation between the output and a white noise input has also been studied by Sinha *et al.* [25]. Niederlinski and Hajdasinski [26] survey the related issues.

However, in CT situation, the reference to Markov parameters is rare. This is because of the natural but difficult-to-compute form (23) in which Markov parameters are defined for CT systems.

The work of Dhawan *et al.* [27] is the first attempt at the use of MP models for SISO CT model identification. The MP model (22) is transformed into an integral equation in which the integrals are realized using block-pulse functions [28], thereby avoiding the derivative route to the realization of Markov parameters. However, truncation of the MP model as in Eq. (22) often leads to poor approximation, due to which the estimation may fail to converge. A simple generalization of the original MP model to ensure convergent approximations may be found in Dhawan *et al.* [27] and Küper [29]. Further generalization of the basis leading to flexible and well-behaved approximations was suggested by Subrahmanyam and Rao [30] and have been extended to MIMO systems by Subrahmanyam *et al.* [31].

A. Estimation of Moving-Average Models

Consider a ν_i-input, ν_o-output MIMO system having a transfer characteristic $\mathbf{G}^0(p)$, and an input–output relationship

$$\mathbf{y}(t) = \mathbf{G}^0(p)\,\mathbf{u}(t) + \mathbf{v}(t) \tag{28}$$

where $\mathbf{y} \in \mathbf{R}^{\nu_o}$, $\mathbf{u} \in \mathbf{R}^{\nu_i}$, and $\mathbf{v} \in \mathbf{R}^{\nu_o}$. CT Markov parameters of this system are defined as the coefficients of the power series

$$\mathbf{G}^0(p) = \sum_{l=1}^{\infty} \mathbf{H}_l p^{-l}$$

where $\{\mathbf{H}_l\}$ is the Markov parameter sequence (MPS). Denote

$$H_l = \begin{bmatrix} h_{l,11} & h_{l,12} & \cdots & h_{l,1\nu_i} \\ h_{l,21} & h_{l,22} & \cdots & h_{l,2\nu_i} \\ \vdots & \vdots & \ddots & \vdots \\ h_{l,\nu_o 1} & h_{l,\nu_o 2} & \cdots & h_{l,\nu_o \nu_i} \end{bmatrix}$$

In terms of the MPS,

$$\mathbf{y}(t) = \sum_{l=1}^{\infty} \mathbf{H}_l \mathbf{u}^l(t) + \mathbf{v}(t),$$

where $\mathbf{u}^l(t)$ is the lth integral of $\mathbf{u}(t)$. Assuming absolute convergence of the MPS and thus uniform convergence of partial sums, a truncated MP model is obtained as

$$\mathbf{y}(t) = \sum_{l=1}^{n} \mathbf{H}_l \mathbf{u}^l(t) + \mathbf{e}(t) \tag{29}$$

where $\mathbf{e}(t)$ includes the truncation (of the MPS) error and the contribution of unknown initial conditions in addition to the usual noise term $\mathbf{v}(t)$. This model is valid only when the power series expansion of $\mathbf{G}^0(p)$ is absolutely convergent. Note that, when the system is represented in the sampled domain as $\mathbf{G}^0(q^{-1})$, the resulting DT MPS is the impulse response sequence of the system. For asymptotically stable systems, the DT MPS is absolutely convergent. But, when represented in the CT domain, even stable systems may have diverging MP sequences. To ensure absolute convergence and to increase the rate of convergence of the approximation, a more general version of Markov parameters, called Markov–Poisson parameters, was suggested by Subrahmanyam and Rao [30]. In terms of these, $\mathbf{G}^0(p)$ is expanded as

$$\mathbf{G}^0(p) = \sum_{l=1}^{\infty} \bar{\mathbf{H}}_l \left(\frac{\beta}{p+\lambda}\right)^l \mathbf{u}(t) \tag{30}$$

The Markov–Poisson parameters $\{\bar{\mathbf{H}}_l\}$ are related to the Markov parameters $\{\mathbf{H}_l\}$ as

$$\bar{\mathbf{H}}_l = \frac{1}{\beta^l} \sum_{i=1}^{l} {}^{l-1}C_{i-1} \lambda^{l-i} \mathbf{H}_i, \quad l = 1, \ldots; \beta > 1 \tag{31}$$

Thus, the model is

$$\mathbf{y}(t) = \sum_{l=1}^{n} \bar{\mathbf{H}}_l \left(\frac{\beta}{p+\lambda}\right)^l \mathbf{u}(t) + \mathbf{e}(t)$$

implying the basis

$$\mathcal{B}_{\mathrm{PF}}(p) = \left[\frac{\beta}{p+\lambda}, \left(\frac{\beta}{p+\lambda}\right)^2, \ldots, \left(\frac{\beta}{p+\lambda}\right)^n\right]$$

The elements of $\mathcal{B}_{\mathrm{PF}}(p)$ are the well-known Poisson filters [32] of increasing order in which λ and β are tunable parameters. This generalization improves the low-frequency predictive ability of the model. The choice of the filter parameter λ has to be made according to the *a priori* knowledge of the poles of the system. In general, a $\lambda > 0$ is well suited for overdamped systems with poles not very close to the imaginary axis of the s-plane. On the other hand, a $\lambda < 0$ with a large β is appropriate when the poles (complex) of the system are arbitrarily close to the imaginary axis.

Parameter estimation may now be carried out by decomposing the problem into ν_o MISO subproblems and considering one subproblem at a time or in parallel. In the sequel, one such MISO problem is considered and the subscript i, which denotes the row index, is dropped mainly for notational simplicity. Further, only n_j parameters are considered for the jth element of the MISO problem. Approximating the derivative operator by δ in the parameter estimation equation we get

$$\bar{y}(k) = \Phi(k)^T\theta$$

where

$$\Phi(k) = [\Phi_1(k), \Phi_2(k), \ldots, \Phi_{\nu_i}(k)]^T$$
$$\Phi_j(k) = [\mathcal{F}_1(\delta)u_j(k), \mathcal{F}_2(\delta)u_j(k), \ldots, \mathcal{F}_{n_j}(\delta)u_j(k)]^T, \quad j = 1, \ldots, \nu_i$$
$$\theta = [\bar{h}_{1,1}, \ldots \bar{h}_{n_1,1}| \ldots |\bar{h}_{1,\nu_i}, \ldots \bar{h}_{n_{\nu_i},\nu_i}]^T$$

and

$$\mathcal{F}_l(\delta) = \left(\frac{\beta}{\delta + \lambda}\right)^l$$

Next, we define the cost function as

$$J(\theta) = [\theta - \hat{\theta}(0)]^T\mathbf{P}(0)^{-1}[\theta - \hat{\theta}(0)] + \sum_{k=1}^{N} [y(k) - \Phi(k)\,\theta]^2$$

The LS estimate that minimizes $J(\theta)$ is

$$\hat{\theta}(N) = \left[\mathbf{P}(0)^{-1} + \sum_{k=1}^{N} \Phi(k)\Phi(k)^T\right]^{-1}\left[\mathbf{P}(0)^{-1}\hat{\theta}(0) + \sum_{k=1}^{N} \Phi(k)y(k)\right]$$

provided the inverse exists. This estimate may be calculated using the conventional recursive least squares algorithm.

B. Irreducible ARMA Model Realization

Given the estimates of $\bar{\mathbf{H}}_l, l = 1, \ldots, n$, the first step towards realization of an irreducible ARMA TFM model is to examine the columns or rows of the Hankel matrix $\mathcal{H}(p, q)$ formed from the estimates as

$$\mathcal{H}(p, q) = \begin{bmatrix} \bar{\mathbf{H}}_1 & \bar{\mathbf{H}}_2 & \cdots & \bar{\mathbf{H}}_q \\ \bar{\mathbf{H}}_2 & \bar{\mathbf{H}}_3 & \cdots & \bar{\mathbf{H}}_{q+1} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\mathbf{H}}_p & \bar{\mathbf{H}}_{p+1} & \cdots & \bar{\mathbf{H}}_{p+q} \end{bmatrix} \tag{32}$$

for predecessor independence [33,34]. In view of the MISO decomposition,

$$\bar{\mathbf{H}}_l = [\bar{h}_{l,1}, \bar{h}_{l,2}, \ldots, \bar{h}_{l,\nu_i}]$$

Interchanging columns, Eq. (32) may be written as

$$\mathcal{H}(p,q) = [\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_{v_i}],$$

where $\mathcal{H}_j, j = 1, 2, \ldots, v_i$ are the $p \times q$ Hankel matrices of the SISO elements of the MISO submodel.

Thus, the problem of structural identification of the MISO model is also decomposed into equivalent problems of finding ranks of Hankel matrices of individual elements over a row. Singular value decomposition may be used for this purpose.

According to the "partial realization theory" of [35] and [36], given a finite sequence of Markov parameters, it is possible to find a finite dimensional realization whose first few Markov parameters are correspondingly equal to the given finite sequence of Markov parameters. Accordingly, given a finite estimate Markov–Poisson parameter sequence, irreducible TFM models can be derived solving the following equations together with Eq. (31).

$$h_{l,ij} = b_{l,ij} - \sum_{r=0}^{l-1} h_{r,ij} a_{l-r,ij}, \quad l = 1, 2, \ldots, n_{ij}$$

$$h_{l+n_{ij},ij} = -\sum_{r=1}^{n_{ij}} h_{n_{ij}+l-r,ij} a_{r,ij}, \quad l = 1, 2, \ldots$$

where the *ij*th element of the TFM is considered to be of the form

$$G_{ij}(s) = \frac{b_{1,ij}s^{n_{ij}-1} + \ldots + b_{n_{ij},ij}}{s^{n_{ij}} + a_{1,ij}s^{n_{ij}-1} + \ldots + a_{n_{ij},ij}}$$

Supposing the system is of this ARMA form, some insight may be given regarding the nature of the MPS.

- The MPS is convergent when all the poles of all the elements of the TFM are inside the unit circle centered at the origin of the s-plane. Equivalently, the Markov–Poisson parameter sequence (MPPS) is convergent when all the poles of all the elements of the TFM are inside the circle of radius β centered at $(-\lambda, 0)$ of the pole-zero plot. This circle may be termed the zone of convergence of the sequence.
- The MPS (or MPPS) is finite if and only if all the poles of all the elements of the TFM lie at the origin of the zone of convergence.

C. Conversion of MPS into Finite Form

The usual infinite-length MPS is finite (with length $\mathbf{max}_{i,j}\{n_{ij}\}$ when no CD is assumed, or $\mathbf{max}_j\{n_j\}$ when column-wise CD is assumed) only when the poles of each subsystem of the TFM lie at the origin of the convergence zone. For a known system, all poles can be placed at the center by state feedback. Then such a modified system will have a finite MPS. In the identification problem, since such state feedback cannot be introduced as the system itself is unknown, it is possible to introduce the effect of pole-placement on the input–output measurement data, by some iterative pole-placement algorithm. For the sake of simplicity, the SISO case is considered in the following.

Consider the state equation of $G^0(p)$ in its controller form

$$\dot{x}(t) = Ax(t) + bu(t)$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_n & -a_{n-1} & -a_{n-1} & \dots & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

The matrix A can be written as

$$A = A_0 - bk$$

where

$$A_0 = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}, \qquad k = [a_n, a_{n-1}, \dots, a_1]$$

Therefore we have the state equation as

$$\dot{x}(t) = A_0 x(t) + b\overline{u}(t),$$

where $\overline{u}(t) = u(t) - kx(t)$ is the filtered input signal. The fictitious system described as above by the signal pair $\overline{u}(t)$ and $x(t)$ has a finite MPS, as the eigenvalues of A_o are all at zeros. Therefore, by transforming the original system into that described by the above, the approximation error due to truncation of the MPS can be made to vanish. This is equivalent to placing the poles of the system at the origin of the convergence zone. Based on this, a time-recursive and iterative algorithm was initially proposed by [30], which was later extended to tackle multivariable systems by [31]. For a detailed description and analysis see this latter reference.

VII. TIME MOMENT MODELS

Like Markov parameters, time moments also play an important role in the field of reduced-order modeling [37,38]. Despite the wealth of other mathematically sound methods available for reduced order modeling, the moment matching method is still considered as the simplest and is widely used. In the field of system identification, an approach for multivariable system identification has been recently proposed by Subrahmanyam *et al.* [39].

A. Estimation of Moving-Average Models

The TFM may be written in terms of the time moments which are related to the impulse response as

$$\mathbf{G}^0(p) = \int_0^\infty \mathbf{g}^0(t)\mathrm{e}^{-pt}\mathrm{d}t = \sum_{l=0}^{\infty} \mathbf{M}_l p^l$$

where

$$\mathbf{M}_l = \frac{(-1)^l}{l!}\int_0^\infty t^l \mathbf{g}^0(t)\,\mathrm{d}t, \qquad l = 1,2$$

happen to be the normalized time moments of the impulse response. Define

$$H_l = \begin{bmatrix} m_{l,11} & m_{l,12} & \cdots & m_{l,1v_i} \\ m_{l,21} & m_{l,22} & \cdots & m_{l,2v_i} \\ \vdots & \vdots & \ddots & \vdots \\ m_{l,v_o1} & m_{l,v_o2} & \cdots & m_{l,v_ov_i} \end{bmatrix}$$

In terms of the time moment sequence (TMS) $\{\mathbf{M}_l\}$, the system input–output relation becomes

$$\mathbf{y}(t) = \sum_{l=0}^{\infty} \mathbf{M}_l \mathbf{u}^{(l)}(t) + \mathbf{v}(t)$$

where $\mathbf{u}^{(l)}(t)$ is the lth derivative of $\mathbf{u}(t)$. Assuming absolute convergence of TMS and thus uniform convergence of partial sums, similar to the case of MP modeling, the truncated TM model is

$$\mathbf{y}(t) = \sum_{l=0}^{n} \mathbf{M}_l \mathbf{u}^{(l)}(t) + \mathbf{e}(t) \tag{33}$$

To validate the use of the above model even for systems with diverging

TMS, additional exponential scaling of the series will be necessary to ensure convergence.

To avoid the direct use of derivatives, Eq. (33) is operated on both sides by an $(n+1)$th order Poisson filter operator $\beta^{n+1}/(p+\lambda)^{n+1}$ [32]. Denoting

$$\mathcal{F}_{l,n+1}(p) = \beta^{n+1}\frac{p^l}{(p+\lambda)^{n+1}}, \qquad l = 0, 1, \ldots, n$$

the time moment (TM) model is

$$\mathcal{F}_{0,n+1}(p)\,\mathbf{y}(t) = \sum_{l=0}^{n} \mathbf{M}_l \mathcal{F}_{l,n+1}(p)\,\mathbf{u}(t) + \mathbf{e}(t) \tag{34}$$

For the ith row of Eq. (34) (dropping the subscript i in all relevant symbols), taking into account n_j time moments of the jth MISO subsystem, and letting $n = \max_j\{n_j\}$, the parameter estimation equation in discrete-time is obtained as

$$\bar{y}(k) = \Phi(k)^T\theta$$

where

$$\begin{aligned}
\Phi(k) &= [\Phi_1(k), \Phi_2(k), \ldots, \Phi_{\nu_i}(k)]^T \\
\Phi_j(k) &= [\mathcal{F}_{0,n+1}(\delta)u_j(k), \mathcal{F}_{1,n+1}(\delta)u_j(k), \ldots, \mathcal{F}_{n_j,n+1}(\delta)u_j(k)]^T, \\
&\quad j = 1, \ldots, \nu_i \\
\bar{y}(k) &= \mathcal{F}_{0,n+1}(\delta)y(k) \\
\theta &= [m_{0,1}, \ldots, m_{n,1}|\ldots|m_{0,\nu_i}, \ldots, m_{n_{\nu_i},\nu_i}]^T
\end{aligned}$$

Parameter estimation may now be carried out with the usual least squares algorithm.

B. Irreducible ARMA Model Realization

Given the estimates of $\mathbf{M}_l$, $l = 1, \ldots, n$, an irreducible ARMA TFM model can be realized in a manner similar to the case of Markov parameter models. Let

$$\begin{aligned}
A_{ij}(s) &= 1 + a_{1,ij}s + \ldots + a_{n_{ij},ij}s^{n_{ij}} \\
B_{ij}(s) &= b_{0,ij} + b_{1,ij}s + \ldots + b_{n_{ij}-1,ij}s^{n_{ij}-1}
\end{aligned}$$

and

$$\mathbf{M}_l = \{m_{l,ij}; i = 1, \ldots, \nu_o, j = 1, \ldots, \nu_i\}$$

Given the estimates of $\mathbf{M}_l$, $l = 1, \ldots, n$, the TFM elements can be obtained by solving the following equations:

$$m_{l,ij} = b_{l,ij} - \sum_{r=0}^{l-1} m_{r,ij} a_{l-r,ij}, \qquad l = 0, \ldots, n_{ij} - 1,$$

and

$$m_{l+n_{ij},ij} = -\sum_{r=1}^{n_{ij}} m_{n_{ij}+1-r,ij} a_{l,ij}, \qquad l = 1, 2, \ldots$$

Supposing the system is of this ARMA form, the following remarks are in order.

1. The TMS is convergent if all the poles of all the elements of the TFM are outside the unit circle centered at the origin of the s-plane. This circle is the zone of convergence of the sequence.
2. The TMS is finite if and only if all the elements of the TFM are denominator free (i.e., have denominator 1).

C. Conversion of TMS into Finite Form

The TMS is finite when all the subsystems of the TFM are denominator free, in which case the length of the TMS is $\max_j\{n_j\}$ and modeling will not involve unmodelled dynamics. This situation can be met by adding fictitious zeros to each subsystem, to cancel their respective denominators. In an identification experiment, this is achievable for ARMA systems as illustrated below for the SISO case

$$y(t) = \frac{B(p)}{A(p)} u(t)$$

If the denominator $A(p)$ is known, we can write

$$y(t) = \frac{B(p)}{A(p)} u(t) = \sum_{i=1}^{n-1} b_i p^i \bar{u}(t),$$

where $\bar{u}(t) = [1/A(p)]\, u(t)$. Thus the model between $\bar{u}(t)$ and $y(t)$ has a finite TMS. Therefore, by estimating the denominators and then canceling them in an iterative way, it is possible to convert the TMS into a finite form, so as to remove the truncation error. For such an iterative algorithm with detailed analysis, see [39].

VIII. CHOICE OF PARAMETRIC FORM

When physical phenomena, which can be captured only by complex relationships, are represented by simplified models, the probability that a given model set includes the precise model of the observed process is zero. Modeling error is inevitable and the performance (viz. predictive ability) of the estimated models depends on the choice of model structure and the prior knowledge embedded into the chosen model structures, for a given model order.

In the proposed class of GMAMs, the following parametrizations are considered for CT system modeling.

- Motivated by Markov–Poisson parameter models, with a Poisson filter chain:

$$\mathscr{B}_{\mathrm{PF}}(\delta) = \left[\frac{\beta}{\delta+\lambda}, \left(\frac{\beta}{\delta+\lambda}\right)^2, \ldots, \left(\frac{\beta}{\delta+\lambda}\right)^n\right] \tag{35}$$

  and,
- Motivated by TM models, with a state-variable filter (SVF)

$$\mathscr{B}_{\mathrm{SVF}}(\delta) = \left[\frac{1}{E(\delta)}, \frac{\delta}{E(\delta)}, \ldots, \frac{\delta^{n-1}}{E(\delta)}\right] \tag{36}$$

  where $1/E(\delta)$ is a nth order stable filter. A typical choice is an nth-order Poisson filter.

The following issues are now studied via numerical examples.

- Predictive ability. The above two choices $\mathscr{B}_{\mathrm{PF}}(\delta)$ and $\mathscr{B}_{\mathrm{SVF}}(\delta)$ are related through a linear nonsingular transformation (for $\lambda \neq 0$), e.g., for $n = 4$ and $\beta = 1$.

$$\mathscr{B}_{\mathrm{SVF}}(\delta) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -\lambda \\ 0 & 2 & -2\lambda & \lambda^2 \\ 1 & -3\lambda & 3\lambda^2 & \lambda^3 \end{bmatrix} \mathscr{B}_{\mathrm{PF}}(\delta)$$

  Hence for a given model order, models based on these two sets will have the same predictive ability.
- The numerical behavior of the estimation algorithm is dictated by the condition number of the matrix

$$\mathbf{R} = \sum_{k=1}^{N} \Phi(k)\Phi(k)^T$$

  It has been pointed out by [30] that use of $\mathscr{B}_{\mathrm{PF}}(\delta)$ results in high condition numbers of the above matrix, as these functions are overlapping and nonorthogonal. On the other hand, the second set $\mathscr{B}_{\mathrm{SVF}}(\delta)$ is near-orthogonal [24], which improves the condition number.

- Numerical conditioning may be improved if an intelligently chosen linear transformation of these sets of basis functions is made before parameter estimation commences. When such transformation results in an orthogonal set, the numerical properties of the algorithm will be significantly improved. A popular orthogonal basis is in terms of Laguerre filters:

$$\mathscr{B}_{\mathrm{LAG}}(s) = \left[\frac{1}{\delta+\lambda}, \frac{1}{\delta+\lambda}\left(\frac{\delta-\lambda}{\delta+\lambda}\right), \ldots, \frac{1}{\delta+\lambda}\left(\frac{\delta-\lambda}{\delta+\lambda}\right)^{n-1}\right]^T$$

The required linear transformations are for ($n = 4$ and $\beta = 1$),

$$\mathscr{B}_{\mathrm{LAG}}(\delta) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & -2\lambda & 0 & 0 \\ 1 & -4\lambda & 4\lambda^2 & 0 \\ 1 & -6\lambda & 12\lambda^2 & -8\lambda^3 \end{bmatrix} \mathscr{B}_{\mathrm{PF}}(\delta)$$

and

$$\mathscr{B}_{\mathrm{LAG}}(\delta) = \begin{bmatrix} 1 & 3\lambda^2 & 3\lambda & 1 \\ -\lambda^3 & -\lambda^2 & \lambda & 1 \\ \lambda^3 & -\lambda^2 & \lambda & 0 \\ 1 & 3\lambda^2 & 3\lambda & 1 \end{bmatrix} \mathscr{B}_{\mathrm{SVF}}(\delta)$$

Example 1. Consider a CT system

$$G^0(s) = \frac{2s+6}{s^3+7s^2+11s+6}$$

This is simulated using trapezoidal approximation with a Gaussian white input of variance 1.0 and steady-state gain 1.0 for 100 sec at a sampling time 0.1 sec.

Models of various orders are estimated with $\lambda = \beta = 1.0$. Parameter estimates are shown in Table 2. Frequency response plots of the true system and the estimated models are shown in Figs 2, 3, and 4, respectively, for $\mathscr{B}_{\mathrm{PF}}(\delta)$, $\mathscr{B}_{\mathrm{SVF}}(\delta)$, and $\mathscr{B}_{\mathrm{LAG}}(\delta)$. As seen from these figures, the estimation quality improves with increase in model order n. Figure 5 shows that, for a given model order ($n = 4$), all the three sets of basis functions produce models with almost the same frequency response.

Table 3 compares the condition numbers for increasing n. Though the condition numbers with $\mathscr{B}_{\mathrm{PF}}(\delta)$ and $\mathscr{B}_{\mathrm{SVF}}(\delta)$ are larger than those with $\mathscr{B}_{\mathrm{LAG}}(\delta)$, the estimates do not show significant numerical errors (see Fig. 5). However, the same may not be the case with small-word-length

Table 2 Parameter Estimates of Example 1

Basis	d	$\hat{\theta}$
$\mathcal{B}_{\mathrm{PF}}$	1	[0.9827]
	2	[0.1914, 0.8133]
	3	[0.0585, 1.4174, −0.4760]
	4	[0.0556, 1.4320, −0.5038, 0.0163]
	5	[0.0459, 1.5054, −0.7109, 0.2783, −0.1190]
$\mathcal{B}_{\mathrm{SVF}}$	1	[0.9824]
	2	[1.0046, 0.1867]
	3	[0.9999, 1.5233, 0.0562]
	4	[1.0000, 2.5116, 1.5886, 0.0517]
	5	[0.998, 3.5395, 4.0578, 1.6732, 0.0425]
$\mathcal{B}_{\mathrm{LAG}}$	1	[0.9827]
	2	[0.5981, −0.4066]
	3	[0.6482, −0.4707, −0.1190]
	4	[0.6477, −0.4702, −0.1198, −0.0020]
	5	[0.6482, −0.4718, −0.1179, −0.0050, −0.0074]

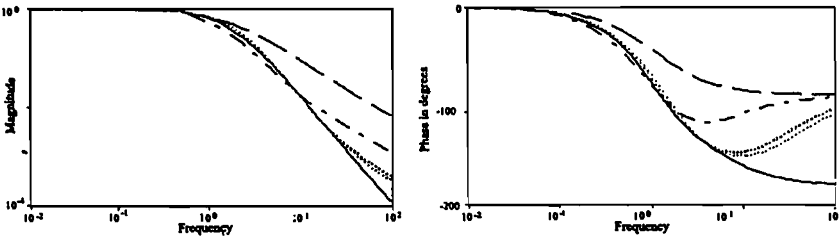


Figure 2 Results of estimation with $\mathcal{B}_{\mathrm{PF}}(\delta)$: (a) —— $G^0(\mathrm{j}\omega)$, (b) ---- $n = 1$, (c) —·—·— $n = 2$, (d) ····· $n = 3$, 4 and 5.

computers, and transformations such as suggested in item 3 above will be helpful.

Example 2. Now consider an underdamped second-order system [40],

$$G^0(s) = \frac{1}{s^2 + 0.2s + 1}$$

The measurement data are obtained with a sampling time of 0.001 sec. A model of the following structure

$$G(\delta, \theta) = \frac{\sum_{i=1}^{n} \theta_i \delta^i}{(\delta^2 + 2\zeta\omega_m + \omega_n^2)^{n/2}}, \quad n \text{ even}$$

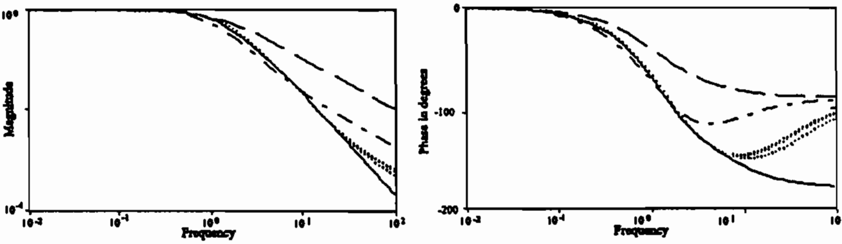


Figure 3 Results of estimation with $\mathscr{B}_{\mathrm{SVF}}(\delta)$: (a) —— $G^0(\mathrm{j}\omega)$, (b) ---- $n = 1$, (c) -·-·- $n = 2$, (d) ····· $n = 3$, 4 and 5.

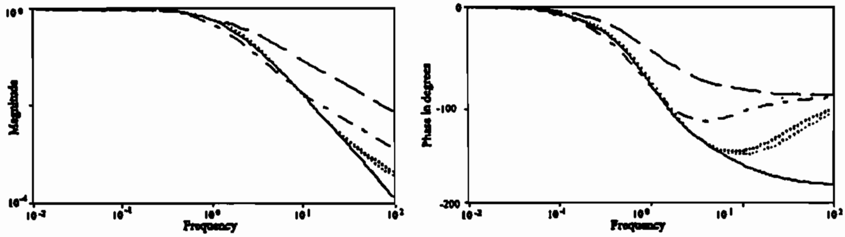


Figure 4 Results of estimation with $\mathscr{B}_{\mathrm{LAG}}(\delta)$: (a) —— $G^0(\mathrm{j}\omega)$, (b) ---- $n = 1$, (c) -·-·- $n = 2$, (d) ····· $n = 3$, 4 and 5.

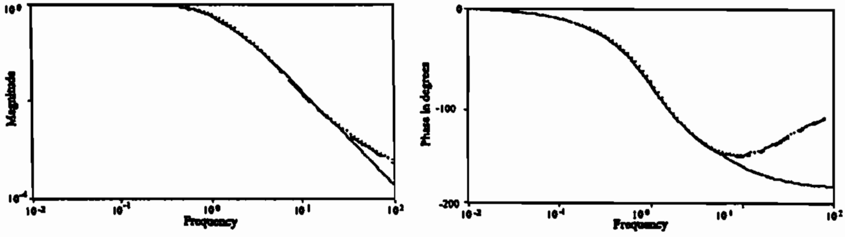


Figure 5 Comparison of results for $n = 4$: (a) —— $G^0(\mathrm{j}\omega)$, (b) ---- $\mathscr{B}_{\mathrm{PF}}(\delta)$, (c) -·-·- $\mathscr{B}_{\mathrm{SVF}}(\delta)$, (d) ····· $\mathscr{B}_{\mathrm{LAG}}(\delta)$.

with $\zeta = 0.2$ and $\omega_n = 1.2$ (for the original system these are 0.1 and 1.0 respectively) is estimated. This may be written in terms of $\mathscr{B}_{\mathrm{SVF}}(\delta)$ with $E(\delta)$ of Eq. (36) as

$$E(\delta) = (\delta^2 + 2\zeta\omega_n\delta + \omega_n^2)^{n/2}$$

Table 3 Comparison of Condition Numbers (Example 2)

n	$\mathcal{B}_{\mathrm{PF}}(\delta)$	$\mathcal{B}_{\mathrm{SVF}}(\delta)$	$\mathcal{B}_{\mathrm{LAG}}(\delta)$
1	1.000000	1.000000	1.000000
2	186.7289	45.64127	45.64652
3	1880.815	159.8336	70.09879
4	11109.74	380.5729	94.07874
5	11698.33	1273.411	120.6597

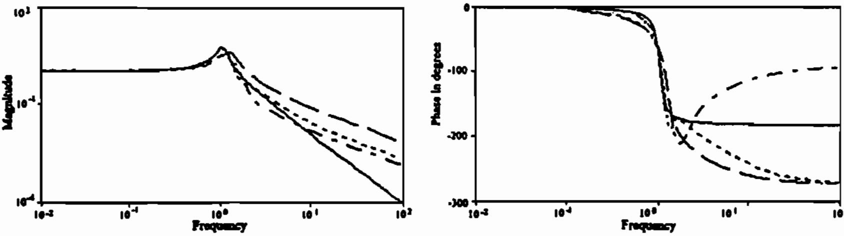


Figure 6 Results of estimation with $\mathcal{B}_{\mathrm{SVF}}(\delta)$ for Example 2: (a) —— $G^0(j\omega)$, (b) ----- $n = 1$, (c) –·–·– $n = 4$, (d) ····· $n = 6$.

This parametrization is in spirit close to the case of Kautz filters. By regrouping terms of these filters, the above parametrization may be obtained.

Models of orders 2, 4, and 6 are estimated and the parameter estimates are shown in Table 4. Frequency responses of these estimated models are shown in Fig. 6. This example shows how prior knowledge of system dynamics may be embedded into the basis to model even complex systems with a small number of parameters.

These illustrative examples suggest the use of $\mathcal{B}_{\mathrm{SVF}}(\delta)$ with the filter polynomial $E(\delta)$ chosen to represent prior knowledge of poles, real or complex in general, as

$$E(\delta) = \prod_{i=1}^{n_r} (\delta + \lambda)^{n_i} \prod_{i=1}^{n_c} (\delta^2 + a\delta + b)^{n_i}. \tag{37}$$

The advantage of regrouping terms of the resultant parametrization in terms of Laguerre/Kautz filters is to have a numerically well-conditioned estimation. Otherwise they provide the same level of approximation.

Table 4 Parameter Estimates Example 2

d	θ
2	[1.4143, −0.6455]
4	[2.1353, −0.1078, 1.0653, −0.1991]
6	[3.0206, 1.8372, 3.0559, 1.3535, 0.6472, 0.1282]

IX. CONCLUSIONS

One of the crucial phases in the procedure for parametric system identification is the parametrization of the model structure. Traditionally, models are parametrized as rational transfer functions. Such descriptions (being inherently nonlinear), despite their clarity of dynamics with poles and zeros, require nonlinear estimation. On the other hand, linearly parametrized model structures simplify the problem, with linear estimation as the main advantage. A class of CT generalized moving average models that are linear-in-parameters is proposed in this chapter.

Among various possible GMAMs, CT models parametrized with "Markov parameters" and "time moments" have been studied in this chapter, for irreducible model identification of MIMO systems. An advantage of these parametrizations is that, irreducible rational transfer function (matrix) or minimal state-space descriptions of models are realizable even from truncated sequences of Markov parameters and time moments. Thus, the sense of poles and zeros is preserved. For a limited class of finite-dimensional systems, for which such zeros and plots do exist, [31] and [39] showed that it is possible to find them in an iterative fashion. See also [41].

Due to the linearity of the suggested parametrizations, the estimation is robust to zero-mean white/colored disturbances. Monte Carlo simulation results confirming this may be found in [30,31] and [39].

The problem of inflated models for MIMO system has been avoided since the chosen parametrizations do not involve unknown denominators. Recall that in the case of rational TFM models with EE minimization, formation of a common denominator inflates the size of the unknown parameter vector.

Another feature of the methodology proposed here is that, by embedding prior knowledge of the dynamics of the system in the choice of a basis for the model, it is possible to estimate good lower-order approximations of complex systems. In such cases, the parametrization in terms of $\mathscr{E}_{\mathrm{SVF}}(p)$ with the filter polynomial $E(p)$ as in Eq. (37) seems to be a natural choice.

REFERENCES

1. P. Eykhoff. System Identification. New York: Wiley, 1974.
2. L Ljung. System Identification: Theory for the User. Englewood Cliffs: Prentice Hall, 1987.
3. T Söderström, P Stoica. System Identification. Hemel Hempstead, UK: Prentice Hall, 1989.
4. JP Norton. An Introduction to Identification. New York: Academic Press, 1986.
5. H Unbehauen, GP Rao. Identification of Continuous-time Systems. Amsterdam: North Holland, 1987.
6. NK Sinha, GP Rao (eds). Identification of Continuous-time Systems: Methodology and Computer Implementation. Dordrecht: Kluwer Academic Publishers, 1991.
7. RH Middleton, GC Goodwin. Digital Control and Estimation: A Unified Approach. Prentice-Hall, 1990.
8. S Mukhopadhyay. Continuous-time Models and Approaches for Estimation and Control of Linear Systems. PhD Thesis, Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India, 1991.
9. PJ Gawthrop. Continuous-time Self-tuning Control. London: IEE Publication series, 1987.
10. S Mukhopadhyay, A Patra, GP Rao. Int J Control 55:1161–1187, 1992.
11. GEP Box, GW Jenkins. Time Series Analysis: Forecasting and Control. 2nd Ed. San Francisco: Holden Day, 1970.
12. H Unbehauen, GP Rao. Automatica 26:23–35, 1990.
13. ZY Zhao, S Sagara, K Wada. Int J Control 53:445–461, 1991.
14. ZY Zhao, S Sagara, M Tomizuka. Int J Control 56:1441–1452, 1992.
15. ZJ Yang, S Sagara, K Wada. Control Theory and Advanced Technol 9:53–75, 1993.
16. S Mukhopadhyay, A Patra, GP Rao. Int J Control 53:223–253, 1991.
17. K Diekmann, H Unbehauen. Recursive identification of multiple-input multiple-output systems. In: Proceedings of IFAC Symposium on Identification and System Parameter Estimation, Darmstadt, 1979, pp 423–429.
18. GP Rao, K Diekmann, H Unbehauen. Parameter estimation in large-scale interconnected systems. In: Proceedings of IFAC Symposium on Identification and System Parameter Estimation, Budapest, Hungary, 1984, pp 729–733.
19. S Mukhopadhyay, GP Rao. Proc IEE, Part D, 138:93–102, 1991.
20. B Wahlberg, L Ljung. IEEE Trans Automatic Control AC-31:134–144, 1986.
21. VN Bapat. Some Extensions to Poisson Moment Functional Based Estimation of Continuous-time Models of Dynamical Systems. PhD Thesis,

Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India, 1993.
22. B Wahlberg. Laguerre and Kautz models. In: Preprints of SYSID'94 10th IFAC Symposium on Identification and System Parameter Estimation, Copenhagen, Denmark, 1994.
23. GC Goodwin, B Ninness, V Poor. Choice of basis functions for continuous and discrete system modeling. In: Preprints of IFAC Symposium on Identification and System Parameter Estimation, Budapest, Hungary, 1991, pp 1179–1184.
24. GC Goodwin, M Gevers, DQ Mayne. Bias and variance distribution in transfer function estimation. In: Preprints of IFAC Symposium on Identification and System Parameter Estimation, Budapest, Hungary, 1991, pp 952–957.
25. NK Sinha, AK Mahalanabis, H El-Sherief. Int J Syst Sci 8:425–430, 1978.
26. A Niederlinski, A Hajdasinski. Multivariable system identification—A survey. In: Proceedings of IFAC Symposium on Identification and System Parameter Estimation, 1979, pp 43–76.
27. RK Dhawan, A Sahai, DV Nishar, GP Rao. Recursive estimation of Markov parameters in linear continuous-time SISO systems via block-pulse functions. In: Proceedings of IFAC Symposium on Identification and System Parameter Estimation, Budapest, Hungary, 1991.
28. GP Rao. Piecewise Constant Orthogonal Functions and their Applications to Systems and Control. Lecture Notes in Control and Information Sciences-55. Berlin: Springer-Verlag, 1983.
29. P Küper. Identikation kontinuierlicher dynamischer Systeme mit hifte von Markov-Parameters. Studienarbeit ESR-9138, Lehrstuhl für Elektrische Steuerung and Regelung, Ruhr Universität, Bochum, 1992.
30. AVB Subrahmanyam, GP Rao. Proc IEE, Part D, 140:1–10, 1993.
31. AVB Subrahmanyam, DC Saha, GP Rao. Automatica 32:249–253, 1996.
32. DC Saha, GP Rao. Identification of Continuous-time Systems—A Poisson Moment Functional Approach. Lecture Notes in Control and Information Sciences-56. Berlin: Springer-Verlag, 1983.
33. FR Gantmacher. Applications of the Theory of Matrices. New York: Interscience, 1959.
34. W A Wolovich. Linear Multivariable Systems. New York: Springer-Verlag, 1974.
35. A Techer. Trans IEEE Autom Control 17:427–436, 1970.
36. RE Kalman. Aspects of Network and Systems Theory. New York: Holt, Reinhart and Winston, 1970.
37. M Jamshidi. Large-Scale Systems: Modeling and Control. Amsterdam: North-Holland, 1983.
38. A Bultheel, M van Barel. J Comput Appl Math 14:401–438, 1986.
39. AVB Subrahmanyam, DC Saha, GP Rao. Control Theory and Advanced Technol 10:1339–1378, 1995.
40. B Wahlberg. Identification of resonant systems using Kautz filters. In:

Proceedings of IEEE Conference on Decision and Control, Brighton, UK, 1991.

41. AVB Subrahmanyam. Linear and Robust Estimation of Continuous-time Models. PhD Thesis, Department of Electrical Engineering, Indian Institute of Technology, Kharagpur, India, 1995.

5

Multiresolution Approach to Identification of System Impulse Response

Zi-Jiang Yang

Kyushu Institute of Technology, Fukuoka, Japan

Setsuo Sagara

Fukuoka Institute of Technology, Fukuoka, Japan

Teruo Tsuji

Kyushu Institute of Technology, Kitakyushu, Japan

I. INTRODUCTION

System identification usually implies the modeling of an unknown system from its input–output data. The system model may be either nonparametric or parametric [1]. Usually, in the time domain, a linear system can be described by its nonparametric model, i.e., its impulse response model. An important benefit of impulse response identification is that no *a priori* knowledge about system order and dead time of the process is necessary but only a rough estimate of the settling time. Recently, it has been found that realization algorithms based on the system impulse response (Markov parameters) can be effectively applied to the problem of state-space model identification. Such algorithms include recent improvements based on the singular value decomposition techniques such as the eigensystem realization algorithm (ERA), the eigensystem realization algorithm using data correlations (ERADC), etc. [2]. In such methods, the Markov parameters are used to determine a balanced state-space model, and the model order is determined by the Hankel singular values. This leads to a clear trade-off between model order and identification quality in terms of a singular value plot. It is obvious that such realization algorithms require accurate estimates of the Markov parameters. Therefore how to identify the system impulse response effectively is viewed as an important task in system identification theories and techniques.

Since in most cases the input–output data are sampled in discrete time, impulse response identification typically implies that we estimate the discrete-time impulse response (unit impulse response) model (DTIRM). It should be noted that especially in the case of a small sampling interval, the input correlation matrix may have a large size; therefore, it is usually required that the input signal should include sufficient frequency components to make the input correlation matrix nonsingular. A white noise may be a good choice for this purpose. However, from the practical viewpoint, it is undesirable to let the input cause the system under study to fluctuate strongly due to the possible resulting damage; therefore, in many practical situations, only band-limited (smoothed) input signals are allowable. This often makes the identification problem ill-conditioned and hence makes the impulse response estimate very sensitive to the measurement noise. In contrast to the common DTIRM identification, the Laguerre/Kautz model approach can reduce greatly the number of parameters needed to obtain useful approximations of the processes of interest, and is therefore numerically efficient [3,4]. However, this approach requires *a priori* knowledge of the dominant poles of the system, to achieve accurate approximations.

So far, many researchers have been attacking the ill-conditioned problem encountered in impulse response identification via various approaches. A popular one is the regularization approach with single or multiple regularization parameters, which regularize the inverse of the input correlation matrix [5–8]. Usually, the regularization parameters are determined by trial and error, or determined optimally through a very sophisticated iterative estimation scheme [8]. Most recently, Sano and Tsuji [9] proposed the decimation and interpolation approach to smooth DTIRM identification in the ill-conditioned cases of band-limited input, fast sampling rate, and short data length. Similarly, Yang *et al.* [10] tried to approximate the impulse response using the generalized radial basis function network [11] in which the Gaussian basis functions are equidistantly located in the time domain and have the same width. Since the center distance of the basis functions is usually chosen as a multiple of the sampling interval, far fewer parameters require to be estimated. This leads to the improvement of the nonsingularity of the input correlation matrix and therefore yields smooth estimate results. However, so far accurate identification of the impulse response which includes high-frequency components locally has not been discussed enough. Such impulse responses can be found in systems with fast or slowly decaying poles, or systems of nonminimum phase, which are frequently encountered

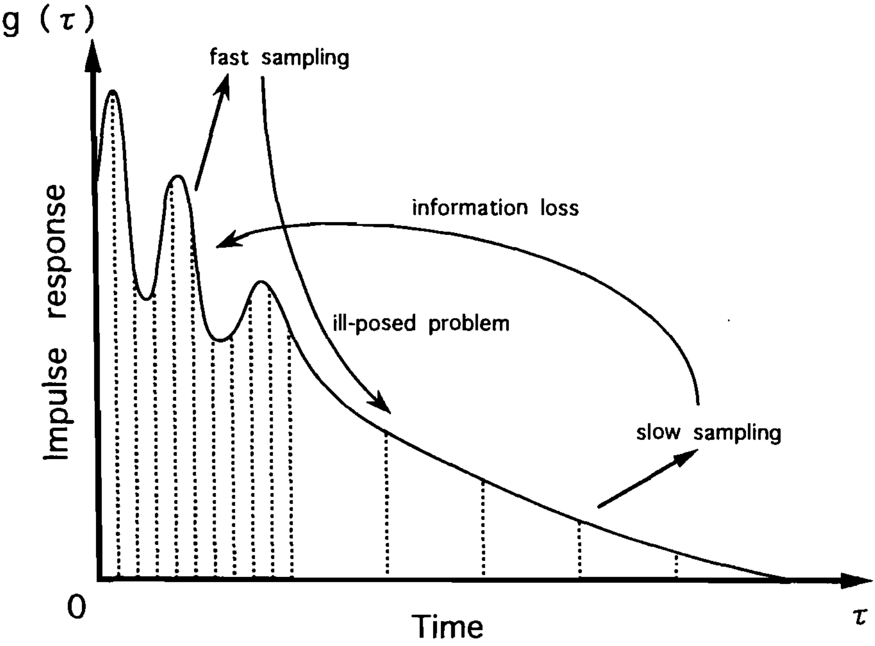


Figure 1 An example of the impulse response which includes high-frequency components locally.

in practice. For the purpose of illustration, an example of the system impulse response which includes high-frequency components locally is shown in Fig. 1. The high-frequency components require a small sampling period which may tend to cause ill-conditioned problems in the cases of band-limited input and short data length [9], whereas a large sampling period may cause some information loss of the high-frequency components of the impulse response. Therefore, it is difficult to identify the impulse response shown in Fig. 1 accurately via the conventional methods with uniform sampling rate.

In this work, the authors propose a new approach to identification of the DTIRM for a linear system from sampled input–output data using multiresolution approximation theory. Our attention is especially focused on systems whose impulse responses include high-frequency components locally.

The wavelet transformation, which has received great attention recently, is a new tool for time–frequency analysis. The wavelet transformation of a signal is an expansion of the signal into a special family of functions called wavelets [12,13]. A family of wavelets is generated from a single prototype wavelet by dilation of the timescale and shift along the time axis. Since the wavelets exist locally both in the time domain and the frequency domain, one can extract some information about a signal with respect to time and frequency. The wavelet transformation provides sharp time resolution for rapidly changing signal components and fine frequency resolution for slowly varying components. The multiresolution analysis which decomposes a signal into orthogonal subspaces of different frequency bands by the wavelet transformation, using the wavelets which are dilated at corresponding timescales, is capable of analyzing both the local and global properties of the signal [12–15]. Some studies of system identification using the wavelet transformation have been carried out by Maeda *et al.* [16], Tabaru *et al.* [17], Benveniste *et al.* [18], and Safavi *et al.* [19].

In this work, the continuous-time impulse response (CTIR) of the system under study is approximated by a multiresolution neural network (MRNN) composed of the scaling and wavelet functions which are shifted and dilated based on multiresolution approximation theory. The scaling and wavelet functions which are shifted and dilated based on multiresolution approximation theory are employed as the basis functions of the network. Then the system under study can be viewed as a weighted sum of a group of subsystems in which the shifted and dilated scaling functions and wavelet functions are interpreted as their impulse responses respectively. It should be noticed here that due to the excellent time–frequency localization properties of the basis functions, it is not necessary to include all the basis functions into the network. For rapidly changing components, we choose high-frequency basis functions, and for slowly varying components we choose low-frequency basis functions. That is, some redundant basis functions are removed. Therefore, the number of parameters needed to approximate the impulse response of interest can be reduced greatly. In this work, a heuristic identification procedure with the aid of the popular orthogonal least squares (OLS) method [20,21] and AIC is proposed to select significant subsystems such that only moderate parameters are required to be estimated, in contrast to the conventional method. The algorithm is performed in a hierarchical manner. The impulse response is estimated from a coarse resolution level (low-frequency subspace) to a fine resolution level (high-frequency subspace) successively via the OLS method. The AIC is utilized to select significant subsystems at each resolution level, such that some redundant subsystems in the high-frequency domain which are sensitive to noise effects in the case of band-limited input signal are discarded. This leads to a parsimonious model of the impulse response and hence only moderate parameters are required to be estimated, in contrast to the conventional method. It is shown through simulation study that the proposed method yields accurate estimate of the impulse response which includes high-frequency components locally, even in the ill-conditioned cases of band-limited input, fast sampling rate, short data length, and significant measurement noise.

II. CONVENTIONAL IDENTIFICATION METHOD FOR DTIRM

In this section, we review briefly the traditional DTIRM identification method of a linear system. A linear time-invariant continuous system can be described by its CTIR response $g(\tau)$ as

$$x(t) = \int_0^t g(\tau)u(t-\tau)\,\mathrm{d}\tau \tag{1}$$

where $u(t)$ and $x(t)$ are the input and output signals, respectively. It is well known that the impulse response is a complete characterization of the system.

In practice, it is usually the situation that the signals are observed at the signal sampling instants kT $(k = 1, 2, \ldots)$:

$$x(kT) = \int_0^{kT} g(\tau)u(kT-\tau)\,\mathrm{d}\tau \tag{2}$$

where T is the sampling interval. In this work, we will denote $x(kT)$, $u(kT)$ as $x(k)$, $u(k)$, respectively.

The sampled measurement of the output is described as

$$y(k) = x(k) + v(k) \tag{3}$$

where $v(k)$ denotes the measurement noise, which is assumed to be zero-mean, Gaussian and white.

Typically, in computer control applications, the input signal $u(t)$ is kept constant between the sampling instants:

$$u(t) = u(k), \quad kT \le t < (k+1)T \tag{4}$$

Equations (2), (3), and (4) lead to

$$\begin{aligned} y(k) &= \sum_{i=1}^{k} u(k-i)\int_{(i-1)T}^{iT} g(\tau)\,\mathrm{d}\tau + v(k) \\ &\approx \sum_{i=1}^{n+1} h_i u(k-i) + v(k) \end{aligned} \tag{5}$$

where n is a natural number such that $(n+1)T$ is longer than the settling time.

Usually, impulse response identification means that we estimate the unit impulse response $h_i(i = 1, 2, \ldots, n+1)$ from sampled data, instead of the original CTIR $g(\tau)$. Equation (5) can be written in vector form:

$$\begin{aligned} y(k) &= \boldsymbol{p}^T(k)\boldsymbol{h} + v(k) \\ \boldsymbol{p}^T(k) &= [u(k-1), u(k-2) \ldots, u(k-n-1)] \\ \boldsymbol{h}^T &= [h_1, h_2, \ldots, h_{n+1}] \end{aligned} \tag{6}$$

and the vector $\boldsymbol{h}$ can be estimated by the least squares (LS) method as

$$\hat{\boldsymbol{h}} = \left[\frac{1}{N}\sum_{k=n+1}^{n+N} \boldsymbol{p}(k)\boldsymbol{p}^T(k)\right]^{-1} \cdot \left[\frac{1}{N}\sum_{k=n+1}^{n+N} \boldsymbol{p}(k)y(k)\right] \tag{7}$$

provided that the inverse matrix of

$$\mathcal{R} = \left[\frac{1}{N}\sum_{k=n+1}^{n+N} \boldsymbol{p}(k)\boldsymbol{p}^T(k)\right]$$

exists.

It has been derived that the mean square error of the estimated impulse response is inversely proportional to the data length, proportional to the noise variance, and proportional to the sum of the inverses of the eigenvalues of $\mathcal{R}$ [8,9]. From a practical viewpoint, since we cannot fluctuate the system strongly due to the possible resulting damage, it is often required that a band-limited (smoothed) input signal is applied. In this case, it is well known that the input correlation matrix $\mathcal{R}$ may become nearly singular (some of its eigenvalues tend to zero), and hence the estimated impulse response tends to oscillate significantly due to noise effects. Moreover, in the case where the impulse response includes high-frequency components locally, one has to choose a very small sampling interval than ordinarily required to avoid aliasing effects. In the case of a fast sampling rate, $\mathcal{R}$ may also become nearly singular and thus makes the estimated impulse response unacceptable [9].

III. MULTIRESOLUTION APPROXIMATION OF CTIR

A multiresolution analysis of $L^2(R)$ is a sequence $V_i(i \in Z)$ of closed subspaces of $L^2(R)$ such that the following hold [12–15]:

1.

$$\ldots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \ldots;$$

$$\longleftarrow \text{Coarser} \qquad \text{Finer} \longrightarrow$$

2.

$$\bigcap_{m=-\infty}^{+\infty} V_m = \{0\}, \qquad \bigcup_{m=-\infty}^{+\infty} V_m \text{ is dense in } L^2(R);$$

3.

$$g(t) \in V_m \Leftrightarrow g(2t) \in V_{m-1}, \qquad g(t) \in L^2(R);$$

4.

$$g(t) \in V_0 \Rightarrow g(t - iT) \in V_0, \qquad g(t) \in L^2(R);$$

5.

$$\exists \phi(t) \in V_0 \text{s.t.} \{\phi(t - iT) i \in Z\} \text{ is an orthonormal basis in } V_0.$$

Based on multiresolution approximation theory, the space V_0 can be decomposed as

$$\begin{aligned} V_0 &= W_1 \oplus V_1 \\ &= W_1 \oplus W_2 \oplus V_2 \\ &\vdots \\ &= W_1 \oplus W_2 \oplus \ldots \oplus W_L \oplus V_L \end{aligned} \tag{8}$$

where

$$\begin{aligned} V_m &= \operatorname{span}\{2^{-m/2}\phi(2^{-m}\tau - iT)\} \\ W_m &= \operatorname{span}\{2^{-m/2}\psi(2^{-m}\tau - iT)\} \end{aligned} \tag{9}$$

for $m \in Z$, $\phi(\tau)$ and $\psi(\tau)$ are the scaling function and the wavelet, respectively. It should be noticed here that the subspaces $W_1, W_2, \ldots, W_L, V_L$ are mutually orthogonal [12–15]. In the frequency domain, considering the positive frequencies, V_L is the space of functions with frequencies concentrated in the interval $[0, \pi/2^L)$, while W_m is the space of functions with frequencies concentrated in the interval $[\pi/2^m, \pi/2^{m-1})$ [22]. The relations of the subspaces in the frequency domain are shown in Fig. 2.

Various wavelets and corresponding scaling functions have been found and investigated in the literature [12]. The widely reported Meyer scaling function and wavelet are infinitely continuously differentiable and symmetric, and it is therefore easy to investigate their time–frequency localization properties through visualization via computer graphics. In this study we use the Meyer scaling function and wavelet. Normalizing the

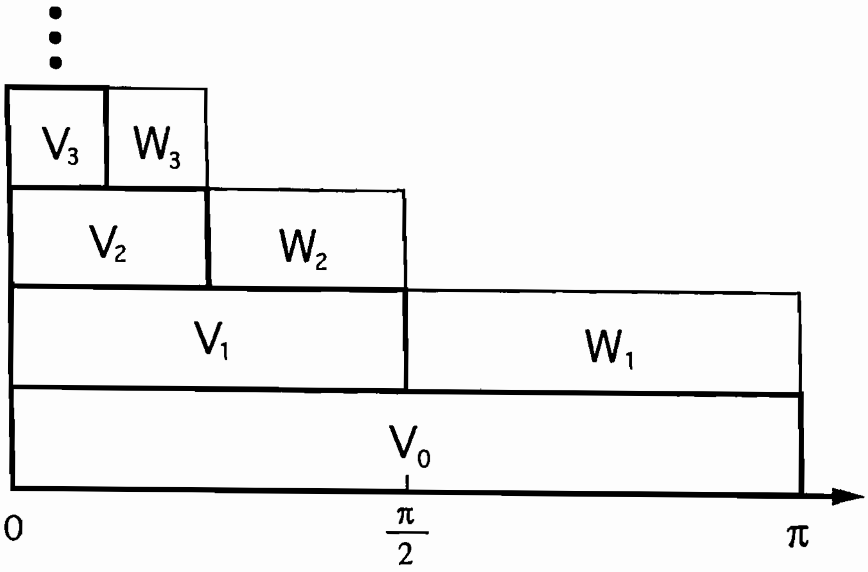


Figure 2 Relations of subspaces in the frequency domain.

sampling period to unity, we have the Fourier transformations of the Meyer scaling function and wavelet respectively [12]:

$$\Phi(\omega) = \begin{cases} 1 & |\omega| \le 2\pi/3 \\ \cos\left[\frac{\pi}{2}\nu\left(\frac{3}{2\pi}|\omega| - 1\right)\right] & 2\pi/3 \le |\omega| \le 4\pi/3 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$\Psi(\omega) = \begin{cases} e^{-i\omega/2}\sin\left[\frac{\pi}{2}\nu\left(\frac{3}{2\pi}|\omega| - 1\right)\right] & 2\pi/3 \le |\omega| \le 4\pi/3 \\ e^{-i\omega/2}\cos\left[\frac{\pi}{2}\nu\left(\frac{3}{4\pi}|\omega| - 1\right)\right] & 4\pi/3 \le |\omega| \le 8\pi/3 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\omega(\tau)$ is a C^n (n an arbitrary natural number) function satisfying

$$\nu(\tau) = \begin{cases} 0 & \tau \le 0 \\ 1 & \tau \ge 1 \end{cases} \tag{12}$$

with the additional property

$$\nu(\tau) + \nu(1 - \tau) = 1 \tag{13}$$

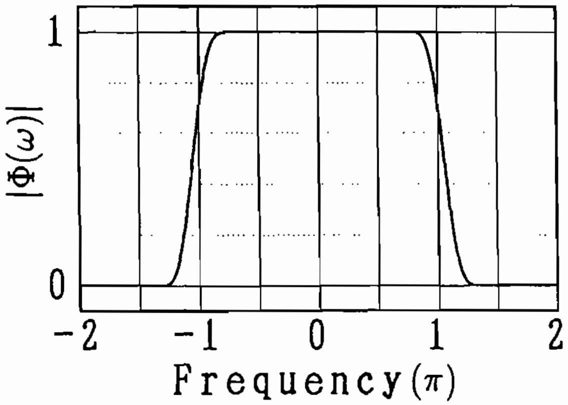


Figure 3 Amplitude of $\Phi(\omega)$.

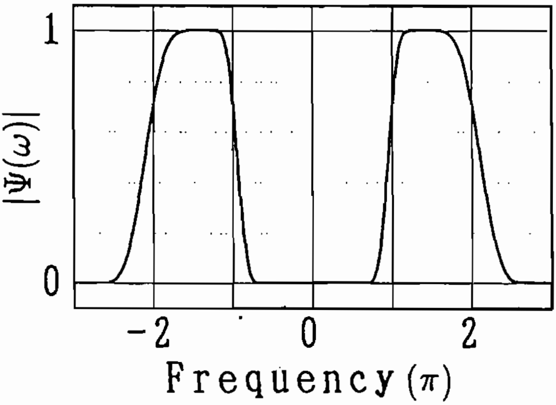


Figure 4 Amplitude of $\Psi(\omega)$.

The amplitudes of $\Phi(\omega)$ and $\Psi(\omega)$ are shown in Figs 3–4, and the inverse Fourier transformations $\phi(\tau)$ and $\psi(\tau)$ are shown in Figs 5–6.

Considering the causality of the impulse response and the fact that the impulse response of a stable system exists in a finite time interval, we have to approximate the function $g(\tau)$ over the finite domain $\mathcal{D} = [0, (n+1)T]$. Then $g(\tau)$ can be approximated in the space V_0 by a wavelet neural network (WNN) of the form [18,23,24].

$$g(\tau) = \sum_{i=0}^{n} c_{0,i}\phi(\tau - iT) \tag{14}$$

with a finite number of nodes.

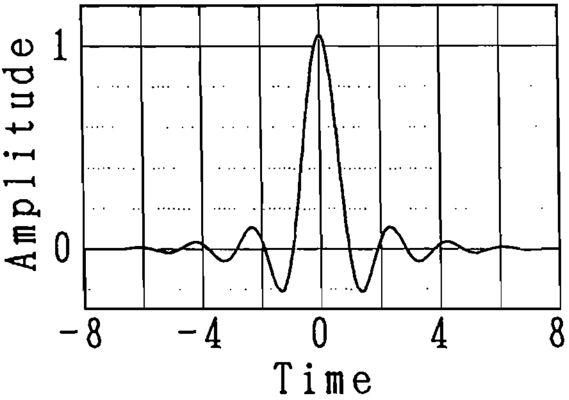


Figure 5 The Meyer scaling function $\phi(\tau)$.

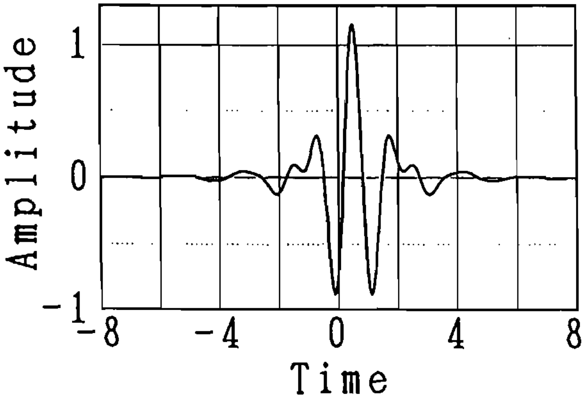


Figure 6 The Meyer wavelet function $\psi(\tau)$.

Detailed discussions on the WNN can be found in Benveniste *et al.* [18], Zhang and Benveniste [23], and Zhang *et al.* [24]. Some practical considerations should be, however, remarked upon here.

Firstly, the sampling period T should be chosen such that it satisfies the Nyquist rate, i.e., the resolution level of the network should be sufficiently fine [24].

Secondly, the structure of the WNN is similar to that of the radial basis function (RBF) network [11]. For practical purposes, the scaling functions need not have compact support and hence the fast-decaying functions such as the Meyer scaling function are also acceptable [24].

Thirdly, since we are only interested in approximating $g(\tau)$ for $\tau \in \mathcal{D}$, the scaling functions are truncated to zero when $\tau \notin \mathcal{D}$. In this case, the coefficients in Eq. (14) should be solved by the LS method or the gradient-descent method based on the LS criterion when the samples of $g(\tau)$ are obtained in $\mathcal{D}$. The common empirical wavelet transformation will lead to poor approximation [24]. Moreover, strict attention should be paid when the scaling functions are truncated, i.e., the dominant part (or the peak value) of the scaling functions should be kept in $\mathcal{D}$ [23]. Otherwise, the approximation may be ill conditioned and hence the solution of the coefficients may not be unique. Empirically, the symmetric Meyer scaling function is recommendable for this purpose since it is fast-decaying and the peak of the main lobe is much more significant compared to the other parts.

Fourthly, it is trivial to understand the fact that in the conventional DTIRM identification method mentioned in Section II, $g(\tau)$ is expressed in the space V_0 by the Haar scaling function (piecewise constant approximation). The Haar scaling function and wavelet, while having compact support, are not continuous and hence do not have good time–frequency localization, which is a major inconvenience for multiresolution approximations.

In this chapter, to extract some information about the impulse response with respect to time and frequency, the WNN named by Zhang *et al.* [24] is further developed to an MRNN in which the scaling and wavelet functions are shifted and dilated based on multiresolution approximation theory. This idea has also been mentioned in Zhang *et al.* [24].

Decompose $g(\tau) \in V_0$ successively such that

$$g(\tau) = g_w^1(\tau) + g_w^2(\tau) + \ldots + g_w^L(\tau) + g_v^L(\tau) \tag{15}$$

where $g_w^m(\tau) \in W_m$ $(m = 1, 2, \ldots, L)$, $g_v^L(\tau) \in V_L$. Equation (15) shows that $g(\tau)$ can be represented as a low-pass approximation at scale L plus the sum of L detail (wavelet) components at different resolutions.

Using the scaling functions and the wavelets as basis functions, $g(\tau)$ can be expressed via the following MRNN:

$$g(\tau) = \sum_{m=1}^{L} \sum_{i=0}^{(n/2^m)-1} d_{m,i}\psi_{m,i}(\tau) + \sum_{i=0}^{n/2^L} c_{L,i}\phi_{L,i}(\tau) \tag{16}$$

where $\phi_{L,i}(\tau)$ and $\psi_{m,i}(\tau)$ are the dilated and shifted scaling function and wavelet respectively:

$$\begin{aligned}\psi_{m,i}(\tau) &= 2^{-m/2}\psi(2^{-m}\tau - iT)\\ \phi_{L,i}(\tau) &= 2^{-L/2}\phi(2^{-L}\tau - iT)\end{aligned} \tag{17}$$

Equation (16) implies that by dilation and shift, the CTIR is decomposed in the time–frequency plane by a family of basis functions which are locally receptive in both the time and frequency domains. The number of parameters in the MRNN (16) is also $n+1$, the same as that in Eq. (5). However, owing to the excellent time–frequency localization properties of the basis functions, it is not necessary to choose all the basis functions in Eq. (16), i.e., some of the basis functions which are redundant for approximation of the impulse response can be neglected, according to the time–frequency characteristics of the impulse response under study.

As mentioned previously, since we are only interested in approximating $g(\tau)$ for $\tau \in \mathcal{D}$, the basis functions are truncated to zero when $\tau \notin \mathcal{D}$. In this case, the coefficients of the MRNN (16) should be solved by the LS method when the samples of $g(\tau)$ are obtained in $\mathcal{D}$. The common empirical wavelet transformation will lead to poor approximation [24]. Moreover, the constraints on the adjustable parameters n and L should be noticed here. n and L should be chosen such that the dominant part (or the peak value) of all the dilated and shifted basis functions in the MRNN (16) are within $\mathcal{D}$. Otherwise, the problem of solving the coefficients of the MRNN (16) may be ill conditioned. Additionally, it is required usually that n and L are chosen such that $n/2^L$ is also a natural number.

In our case, however, the samples of the impulse response $g(\tau)$ are not measurable directly, therefore we have to embed the derived MRNN into the system equation (1) and solve the coefficients from the sampled input–output data.

IV. IDENTIFICATION ALGORITHM

A. Decomposition of the System Model for Identification

Using the MRNN (16), the system equation (1) becomes

$$x(t) = \sum_{m=1}^{L} \sum_{i=0}^{(n/2^m)-1} d_{m,i} x_{m,i}^{w}(t) + \sum_{i=0}^{n/2^L} c_{L,i} x_{L,i}^{v}(t) \tag{18}$$

where

$$\begin{aligned} x_{m,i}^{w}(t) &= \int_0^t \psi_{m,i}(\tau) u(t-\tau)\, \mathrm{d}\tau \\ x_{L,i}^{v}(t) &= \int_0^t \phi_{L,i}(\tau) u(t-\tau)\, \mathrm{d}\tau \end{aligned} \tag{19}$$

Equation (18) implies that the unknown linear system can be divided into $n+1$ subsystems, in which $\psi_{m,i}(\tau)$ and $\phi_{L,i}(\tau)$ $(\tau \in \mathcal{D})$ are interpreted as their known impulse responses respectively.

As mentioned in the previous section, since we are only interested in approximating $g(\tau)$ for $\tau \in \mathcal{D}$ via the MRNN, the basis functions $\psi_{m,i}(\tau)$ and $\phi_{L,i}(\tau)$ are truncated to zero when $\tau \notin \mathcal{D}$. Keeping this fact in mind, we have the following result from Eq. (2) and (18).

$$
\begin{aligned}
y(k) &= \sum_{i=1}^{n+1} u(k-1) \int_{(i-1)T}^{iT} g(\tau)\,\mathrm{d}\tau + v(k) \\
&= \sum_{m=1}^{L} \sum_{i=0}^{(n/2^m)-1} d_{m,i} x^{w}_{m.i}(k) + \sum_{i=0}^{n/2^L} c_{L.i} x^{v}_{L.i}(k) + v(k)
\end{aligned}
\tag{20}
$$

where

$$
\begin{aligned}
x^{v}_{L,i}(k) &= \sum_{l=1}^{n+1} u(k-l) \int_{(l-1)T}^{lT} \phi_{L,i}(\tau)\,\mathrm{d}\tau \\
&= p^T(k) h^{v}_{L,i} \\
x^{w}_{m,i}(k) &= \sum_{l=1}^{n+1} u(k-l) \int_{(l-1)T}^{lT} \psi_{m.i}(\tau)\,\mathrm{d}\tau \quad (m = 1, 2, \ldots, L) \\
&= p^T(k) h^{w}_{m.i}
\end{aligned}
\tag{21}
$$

and

$$
\begin{aligned}
p^T(k) &= [u(k-1), u(k-2) \ldots, u(k-n-1)] \\
h^{vT}_{L,i} &= [h^{v}_{L,i,1}, h^{v}_{L,i,2}, \ldots, h^{v}_{L,i,n+1}] \\
h^{wT}_{m,i} &= [h^{w}_{m,i,1}, h^{w}_{m.i.2}, \ldots, h^{w}_{m,i,n+1}] \quad (m = 1, 2, \ldots, L)
\end{aligned}
\tag{22}
$$

Usually, the error of the identified model includes mainly two components: the bias error (modeling error) and the variance error (effects of noise) [25]. The first component increases when the model complexity decreases, while the second component increases when the model complexity increases. Therefore, a trade-off should be taken between the two antagonistic trends.

It is obvious that the number of unknown parameters in Eq. (20) is also $n+1$, the same as that in Eq. (5). Owing to the excellent time–frequency localization properties of the basis functions, however, it is not necessary to choose all the subsystems in Eq. (20), i.e., some of the subsystems which are redundant can be discarded with little deterioration in modeling accuracy; and the number of parameters needed to approximate the impulse response of interest can be reduced greatly in contrast to the conventional impulse response identification method. Therefore, the key point in the identification algorithm is to determine an estimation model as parsimonious as possible, taking the bias/variance trade-off into account. Here the well-known AIC is utilized as a criterion:

$$\mathrm{AIC}(\rho) = N \ln V + \rho M \tag{23}$$

where

$$V = \frac{1}{N} \sum_{k=n}^{n+N-1} (y(k) - \hat{x}(k))^2 \tag{24}$$

and $\hat{x}(k)$ is the estimated output, and M is the number of the parameters. In this study, ρ is chosen as $\rho = 4$. For detailed discussions on this choice, the reader is referred to [20,21,26,27].

However, in many cases, the total number of subsystems in Eq. (20), $n + 1$, may be tens or hundreds, therefore, in practice it is a difficult task to determine which terms are significant and which terms can be discarded, in the absence of enough *a priori* knowledge. Here, without loss of generality, the problem is simplified based on the following policy.

1. Usually, the coarsest approximation of the impulse response at scale L exists in the whole interval $0 \leq \tau \leq (n+1)T$ and is expressed by a moderate number of basis functions in the space V_L. Therefore, all the $M_{vL} = n/2^L + 1$ basis functions in V_L are selected. It has been reported that in the case of band-limited input, identification of the low-pass approximation of the impulse response is not ill conditioned [9,10].
2. In most cases, for the stable systems, the wavelet spectra of the impulse response in each frequency band are large at short τ and decay when τ becomes longer. Therefore, if we determine the impulse response length, i.e., the number of selected wavelet basis functions in each of the spaces $W_1, W_2, \ldots, W_L$ appropriately, a parsimonious model which is less sensitive to noise can be obtained.

According to the above discussions, the system model (20) becomes

$$\begin{aligned} y(k) &= \sum_{i=1}^{n+1} u(k-i) \int_{(i-1)T}^{iT} g(\tau)\,\mathrm{d}\tau + v(k) \\ &= \sum_{m=1}^{L} \sum_{i=0}^{M_{wm}-1} d_{m,i} x^{w}_{m,i}(k) + \sum_{i=0}^{M_{vL}-1} c_{L,i} x^{v}_{L,i}(k) + v(k) \end{aligned} \tag{25}$$

where M_{wm} denotes the number of the selected wavelet basis functions in W_m. If $M_{wm} = 0$, then none of the wavelet basis functions in W_m is chosen.

The above equation can be rewritten in vector form:

$$\begin{aligned} y(k) &= z^T(k)\boldsymbol{\theta} + v(k) \\ z^T(k) &= [z_{vL}^T(k), z_{w1}^T(k), z_{w2}^T(k), \ldots, z_{wL}^T(k)] \\ \boldsymbol{\theta}^T &= [\boldsymbol{\theta}_{vL}^T, \boldsymbol{\theta}_{w1}^T, \boldsymbol{\theta}_{w2}^T, \ldots, \boldsymbol{\theta}_{wL}^T] \end{aligned} \tag{26}$$

where

$$\begin{aligned} z_{vL}^T(k) &= [x_{L,0}^v(k), x_{L,1}^v(k), \ldots, x_{L,(M_{vL}-1)}^v(k)] \\ z_{wm}^T(k) &= [x_{m,0}^w(k), x_{m,1}^w(k), \ldots, x_{m,(M_{wm}-1)}^w(k)] \quad (m = 1, 2, \ldots, L) \end{aligned} \tag{27}$$

$$\begin{aligned} \boldsymbol{\theta}_{vL}^T &= [c_{L,0}, c_{L,1} \ldots, c_{L,(M_{vL}-1)}] \\ \boldsymbol{\theta}_{wm}^T &= [d_{m,0}, d_{m,1} \ldots, d_{m,(M_{wm}-1)}] \quad (m = 1, 2, \ldots, L) \end{aligned} \tag{28}$$

and the total number of parameters are denoted as $M = M_{vL} + M_{w1} + M_{w2} + \ldots + M_{wL} \leq n + 1$.

B. Brief View of the OLS Method

The preceding discussions imply that the key point of the identification problem is to choose suitable $M_{w1}, M_{w2}, \ldots, M_{wL}$, which determine the lengths of the decomposed components of the impulse response in the spaces $W_1, W_2, \ldots, W_L$, respectively. In our previous work [28], the authors used the genetic algorithm (GA) to determine suitable $M_{w1}, M_{w2}, \ldots, M_{wL}$. The GA is a parallel, global, probabilistic search procedure based on the mechanics of natural selection and natural genetics [29]. Because the GA simultaneously evaluates many points in the search space, it can in effect search many local optima and thereby increases the likelihood of finding the global optimum. Although the method with the aid of the GA proposed by the authors leads to efficient results, a drawback of this method is that it is very computationally demanding due to the use of the GA.

The OLS method, which is a very powerful method that combines model structure selection and parameter estimation via orthogonalization, has been studied widely in the field of nonlinear system identification [20,21,27] in the last decade. Various simulation studies and practical applications have shown that this algorithm provides a simple and powerful means of fitting parsimonious models to real systems. In this study, a heuristic identification procedure using the OLS method is proposed to select suitable $M_{w1}, M_{w2}, \ldots, M_{wL}$, from a large set of candidates.

Solving LS problems by an orthogonal decomposition of the regression matrix is a well-developed technique. Arranging (20) from $k = n + 1$ to $n + N$ yields the following matrix form:

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{P\theta} + \boldsymbol{V} \\ \boldsymbol{y} &= [y(n+1), y(n+2), \ldots, y(n+N)]^T \\ \boldsymbol{V} &= [v(n+1), v(n+2), \ldots, v(n+N)]^T \\ \boldsymbol{P} &= [\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_M] \\ &= [\mathbf{Z}_{vL}, \mathbf{Z}_{w1}, \mathbf{Z}_{w2}, \ldots, \mathbf{Z}_{wL}] \end{aligned} \tag{29}$$

where

$$\begin{aligned} \boldsymbol{p}_i &= [p_i(n+1), p_i(n+2), \ldots, p_i(n+N)]^T \quad (i = 1, 2, \ldots, M) \\ \mathbf{Z}_{vL} &= [\mathbf{Z}_{vL,1}, \mathbf{Z}_{vL,2}, \ldots, \mathbf{Z}_{vL,M_{vL}}] \\ \mathbf{Z}_{vL,i} &= [x^v_{L,(i-1)}(n+1), x^v_{L,(i-1)}(n+2) \ldots, x^v_{L,(i-1)}(n+N)]^T \\ &\qquad (i = 1, 2, \ldots, M_{vL}) \\ \mathbf{Z}_{wm} &= [\mathbf{Z}_{wm,1}, \mathbf{Z}_{wm,2}, \ldots, \mathbf{Z}_{wm,M_{wm}}] \quad (m = 1, 2, \ldots, L) \\ \mathbf{Z}_{wm,i} &= [x^w_{m,(i-1)}(n+1), x^w_{m,(i-1)}(n+2) \ldots, x^w_{m,(i-1)}(n+N)]^T \\ &\qquad (i = 1, 2, \ldots, M_{wm}) \end{aligned} \tag{30}$$

The OLS method involves the transformation of the set of $\boldsymbol{p}_i$ into a set of orthogonal basis vectors, and thus makes it possible to calculate the individual contribution to the desired output energy from each basis vector. The regression matrix $\boldsymbol{P}$ can be decomposed into

$$\boldsymbol{P} = \boldsymbol{WA} \tag{31}$$

where $\boldsymbol{A}$ is an $M \times M$ triangular matrix:

$$\boldsymbol{A} = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \ldots & \alpha_{1M} \\ 0 & 1 & \alpha_{23} & \ldots & \alpha_{2M} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & \ddots & 1 & \alpha_{(M-1)M} \\ 0 & \ldots & 0 & 0 & 1 \end{bmatrix} \tag{32}$$

and $\boldsymbol{W}$ is an $N \times M$ matrix with orthogonal columns $\boldsymbol{w}_i$ such that

$$\boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{D} \tag{33}$$

where D is an $M \times M$ diagonal matrix with diagonal elements d_i:

$$\begin{aligned} d_i &= w_i^T w_i \\ &= \sum_{k=n+1}^{n+N} w_i(k) w_i(k) \quad (i = 1, 2, \ldots, M) \end{aligned} \tag{34}$$

The space spanned by the set of orthogonal basis vectors w_i is the same space spanned by the set of p_i, and (29) can be rewritten as

$$y = Wg + V \tag{35}$$

The orthogonal LS solution $\hat{g}$ is given by

$$\hat{g} = D^{-1} W^T y \tag{36}$$

or

$$\hat{g}_i = \frac{w_i^T y}{w_i^T w_i} \quad (i = 1, 2, \ldots, M) \tag{37}$$

Since the w_i are mutually orthogonal, adding some other regressors does not influence the values of the determined elements of $\hat{g}$. This advantage is very important in forward regression procedures of model structure determination.

The quantities $\hat{g}$ and $\hat{\theta}$ satisfy the triangular system

$$A\hat{\theta} = \hat{g} \tag{38}$$

and the parameter estimate $\hat{\theta}$ is readily computed from the above relation using backward substitution. Various orthogonal techniques can be applied to derive (31) [21]. The well-known classical Gram–Schmidt method which is most used for the OLS method computes one column of A at a time and orthogonalizes P as follows: at the jth stage make the jth column orthogonal to each of the $j-1$ previously orthogonalized columns and repeat the operation for $j = 2, 3, \ldots, M$. The computational procedure can be represented as

$$\left.\begin{aligned} & w_1 = p_1 \\ & \alpha_{ij} = \frac{w_i^T p_j}{w_i^T w_i} \qquad (1 \le i \le j) \\ & w_j = p_j - \sum_{i=1}^{j-1} \alpha_{ij} w_i \end{aligned}\right\} \quad j = 2, \ldots, M \tag{39}$$

The OLS method has superior numerical properties compared with the ordinary LS method. Our interest in the OLS method, however, is to use it for model structure determination. In studies of nonlinear system identification, the regressor selection procedure is based on the individual contribution to the desired output energy from each basis vector, and some statistical criteria such as the Akaike-type criteria can be employed to terminate the selection [20,21,27]. Practically, however, the problem of impulse response is often seriously ill conditioned especially in the case of fast sampling rate and band-limited input, and hence the common regressor selection procedure does not yield satisfying results in many cases. In this study, as mentioned above we simplify the problem to the determination of the length of the impulse response in each subspace, i.e., to choose suitable $M_{w1}, M_{w2}, \ldots, M_{wL}$, which determine the lengths of the decomposed components of the impulse response in spaces $W_1, W_2, \ldots, W_L$ respectively.

C. Hierarchical Identification Procedure

The heuristic identification procedure is performed in a hierarchical manner. The impulse response is estimated from a coarse resolution level (low-frequency subspace) to a fine resolution level (high-frequency subspace) successively via the OLS method. At each resolution level, the AIC is utilized to select an appropriate length of the impulse response (the number of wavelet basis functions in the corresponding subspace).

The algorithm of the proposed multiresolution identification method for the system impulse response is described as follows.

Step 1. Identify the impulse response in V_L space. As mentioned previously, all the $M_{vL} = n/2^L + 1$ basis functions in V_L are selected. The OLS solution and AIC in V_L space are calculated as follows:

$$w_1 = Z_{vL,1}$$

$$\hat{g}_1 = \frac{w_1^T y}{w_1^T w_1}$$

$$\left.\begin{aligned} \alpha_{ij} &= \frac{w_i^T Z_{vL,j}}{w_i^T w_i} \qquad (1 \le i < j) \\ w_j &= Z_{vL,j} - \sum_{i=1}^{j-1} \alpha_{ij} w_i \\ \hat{g}_j &= \frac{w_j^T y}{w_j^T w_j} \end{aligned}\right\} \quad j = 2, \ldots, M_{vL} \tag{40}$$

$$\mathrm{AIC}_{M_{vL}}(4) = N \ln V + 4M_{vL} \tag{41}$$

Step 2. Identify the impulse response in spaces $W_L, W_{(L-1)}, \ldots, W_1$ successively:

Step 2a. Set the number of selected basis functions to $M = M_{vL}$.

Step 2b. Set the space (resolution level) index to $m = L$.

Step 2c. Calculate

$$\left.\begin{aligned} \alpha_{ij} &= \frac{w_i^T Z_{wm,j}}{w_i^T w_i} \qquad (1 \le i < j) \\ w_j &= Z_{wm,j} - \sum_{i=1}^{j-1} \alpha_{ij} w_i \\ \hat{g}_j &= \frac{w_j^T y}{w_j^T w_j} \end{aligned}\right\} \quad j = M+1, \ldots, M + \hat{M}_{wm} \tag{42}$$

where $\hat{M}_{wm}(1 \le \hat{M}_{wm} \le n/2^m)$ satisfies

$$\mathrm{AIC}_{M+\hat{M}_{wm}}(4) < \mathrm{AIC}_{M+l}(4) \quad \text{for } \hat{M}_{wm} \neq l \tag{43}$$

and $1 \le l \le n/2^m$.

Step 2d. Set the number of selected basis functions to $M = M + \hat{M}_{wm}$.

Step 2e. Decrease the space index to $m = m - 1$ and go to step 2c until $m = 1$.

Step 3. Check the redundancy of the selected subsystems in high-frequency spaces, and remove those subsystems which are considered to be redundant. That is, find a suitable space index m from $1, 2, \ldots, L$ such that $AIC_M(4)$ reaches its minimal value, where the number of subsystems M is given by $M = M - \hat{M}_{w1} - \hat{M}_{w2} - \ldots - \hat{M}_{wm}$.

Step 4. Compute the estimates of the original parameters using backward substitution from

$$\hat{\theta} = \hat{g} - (A - I)\hat{\theta} \tag{44}$$

Step 5. Synthesize the estimated discrete-time impulse response of the system as

$$\hat{h} = \sum_{m=1}^{L} \sum_{i=0}^{\hat{M}_{wm}-1} \hat{d}_{m,i} h_m^w + \sum_{i=0}^{M_{vL}-1} \hat{c}_{L,i} h_L^v \tag{45}$$

V. SIMULATION RESULTS

To illustrate the effectiveness of the proposed identification algorithm, some numerical examples are shown in this section.

A. System 1

We consider the following system with a zero-order hold in the input:

$$x(t) = \frac{3.98p^2 - 10.3p + 1880}{p^3 + 17p^2 + 980p + 962} u(t) \tag{46}$$

where p denotes the differential operator.

Since the impulse response of this nonminimum phase system includes high-frequency components locally (the reverse action in the first part), a small sampling period $T = 0.04$ (sec) is chosen. The length of the impulse response is assumed to be $(n+1)T = 129 \times 0.04 = 5.16$ (sec), and a short data length for identification is given as $N = 300$.

The band-limited input is the output of a third-order Butterworth filter

$$u(t) = \frac{1}{(p/\omega_c)^3 + 2(p/\omega_c)^2 + 2(p/\omega_c) + 1} \eta(t) \tag{47}$$

where $\eta(t)$ is a discrete-time white noise:

$$\eta(t) = \eta(k), \quad kT \leq t < (k+1)T \tag{48}$$

The filter is implemented by the bilinear transformation at a smaller interval $T' = T/10$.

The simulation results in the case of noise/signal ratio (NSR) $\approx 10\%$ for various values of the cut-off frequency ω_c of the input shaping filter such as $0.4\pi/T$, $0.6\pi/T$, $0.8\pi/T$, $1.0\pi/T$ are shown in Table 1 and Figs 7–8. Here the NSR is defined as $\sigma_v/\sigma_x\%$ where σ_v and σ_x denote the standard deviation of the sampled measurement noise and true system output, respectively. In Figs 7–8, the dashed line and the solid line represent the true impulse response and its estimate, respectively. In Table 1, Conv. and Prop. denote the conventional method and the proposed method respectively, and the relative mean square error (RMSE) is defined as

$$\text{RMSE} = \frac{\|h - \hat{h}\|^2}{\|h\|^2} \tag{49}$$

From Table 1 and Figs 7–8, it can be found that in the ill-conditioned cases of band-limited input, fast sampling rate, and short data length, the

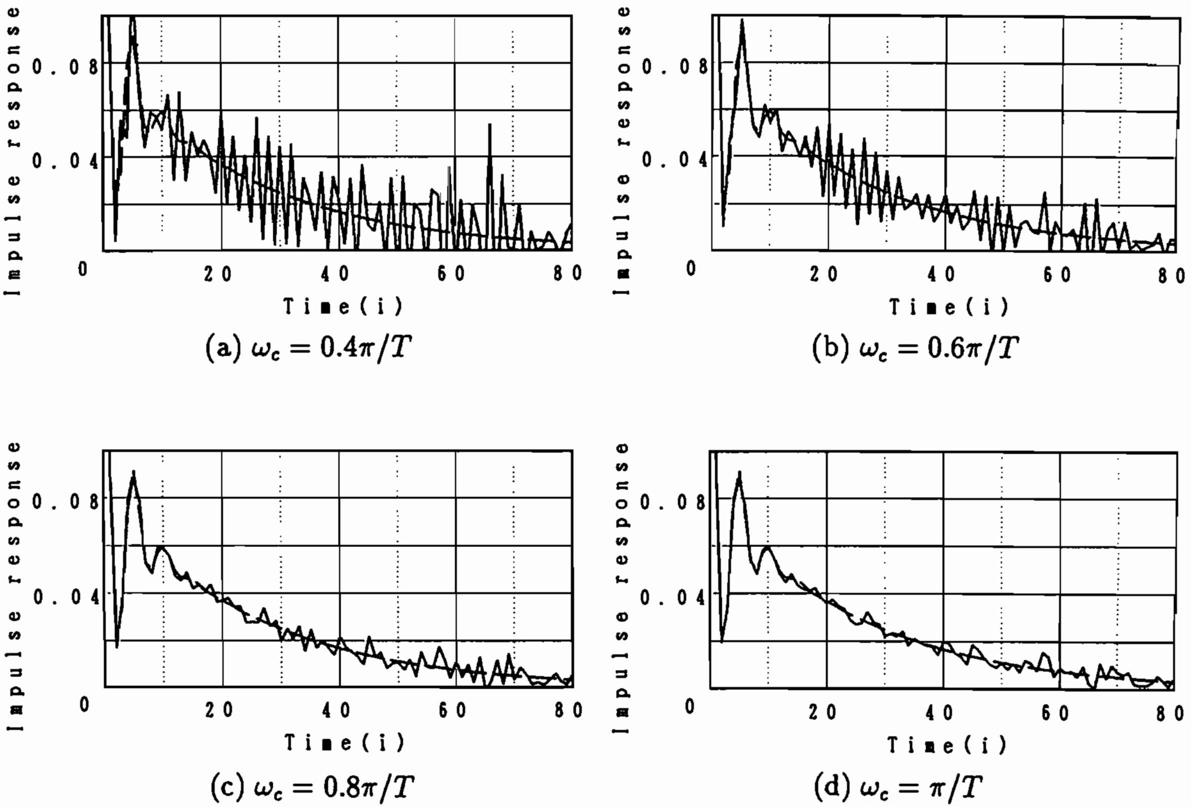
Table 1 Results of System 1

	RMSE		Parameter number in each subspace					Total
ω_c	Conv.	Prop.	$\hat{M}_{w1}$	$\hat{M}_{w2}$	$\hat{M}_{w3}$	$\hat{M}_{w4}$	M_{v4}	$\hat{M}$
$0.4\pi/T$	0.3913	0.0069	0	1	2	2	9	14
$0.6\pi/T$	0.1237	0.0032	2	4	1	2	9	18
$0.8\pi/T$	0.0199	0.0030	2	4	1	2	9	18
$1.0\pi/T$	0.0102	0.0045	2	4	1	1	9	17

estimated impulse response by the conventional method oscillates roughly due to the effects of the measurement noise. However, the proposed method yields a very accurate estimate of the impulse response, and the rapidly changing components are also estimated correctly. Additionally, the conventional method requires the estimation of 129 parameters whereas our proposed method requires the estimation of only 14–17 parameters. As an example, the frequency responses of the estimated impulse responses by the conventional method and the proposed method respectively when $\omega_c = 0.4\pi/T$ are shown in Fig. 9, where the dashed line and the solid line represent the true frequency response and its estimate, respectively. By the proposed method, the redundant subsystems in the high-frequency domain which are sensitive to noise effects in the case of band-limited signal are discarded. Therefore, it can be verified that our method is quite superior to the conventional one, especially in the high-frequency domain.

It can also be found that when the cut-off frequency of the input shaping filter becomes smaller, ie., the identification problem gets more ill conditioned, the proposed algorithm tends to neglect more subsystems in the high-frequency domain, and hence the system model becomes more parsimonious. This fact implies that the proposed algorithm makes a trade-off between the bias error (modeling error) and the variance error (effects of noise) appropriately.

In this example, we only show the results when $L = 4$, i.e., the coarsest approximation of the impulse response is given at scale $L = 4$. It has been verified empirically that when $L = 3, 4, 5, 6$, the results are quite similar, that is, the results are not sensitive to the selection of L. According to the discussions in Section IV, a suitable L is usually chosen such that identification of the low-pass approximation of the impulse response at scale L is not ill conditioned.



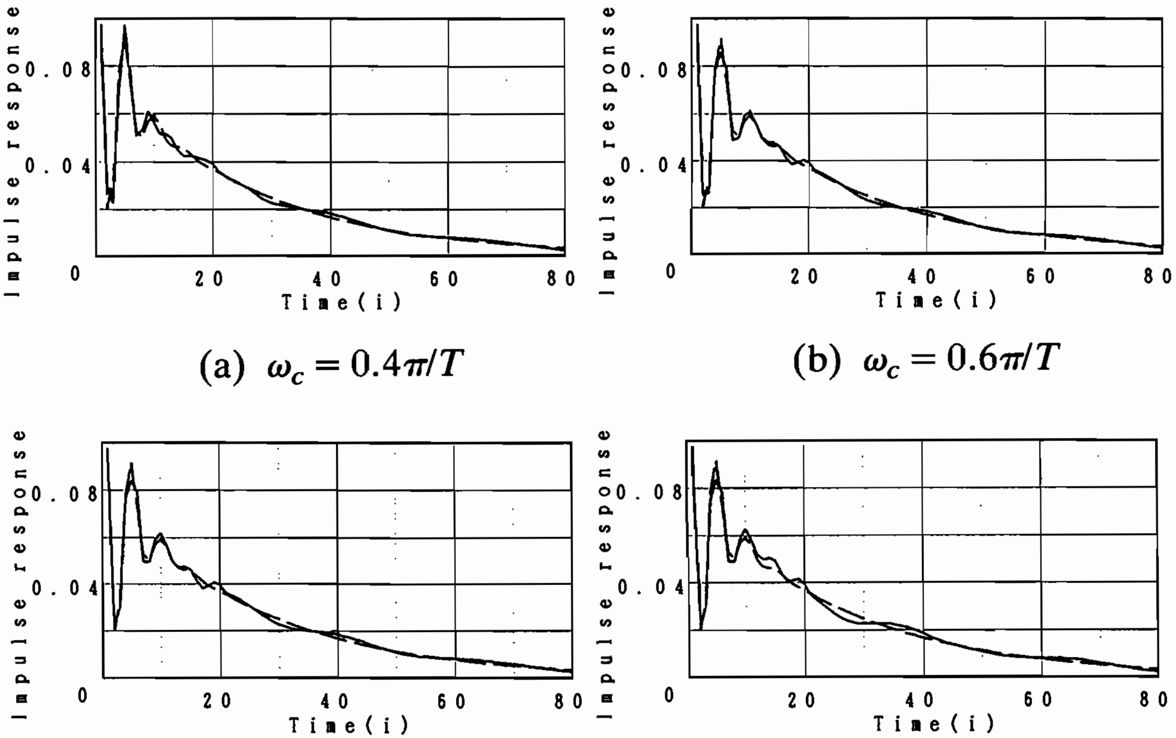
(a) $\omega_c = 0.4\pi/T$

(b) $\omega_c = 0.6\pi/T$

(c) $\omega_c = 0.8\pi/T$

(d) $\omega_c = \pi/T$

Figure 7 Estimated impulse response of System 1 (conventional method).



(a) $\omega_c = 0.4\pi/T$

(b) $\omega_c = 0.6\pi/T$

(c) $\omega_c = 0.8\pi/T$

(d) $\omega_c = \pi/T$

Figure 8 Estimated impulse response of System 1 (proposed method).

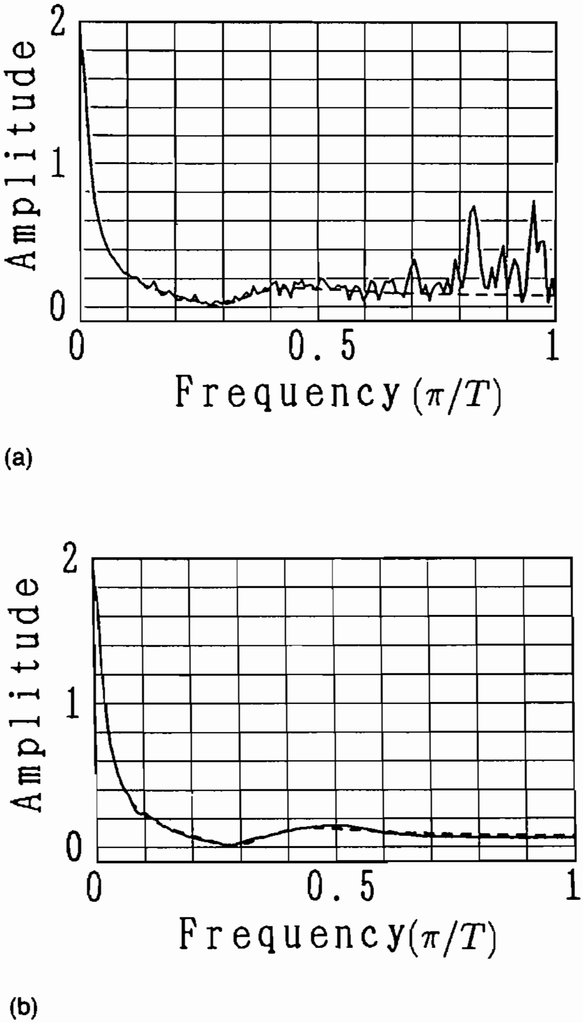


Figure 9 Estimated frequency response of System 1 ($\omega_c = 0.4\pi/T$): (a) conventional method; (b) proposed method.

B. System 2

The second example is the following oscillatory system with a zero-order hold in the input:

$$x(t) = \frac{0.25}{p^2 + 0.2p + 0.25} u(t) \tag{50}$$

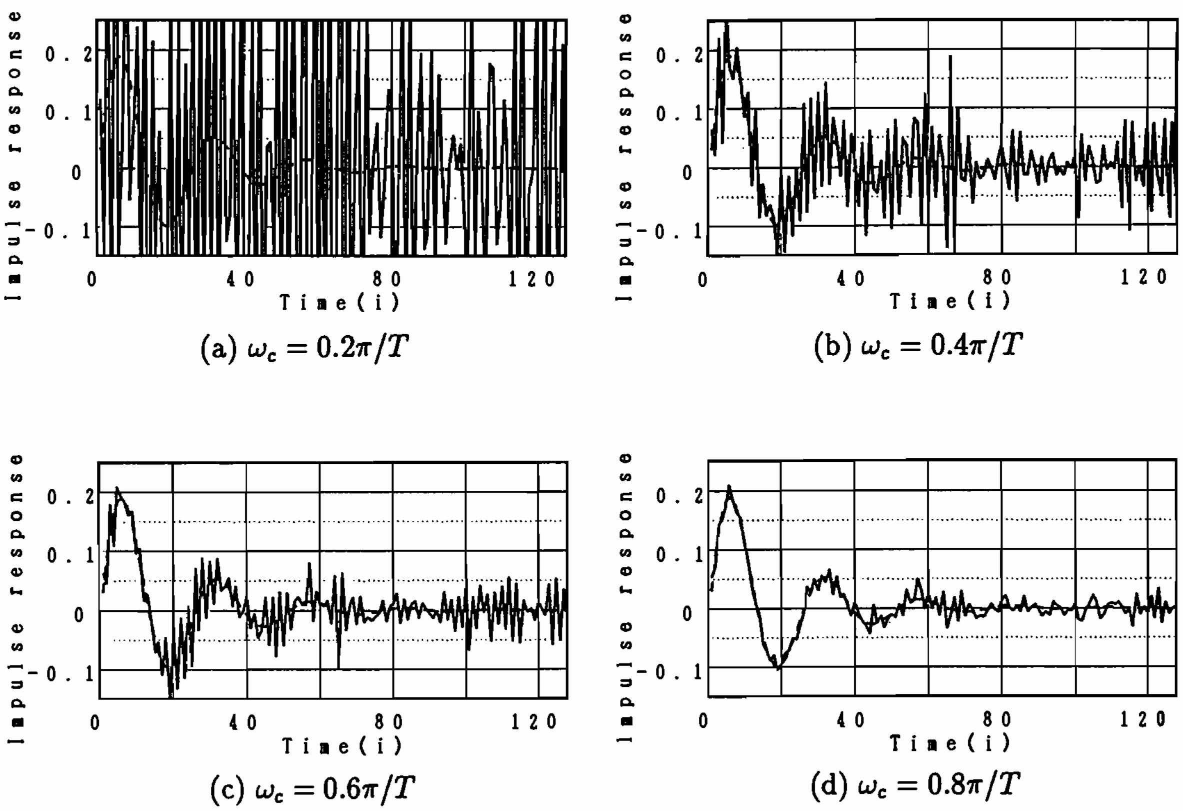
Table 2 Results of System 2

	RMSE		Parameter number in each subspace					Total
ω_c	Conv.	Prop.	$\hat{M}_{w1}$	$\hat{M}_{w2}$	$\hat{M}_{w3}$	$\hat{M}_{w4}$	M_{v4}	$\hat{M}$
$0.2\pi/T$	57.148	0.0043	0	0	3	5	9	17
$0.4\pi/T$	1.5694	0.0033	0	0	4	5	9	18
$0.6\pi/T$	0.4968	0.0037	0	0	4	5	9	18
$0.8\pi/T$	0.0798	0.0039	0	0	4	5	9	18

The sampling period is chosen as $T = 0.5$ (sec), and the length of the impulse response is determined as $(n+1)T = 129 \times 0.5 = 64.5$ (sec). The data length and the input shaping filter, etc., are the same as those for System 1. The results when NSR $\approx 20\%$ are shown in Table 2 and Figs 10–11. Since the impulse response of this system does not include high-frequency components, the subsystems in the high-frequency domain which are sensitive to noise effects in the case of band-limited input signal are discarded. This improves the numerical properties of the impulse response estimate. Since the wavelet basis functions that approximate the impulse response in spaces W_1, W_2 are omitted, it can be found that our method is also very effective in the case where the sampling period is shorter than ordinarily required, which may lead to an ill-conditioned problem.

VI. CONCLUSIONS

In this work, the authors have proposed a new identification method for the discrete-time impulse response model of a linear system from sampled input–output data using multiresolution approximation theory. Our attention is especially focused on systems whose impulse responses have locally rapidly changing components. The CTIR of the system under study is approximated by an MRNN based on multiresolution approximation theory. Hence the system under study can be viewed as weighted summation of a group of subsystems in which the shifted and dilated scaling functions and wavelet functions are interpreted as their impulse responses respectively. Then a heuristic identification procedure with the aid of the OLS method and AIC is proposed. The impulse response is estimated from a coarse resolution level (low-frequency subspace) to a fine resolution level (high-frequency subspace) successively via the OLS



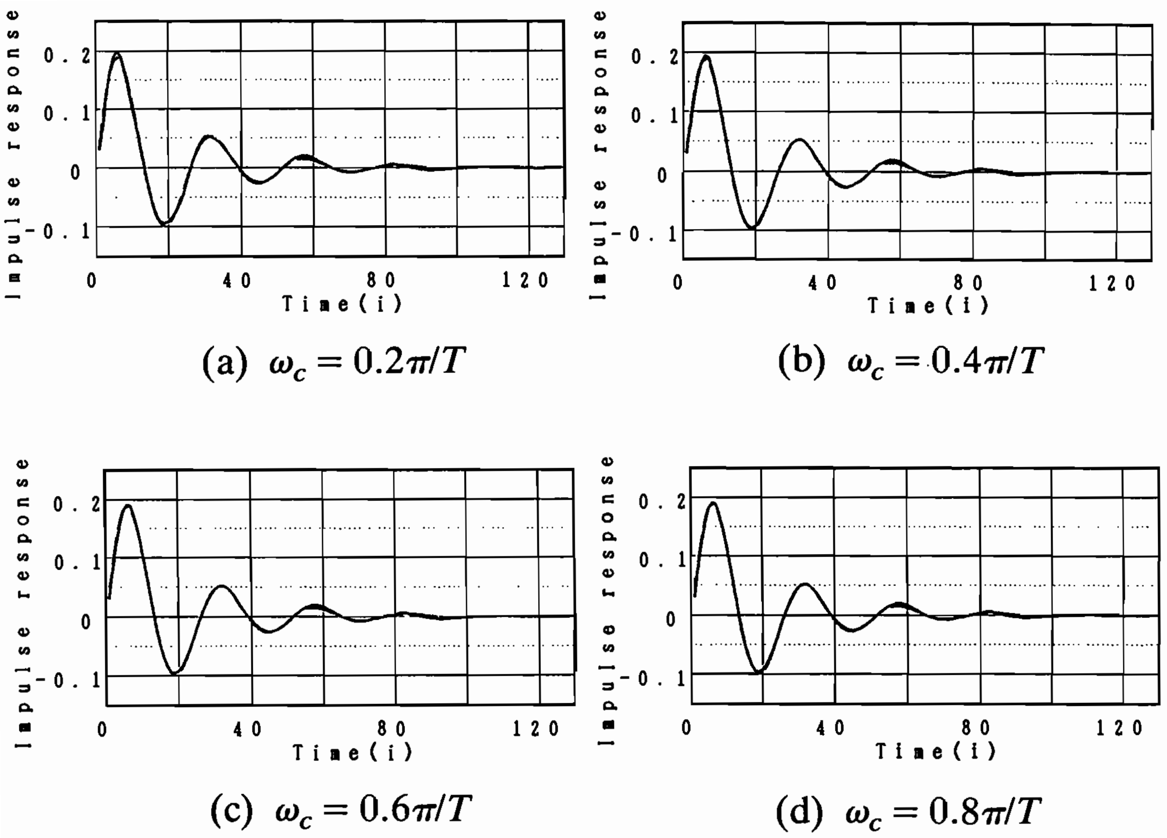
(a) $\omega_c = 0.2\pi/T$

(b) $\omega_c = 0.4\pi/T$

(c) $\omega_c = 0.6\pi/T$

(d) $\omega_c = 0.8\pi/T$

Figure 10 Estimated impulse response of System 2 (conventional method).



(a) $\omega_c = 0.2\pi/T$

(b) $\omega_c = 0.4\pi/T$

(c) $\omega_c = 0.6\pi/T$

(d) $\omega_c = 0.8\pi/T$

Figure 11 Estimated impulse response of System 2 (proposed method).

method. At each resolution level, the AIC is utilized to select an appropriate length of the impulse response (the number of wavelet basis functions in the corresponding subspace). Simulation results show that the proposed method yields an accurate estimate of the impulse response with locally rapidly changing components, even in the ill-posed cases of band-limited input, fast sampling rate, and significant measurement noise.

REFERENCES

1. L Ljung. System identification: Theory for the User. Prentice-Hall, 1987.
2. JN Juang. Applied System Identification. Prentice-Hall, 1994.
3. B Wahlberg. IEEE Trans on Automatic Control 36:551–562, 1991.
4. B Wahlberg, EJ Hannan. Estimation of noise models by means of discrete Laguerre/Kautz filters. In: the 9th IFAC/IFORS Sympo. on Identification and System Parameter Estimation. Budapest, July 1991, pp 1236–1241.
5. AE Hoer. Technometrics 12:55–67, 1970.
6. GH Golub, M Heath, G Wahba. Technometrics 21:215–223, 1979.
7. Y Miyasato. System identification in consideration of prior information. Proc. SIC 89, Matsuyama, Japan, July, 1989, pp 1077–1080.
8. A Sano, H Ohmori, M Kamegai. Stabilized identification via GSVD optimized based on Bayesian information theoretic criterion and minimum mean squares error. In: the 9th IFAC/IFORS Sympo on Identification and System Parameter Estimation, Budapest, July, 1991, pp 1536–1541.
9. A Sano, H Tsuji. Optimal sampling rate for system identification based on decimation and interpolation. In: the 12th IFAC World Congress, Sydney, vol. IV, 1993, pp 219–222.
10. ZJ Yang, T Tsuji, S Sagara. Impulse response identification of continuous systems using generalized radial basis function networks. In: the 10th IFAC/IFORS Sympo on Identification and System Parameter Estimation, Copenhagen, vol. 3, July, 1994, pp 465–470.
11. T Poggio, F Girosi. Proc. IEEE 78:1481–1497, 1990.
12. I Daubechies. Ten Lectures on Wavelets. Philadelphia: Society for Industrial and Applied Mathematics, 1992.
13. AN Akansu, RA Haddad. Multiresolution Signal Decomposition Transforms, Subbands and Wavelets. New York: Academic Press, 1992.
14. SG Mallat. Trans Amer Math Soc 315:69–87, 1989.
15. SG Mallat. IEEE Trans on Pattern Analysis and Machine Intelligence 11: 674–693, 1989.
16. H Maeda, T Kitagawa, S Kodama. Trans IEICE. J76-A:1565–1570, 1993 (in Japanese).
17. T Tabaru, S Shin, T Kitamori. Identification of frequency response based on multiresolution analysis. In: Proceedings of the Asian Control Conference, vol. 2, Tokyo, July, 1994, pp 413–416.

18. A Benveniste, A Juditsky, B Delyon, Q Zhang, P-Y Glorennec. Wavelets in identification. In: the 10th IFAC/IFORS Sympo. on Identification and System Parameter Estimation, vol. 2, Copenhagen, July, 1994, pp 27–48.
19. AA Safavi, GW Barton, JA Romagnoli. On the choice of wavelet in a wave-net for system identification. In: Proceedings of the Asian Control Conference, vol. 1, Tokyo, July, 1994, pp 77–80.
20. S Chen, SA Billings, W Luo. Int J Control 50:1873–1896, 1989.
21. S Chen, SA Billings, CFN Cowan, PM Grant. Int J Syst Sci 21:2513–2539, 1990.
22. M Vetterli, C Herley. IEEE Trans Signal Processing 40:2207–2232, 1992.
23. Q Zhang, A Benveniste. IEEE Trans on Neural Networks 3:889–898, 1992.
24. J Zhang, GG Walter, WN Lee. IEEE Trans Signal Processing 43:1485–1497, 1995.
25. B Ninness, G Goodwin. Estimation of model quality. In: the 10th IFAC/IFORS Sympo. on Identification and System Parameter Estimation, vol. 1, Copenhagen, July, 1994, pp 25–44.
26. TG Freeman. Automatica 21:361–370, 1985.
27. M Pottmann, H Unbehauen, DE Seborg. Int J Control 57:97–120, 1993.
28. ZJ Yang, S Sagara, T Tsuji. Impulse response identification using multiresolution analysis. In: the 5th IFAC Sympo on Adaptive Systems in Control and Signal Processing, Budapest, June, 1995, pp 515–520.
29. DE Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, 1989.

This Page Intentionally Left Blank

6
Comparative Study of Rank Test Methods for ARMA Order Estimation

Joakim Sorelius, Torsten Söderström, Petre Stoica, and Mats Cedervall

Uppsala University, Uppsala, Sweden

I. INTRODUCTION

Autoregressive moving average (ARMA) models are extensively used in signal processing, automatic control, and system identification to describe signals generated from a physical system. The reason is that a vast class of stationary signals can, as far as second-order moments are concerned, be described as the output of a stable rational filter, driven by white noise, i.e., these signals can be described by ARMA models. This follows from the famous theorem by Weierstrass, which implies that we can approximate the signal arbitrarily well by increasing the order of the filter. However, for practical reasons, it is not feasible to use models of a too high degree. Except for the computational burden it may induce, many algorithms used in signal processing become numerically unreliable as the poles and zeros of the system approach one another. The conclusion is that we must determine a finite, but not too large, order of the ARMA process. Yet the order should be chosen so that the ARMA model describes the signal under study with an accuracy that is sufficient for the application at hand.

There is a large number of methods that can be used for order estimation of both scalar and multivariable ARMA processes. The literature up to 1985 is reviewed and commented on in [1], and a more recent reference is [2]. Many classical order estimation schemes, such as the final prediction error (FPE) and Akaike's information criterion (AIC) [3,4], or improved versions of these methods, can be difficult to use because of their computational complexity and the multivariable search they require. In brief, these methods evaluate a criterion function, which depends on the parameters of the ARMA model, for many different orders of the process. The criterion function is then inspected, and the order that gives its minimum value, or the order at which the decrease in the criterion function becomes insignificant, is chosen as the relevant order of the process under study.

Another class of ARMA order estimators is formed by the so-called *rank test methods*. These methods rely on the fact that a certain Hankel matrix formed from the true covariances of the studied signal has rank equal to the system order, and statistical methods can then be applied to estimate that rank from the data. As is the case of the visual order estimators, the rank tests are related to parameter estimation. Indeed, the recently developed methods for subspace-based state space system identification (4SID) operate on the same Hankel covariance matrix (and its singular value decomposition) as the rank tests, see, e.g., [5–7]. These estimation methods have the advantage over other system identification techniques that they do not require any special canonical parametrization of the system, or estimation of its so-called Kronecker invariants. In fact, they only need the order of a minimal state space realization of the process to provide consistent estimates of the system dynamics from observed data. This order is exactly what is provided by the rank test methods.

In the following sections of this chapter we will describe and compare some recently proposed tests for the estimation of the order of a possibly multivariable ARMA process. An outline of the presentation is as follows. Section II introduces the necessary notation and concepts and gives a proof of the main result concerning the rank properties of the ARMA covariance matrix. In Section III we describe three different rank test methods for ARMA order estimation. They are compared and evaluated in Section IV by means of some numerical examples, and Section V finally states some conclusions from the experiments.

II. AUTOREGRESSIVE MOVING AVERAGE PROCESSES

The general form of an autoregressive moving average model considered here is defined as follows:

$$A(q^{-1})y(t) = C(q^{-1})e(t) \tag{1}$$

In Eq. (1), q^{-1} denotes the backward shift operator, i.e., $q^{-k}y(t) = y(t-k)$ (for $k = 0, 1, 2, \ldots$), $y(t)$ is the ny-dimensional output at time t, and $e(t)$ is the input, sometimes called the driving noise vector, which is assumed to be of the same dimensions as $y(t)$. The left coprime $(ny \times ny)$-dimensional polynomial matrices $A(q^{-1})$ and $C(q^{-1})$ are defined through the relations

$$A(q^{-1}) = I + A_1 q^{-1} + \cdots + A_{na} q^{-na}, \qquad A_{na} \neq 0 \tag{2}$$

$$C(q^{-1}) = I + C_1 q^{-1} + \cdots + C_{nc} q^{-ne}, \qquad C_{ne} \neq 0 \tag{3}$$

and it is assumed that

$$\det A(z) \neq 0 \quad |z| \leq 1$$

$$\det C(z) \neq 0 \quad |z| \leq 1$$

so that $A(z)$ and $C(z)$ have all zeros strictly outside the unit circle. Further, $e(t)$ is a white noise process, i.e., $e(t)$ is a sequence of independent and identically distributed random variables of zero mean and with a positive definite covariance matrix, given by

$$Ee(t)e^T(s) = \Lambda_e \delta_{t,s} \tag{4}$$

where E denotes the expectation operator and $\delta_{t,s}$ is the Kronecker delta. In the model (1)–(4) we have chosen $A_0 = C_0 = I$, and the resulting ARMA is called monic. Note that this choice of A_0 and C_0 introduces no large restriction. Indeed, any nonsingular A_0 or $C_0 \neq I$ can be factorized out of (1) and incorporated in the covariance matrix Λ_e. In what follows we will assume that $e(t)$ is normally distributed, which will simplify the treatment in the coming sections.

The input/output relation (1) can be written as

$$y(t) = \mathcal{G}(q^{-1})e(t) \tag{5}$$

The transfer function $\mathcal{G}(q^{-1})$, of dimension $ny \times ny$, given by

$$\mathcal{G}(q^{-1}) = A^{-1}(q^{-1})C(q^{-1}) \tag{6}$$

is stable, has a stable inverse, and fulfills $\mathcal{G}(0) = I$. Conversely, given a rational strictly positive definite spectral density $\Phi(\omega)$ there exists a unique rational filter $\mathcal{G}(q^{-1})$ and a covariance matrix Λ_e such that

$$\Phi(\omega) = \frac{1}{2\pi} \mathcal{G}(\mathrm{e}^{-\mathrm{i}\omega})\Lambda_e \mathcal{G}^T(\mathrm{e}^{\mathrm{i}\omega}) \tag{7}$$

$$\mathcal{G}(q^{-1}) \quad \text{and} \quad \mathcal{G}^{-1}(q^{-1}) \quad \text{are asymptotically stable,} \tag{8}$$

and $\mathcal{G}(0) = I$. This is the spectral factorization theorem [8] and implies that a signal with rational spectral density $\Phi(\omega)$ can be described by the model (4)–(5). This signal can be represented as a monic ARMA process described by (1)–(3) (see, e.g., [9]).

The general ARMA process (1) can also be described through the following state-space representation of order r:

$$\begin{aligned} x(t+1) &= Fx(t) + Ge(t) \\ y(t) &= Hx(t) + e(t) \end{aligned} \tag{9}$$

where $x(t)$ is the r-dimensional state vector and $e(t)$ is the previously introduced white noise. The representation (9) is assumed to be minimal, i.e., it uses the minimal number of states. If F, G and H are such that the transfer function $\mathcal{G}(q^{-1}) = I + H(qI - F)^{-1}G$ and its inverse $\mathcal{G}^{-1}(q^{-1}) = I - H(qI - F + GH)^{-1}G$ both are asymptotically stable, it is usually referred to as the innovation representation [10]. The model (9) will be used in the sequel for the derivation of the main result for the rank test methods. The order of the ARMA process represented by (9) is defined as follows.

Definition 1. The order r of a (possibly multivariable) ARMA process, described by (1)–(4), is defined as the order of its minimal state space realization (9). ■

As a consequence of Definition 1, it holds that for a scalar ARMA process,

$$r = \max(na, nc) \tag{10}$$

with na and nc as defined in (2)–(3). For a multivariable ARMA, the relationship between the ARMA order r and the degrees of the $A(q^{-1})$ and $C(q^{-1})$ polynomials is not so straightforward (see, e.g., [11]).

For the order estimation problem it is not necessary to specify a candidate state space or ARMA model; we shall see that the rank tests for ARMA order estimation only operate on different versions of the covariance matrix of the ARMA process under study. Let

$$R(\tau) = Ey(t)y^T(t-\tau) \quad (ny \times ny) \tag{11}$$

denote the autocovariance at lag τ. Then define the general covariance matrix

$$\mathcal{R}_{p,q} \triangleq \begin{pmatrix} R(q) & R(q-1) & \dots & R(q-p+1) \\ R(q+1) & R(q) & \dots & \\ R(q+2) & & & \\ \vdots & & \ddots & \\ R(q+p-1) & & & R(q) \end{pmatrix} \tag{12}$$

$$= EY_p(t)Y_p^T(t-q) \tag{13}$$

where

$$Y_k(t) \triangleq (y^T(t)y^T(t+1) \cdots y^T(t+k-1))^T \qquad (k\,ny \times 1) \tag{14}$$

The square matrix $\mathcal{R}_{p,q}$ is block Toeplitz and is of dimension $p\,ny \times p\,ny$. Often we will work with the block Hankel counterpart of (12), obtained by permuting the block columns of $\mathcal{R}_{p,q}$:

$$\mathcal{H}_{p,q} \triangleq \mathcal{R}_{p,q}J \tag{15}$$

$$= \begin{pmatrix} R(q-p+1) & R(q-p+2) & \dots & R(q) \\ R(q-p+2) & R(q-p+3) & \dots & \\ R(q-p+3) & & & \\ \vdots & & \ddots & \\ R(q) & & & R(q+p-1) \end{pmatrix} \tag{16}$$

where J is the $p\,ny \times p\,ny$ permutation matrix

$$J = \begin{pmatrix} 0 & & & I \\ & & I & \\ & \cdot & & \\ \cdot & & & \\ I & & & 0 \end{pmatrix} \tag{17}$$

i.e., J is the matrix with a number of p identity matrices of dimension $ny \times ny$ placed along its main (block) antidiagonal. We note that the block Hankel matrix $\mathcal{H}_{p,q}$ and the block Toeplitz matrix $\mathcal{R}_{p,q}$ share the same rank properties, since the rank of a (block) matrix is unaffected by permutation of its (block) columns. We now turn to the main result concerning the rank of the matrix $\mathcal{R}_{p,q}$, or equivalently $\mathcal{H}_{p,q}$, which has a potential for ARMA order estimation. We formulate this result as a theorem.

Theorem 1. *It holds that*

$$\operatorname{rank} \mathcal{H}_{p,p} = r \quad \text{for } p \geq r \tag{18}$$

Proof. We first obtain a well-known factorization of $\mathcal{H}_{r,r}$. From (9), we have that

$$x(t+k) = F^k x(t) + \sum_{l=0}^{k-1} F^{k-l-1} Ge(t+l) \tag{19}$$

and hence

$$Ex(t+k)x^T(t) = F^k P \quad \text{for } k \geq 0 \tag{20}$$

where $P = Ex(t)x^T(t)$ is the (positive definite) covariance matrix of the state vector, and also

$$Ex(t+k)e^T(t) = F^{k-1}G\Lambda_e \tag{21}$$

It follows from (9) that

$$\begin{aligned} R(\tau) &= HF^{\tau}PH^T + HF^{\tau-1}G\Lambda_e \\ &= HF^{\tau-1}D \quad \text{for } \tau \geq 1 \end{aligned} \tag{22}$$

with $D \triangleq FPH^T + G\Lambda_e$ $(r \times ny)$. According to the above calculations we have

$$\mathcal{H}_{r,r} = \begin{pmatrix} H \\ HF \\ \vdots \\ HF^{r-1} \end{pmatrix} (D\, FD \cdots F^{r-1}D) \tag{23}$$

$$\triangleq \mathcal{O}(H,F)\mathcal{C}(F,D) \tag{24}$$

where $\mathcal{O}(H,F)$ and $\mathcal{C}(F,D)$ are the observability and controllability matrices of a state space model associated with F, H and D. By the Cayley–Hamilton theorem [12], each matrix F satisfies its own characteristic equation

$$\det(qI - F) = q^r + f_1 q^{r-1} + \cdots + f_r = 0 \tag{25}$$

so that

$$F^r = -f_1F^{r-1} - f_2F^{r-2} - \cdots - f_rI \tag{26}$$

This enables us to claim that for $p \geq r$, we can increase the dimension of $\mathcal{H}_{p,p}$ without changing its rank properties. Indeed, with the obvious changes of the number of block columns and block rows of $\mathcal{O}(H,F)$ and $\mathcal{C}(F,D)$ the factorization (23) holds true for any $\mathcal{H}_{p,p}$. For $p > r$, each of the $p-r$ last blocks in $\mathcal{O}$ and in $\mathcal{C}$ can be written as a linear combination of the r first blocks, according to (26), and thus the rank of $\mathcal{H}_{p,p}$ is unaffected. Hence, for $p \geq r$, $\mathcal{H}_{p,p}$ has the same rank as

$$\mathcal{H}_{\infty,\infty} = E\begin{pmatrix} y(t) \\ y(t+1) \\ \vdots \end{pmatrix} (y^T(t-1)y^T(t-2) \cdots) \tag{27}$$

$$= \begin{pmatrix} R(1) & R(2) & R(3) & . & . \\ R(2) & R(3) & . & . & \\ R(3) & . & . & & \\ . & . & & & \end{pmatrix} \tag{28}$$

which is the infinite dimensional block Hankel covariance matrix. From (19), we have that

$$y(t+k) = HF^k x(t) + (HF^{k-1}G \dots HG\,I)\begin{pmatrix} e(t) \\ \vdots \\ e(t+k) \end{pmatrix} \tag{29}$$

which, in turns, yields the following equation

$$\begin{pmatrix} y(t) \\ y(t+1) \\ y(t+2) \\ \vdots \end{pmatrix} = \begin{pmatrix} H \\ HF \\ HF^2 \\ \vdots \end{pmatrix} x(t) + \begin{pmatrix} I & & & \\ HG & I & & 0 \\ HFG & HG & I & \\ & & & \ddots \end{pmatrix}\begin{pmatrix} e(t) \\ e(t+1) \\ e(t+2) \\ \vdots \end{pmatrix} \tag{30}$$

Post-multiplying both sides of (30) by $(y^T(t-1)y^T(t-2)\dots)$ and taking expectation, we obtain

$$\mathcal{H}_{\infty,\infty} = \begin{pmatrix} H \\ HF \\ HF^2 \\ \vdots \end{pmatrix} Ex(t)(y^T(t-1)y^T(t-2)\dots) \tag{31}$$

$$= \mathcal{O}_\infty(H,F)\Gamma \tag{32}$$

where $\mathcal{O}_\infty(H,F) \triangleq (H^T F^T H^T \dots)^T$ is the extended (infinite dimensional) observability matrix. The matrix $\Gamma \triangleq Ex(t)(y^T(t-1)y^T(t-2)\dots)$ can be factorized as follows. From (9), we can write

$$x(t) = (G \;\; FG \;\; F^2G \dots)\begin{pmatrix} e(t-1) \\ e(t-2) \\ \vdots \end{pmatrix} \tag{33}$$

and introducing the extended (infinite dimensional) controllability matrix

$$\mathcal{C}_\infty(F,G) \triangleq (G \;\; FG \;\; F^2G \dots) \tag{34}$$

we have

$$\Gamma = \mathcal{C}_\infty(F,G)E\begin{pmatrix} e(t-1) \\ e(t-2) \\ \vdots \end{pmatrix}(y^T(t-1)y^T(t-2)\dots)$$

elements have zero mean and variance of order N^{-1}. This follows from the central limit theorem and the fact that the estimates (41) that constitute the entries of $\hat{\mathscr{H}}_{p,p}$ are consistent and asymptotically normal, see, e.g., [8,13]. The above result implies that

$$\tilde{\mathscr{H}}_{p,p} = O(N^{-1/2}) \tag{43}$$

where $O(N^{-\alpha})$ (for some $\alpha \geq 0$) thus denotes a term whose elements have standard deviation of the order of $N^{-\alpha}$. We can also write

$$\sqrt{N}\,\mathrm{vec}(\hat{\mathscr{H}}_{p,p} - \mathscr{H}_{p,p}) \sim \mathscr{N}(0, V) \tag{44}$$

In (44), vec $(\cdot)$ denotes the operator that stacks the columns of a matrix $(\cdot)$ on top of each other. The matrix V can be viewed as the asymptotic, normalized, covariance matrix of vec $\hat{\mathscr{H}}_{p,p}$. Since $\tilde{\mathscr{H}}_{p,p}$ is a random matrix, $\hat{\mathscr{H}}_{p,p}$ has in general full rank for any p, and the result of Theorem 1 is thus valid only asymptotically in N. This motivates the development of a statistical test which would enable us to determine the rank of $\mathscr{H}_{p,p}$, given the estimate $\hat{\mathscr{H}}_{p,p}$. The rank tests for ARMA order estimation to be presented here typically determine an m-vector $\hat{\mu}$ from a factorization of the sample covariance matrix. Under the null hypothesis

$$H_0 : \mathrm{rank}\, \mathscr{H}_{p,p} = r \tag{45}$$

$\hat{\mu}$ is shown to satisfy

$$\sqrt{N}\hat{\mu} \sim \mathscr{N}(0, P) \tag{46}$$

for some positive definite matrix P, which depends on V in (44). It then holds true that the test quantity

$$\hat{\eta} \triangleq N\hat{\mu}^T \hat{P}^{-1} \hat{\mu} \sim \chi^2(m) \tag{47}$$

where $\hat{P}$ is a consistent estimate of P. According to (47), $\hat{\eta}$ is asymptotically χ^2-distributed with m degrees of freedom, and the following significance test can be used to decide upon the rank of $\mathscr{H}_{p,p}$:

$$\begin{cases} \hat{\eta} \leq \chi^2_\alpha(m) \Rightarrow \mathrm{rank}\, \mathscr{H}_{p,p} = r & (H_0 \text{ is true}) \\ \hat{\eta} > \chi^2_\alpha(m) \Rightarrow \mathrm{rank}\, \mathscr{H}_{p,p} > r & (\text{reject } H_0) \end{cases} \tag{48}$$

The significance level α is defined as

$$\alpha = \mathrm{prob}(u > \chi^2_\alpha(m) \,|\, u \sim \chi^2(m)) \tag{49}$$

The parameter α is called the probability of false alarm; it is the probability of declaring rank $\mathscr{H}_{p,p} > r$ when in fact it holds true that rank $\mathscr{H}_{p,p} = r$. The threshold $\chi^2_\alpha(m)$ for different values of m and α can be read from a table of the χ^2 distribution, see, e.g., [14].

III. RANK TEST METHODS

A. A Rank Test Based on the Eigenvalue Decomposition

In this section a rank test based on the eigenvalue decomposition (EVD) of the Hankel matrix (15) is described. This test was first developed by J.-J Fuchs [15,16]. We first treat the estimation of the order of a scalar ARMA process, and then give a brief indication of how to extend the test to include the more complicated multivariable case.

1. Scalar Case

Consider the ARMA process given by (1), where the polynomials $A(q^{-1})$ and $C(q^{-1})$ are assumed to be scalars. In this simple case it is possible to do more than merely estimating $r = \max(na, nc)$, the order of the minimal state space realization of the ARMA. We shall see that we can estimate both $na = \deg A$ and $nc = \deg C$ from a sample of N measurements of $y(t)$, by using the more general Hankel matrix (16) rather than the one discussed in Theorem 1. Indeed, consider the $\hat{na} \times \hat{na}$ Hankel matrix

$$\mathcal{H}_{\hat{na},\hat{nc}} = EY_{\hat{na}}(t)Y^T_{\hat{na}}(t - \hat{nc})J \tag{50}$$

$$= \begin{pmatrix} R(\hat{nc} - \hat{na} + 1) & \dots & R(\hat{nc}) \\ \vdots & \ddots & \\ R(\hat{nc}) & & R(\hat{nc} + \hat{na} - 1) \end{pmatrix} \tag{51}$$

$$= \mathcal{R}_{\hat{na},\hat{nc}}J. \tag{52}$$

When $R(\tau)$ appearing above is a scalar, the rank properties of $\mathcal{H}_{\hat{na},\hat{nc}}$, or equivalently those of $\mathcal{R}_{\hat{na},\hat{nc}}$, are well known [17]. Introduce the integer

$$r^* = \min(\hat{na} - na, \hat{nc} - nc) \tag{53}$$

Consider the case when $\hat{na} = \hat{nc}$. From Theorem 1, it follows that $r^* \leq 0$ (which means that $\hat{na} \leq na$) gives an $\mathcal{H}_{\hat{na},\hat{na}}$ matrix of full rank, and that $r^* > 0$ gives a singular $\mathcal{H}_{\hat{na},\hat{na}}$. When $\hat{na} \neq \hat{nc}$, we can show that the determinant of $\mathcal{H}_{\hat{na},\hat{nc}}$ satisfies

$$\det \mathcal{H}_{\hat{na},\hat{nc}} \begin{cases} = 0 & \text{for } r^* > 0 \\ \neq 0 & r^* = 0 \\ \text{generically} \neq 0 & r^* < 0 \end{cases} \tag{54}$$

By "generically" above we understand that systems which give a singular

$\mathcal{H}_{\hat{na},\hat{nc}}$ for some $r^* < 0$ can occur, but this happens with probability zero since the parameters of such a system must fulfill a number of nontrivial relations. A discussion of this matter, as well as a formal proof of (54), is given in [17].

The properties (54) of $\mathcal{H}_{\hat{na},\hat{nc}}$ can be used to estimate the degrees *na* and *nc*, respectively. Indeed, testing the singularity of $\mathcal{H}_{\hat{na},\hat{nc}}$ for different combinations of $\hat{na}$ and $\hat{nc}$ will reveal how the parameter r^* depends on $\hat{na}$ and $\hat{nc}$ as described in detail below.

Algorithm 1. Scalar Fuchs Test

1. Test $\mathcal{H}_{\hat{na},\hat{na}}$ for singularity for increasing $\hat{na} = 2, 3, \ldots$. Stop at $\hat{na} = j$, with $\mathcal{H}_{j,j}$ the first singular matrix in the sequence.
   Set $r = j - 1$ and go to Step 2.
2. If $\mathcal{H}_{r+1,r}$ is nonsingular go to Step 3.
   Test $\mathcal{H}_{r+1,r-i}$ for singularity for increasing $i = 1, 2, \ldots$. Step at the first nonsingular matrix $\mathcal{H}_{r+1,r-i}$ in the sequence. We then have $na = r$ and $nc = r - i$. END
3. Test $\mathcal{H}_{r-i,r+1}$ for singularity for increasing $i = 0, 1, 2, \ldots$. Stop at the first nonsingular matrix $\mathcal{H}_{r-i,r+1}$ in the sequence. We then have $na = r - i$ and $nc = r$. END ■

Note that in order to obtain r, the order of the ARMA under study, just Step 1 needs to be performed.

The test is constructed in such a way that $\mathcal{H}_{\hat{na},\hat{nc}}$ admits at most one zero eigenvalue $\lambda_{\hat{na},\hat{nc}}$, and we will always have $r^* \geq 0$. The corresponding sample covariance matrix, $\hat{\mathcal{H}}_{\hat{na},\hat{nc}}$, will generically have full rank and the results (54) are valid only asymptotically in N. Let $\hat{\lambda}_{\hat{na},\hat{nc}}$ denote the smallest eigenvalue of $\hat{\mathcal{H}}_{\hat{na},\hat{nc}}$, and introduce the null hypothesis

$$H_0: \quad \lambda_{\hat{na},\hat{nc}} = 0 \quad (\mathcal{H}_{\hat{na},\hat{nc}} \text{ singular}) \tag{55}$$

We now turn to the derivation of the statistical properties of $\hat{\lambda}_{\hat{na},\hat{nc}}$, which will enable us to develop a statistical test for deciding whether $\mathcal{H}_{\hat{na},\hat{nc}}$ is to be declared singular or not. First observe that, under the null hypothesis (i.e., $r^* > 0$), the exact covariance matrix $\mathcal{H}_{\hat{na},\hat{nc}}$ appearing in the test is always of one of two possible types:

$$\begin{aligned} &\text{Type 1: } \mathcal{H}_{\hat{na},\hat{nc}};\ \hat{na} = na + 1,\ \hat{nc} = nc + i,\ 1, 2, \ldots, na - nc + 1 \\ &\text{Type 2: } \mathcal{H}_{\hat{na},\hat{nc}};\ \hat{nc} = nc + 1,\ \hat{na} = na + i,\ 1, 2, \ldots, nc - na + 1 \end{aligned} \tag{56}$$

The following lemma holds concerning the eigenvectors of the Type 1 and Type 2 matrices:

Lemma 1. *Under the null hypothesis (55), the eigenvector of $\mathcal{H}_{\hat{na},\hat{nc}}$ corresponding to the distinct zero eigenvalue is*

$$g_1 = (a_{na} \cdots a_1 \ 1)^T \tag{57}$$

or

$$g_2 = (0 \cdots 0 \ a_{na} \cdots a_1 \ 1)^T \tag{58}$$

for a Type 1 and Type 2 matrix, respectively.

Proof. The proof of Lemma 1 is based on the Yule–Walker equations, which state that the covariance function $R(\tau)$ of an ARMA process (1) satisfies the relation

$$R(k) = -\sum_{j=1}^{na} a_j R(k+j) \quad k > nc \tag{59}$$

First consider matrices of Type 1: the lth row of $\mathcal{H}_{na+1,nc+i}g_1$ is

$$\begin{aligned}\{\mathcal{H}_{na+1,nc+i}g_1\}_l &= a_{na}R(nc+i-na+l-1) + a_{na-1}R(nc+i-na+l) \\ &\quad + \cdots + R(nc+i+l-1) \\ &= R(nc+i+l-1) + \sum_{j=1}^{na} a_j R(nc+i+l-1+j) \\ &= 0 \quad \text{for } l = 1, 2, \ldots, na+1\end{aligned} \tag{60}$$

In the same manner, for Type 2 matrices, the lth row of $\mathcal{H}_{na+i,nc+1}g_2$ is

$$\begin{aligned}\{\mathcal{H}_{na+i,nc+1}g_2\}_l &= a_{na}R(nc-na+l) + a_{na-1}R(nc-na+l+1) \\ &\quad + \cdots + R(nc+l) \\ &= R(nc+l) + \sum_{j=1}^{na} a_j R(nc+l+j) \\ &= 0 \quad \text{for } l = 1, 2, \ldots, na+i\end{aligned} \tag{61}$$

which completes the proof. ■

In order to establish the statistical distribution of the smallest eigenvalue of $\hat{\mathcal{H}}_{\hat{na},\hat{nc}}$ needed for the singularity test we must distinguish between Type 1 and Type 2 matrices. Here we will focus on Type 2 matrices, but the arguments are easily modified to the case of $\hat{\mathcal{H}}_{\hat{na},\hat{nc}}$ being of Type 1, and the final results are the same for both cases. (The derivation for Type 1 matrices is described in detail in [15].)

We partition the EVD of a Type 2 matrix $\mathcal{H}_{na+i,nc+1}$, of dimension $(na+i)\times(na+i)$, as

$$\mathcal{H}_{na+i,nc+1} = (S\ \bar{g}_2)\begin{pmatrix}\Lambda & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}S^T\\ \bar{g}_2^T\end{pmatrix} \tag{62}$$

where the $na+i-1$ columns of the matrix S contain the normalized eigenvectors corresponding to the nonzero eigenvalues of $\mathcal{H}_{na+i,nc+1}$ and the vector $\bar{g}_2 \triangleq g_2/\|g_2\|$ is the normalized eigenvector corresponding to the distinct zero eigenvalue. Also, $\bar{g}_2$ and the columns of S are orthogonal. The corresponding decomposition of a sample Type 2 matrix is

$$\hat{\mathcal{H}}_{na+i,nc+1} = (\hat{S}\ \hat{\bar{g}}_2)\begin{pmatrix}\hat{\Lambda} & 0\\ 0 & \hat{\lambda}_{na+i,nc+1}\end{pmatrix}\begin{pmatrix}\hat{S}^T\\ \hat{\bar{g}}_2^T\end{pmatrix} \tag{63}$$

where $\hat{S}$ and $\hat{\bar{g}}_2$ are defined similarly to S and $\bar{g}_2$. Standard results from matrix perturbation theory [18] give that $\bar{g}_2$ and $\hat{\bar{g}}_2$ are related as

$$\tilde{g}_2 \triangleq \hat{\bar{g}}_2 - \bar{g}_2 = O(N^{-1/2}) \tag{64}$$

Thus, using the orthogonality properties of $\hat{\bar{g}}_2$ and $\hat{S}$ we obtain (note that $\hat{\mathcal{H}}_{na+i,nc+1}\bar{g}_2 = (\hat{\mathcal{H}}_{na+i,nc+1} - \mathcal{H}_{na+i,nc+1})\bar{g}_2 = O(N^{-1/2})$)

$$\begin{aligned}\hat{\lambda}_{na+i,nc+1} &= \hat{\bar{g}}_2^T\hat{\mathcal{H}}_{na+i,nc+1}\hat{\bar{g}}_2\\ &= (\bar{g}_2 + \tilde{g}_2)^T\hat{\mathcal{H}}_{na+i,nc+1}(\bar{g}_2 + \tilde{g}_2)\\ &= \bar{g}_2^T\hat{\mathcal{H}}_{na+i,nc+1}\bar{g}_2 + \bar{g}_2^T\hat{\mathcal{H}}_{na+i,nc+1}\tilde{g}_2\\ &\quad + \tilde{g}_2^T\hat{\mathcal{H}}_{na+i,nc+1}\bar{g}_2 + \tilde{g}_2^T\hat{\mathcal{H}}_{na+i,nc+1}\tilde{g}_2\\ &= \bar{g}_2^T\hat{\mathcal{H}}_{na+i,nc+1}\bar{g}_2 + O(N^{-1})\end{aligned} \tag{65}$$

By using (41) and (50) we can write

$$\hat{\mathcal{H}}_{na+i,nc+1} = \frac{1}{N}\sum_{t=1}^{N} Y_{na+i}(t)Y_{na+i}^T(t-nc-1)J \tag{66}$$

Combining (65) and (66) yields

$$\begin{aligned}\hat{\lambda}_{na+i,nc+1} &= \frac{1}{N}\sum_{t=1}^{N}\bar{g}_2^T Y_{na+i}(t)Y_{na+i}^T(t-nc-1)J\bar{g}_2 + O(N^{-1})\\ &= \frac{1}{N}\sum_{t=1}^{N} v(t)w(t) + O(N^{-1})\end{aligned} \tag{67}$$

In (67) we have introduced the two stationary sequences $v(t)$ and $w(t)$ defined as follows:

$$\begin{aligned} v(t) &\triangleq \bar{g}_2^T Y_{na+i}(t) \\ &= \frac{1}{\|g_2\|}(a_{na}y(t+i-1)+\cdots+y(t+i-1+na)) \\ &= \bar{g}_1^T Y_{na+1}(t+i-1) \end{aligned} \tag{68}$$

with $\bar{g}_1 \triangleq g_1/\|g_1\|$ and where we have made use of the fact that, from (57)–(58), $\|g_1\| = \|g_2\|$. Furthermore,

$$\begin{aligned} w(t) &\triangleq Y_{na+i}^T(t-nc-1)J\bar{g}_2 \\ &= \frac{1}{\|g_2\|}(a_{na}y(t-nc-1+na)+\cdots+y(t-nc-1)) \\ &= \bar{g}_1^T J Y_{na+1}(t-nc-1) \end{aligned} \tag{69}$$

We now use the expressions (67)-(69) to evaluate the mean and the variance of $\lambda_{na+i,nc+1}$. To that end we first determine the autocovariance and cross-covariance sequences of $v(t)$ and $w(t)$. The autocovariance sequence for $v(t)$ is

$$\begin{aligned} \gamma_k &\triangleq Ev(t)v(t-k) \\ &= \bar{g}_1^T E\{Y_{na+1}(t+i-1)Y_{na+1}^T(t+i-1-k)\}\bar{g}_1 \\ &= \bar{g}_1^T \mathcal{R}_{na+1,k}\bar{g}_1. \end{aligned} \tag{70}$$

Similarly, for $w(t)$ we get

$$\begin{aligned} Ew(t)w(t-k) &= \bar{g}_1^T JE\{Y_{na+1}(t-nc-1)Y_{na+1}^T(t-nc-1-k)\} J\bar{g}_1 \\ &= \bar{g}_1^T J\mathcal{R}_{na+1,k} J\bar{g}_1 \\ &= \bar{g}_1^T \mathcal{R}_{na+1,k}\bar{g}_1 \\ &= \gamma_k \end{aligned} \tag{71}$$

where, in the third equality, $J\mathcal{R}J = \mathcal{R}$ follows from (11)–(12). If k in (70) and (71) is larger than nc, $\mathcal{R}_{na+1,k}$ is of Type 1 and it follows from Lemma 1 that $\gamma_k = 0$ for $k > nc$. By the stationarity of $w(t)$, $\gamma_{-k} = \gamma_k$ and hence

$$\gamma_k = 0, \quad |k| > nc \tag{72}$$

The cross-covariances between $v(t)$ and $w(t)$ are given by

$$\begin{aligned} \nu_k &\triangleq Ev(t)w(t-k) \\ &= \bar{g}_1^T E\{Y_{na+1}(t+i-1)Y_{na+1}^T(t-nc-1-k)J\}\bar{g}_1 \\ &= \bar{g}_1^T \mathcal{H}_{na+1,nc+i+k}\bar{g}_1 \end{aligned} \tag{73}$$

For $k > -i$, the covariance matrix $\mathscr{H}_{na+1,nc+i+k}$ is of Type 1, so from Lemma 1 we have that $\nu_k = 0$, $k > -i$, and particularly

$$\nu_k = 0, \quad \text{for } k \geq 0 \tag{74}$$

The above calculations together with the central limit theorem thus give that $\hat{\lambda}_{na+1,nc+i}$ as expressed in (67) can be viewed as an asymptotically Gaussian distributed, zero mean, random variable [13,19] with variance given by

$$\begin{aligned}
E\hat{\lambda}^2_{na+i,nc+1} &= E\left(\frac{1}{N}\sum_{t=1}^{N} v(t)w(t) + O(N^{-1})\right)^2 \\
&= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N} Ev(t)w(t)v(s)w(s) + O(N^{-2}) \\
&= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N} \{Ev(t)w(t)Ev(s)w(s) + Ev(t)v(s)Ew(t)w(s) \\
&\quad + Ev(t)w(s)Ew(t)v(s)\} + O(N^{-2}) \\
&= \frac{1}{N^2}\sum_{k=-N}^{N}(N - |k|)\{\nu_0^2 + \gamma_k^2 + \nu_k\nu_{-k}\} + O(N^{-2}) \\
&= \frac{1}{N}\left(\gamma_0^2 + 2\sum_{k=1}^{nc}\gamma_k^2\right) + O(N^{-2})
\end{aligned} \tag{75}$$

where we have used (72) and (74) and the formula for the expectation of a product of four Gaussian random variables [20].

Note that the results on the distribution of the smallest eigenvalue of $\hat{\mathscr{H}}_{na+i,nc+1}$ obtained above are valid also for matrices of Type 1 [15]. Denote by $\hat{g}$ the eigenvector obtained from the EVD (63) of $\hat{\mathscr{H}}_{\widehat{na},\widehat{nc}}$ of either type and let $\hat{\sigma}^2_{\widehat{na},\widehat{nc}}$ be the estimate of the asymptotic variance of its smallest eigenvalue $\hat{\lambda}_{\widehat{na},\widehat{nc}}$. Form the following test quantity:

$$\hat{\eta} = \frac{\hat{\lambda}^2_{\widehat{na},\widehat{nc}}}{\hat{\sigma}^2_{\widehat{na},\widehat{nc}}} \tag{76}$$

To obtain $\hat{\sigma}^2_{\widehat{na},\widehat{nc}}$, replace all quantities in (75) with their sample estimates (note that $nc + 1$ in (75) is replaced by $\widehat{nc}$) so that

$$\hat{\sigma}^2_{\widehat{na},\widehat{nc}} = \frac{1}{N}\left(\hat{\gamma}_0^2 + 2\sum_{k=1}^{\widehat{nc}-1}\hat{\gamma}_k^2\right) \tag{77}$$

The estimates $\{\hat{\gamma}_k\}_{k=0}^{\widehat{nc}-1}$ are obtained from (70) by replacing $na+1$ by $\widehat{na}$ (see also (67)–(68)) which yields

$$\hat{\gamma}_k = \hat{\bar{g}}^T \hat{\mathcal{R}}_{\widehat{na},k} \hat{\bar{g}} \tag{78}$$

It follows from the statistical properties of $\hat{\lambda}_{\widehat{na},\widehat{nc}}$ that under H_0 of (55) and asymptotically in N, $\hat{\eta}$ is χ^2 (1) distributed. Thus, the following test can be used to decide upon the rank of $\mathcal{H}_{\widehat{na},\widehat{nc}}$:

$$\begin{cases} \hat{\eta} \leq \chi^2_\alpha(1) \Rightarrow \mathcal{H}_{\widehat{na},\widehat{nc}} \text{ singular} \\ \hat{\eta} > \chi^2_\alpha(1) \Rightarrow \mathcal{H}_{\widehat{na},\widehat{nc}} \text{ nonsingular} \end{cases} \tag{79}$$

where the threshold $\chi^2_\alpha(\cdot)$ is defined in (49).

2. Generalization to the Multivariable Case

Now consider the case of a multivariable ARMA, a state space realization of which is assumed to be given by (9). Define $\mathcal{H}_{p,p}(S_k)$ as the matrix of dimension $k \times p\,ny$ built from all the p block columns of $\mathcal{H}_{p,p}$ in (15) and the k rows with indices $i_1, i_2, \ldots, i_k$, where

$$S_k \triangleq (i_1, i_2, \ldots, i_k) \tag{80}$$

is a vector containing the above indices. We assume that the first block row of $\mathcal{H}_{p,p}$ has full row rank, which means that the elements of y are linearly independent, and that p is chosen larger than or equal to r, so that Theorem 1 holds. Then the following test strategy is proposed to determine the order of the ARMA.

Algorithm 2. Multivariable Fuchs Test

1. Start with $k = ny$, $S_k = (1, 2, \ldots, ny)$.
2. Add the next row of $\mathcal{H}_{p,p}$, with index denoted by l, to the matrix $\mathcal{H}_{p,p}(S_k)$ and test the resulting matrix for singularity.
   - if singular, drop the row from $\mathcal{H}_{p,p}(S_k)$ as well as all rows in $\mathcal{H}_{p,p}$ with index $l + j\,ny$, $j = 0, 1, 2, \ldots$.
   - if nonsingular, add index l to S_k to obtain S_{k+1}, and set $k = k + 1$.
3. If there are untested rows left in $\mathcal{H}_{p,p}$, return to the beginning of Step 2; else the test is terminated, and the order of the ARMA is $r = k$.

In the multivariable case, the singularity test is somewhat more involved than in the simple scalar case previously treated because of the row manipulations described above. As before, the test operates on the sample covariance matrix $\hat{\mathcal{H}}_{p,p}(S_k)$, obtained from $\mathcal{H}_{p,p}(S_k)$ with the exact covariances replaced by their estimates (41). In general, this matrix is not square, and the EVD cannot be used. Instead we determine the singular value decomposition (SVD) of $\hat{\mathscr{H}}_{p,p}(S_k)$ and partition it as follows:

$$\hat{\mathscr{H}}_{p,p}(S_k) = (\hat{U}\,|\,\hat{u}_k)\left(\begin{array}{c|cc} \hat{\Sigma} & 0 & 0 \\ \hline 0 & \hat{\sigma}_k & 0 \end{array}\right)\left(\begin{array}{c} \hat{V}_1^T \\ \hline \hat{V}_2^T \end{array}\right) \tag{81}$$

In (81), $(\hat{U}\,\hat{u}_k)$ and $(\hat{V}_1\,\hat{V}_2)$, of dimensions $k \times k$ and $pny \times pny$ respectively, are orthogonal matrices, and $\hat{u}_k$ is the normalized k-dimensional eigenvector corresponding to the smallest eigenvalue, denoted by $\hat{\sigma}_k$. The test quantity is then defined as

$$\hat{\eta} = N\hat{z}^T\hat{Q}^{-1}\hat{z} \tag{82}$$

where the vector $\hat{z}$ and the matrix $\hat{Q}$ are given by

$$\hat{z}^T = \hat{u}_k^T\hat{\mathscr{H}}_{p,p}(S_k)\hat{V}_2 = (\hat{\sigma}_k\,0\,0\,\ldots\,0) \qquad (m \times 1) \tag{83}$$

$$\hat{Q} = \hat{V}_2^T\left\{\hat{\gamma}(0)\hat{T}_\nu(0) + \sum_{l=1}^{M-1}\hat{\gamma}(l)(\hat{T}_\nu(l) + \hat{T}_\nu^T(l))\right\}\hat{V}_2 \quad (m \times m) \tag{84}$$

with $m \triangleq pny - k + 1$. Note the similarity between (77) and (84). In (83)–(84) the following definitions are used:

$$\hat{\gamma}(l) = \hat{u}_k^T\hat{T}_u(l)\hat{u}_k \tag{85}$$

$$\hat{T}_u(l) = \frac{1}{N}\sum_{t=1}^{N}\begin{pmatrix} y(t+i_1) \\ y(t+i_2) \\ \vdots \\ y(t+i_k) \end{pmatrix}(y^T(t-l+i_1)\ \cdots\ y^T(t-l+i_k)) \qquad (kny \times kny) \tag{86}$$

$$\hat{T}_\nu(l) = \frac{1}{N}\sum_{t=1}^{N}\begin{pmatrix} y(t-1) \\ y(t-2) \\ \vdots \\ y(t-p) \end{pmatrix}(y^T(t+l-1)\ \cdots\ y^T(t+l-p))$$
$$= \hat{\mathscr{H}}_{p,p-l-1} \quad (pny \times pny) \tag{87}$$

The integer M appearing in the summation of (84) is defined as the number of block rows in $\hat{\mathscr{H}}_{p,p}$ that have contributed with at least one row to the set specified by S_k. When $y(t)$ is a scalar, (86) reduces to $\hat{T}_u(l) = \hat{\mathscr{R}}_{k,l}$ so that $\hat{\gamma}(l) = \hat{u}_k^T\hat{\mathscr{R}}_{k,l}\hat{u}_k$ which is the quantity (77) defined for the scalar case. Since $\hat{\mathscr{H}}_{p,p}(S_k)$ is not square in general, we must require that $k = p = r + 1$ (H_0 of the scalar case), for (83)–(84) to reduce exactly to (76)–(77).

It can be verified, with calculations similar to those that were carried out for the scalar case (equations (67)–(75)), that with the above definitions, asymptotically in N, we have

$$\hat{\eta} \sim \chi^2(m) \tag{88}$$

and the following singularity test can be applied:

$$\begin{cases} \hat{\eta} \le \chi^2_\alpha(m) \Rightarrow \mathcal{H}_{p,p}(S_k) \text{ is singular (has not full row rank)} \\ \hat{\eta} > \chi^2_\alpha(m) \Rightarrow \mathcal{H}_{p,p}(S_k) \text{ is nonsingular (has full row rank)} \end{cases} \tag{89}$$

Clearly, one of the difficulties with the proposed test is to choose the value of p large enough so that one can be sure that a zero singular value obtained in the test is due to the newly introduced row. Often, some *a priori* knowledge of an upper bound of the order of the ARMA under study can be used to determine a suitable value of p.

The results on the multivariable rank test briefly outlined above are described in detail in [16].

B. A Rank Test Based on the LDU Decomposition

In this section, a rank test based on Gaussian lower triangular–diagonal–upper triangular (LDU) decomposition of $\hat{\mathcal{H}}_{p,p}$ (with $p \ge r$) as defined in (15) is presented. The test, which was first presented by Gill and Lewbel [21], directly copes with the case of a general multivariable ARMA process.

As the name suggests, the LDU-decomposition partitions the matrix $\hat{\mathcal{H}}_{p,p}$ into a praoduct of three matrices: L, which is lower triangular, D, which is diagonal, and U, which is upper triangular. The decomposition is usually performed by successive Gaussian elimination. Some kind of pivoting operation is necessary to ensure numerical stability of the decomposition procedure. We shall see that the test to be developed below requires the pivoting to be complete. With complete pivoting we mean that at each step of the Gaussian elimination the current submatrix is searched for its largest element (in absolute magnitude), which is shifted to the stop left corner by column and row interchanges (this is in contrast to partial pivoting which only shifts the largest element in the first column of the submatrix). Pivoting is discussed, e.g., in [18] and the numerical implementation of the LDU-decomposition is treated in detail in [22].

LDU-decomposition with complete pivoting of $\mathcal{H}_{p,p}$ yields

$$P\mathcal{H}_{p,p}Q = LDU \tag{90}$$

where P and Q are permutation matrices corresponding to the pivoting and D is a diagonal matrix which will have a certain structure due to the row and column pivoting as will be explained below. As mentioned before, matrices L and U^T are lower triangular, and they are normalized to have ones along the diagonals. If the $p\,ny \times p\,ny$ matrix $\mathscr{H}_{p,p}$ has rank r, then the LDU-decomposition with complete pivoting can be partitioned as

$$LDU = \begin{pmatrix} L_{11} & 0 \\ L_{21} & I_{p\,ny-r} \end{pmatrix} \left(\begin{array}{c|c} D_1 & 0 \\ \hline 0 & \begin{matrix} 0 & & \\ & \ddots & \\ & & 0 \end{matrix} \end{array} \right) \begin{pmatrix} U_{11} & U_{12} \\ 0 & 0 \end{pmatrix} \tag{91}$$

and the decomposition (91) is unique [24]. The corresponding LDU-decomposition for $\hat{\mathscr{H}}_{p,p}$ is denoted

$$\hat{P}\hat{\mathscr{H}}_{p,p}\hat{Q} = \hat{L}\hat{D}\hat{U}, \tag{92}$$

and is partitioned as

$$\hat{L}\hat{D}\hat{U} = \begin{pmatrix} \hat{L}_{11} & 0 \\ \hat{L}_{21} & \hat{L}_{22} \end{pmatrix} \begin{pmatrix} \hat{D}_1 & 0 \\ 0 & \hat{D}_2 \end{pmatrix} \begin{pmatrix} \hat{U}_{11} & \hat{U}_{12} \\ 0 & \hat{U}_{22} \end{pmatrix} \tag{93}$$

In (91) and (93) the row as well as the column partition is by r and $p\,ny - r$. An important property of the complete pivoting is that it ensures that the $r \times r$ diagonal matrix D_1 in (91) is nonsingular and that the diagonal of D contains exactly $p\,ny - r$ zeros, placed as indicated in (91). Also, it follows from (90) that L_{11}, $\hat{L}_{11}$, U_{11}^T, $\hat{U}_{11}^T$ are unit lower triangular, and since $\hat{\mathscr{H}}_{p,p} \neq \mathscr{H}_{p,p}$, $\hat{L}_{22}$ as well as $\hat{U}_{22}^T$ will be unit lower triangular but different from their exact counterparts. It is shown in [21] that as N grows to infinity, the roofed quantities in the sample LDU-decomposition (93) converge in probability to the true ones, given by (91). In other words,

$$\hat{P}\hat{\mathscr{H}}_{p,p}\hat{Q} = \hat{L}\hat{D}\hat{U} \xrightarrow{P} P\mathscr{H}_{p,p}Q = LDU \tag{94}$$

In particular, $\hat{D}_2 \xrightarrow{P} 0$ and we shall see that under the null hypothesis,

$$H_0\colon \text{rank } \mathscr{H}_{p,p} = r \tag{95}$$

we can derive the statistical properties of the $p\,ny - r$ vector

$$\hat{d}_2 \triangleq \text{diag}(\hat{D}_2) \tag{96}$$

which will enable us to develop the test for determining the rank of $\mathscr{H}_{p,p}$. Let Δ_i be a $(p\,ny - r) \times (p\,ny - r)$ matrix that has 1 as its (i, i)-element and zeros elsewhere. Also define

$$\Delta = (\Delta_1 \Delta_2 \cdots \Delta_{p\,ny-r})^T \tag{97}$$

It is easily verified that the $(p\,ny - r) \times (p\,ny - r)^2$ matrix Δ satisfies the orthogonality property

$$\Delta^T \Delta = I_{p\,ny-r} \tag{98}$$

and also that

$$\operatorname{vec} \hat{D}_2 = \Delta \hat{d}_2 \tag{99}$$

with $\hat{d}_2$ as defiend in (96). Introduce the test quantity

$$\hat{\eta} = N \hat{d}_2^T \hat{W}^{-1} \hat{d}_2 \tag{100}$$

with $\hat{W}$ defined as

$$\hat{W} = \Delta^T (\hat{K}^T \otimes \hat{H})(\hat{Q}^T \otimes \hat{P}) \hat{V} (\hat{Q} \otimes \hat{P}^T)(\hat{K} \otimes \hat{H}^T) \Delta \tag{101}$$

In (101), $\otimes$ denotes the Kronecker product and $\hat{V}$ is a consistent estimate of the covariance matrix of the limiting distribution of $\sqrt{N} \operatorname{vec}(\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p})$. One way of obtaining $\hat{V}$ from the data will be given later on, more specifically in equation (116). The matrices $\hat{H}$ and $\hat{K}$ are defined by

$$\hat{H} = \begin{pmatrix} -\hat{L}_{22}^{-1} \hat{L}_{21} \hat{L}_{11}^{-1} & \hat{L}_{22}^{-1} \end{pmatrix} \tag{102}$$

$$\hat{K} = \begin{pmatrix} -\hat{U}_{11}^{-1} \hat{U}_{12} \hat{U}_{22}^{-1} \\ \hat{U}_{22}^{-1} \end{pmatrix} \tag{103}$$

The following theorem will be used to construct the LDU-decomposition rank test.

Theorem 2. *Under the assumptions made, it holds true that the test quantity (100) satisfies*

$$\hat{\eta} \sim \chi^2(p\,ny - r) \tag{104}$$

i.e., $\hat{\eta}$ *is asymptotically* χ^2*-distributed with* $p\,ny - r$ *degrees of freedom.*

Proof. The proof is patterned from the one in [21]. To verify (104) it is, according to (46), sufficient to show that

$$\sqrt{N} \hat{d}_2 \sim \mathcal{N}(0, W) \tag{105}$$

where W is the asymptotic counterpart of (101). From (94), we infer that

$$\sqrt{N} \hat{P} (\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p}) \hat{Q} \xrightarrow{p} \sqrt{N} (\hat{L} \hat{D} \hat{U} - LDU) \tag{106}$$

For notational convenience, introduce the matrices

$$\hat{L}_1 = \begin{pmatrix} \hat{L}_{11} \\ \hat{L}_{21} \end{pmatrix} \qquad \hat{L}_2 = \begin{pmatrix} 0 \\ \hat{L}_{22} \end{pmatrix}$$

$$\hat{U}_1 = (\hat{U}_{11} \;\; \hat{U}_{12}) \quad \hat{U}_2 = (0 \;\; \hat{U}_{22}) \tag{107}$$

and similarly define L_1, L_2, U_1, and U_2. Then it is easy to see that under H_0

$$\begin{aligned}\sqrt{N}(\hat{L}\hat{D}\hat{U} - LDU) &= \sqrt{N}(\hat{L}_1\hat{D}_1\hat{U}_1 + \hat{L}_2\hat{D}_2\hat{U}_2 - L_1D_1U_1) \\ &= \sqrt{N}(\hat{L}_1 - L_1)\hat{D}_1\hat{U}_1 + L_1\sqrt{N}(\hat{D}_1 - D_1)\hat{U}_1 \\ &\quad + L_1D_1\sqrt{N}(\hat{U}_1 - U_1) + \hat{L}_2\sqrt{N}\hat{D}_2\hat{U}_2 \end{aligned} \tag{108}$$

and using (94) and (106) we obtain, asymptotically in N, that

$$\begin{aligned}\sqrt{N}\hat{P}(\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p})\hat{Q} &= \sqrt{N}(\hat{L}_1 - L_1)\hat{D}_1\hat{U}_1 + \hat{L}_1\sqrt{N}(\hat{D}_1 - D_1)\hat{U}_1 \\ &\quad + \hat{L}_1\hat{D}_1\sqrt{N}(\hat{U}_1 - U_1) + \hat{L}_2\sqrt{N}\hat{D}_2\hat{U}_2 \end{aligned} \tag{109}$$

Now, the matrices $\hat{H}$ and $\hat{K}$ introduced in (102)–(103) above satisfy the readily verified properties

$$\begin{aligned}&\hat{H}\hat{L}_1 = 0, \quad \hat{H}\hat{L}_2 = I_{p\,ny-r} \\ &\hat{U}_1\hat{K} = 0, \quad \hat{U}_2\hat{K} = I_{p\,ny-r}\end{aligned} \tag{110}$$

so that pre- and postmultiplying equation (109) with $\hat{H}$ and $\hat{K}$, respectively, yields

$$\sqrt{N}\hat{D}_2 = \hat{H}\hat{P}\sqrt{N}(\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p})\hat{Q}\hat{K} \tag{111}$$

and by (96) and (98)–(99) we have that

$$\begin{aligned}\sqrt{N}\hat{d}_2 &= \sqrt{N}\Delta^T\operatorname{vec}(\hat{D}_2) \\ &= \sqrt{N}\Delta^T\operatorname{vec}(\hat{H}\hat{P}(\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p})\hat{Q}\hat{K}) \\ &= \sqrt{N}\Delta^T(\hat{K}^T\hat{Q}^T \otimes \hat{H}\hat{P})\operatorname{vec}(\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p}) \\ &= \Delta^T(\hat{K}^T \otimes \hat{H})(\hat{Q}^T \otimes \hat{P})\sqrt{N}\operatorname{vec}(\hat{\mathcal{H}}_{p,p} - \mathcal{H}_{p,p})\end{aligned} \tag{112}$$

where use was made of the formulas (see, e.g., [8]) $\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec} B$ and $(AC \otimes BD) = (A \otimes B)(C \otimes D)$. Now, from (44) we finally obtain

$$\sqrt{N}\hat{d}_2 \sim \mathcal{N}(0, W) \tag{113}$$

with W being the asymptotic counterpart of (101). ■

The estimate $\hat{V}$ needed to calculate $\hat{W}$ in (101) above can be obtained as follows. Define the vector h through

$$\begin{aligned}h &\triangleq \operatorname{vec}(\mathcal{H}_{p,p}) \\ &= E\operatorname{vec}(Y_p(t)Y_p^T(t-p)J) \\ &= E\{JY_p(t-p) \otimes Y_p(t)\} \quad (p\,ny)^2 \times 1\end{aligned} \tag{114}$$

and similarly define $\hat{h}$ by replacing the expectation operator in (114) by $1/N\Sigma_{t=1}^{N}$. Then

$$\begin{aligned} V &= NE(\hat{h}-h)(\hat{h}-h)^T \\ &= NE\hat{h}\hat{h}^T - Nhh^T \\ &= E\left\{\frac{1}{N}\sum_{t=1}^{N}\sum_{s=1}^{N} JY_p(t-p)Y_p^T(s-p)J \otimes Y_p(t)Y_p^T(s)\right\} - Nhh^T \end{aligned} \tag{115}$$

Defining $\mathcal{H}_{p,p}(\tau) \triangleq EY_p(t)Y_p^T(t-p-\tau)J$ and using the formula for the product of four Gaussian random variables [20] one can show that

$$\begin{aligned} V = \frac{1}{N}\sum_{\tau=-N}^{N} (N-|\tau|)\{(J\mathcal{R}_{p,\tau}J) \otimes \mathcal{R}_{p,\tau} + [\operatorname{vec}\mathcal{H}_{p,p}(\tau)] \\ \times [\operatorname{vec}\mathcal{H}_{p,p}(-\tau)]^T\} \end{aligned} \tag{116}$$

The calculations needed to arrive at (116) are straightforward but notationally complicated, and we choose to omit the details (to derive (116) from (115), consider each block of V separately). A consistent estimate of V is finally obtained by replacing all quantities in (116) by their sample counterparts. However, for a limited number of data points, this will turn out to be quite a bad estimate of V. The reason is that the covariances appearing in (116) will be poorly estimated at large lags τ. The classical remedy is to truncate the sum at some $|\tau| = M \ll N$. This can be done as we know that the true covariance function $R(\tau)$ should be "small" for large $|\tau|$. In practice, a good truncation point may be difficult to determine, as it depends on many factors, such as the order of the process, the number of available data points, and the value of p chosen when constructing $\mathcal{H}_{p,p}$. A rule of thumb can, however, be obtained as follows. The largest lag of the covariance function appearing in $\mathcal{H}_{p,p}(\tau)$ in (116) is $\tau + 2p - 1$. Inspired by results from spectral analysis we assume that the maximum lag we can estimate is $\tau = N/8$ so we choose

$$M = \max\left(0, \left\lceil \frac{N}{8} \right\rceil - 2p + 1\right) \tag{117}$$

where $\lceil x \rceil$ denotes the first integer which is larger than or equal to x.

The result of Theorem 2 suggests the following rank test for ARMA order estimation:

Algorithm 3. LDU Decomposition Rank Test

1. Perform the LDU decomposition (92) of the sample covariance matrix $\hat{\mathcal{H}}_{p,p}$. Set $j = 0$.

2. Partition the decomposition according to (93) to obtain the $(p\,ny-j)$-vector $\hat{d}_2$, and construct the matrix $\hat{W}$ (101).
   - If $\hat{\eta} \leq \chi^2_\alpha(p\,ny-j)$, then rank $\mathcal{H}_{p,p}=j$ and the order of the ARMA is $r=j$.
   - Else, set $j=j+1$. If $j<p\,ny$, go to the beginning of Step 2, otherwise the test is terminated.
3. If the test terminates with $j=p\,ny$, then the rank of $\mathcal{H}_{p,p}$ could not be determined and p must be increased to find the ARMA order. ■

C. A Gramian-Based Rank Test

In this section we will present a Gramian-based (GRAB) order estimation scheme. The test relies on the eigenvalue decomposition of the "Gramian" of the covariance matrix. The discussion here focuses on the case of a scalar ARMA process, but an extension to the general multivariable case is straightforward, as will be briefly indicated below. The test is due to Stoica and Cedervall and was first developed for the detection of the number of signals impinging on an array of sensors [23].

1. Scalar Case

Let the covariance matrix of a scalar ARMA be defined as

$$\mathcal{R} \triangleq E\begin{pmatrix} y(t) \\ y(t+1) \\ \vdots \\ y(t+p-1) \end{pmatrix}(y(t-k)\ \ y(t-k+1)\ \cdots\ y(t-1)) \tag{118}$$

$$= EY_p(t)Y_k^T(t-k) \quad (p \times k) \tag{119}$$

Then it follows from Theorem 1 that

$$\text{rank}\,\mathcal{R} = r \qquad \text{for } p \geq r, \quad k \geq r \tag{120}$$

To see this, note that

$$\mathcal{R} = \left(\begin{array}{cccc} R(k) & R(k-1) & \ldots & R(1) \\ R(k+1) & R(k) & & R(2) \\ \vdots & & \ddots & \\ R(2k-1) & & & R(k) \\ \hline R(2k) & & & R(k+1) \\ \vdots & & & \vdots \\ R(k+p-1) & & & R(p) \end{array}\right)$$

$$= \left(\frac{\mathcal{R}_{k,k}}{EY_{p-k}(t)Y_k^T(t-2k)} \right) \tag{121}$$

Clearly, $\mathcal{R}$ can be factorized in the same manner as shown in (23) of the proof of Theorem 1. Applying the Cayley–Hamilton theorem enables us to claim that

$$\operatorname{rank} \mathcal{R} = \operatorname{rank} \mathcal{R}_{k,k} = r \qquad \text{for } k \geq r, \quad p \geq r \tag{122}$$

which justifies (120). Let

$$\hat{\mathcal{R}} = \frac{1}{N} \sum_{t=1}^{N} Y_p(t) Y_k^T(t-k) \quad (p \times k) \tag{123}$$

denote the sample estimate of the covariance matrix $\mathcal{R}$ in (119). Introduce the null hypothesis as

$$H_0 : k = r - 1 \tag{124}$$

In what follows we will assume that p has been chosen sufficiently large so that $p \geq r$. The main theme in this section is to develop a statistical test based on the smallest eigenvalue of the Gramian matrix

$$\hat{\Psi} \triangleq \hat{\mathcal{R}}^T \hat{\mathcal{R}} \quad (k \times k) \tag{125}$$

Denote this eigenvalue by $\hat{\lambda}$. Under the null hypothesis,

$$\Psi \triangleq \mathcal{R}^T \mathcal{R} \tag{126}$$

the asymptotic counterpart of $\hat{\Psi}$, is singular and admits one and only one zero eigenvalue. As a consequence, we would expect $\hat{\lambda}$ to be small. In what follows we will derive the perturbation properties for the smallest eigenvalue of the sample Gramian matrix $\hat{\Psi}$ under the assumption that H_0 and (120) hold true. To obtain a suitable asymptotic approximation for the smallest eigenvalue of $\hat{\Psi}$, first note that as

$$\hat{\lambda} > 0 \quad \text{for any } N < \infty, \tag{127}$$

a first-order approximation of $\hat{\lambda}$ (which would allow negative values) cannot be appropriate. In fact, owing to (127), the first-order term in an asymptotic expansion of $\hat{\lambda}$ must vanish, and hence the second-order term should be the one of interest. In order to derive the latter term, the following additional notation is introduced. Let

$$\begin{aligned} \Psi &= (S \;\; g) \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} S^T \\ g^T \end{pmatrix} \\ &= S \Lambda S^T + (0) g g^T \end{aligned} \tag{128}$$

denote the EVD of Ψ. Similarly,

$$\hat{\Psi} = \hat{S}\hat{\Lambda}\hat{S}^T + \hat{\lambda}\hat{g}\hat{g}^T \tag{129}$$

is the EVD of $\hat{\Psi}$. The eigenvectors in both EVDs above are orthonormal, and the distinct (by assumption) eigenvalues are arranged in decreasing order. Also, let

$$\mathcal{R} = U\Lambda^{1/2}S^T \tag{130}$$

denote the SVD of $\mathcal{R}$. The matrices Λ and S are as defined previously and U is the $[p \times (k-1)]$-dimensional matrix of the principal left singular vectors of $\mathcal{R}$ (so that $U^TU = I$). It follows that

$$Q \triangleq I - UU^T \tag{131}$$

is an idempotent $(p \times p)$-matrix with the following rank:

$$\operatorname{rank} Q = (p - k + 1) \tag{132}$$

The following theorem gives an asymptotically valid expression of $\hat{\lambda}$.

Theorem 3. *Under H_0,*

$$\hat{\lambda} = g^T\hat{\mathcal{R}}^T Q\hat{\mathcal{R}}g + O(N^{-3/2}) \tag{133}$$

and the vector

$$d \triangleq \hat{\mathcal{R}}g \tag{134}$$

is asymptotically normally distributed with zero mean and covariance matrix given by

$$D = \frac{1}{N^2}\sum_{\tau=-k+1}^{k-1}(N - |\tau|)(\mathcal{R}_{p,\tau}g^T\mathcal{R}_{k,\tau}g) \tag{135}$$

where $\mathcal{R}_{p,\tau}$ is defined in (12).

Proof. We first derive (133). Since the estimate (123) is consistent and asymptotically normal, it can be assumed that the roofed quantities in (129) tend to the corresponding ones in (128) with an error of the order of $N^{-1/2}$ as $N \to \infty$ (see, e.g., [24]). In other words

$$\begin{aligned}
\tilde{\mathcal{R}} &\triangleq \hat{\mathcal{R}} - \mathcal{R} = O(N^{-1/2})\\
\tilde{S} &\triangleq \hat{S} - S \;= O(N^{-1/2})\\
\tilde{g} &\triangleq \hat{g} - g \;= O(N^{-1/2})\\
\tilde{\Lambda} &\triangleq \hat{\Lambda} - \Lambda = O(N^{-1/2})\\
\tilde{\lambda} &\triangleq \hat{\lambda} - 0 \;= O(N^{-1/2})
\end{aligned} \tag{136}$$

Now, note that

$$\hat{\lambda} = g^T \hat{\Psi} g - g^T (\hat{\Psi} - \hat{\lambda} I) g \tag{137}$$

Since $\mathcal{R} g = 0$, the first term in (137) can also be written as

$$g^T \hat{\Psi} g = g^T (\hat{\mathcal{R}} - \mathcal{R})^T (\hat{\mathcal{R}} - \mathcal{R}) g \tag{138}$$

which shows that it is $O(N^{-1})$. Next we deal with the second term in (137). As

$$\begin{aligned}
\hat{\Psi} - \hat{\lambda} I &= \hat{S} \hat{\Lambda} \hat{S}^T + \hat{\lambda} \hat{g} \hat{g}^T - \hat{\lambda} I \\
&= \hat{S} \hat{\Lambda} \hat{S}^T + \hat{\lambda} (I - \hat{S} \hat{S}^T) - \hat{\lambda} I \\
&= \hat{S} (\hat{\Lambda} - \hat{\lambda} I) \hat{S}^T
\end{aligned} \tag{139}$$

and $\hat{g}^T \hat{S} = 0$ because of the orthogonality of $\hat{g}$ and $\hat{S}$, one can write

$$\begin{aligned}
g^T (\hat{\Psi} - \hat{\lambda} I) g &= g^T \hat{S} (\hat{\Lambda} - \hat{\lambda} I) \hat{S}^T g \\
&= (g - \hat{g})^T \hat{S} (\hat{\Lambda} - \hat{\lambda} I) \hat{S}^T (g - \hat{g}) \\
&= (g - \hat{g})^T \hat{S} \hat{\Lambda} \hat{S}^T (g - \hat{g}) - (g - \hat{g})^T \hat{S} \hat{\lambda} \hat{S}^T (g - \hat{g}) \\
&= g^T (S + \tilde{S})(\Lambda + \bar{\Lambda})(S + \tilde{S})^T g - \bar{g}^T (S + \tilde{S}) \tilde{\lambda} (S + \tilde{S})^T \bar{g} \\
&= g^T \hat{S} \Lambda \hat{S}^T g + O(N^{-3/2})
\end{aligned} \tag{140}$$

In order to proceed, an approximation of $g^T \hat{S}$ in terms of the data matrices Ψ or $\mathcal{R}$ is required. It is readily verified that

$$\begin{aligned}
g^T \hat{\Psi} S &= g^T \hat{S} \hat{\Lambda} \hat{S}^T S + g^T \hat{\lambda} \hat{g} \hat{g}^T S \\
&= g^T (S + \tilde{S})(\Lambda + \bar{\Lambda})(S + \tilde{S})^T S + g^T \tilde{\lambda} (g + \bar{g})(g + \bar{g})^T S \\
&= g^T (S + \tilde{S}) \Lambda (S + \tilde{S})^T S + g^T (S + \tilde{S}) \bar{\Lambda} (S + \tilde{S})^T S + \tilde{\lambda} g^T (g + \bar{g}) \bar{g}^T S \\
&= g^T \hat{S} \Lambda + g^T \tilde{S} \Lambda \tilde{S}^T S + g^T \tilde{S} \bar{\Lambda} (S + \tilde{S})^T S + \tilde{\lambda} (1 + g^T \bar{g}) \bar{g}^T S \\
&= g^T \hat{S} \Lambda + O(N^{-1})
\end{aligned} \tag{141}$$

which gives

$$g^T \hat{S} = g^T \hat{\Psi} S \Lambda^{-1} + O(N^{-1}) \tag{142}$$

Insertion of (142) into (140) yields

$$\begin{aligned}
g^T (\hat{\Psi} - \hat{\lambda} I) g &= g^T \hat{\Psi} S \Lambda^{-1} S^T \hat{\Psi} g + O(N^{-3/2}) \\
&= g^T (\mathcal{R} + \bar{\mathcal{R}})^T (\mathcal{R} + \bar{\mathcal{R}}) S \Lambda^{-1} S^T (\mathcal{R} + \bar{\mathcal{R}})^T (\mathcal{R} + \bar{\mathcal{R}}) g \\
&\quad + O(N^{-3/2}) \\
&= g^T \hat{\mathcal{R}}^T \mathcal{R} S \Lambda^{-1} S^T \mathcal{R}^T \hat{\mathcal{R}} g + g^T \bar{\mathcal{R}}^T \mathcal{R} S \Lambda^{-1} S^T \bar{\mathcal{R}}^T \bar{\mathcal{R}} g \\
&\quad + g^T \bar{\mathcal{R}}^T \bar{\mathcal{R}} S \Lambda^{-1} S^T \mathcal{R}^T \bar{\mathcal{R}} g + g^T \bar{\mathcal{R}}^T \bar{\mathcal{R}} S \Lambda^{-1} S^T \bar{\mathcal{R}}^T \bar{\mathcal{R}} g \\
&\quad + O(N^{-3/2}) \\
&= g^T \hat{\mathcal{R}}^T \mathcal{R} S \Lambda^{-1} S^T \mathcal{R}^T \hat{\mathcal{R}} g + O(N^{-3/2})
\end{aligned} \tag{143}$$

Finally, combining (130), (131), (137), and (143) yields

$$\begin{aligned}\hat{\lambda} &= g^T\hat{\Psi}g - g^T\hat{\mathcal{R}}^T\mathcal{R}S\Lambda^{-1}S^T\mathcal{R}^T\hat{\mathcal{R}}g + O(N^{-3/2})\\ &= g^T\hat{\mathcal{R}}^T\hat{\mathcal{R}}g - g^T\hat{\mathcal{R}}^T UU^T\hat{\mathcal{R}}g + O(N^{-3/2})\\ &= g^T\hat{\mathcal{R}}^T Q\hat{\mathcal{R}}g + O(N^{-3/2}) \qquad (144)\end{aligned}$$

which is the same as (133).

We next turn to the statistical properties of the vector d of (134). The asymptotic normality of d follows from the central limit theorem as given, e.g., in [13]. The result

$$Ed = E\hat{\mathcal{R}}g = \mathcal{R}g = 0 \qquad (145)$$

is immediate from (128). Concerning the covariance matrix D we have that

$$\begin{aligned}D &\triangleq E(\hat{\mathcal{R}}g)(\hat{\mathcal{R}}g)^T\\ &= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N} EY_p(t)Y_k^T(t-k)gg^TY_k(s-k)Y_p^T(s)\\ &= \frac{1}{N^2}\sum_{t=1}^{N}\sum_{s=1}^{N}\{\mathcal{R}gg^T\mathcal{R}^T + (EY_p(t)Y_k^T(s-k))gg^T(EY_k(t-k)Y_p^T(s))\\ &\quad + (EY_p(t)Y_p^T(s))g^T(EY_k(t-k)Y_k^T(s-k))g\}\\ &= 0 + \frac{1}{N^2}\sum_{\tau=-N}^{N}(N-|\tau|)\{\mathcal{R}(\tau)gg^T\mathcal{R}^T(-\tau) + \mathcal{R}_{p,\tau}g^T\mathcal{R}_{k,\tau}g\} \qquad (146)\end{aligned}$$

where

$$\mathcal{R}(\tau) = EY_p(t)Y_k^T(t-k-\tau) \qquad (147)$$

($\mathcal{R} = \mathcal{R}(0)$, by notational convention). In the above calculation, use was made of the standard formula for the fourth-order moments of normally distributed random variables [20]. In order to obtain (135) we need to show that the first term in the sum of (146) is zero. To see this, note that by Lemma 1, the normalized eigenvector corresponding to the (distinct) zero eigenvalue of Ψ is given by

$$g^T = (a_{na}a_{na-1} \cdots a_1 \;\; 1)/a \qquad (148)$$

where $a = (1 + a_1^2 + \cdots + a_{na}^2)^{1/2}$. We have that

$$Y_k^T(t-k-\tau)g = \frac{1}{a}A(q)y(t-k-\tau) \qquad (149)$$

where q denotes the forward shift operator, and hence

$$\begin{aligned}\mathcal{R}(\tau)g &= E[Y_p(t)A(q)y(t-k-\tau)]/a \\ &= E[A(q^{-1})Y_p(t)y(t-k-\tau)]/a \\ &= 0 \quad \text{for } \tau \geq 0\end{aligned} \tag{150}$$

The last equality follows from the Yule–Walker equations (59), and we also made use of the easily verified fact that for any scalar filter $A(z)$ it holds that $E[x(t)\{A(q)z(t)\}] = E[\{A(q^{-1})x(t)\}z(t)]$. Similarly

$$\mathcal{R}(-\tau)g = 0 \quad \text{for } \tau \geq 0 \tag{151}$$

so that the first term in the sum of (146) is indeed zero. Proceeding in the same manner with the second term we find

$$\begin{aligned}g^T\mathcal{R}_{k,\tau}g &= E[g^TY_k(t)Y_k^T(t-\tau)g] \\ &= E[A(q)y(t)A(q)y(t-\tau)]/a^2 \\ &= E[A(q^{-1})y(t)A(q^{-1})y(t-\tau)]/a^2 \\ &= E[C(q^{-1})e(t)C(q^{-1})e(t-\tau)]/a^2 \\ &= 0 \quad \text{for } |\tau| > r = k-1\end{aligned} \tag{152}$$

Inserting (150)–(152) into (146) yields the desired result (135). ■

Now, we cannot make any statement on the distribution of

$$\hat{\lambda} \approx d^TQd \tag{153}$$

directly, but we can transform $\hat{\lambda}$ into another random variable which is approximately normally distributed. To see how this can be done, recall that the random vector d in (153) is indeed asymptotically normally distributed with zero mean and covariance matrix D. Define

$$Z = D^{1/2}QD^{1/2} \quad (p \times p) \tag{154}$$

where $D^{1/2}$ denotes a (symmetric for notational convenience) square root of the positive definite matrix D. As mentioned before, the matrix Q is idempotent with rank $(p-k+1)$. Since D can be assumed to have full rank, this implies that

$$\operatorname{rank} Z = p - k + 1 \tag{155}$$

Let

$$Z = KMK^T; \qquad M = \begin{pmatrix} \mu_1 & & 0 \\ & \ddots & \\ 0 & & \mu_{p-k+1} \end{pmatrix}; \qquad K^TK = I \tag{156}$$

denote the EVD of the matrix Z, and let

$$x \triangleq K^T D^{-1/2} d \tag{157}$$

Making use of the notation introduced above we can write (153) as

$$\hat{\lambda} \approx x^T M x \tag{158}$$

The random vector x is asymptotically $\mathcal{N}(0, I)$ distributed. Let

$$\hat{\gamma}_j = \sum_{k=1}^{p-k+1} \hat{\mu}_k^j \quad \text{for } j = 1, 2, 3 \tag{159}$$

where $\{\hat{\mu}_k)$ are the eigenvalues obtained from the EVD of $\hat{Z}$, a consistent estimate of Z. Also set

$$\hat{\nu} = 1 - 2\hat{\gamma}_1 \hat{\gamma}_3/(3\hat{\gamma}_2^2) \tag{160}$$

Then, when N grows to infinity, the following transformed and normalized random variable derived from $\hat{\lambda}$,

$$\hat{T}_k = \hat{\gamma}_1[(\hat{\lambda}/\hat{\gamma}_1)^{\hat{\nu}} - 1 - \hat{\nu}(\hat{\nu} - 1)\hat{\gamma}_2/\hat{\gamma}_1^2]/(2\hat{\nu}^2 \hat{\gamma}_2)^{1/2} \tag{161}$$

is asymptotically normally distributed with zero mean and unit variance [25], so that the test quantity

$$\hat{\eta} \triangleq \hat{T}_k^2 \sim \chi^2(1) \tag{162}$$

i.e., $\hat{\eta}$ is asymptotically $\chi^2(1)$ distributed.

Summarizing the above results leads to the following ARMA order estimation scheme.

Algorithm 4. Scalar GRAB Rank Test

1. Compute the estimates $\{\hat{\mathcal{R}}_{p,\tau}\}$ of $\{\mathcal{R}_{p,\tau}\}$.
   For $k = 1, 2, \ldots$ do:
2. Compute the sample estimates $\{\hat{\mathcal{R}}(\tau)\}$ and $\{\hat{\mathcal{R}}_{k,\tau}\}$ of the corresponding theoretical covariances.
3. Determine $\hat{\lambda}$, $\hat{g}$, and $\hat{U}$ from the SVD of $\hat{\mathcal{R}}$. Calculate an estimate $\hat{D}$ by replacing all quantities in (135) with their estimates. Next determine $\hat{Z}$ and compute its eigenvalues. Finally obtain $\hat{\eta}$ from (161)–(162).
4. If $\hat{\eta} \leq \chi_\alpha^2(1)$ then accept H_0, i.e., the order of the ARMA is found to be $r = k - 1$. Otherwise reject H_0, set $k = k + 1$, and go back to step 2. The threshold $\chi_\alpha^2(1)$ is defined in (49). ■

2. Generalization to the Multivariable Case

To modify the test for the multivariable case, first note that the calculations carried out above are general in the sense that they apply to any covariance matrix $\mathcal{R}$ of rank r, provided that H_0 holds. The only exception is the result (135) because in the calculations (148)–(152) we assumed that the polynomials $A(z)$ and $C(z)$ were scalars. If we replace (135) by (146), the above calculations are perfectly valid for a general multivariable ARMA process. We must, however, change the test strategy to ensure that Ψ has at most one zero eigenvalue. This is done in the same manner as outlined in Algorithm 2 of Section A. Let $\mathcal{R}(S_k)$ be the matrix of dimension $p\,ny \times k$ obtained from all the p block rows of $\mathcal{R}$ and the k columns with indices specified by the vector S_k (80). We obtain the following test:

Algorithm 5. Multivariable GRAB Rank Test

1. Start with $k = ny$, $S_k = (1, 2, \ldots, ny)$.
2. Add the next column of $\mathcal{R}$ with index denoted by l to the matrix $\mathcal{R}(S_k)$.
   - if singular (as determined by Algorithm 4 with D from (146)) drop the column from $\mathcal{R}(S_k)$, as well as all columns in $\mathcal{R}$ with index $l + j\,ny$, $j = 0, 1, 2, \ldots$.
   - if nonsingular, add index l to S_k to obtain S_{k+1}, and set $k = k + 1$.
3. If there are untested rows left in $\mathcal{R}$, return to the beginning of Step 2; else the test is terminated, and the order of the ARMA is $r = k$.

Note that in order to obtain a good estimate of D, needed in the above algorithm, it may be necessary to truncate the sum (146) at some $|\tau| = M \ll N$ (see the end of the previous section for a discussion on this point).

IV. COMPARISON AND EVALUATION OF THE RANK TESTS

In this section we will illustrate the performance of the different rank tests for ARMA order estimation that were presented above. This will be done by means of some simple scalar numerical examples.

A first impression may be that the three rank tests previously treated are very different from each other. However, they all rely on the same rank result, namely Theorem 1, even if the actual covariance matrix used may be slightly different from method to method. The three methods are summarized in Table 1. We see that the methods are in fact very similar. One important difference between the methods is the user parameters. All methods, except for the scalar Fuchs test, require the parameter p to

Table 1 Summary of the Different Rank Test Methods: Covariance Matrix, Test Quantity, and the Limiting Distribution of the Test Quantity ($m = \dim \hat{z}$, $n = \dim \hat{d}_2$)

Method	Covariance matrix	Test quantity	Limiting distribution	User parameters
Fuchs				
scalar	$\mathscr{H}_{\hat{n}a,\hat{n}c}$	$\hat{\eta} = \hat{\lambda}^2_{\hat{n}a,\hat{n}c}/\hat{\sigma}^2_{\hat{n}a,\hat{n}c}$ see (76)	$\chi^2(1)$	(none)
multivariable	$\mathscr{H}_{p,p}(S_k)$	$\hat{\eta} = N\hat{z}^T\hat{Q}^{-1}\hat{z}$ see (82)	$\chi^2(m)$	p
LDU	$\mathscr{H}_{p,p}$	$\hat{\eta} = N\hat{d}_2^T\hat{W}^{-1}\hat{d}_2$ see (100)	$\chi^2(n)$	p, M
GRAB				
scalar	$\Psi = \mathscr{R}^T\mathscr{R}$	$\hat{\eta} = \hat{T}_k^2$ see (162)	$\chi^2(1)$	p
multivariable	$\Psi = \mathscr{R}^T(S_k)\mathscr{R}(S_k)$	$\hat{\eta} = \hat{T}_k^2$ see (162)	$\chi^2(1)$	p, M

be chosen by the user (it is necessary to have $p \geq r$ for Theorem 1 to hold, and the GRAB test even requires $p > r$). We often have some *a priori* information on the order of the ARMA process under study, which can be used to choose p. We might expect that the methods will determine the true order more frequently when p increases, as then more covariances are taken into account in the tests, but we also know that covariances at large lags may be inaccurately estimated because of the limited amount of data points available. The conclusion is that a too large value of p may be computationally cumbersome, but also lead to degraded performance. In addition to the parameter p, the LDU and the multivariable GRAB tests require the truncating point M to be determined. Again, we will have to make a trade-off between how many covariance elements we include in the calculations and how well these can be estimated.

We now turn to the numerical examples. We will use the two different ARMA models listed in Table 2, fed by Gaussian white noise with zero mean and unit variance, as a starting-point for the investigation of the performance of the different methods. The pole–zero distribution and the power spectral densities of the two processes are shown in Figs 1 and 2. We see that for A1, both the poles and the only zero are located well off the unit circle. The covariance function of A1 is thus decaying rapidly, and since there is no risk for pole–zero cancelation, we expect that the

Table 2 The Two Different ARMA Processes Used to Evaluate the Rank Test Methods

Name	True order	AR parameters	MA parameters	Figure
A1	2	1 0.64 0.7	1 0.8	1
A2	4	1 −2.760 3.809 −2.654 0.924	1 −0.2 0.04	2

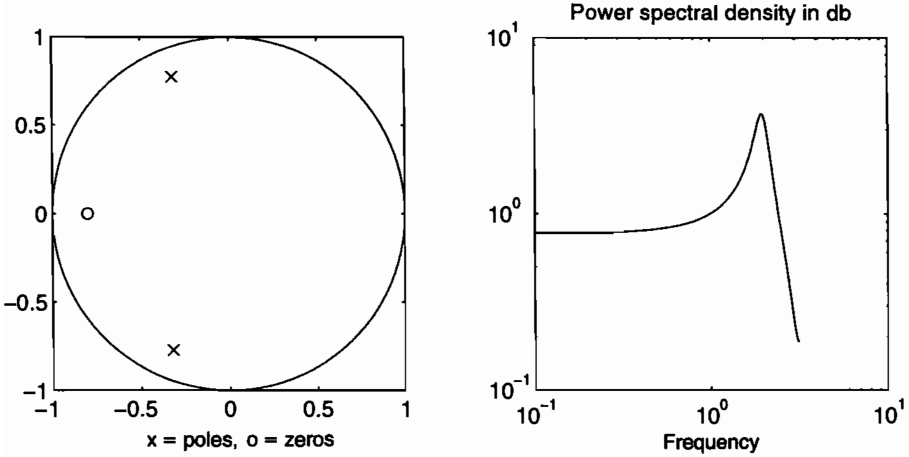


Figure 1 Pole–zero location and power spectral density of the process A1 ($A = [1\,0.64\,0.7]$, $C = [1\,0.8]$), true order $r = 2$.

rank test methods will perform well for this case. Concerning A2, the situation is more delicate. The poles are situated much closer to the unit circle, and we will need a large number of data points to be able to get a good estimate of the covariance function. The difference between the models A1 and A2 can also be stated in terms of the condition number (defined as the ratio of the largest and smallest eigenvalue) of $\hat{\mathcal{R}}_{p,p}$. Indeed, for A2 this condition number is three orders of magnitude larger than the one for A1, and all methods can thus be expected to have more difficulty in correctly estimating the order of the more complex model.

We will assess the performance of the different methods in terms of the probabilities of fitting, overfitting and underfitting. The probability of fitting refers to the percentage of times that the correct order is estimated for a certain number of runs, each on independent realizations of the data. The over- and underfitting probabilities are defined in a similar manner.

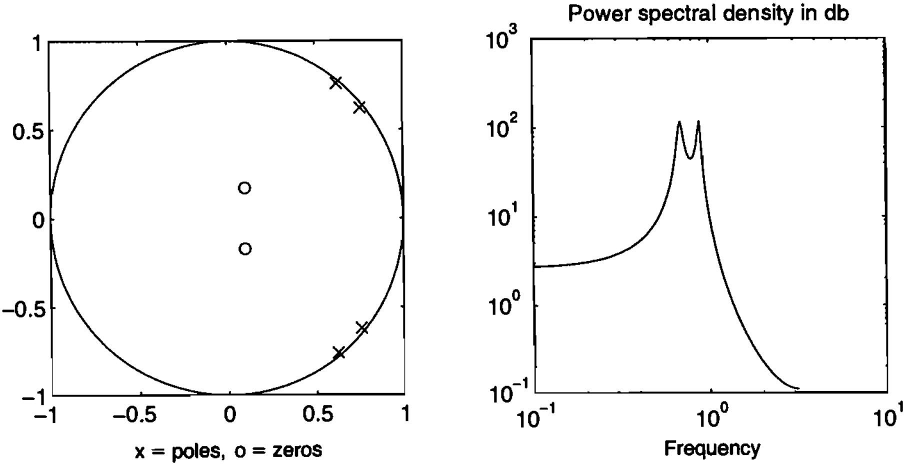


Figure 2 Pole-zero location and power spectral density of the process A2 ($A = [1\ -2.76\ 3.809\ -2.654\ 0.924]$, $C = [1\ -0.2\ 0.04]$), true order $r = 4$.

We will say that a rank test method is "good" when it has a high probability of fitting. In all examples we will use 100 independent trials and choose the false alarm probability to be $\alpha = 0.05$.

Figure 3 shows the performance of the methods for estimating the order of the process A1 as a function of N, the number of data points. For the LDU and GRAB methods, we have chosen $p = 4$, and the truncation point M is determined from (117). Also, the performance of the LDU method without truncation is displayed for comparison. We can see that truncation is important: a much larger N is required to reach the same performance as the truncated LDU. Except for the LDU method without truncation, the methods seem to behave very similarly. For very short data samples, a small difference in performance is visible.

Figure 4 shows the performance for the model A2, with $p = 5$. As expected, a larger amount of data is necessary to estimate the order correctly. Again, the LDU method without truncation has a much lower performance than the other three methods. The LDU and GRAB methods show slightly better small sample properties than the Fuchs method.

As mentioned above, one advantage of the scalar Fuchs method is that it does not require any parameters to be chosen by the user. The LDU and GRAB methods both require the choice of p. Figure 5 shows their performance with respect to p, for different values of N. It is seen that for a short data sample ($N = 300$), the performance of the methods decreases with increasing p. This is expected, because even if more

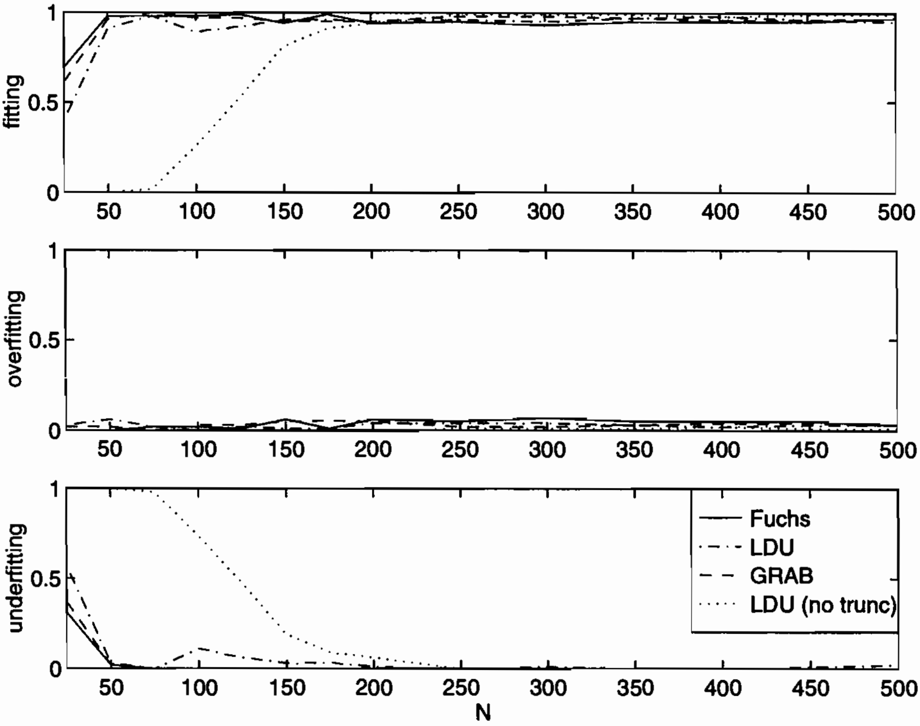


Figure 3 The performance for A1 ($p = 4$, $M = N/8 - 7$).

covariances are taken into account, they are not accurately estimated. We note that the LDU method is more sensitive with respect to p than the GRAB method. This is so because the choice of the truncation point as $M = \lceil N/8 \rceil - 2p + 1$ is not good for large values of p and small values of N, which again illustrates the difficulty in choosing the truncation point for the LDU method. For a large sample ($N = 1000$) both methods are insensitive to the choice of p. We can even see a slightly increased performance for both methods for large values of p.

In Fig. 6 we study the sensitivity of the methods to pole–zero cancelation. We let the poles of the system A1 move on a circle (starting at the original position as shown in Fig. 1) and plot the performance as a function of the angle between the poles and the zero. In this example, we have chosen $N = 300$, and in view of Fig. 5 we have selected $p = 7$, a value that gives good performance both for the LDU and the GRAB methods. Again, M is chosen according to (117). We expect the methods

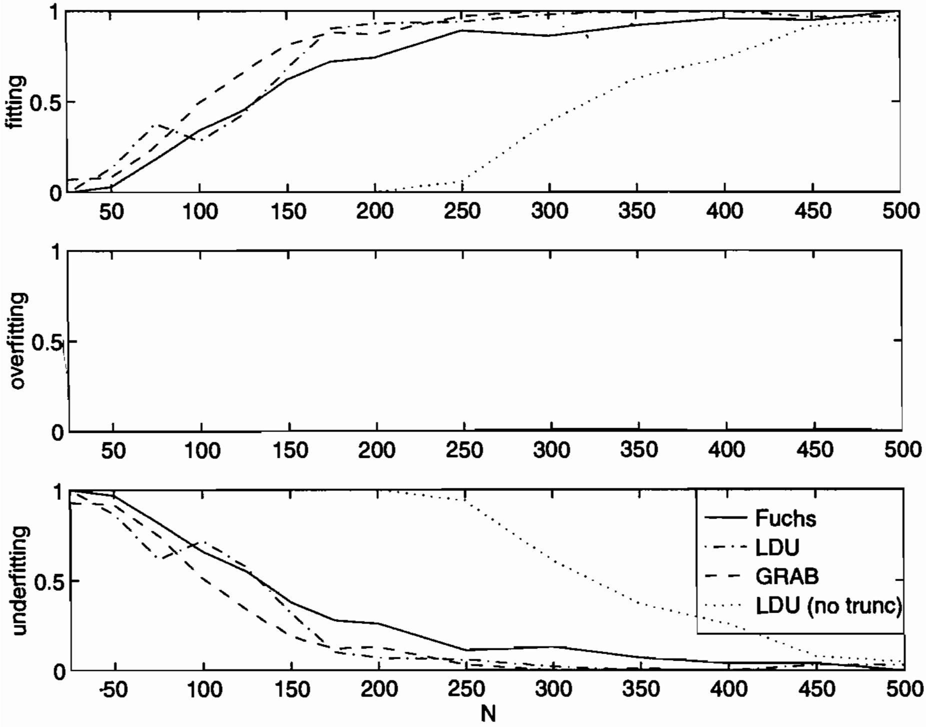


Figure 4 The performance for A2 ($p = 5$, $M = N/8 - 9$).

to underfit when the poles and the zero are closely spaced, since we will then almost have a pole–zero cancelation, and the system will then appear to be of order $r = 1$. We see that this is indeed the case; all methods are sensitive to closely spaced poles and zeros. To realize if this is a property only of the rank test methods, we have included the performance of the so-called F-test in Fig. 6. The F-test, which is an optimal order estimation procedure, uses a function of the estimated prediction errors for increasing model orders as a test quantity. It thus requires the estimation of the model parameters, which necessitates a parametrization of the model and the use of a multidimensional search in the parameter space, see [8]. The F-test has a slightly better performance than the rank test methods when the poles and the zero are closely spaced, but this better performance is obtained at the price of increased complexity and a higher computational burden.

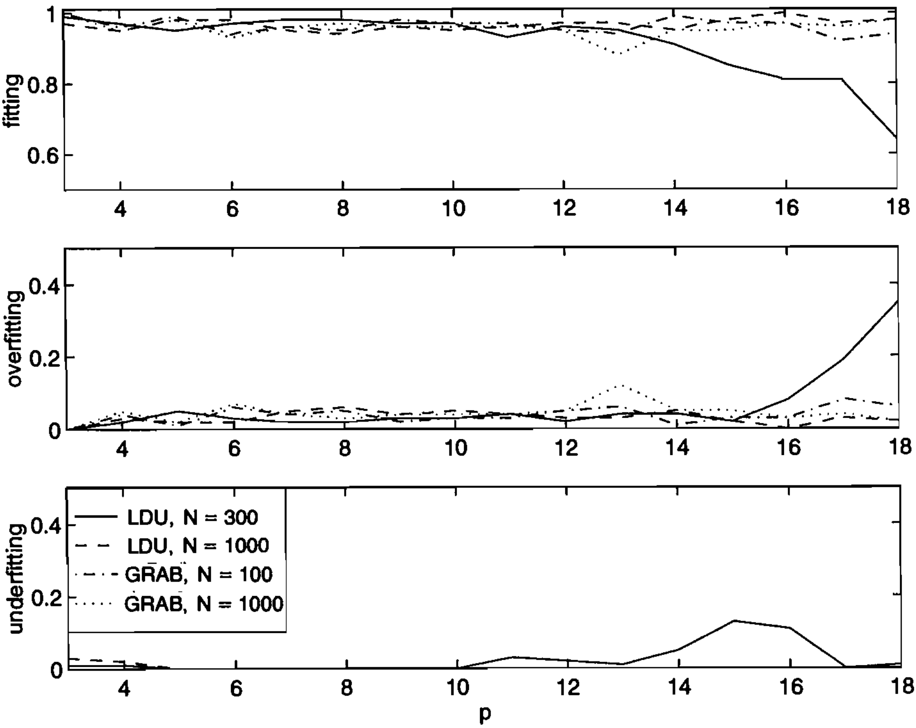


Figure 5 The performance for A1 of GRAB and truncated LDU as a function of p for $N = 300$ and $N = 1000$ (note the different scale).

V. CONCLUSIONS

We have given a detailed presentation of three different methods for ARMA order estimation: the Fuchs test, the LDU test, and the GRAB test. The methods are summarized in Table 1. The performance of the methods is assessed by means of some numerical experiments for scalar ARMA processes.

In the cases we considered, all three methods gave similar results, with a slightly better performance for the GRAB test in the case of a more complicated ARMA process, provided that the user parameter associated with this method was chosen optimally. Our conclusion is that in the case of a scalar ARMA process, the Fuchs method is the easiest one to use, as it involves no user parameters. Since the methods showed a similar performance, it is natural to first try the Fuchs method for an order estimation problem at hand.

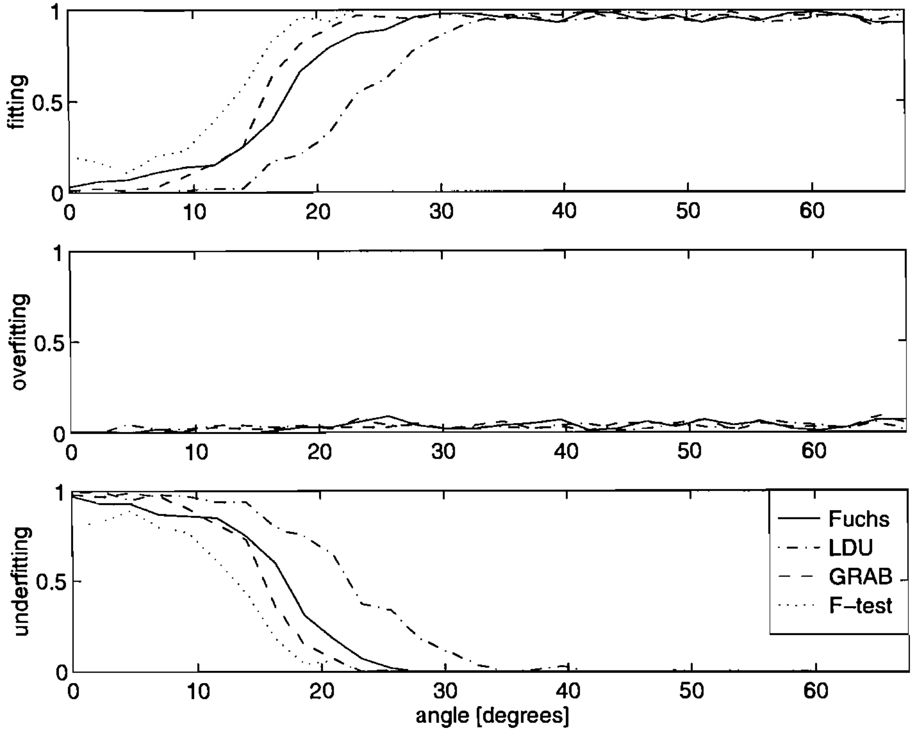


Figure 6 Performance for different pole locations: the angle between the zero and the poles of A1 is varied ($p = 7$, $N = 300$).

The LDU and GRAB tests may provide better results for more complicated model structures, as they allow the use of more information (covariance elements) by suitably changing the user parameters. However, the LDU test may be difficult to use when only a short data sample is available, as it is then very sensitive to the values of these parameters. Indeed, the GRAB test showed better performance than the other methods for closely spaced poles and zeros in the cases we considered.

In many cases, the rank test methods have a somewhat degraded performance as compared to maximum likelihood tests, such as, e.g., the F-test. However, the rank tests are easy to apply as they operate more directly on the output data. Neither a numerical search procedure, nor a parametrization of the ARMA process under study is required, which is a significant advantage of the rank-based order determination methods.

REFERENCES

1. P Stoica, P Eykhoff, P Janssen, T Söderström. Int J Control, 43:1841–1878, 1986.
2. B Choi. ARMA Model Identification. New York: Springer-Verlag, 1992.
3. LD Davisson. IEEE Trans Inf Theory 11:527–532, 1965.
4. H Akaike. IEEE Trans Autom Control 19:716–723, 1974.
5. M Verhaegen. A novel non-iterative MIMO state space model identification technique. In Proceedings of the 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation, Budapest, 1991.
6. P Van Overschee, B De Moor. Automatica 30:75–93, 1994.
7. M Viberg. Automatica 31:1835–1852, 1995.
8. T Söderström, P Stoica. System Identification. Hemel Hempstead: Prentice Hall, 1989.
9. MR Gevers. Int J Control 43:1745–1761, 1986.
10. T Söderström. Discrete-Time Stochastic Systems, Estimation and Control. Hemel Hempstead: Prentice Hall, 1994.
11. MR Gevers, V Wertz. Automatica 20:333–347, 1984.
12. T Kailath. Linear Systems. Englewood Cliffs: Prentice Hall, 1980.
13. L Ljung. System Identification: Theory For the User. Englewood Cliffs: Prentice Hall, 1987.
14. H Cramér. Mathematical Methods of Statistics. Princeton: Princeton University Press, 1966.
15. J-J Fuchs. IEEE Trans Autom Control 32:358–361, 1987.
16. J-J Fuchs. IEEE Trans Autom Control 35:1338–1341, 1990.
17. P Stoica. IEEE Trans Autom Control 26: 572–573, 1981.
18. JH Wilkinson. The Algebraic Eigenvalue Problem. Oxford: Oxford University Press, 1965.
19. TW Anderson. The Statistical Analysis of Time Series. New York: Wiley, 1971.
20. P Janssen, P Stoica. IEEE Trans Autom Control 33:867–870, 1988.
21. L Gill, A Lewbel. J Amer Statist Assoc 87:766–776, 1992.
22. GH Golub, CF Van Loan. Matrix Computations. Maryland, USA: The Johns Hopkins University Press, 1989.
23. P Stoica, M Cedervall. Eigenvalue-based test for detection and rank estimation. In: Proceedings of the 13th IFAC World Congress, San Francisco, 1996.
24. GW Stewart. Introduction to Matrix Computations. New York: Academic Press, 1973.
25. DR Jensen, H Solomon. J Amer Statist Assoc 67:898–902, 1972.

7

A MAP Recursive Nonlinear Filtering

Shin Ichi Aihara

Science University of Tokyo, Suwa College, Nagano, Japan

Arunabha Bagchi

University of Twente, Enschede, The Netherlands

I. SYSTEM AND OBSERVATION MECHANISMS

We consider a simple example of system and observation mechanisms:

$$\frac{\mathrm{d}x^2(t)}{\mathrm{d}t^2} + \mu \frac{\mathrm{d}x(t)}{\mathrm{d}t} - f\left(x(t), \frac{\mathrm{d}x(t)}{\mathrm{d}t}\right) = \text{system noise} \tag{1}$$

$$\begin{cases} y_1(t) = h_1 x(t) + \text{observation noise} \\ y_2(t) = h_2\, \mathrm{d}x(t)/\mathrm{d}t + \text{observation noise} \end{cases} \tag{2}$$

The precise meaning of the above equations can be given by

$$\begin{cases} \mathrm{d}x_1(t)/\mathrm{d}t = x_2(t), \qquad x_1(0) = x_0 \\ \quad \mathrm{d}x_2(t) = -\mu x_2(t)\,\mathrm{d}t + f(x_1, x_2)\,\mathrm{d}t + \mathrm{d}w(t), \qquad x_2(0) = \dot{x}_0 \end{cases} \tag{3}$$

$$y(t) = C(x_1(t)x_2(t))' + e(t), \tag{4}$$

where

$$C = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix} \tag{5}$$

$w(t)$ is a Brownian motion process in R^1 with incremental covariance σ^2 and $e(t)$ is a finitely additive white noise in $L^2(T; R^2)$ independent of w. If the observation noise is modelled by using a Brownian motion, the data are given by the "integral form" and are nowhere differentiable. This causes a serious difficulty when one handles real data. The remarkable advantage of the finitely additive white noise is that the results obtained are always in the form where real data can be directly used. In technical terms we say that the results are always in *robust form*. (See [3] for more general information on finitely additive white noise theory.) The nonlinear filtering problem for general systems has also been studied and the related Zakai equation has been derived in [3].

Here we shall consider a different situation from the usual filtering problem, i.e., a maximum *a posteriori* probability (MAP) state estimation problem is studied. If the observation noise is modeled by a Brownian motion process, the existence of a MAP estimator has already been explored by Zeitouni and Dembo [6]. The main objective of this chapter is twofold. One is to reformulate the MAP state estimate in the finitely additive white noise setup. The other is to derive the recursive estimation equation like the Zakai equation for a nonlinear filtering problem.

In Section II, first we derive an Onsager–Machlup functional to the system equation (3). From this functional, the *a posteriori* probability functional can be derived in the finitely additive white noise setup. In Section III, the basic equation for the recursive MAP estimator is obtained by using a dynamic programming approach. A numerical approximation technique for realizing this basic equation is proposed with the aid of a finite difference scheme in Section IV. We also demonstrate a numerical example. The symbols used in this chapter are listed below:

T: time interval $]0, t_f[$

$L^2(T; R^k)$: space of square integrable functions defined on T with values in R^k; $|\cdot|_H$ and $(\cdot,\cdot)$ denote the norm and inner product, respectively.

$H^1(T; R^2)$: space of functions ϕ defined on T with values in R^2 such that $d\phi/dt \in L^2(T; R^2)$.

C_b^k: space of k-times continuously differentiable functions with bounded derivatives.

II. ONSAGER–MACHLUP FUNCTIONAL AND *A POSTERIORI* PROBABILITY

The signal $x(t) = (x_1(t)x_2(t))'$ is an R^2-valued process on a probability space $(\Omega, \mathcal{A}, \mathcal{P})$. We know that (3) has a unique continuous solution under the Lipschitz condition for f. In (4) e is the Gaussian white noise on $(H, \mathcal{C}, \mu_G)$, where $H = L^2(T; R^2)$, $\mathcal{C}$ is the field of cylinder sets in H, and μ_G is the canonical Gauss measure on $\mathcal{C}$. Now the processes $x(t)$ and $e(t)$ are defined on different probability spaces. To make (4) precise, we have to define x and e on the same probability space. It is clear that the union of such σ-algebras over all possible finite dimensional projections is an algebra, which we denoted by $\mathcal{G}$. We define a finitely additive measure

α on $\mathcal{G}$ by

$$\alpha(C \times A) = \mu_G(C)\mathcal{P}(A), \qquad C \in \mathcal{C}, \quad A \in \mathcal{A}$$

The basic result in signal estimation is the white noise version of the Kallianpur–Striebel formula [3].

Theorem 2.1 *For any integrable g in $(\Omega, \mathcal{A}, \mathcal{P})$,*

$$E_\alpha\{g \mid y\} = \frac{E\{g(x)\exp(-\frac{1}{2}|C(x)|_H^2 + (y, C(x)))\}}{E\{\exp(-\frac{1}{2}|C(x)|_H^2 + (y, C(x)))\}} \tag{6}$$

where $E_\alpha(\cdot)$ and $E(\cdot)$ denote the expectation with respect to the measure α and $\mathcal{P}$, respectively. The likelihood functional is given by

$$LF(y) = E\{\exp(-\tfrac{1}{2}|C(x)|_H^2 + (y, C(x)))\} \tag{7}$$

In order to derive the *a posteriori* probability state estimator, we should modify the Onsager–Machlup functional given by [5], because the system state $x_1(t)$ is not stochastically perturbed.

Theorem 2.2 *Assume that*

$$f \in C_b^2(R^1). \tag{8}$$

Then

$$\lim_{\varepsilon \to 0} \frac{\mathcal{P}\{\sup_t |x_2(t) - \phi_2(t)|_{R^1} \leq \varepsilon \mid x_2(0) = \dot{x}_o, x_1(0) = x_0\}}{\mathcal{P}\{\sup_t |w(t)|_{R^1} < \varepsilon\}}$$
$$= \exp\left\{ -\frac{1}{2\sigma^2} \int_0^{t_f} \left| \frac{d\phi_2(t)}{dt} + \mu\phi_2(t) - f(\phi_1(t), \phi_2(t)) \right|_{R^1}^2 dt \right.$$
$$\left. + \frac{\mu}{2} t_f - \frac{1}{2} \int_0^{t_f} \frac{\partial f(\phi_1, \phi_2)}{\partial x_2} dt \right\}, \tag{9}$$

for $\phi = (\phi_1\ \phi_2)' \in H^1(T; R^2)$ with $\phi_2(0) = \dot{x}_o$ and $\phi_1(t) = x_0 + \int_0^t \phi_2(s)\,ds$.

Proof. It is easy to show that $x_2(t) - \phi_2(t)$ satisfies

$$d(x_2(t) - \phi_2(t)) = -\frac{d\phi_2(t)}{dt} dt - \mu(x_2(t) - \phi_2(t) + \phi_2(t))\,dt$$
$$+ f\left(\int_0^t (x_2(s) - \phi_2(s))\,ds + x_o + \int_0^{t_f} \phi_2(s)\,ds, \right.$$
$$\left. x_2 - \phi_2(t) + \phi_2(t) \right) dt + dw(t)$$
$$= -\frac{d\phi_2(t)}{dt} dt + \tilde{f}(x_2 - \phi_2 + \phi_2)\,dt + dw(t) \quad \text{(say)}$$

By Girsanov's transformation, one has

$$\lim_{\varepsilon\to 0}\frac{\mathcal{P}\{\sup_t|x_2(t)-\phi_2(t)|_{R^1}<\varepsilon|x_2(0)=\dot{x}_o,x_1(0)=x_0\}}{\mathcal{P}\{\sup_t|w(t)|_{R^1}<\varepsilon\}}$$

$$= E\Bigg\{\exp\Bigg\{\int_0^{t_f}\tilde{f}(w+\phi_2)-\frac{\mathrm{d}\phi_2(t)}{\mathrm{d}t}\mathrm{d}w(t)/\sigma^2$$

$$-\frac{1}{2\sigma^2}\int_0^{t_f}\left|\tilde{f}(f(w+\phi_2)-\frac{\mathrm{d}\phi_2(t)}{\mathrm{d}t}\right|^2_{R^1}\mathrm{d}t\Bigg\}$$

$$|\sup_t|w(t)|_{R^1}<\varepsilon,x_2(0)=\dot{x}_o,x_1(0)=x_0\Bigg\}$$

It is easy to show that

$$\int_0^{t_f}\tilde{f}(w+\phi_2)\,\mathrm{d}w(t)=\int_0^{t_f}\tilde{f}(\phi_2)\,\mathrm{d}w(t)$$

$$+\int_0^{t_f}\Bigg[-\mu w(t)+\frac{\partial f(\phi_1(t),\phi_2(t))}{\partial x_1}\int_0^t w(s)\,\mathrm{d}s$$

$$+\frac{\partial f(\phi_1(t),\phi_2(t))}{\partial x_2}w(t)\Bigg]\mathrm{d}w(t)$$

$$+\int_0^{t_f}\left[O(w^2)+O\left(\left(\int_0^t w(s)\,\mathrm{d}s\right)^2\right)\right]\mathrm{d}w(t) \tag{10}$$

By Ito's formula, we have

$$-\mu\int_0^{t_f}w(t)\,\mathrm{d}w(t)=\frac{\mu}{2}\{\sigma^2t_f-|w(t_f)|^2\} \tag{11}$$

and

$$\int_0^{t_f}\frac{\partial f(\phi_1(t),\phi_2(t))}{\partial x_2}w(t)\,\mathrm{d}w(t)=\frac{\partial f(\phi_1(t_f),\phi_2(t_f))}{\partial x_2}\frac{|w(t_f)|^2}{2}$$

$$-\int_0^{t_f}\frac{|w(t)|^2}{2}\Bigg\{\frac{\partial^2 f(\phi_1(t),\phi_2(t))}{\partial x_1\partial x_2}\phi_2(t)$$

$$+\frac{\partial^2 f(\phi_1(t),\phi_2(t))}{\partial x_2^2}\frac{\mathrm{d}\phi_2(t)}{\mathrm{d}t}\Bigg\}\mathrm{d}t$$

$$-\frac{\sigma^2}{2}\int_0^{t_f}\frac{\partial f(\phi_1(t),\phi_2(t))}{\partial x_2}\mathrm{d}t. \tag{12}$$

If w satisfies $\sup_{t \in T} |w(t)|_{R^1} \leq \varepsilon$, then

$$\left| -\mu \int_0^{t_f} w(t)\, \mathrm{d}w(t) - \frac{\mu}{2} \sigma^2 t_f \right| \leq C_1 \varepsilon, \tag{13}$$

and

$$\left| \int_0^{t_f} \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_2} w(t)\, \mathrm{d}w(t) + \frac{\sigma^2}{2} \int_0^{t_f} \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_2} \mathrm{d}t \right| \leq C_2 \varepsilon \tag{14}$$

Integrating by parts, we also have

$$\begin{aligned} &\int_0^{t_f} \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_1} \int_0^t w(s)\, \mathrm{d}s\, \mathrm{d}w(t) \\ &\quad = \int_0^{t_f} \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_1} \mathrm{d}w(t) \int_0^{t_f} w(t)\, \mathrm{d}t \int_0^{t_f} w(t) \int_0^t \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_1} \\ &\quad \times \mathrm{d}w(s)\, \mathrm{d}s\, \mathrm{d}t \end{aligned} \tag{15}$$

If w satisfies $\sup_{t \in T} |w(t)|_{R^1} \leq \varepsilon$, then

$$\left| \int_0^{t_f} \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_1} \mathrm{d}w(t) \right| \leq C_3 \varepsilon \tag{16}$$

and hence

$$\left| \int_0^{t_f} \frac{\partial f(\phi_1(t), \phi_2(t))}{\partial x_1} \int_0^t w(s)\, \mathrm{d}s\, \mathrm{d}w(t) \right| \leq C_4 \varepsilon. \tag{17}$$

From the recent results given by Shepp and Zeitouni [5], we have

$$E\left\{ \exp\left\{ \int_0^{t_f} k(s)\, \mathrm{d}w(s) \right\} \middle| \sup_{t \in T} |w(t)|_{R_1} \leq \varepsilon \right\} \to 1 \quad \text{as } \varepsilon \to 0 \tag{18}$$

for any $k \in L^2(T; R^1)$.

Hence, summarizing the above estimates, we finally obtain the result (9), and the proof is complete.

Now we shall present an *a posteriori* probability under the observation data $y \in L^2(T; R^2)$.

Theorem 2.3 *For* $\phi \in \Phi = \{\phi \in R^2 | \phi_2 \in H^1(T; R^1),\ \phi_1(t) = x_o + \int_0^t \phi_2(s)\,\mathrm{d}s,\ \phi_2(0) = \dot{x}_o\}$,

$$\lim_{\varepsilon \to 0} \frac{\text{Probability}\{\sup_{t\in T}|x_2(t) - \phi_2(t)|_{R^1} < \varepsilon | y(t), t \in T, x_2(0) = \dot{x}_o, x_1(0) = x_o\}}{\mathcal{P}\{\sup_{t\in T}|w(t)|_{R^1} < \varepsilon\}}$$

$$= \frac{1}{LF(y)} \exp\left\{ - \frac{1}{2\sigma^2} \int_0^{t_f} \left| B\left(\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} - F(\phi(t)) \right) \right|^2_{R^1} \mathrm{d}t \right.$$

$$- \frac{1}{2} \int_0^{t_f} B \frac{\partial F(\phi(t))}{\partial \phi_2} \mathrm{d}t - \frac{1}{2} \int_0^{t_f} |C(\phi(t))|^2_{R^2} \mathrm{d}t$$

$$\left. + \frac{1}{2} \int_0^{t_f} C(\phi(t))' y(t)\,\mathrm{d}t \right\}$$

$$= \frac{1}{LF(y)} \exp\left\{ - \int_0^{t_f} L(\phi(t), y(t))\,\mathrm{d}t \right\} \quad \text{(say)},$$

where

$$B = (0 \ \ 1), \tag{19}$$

and

$$F(\phi) = \begin{pmatrix} \phi_2 \\ -\mu\phi_2 + f(\phi_1, \phi_2) \end{pmatrix}. \tag{20}$$

Proof. In Theorem 2.1, we set

$$g(x) = \chi_{\{\omega: \sup|x_2 - \phi_2| < \varepsilon\} \cap \{\omega: x_2(0) = \dot{x}_o x_1(0) = x_o\}}$$

where χ is an indicator function. Hence

$$E_\alpha\{\chi_{\{\omega: \sup|x_2 - \phi_2| < \varepsilon\} \cap \{\omega: x_2(0) = \dot{x}_o x_1(0) = x_o\}} | y(t), t \in T\}$$

$$= \frac{E\{\chi_{\{\omega: \sup|x_2 - \phi_2| < \varepsilon\} \cap \{x_2(0) = \dot{x}_o x_1(0) = x_o\}} \exp(-\frac{1}{2}|C(x)|^2_H + (y, C(x)))\}}{E\{\exp(-\frac{1}{2}|C(x)|^2_H + (y, C(x)))\}}$$

$$= \mathrm{APR}(\phi, y) \quad \text{(say)}. \tag{21}$$

It is easy to show that $\mathrm{APR}(\phi, y)$ is equal to the probability of the event $\{\sup_{t\in T}|x_2(t) - \phi_2(t)|_{R^1} < \varepsilon\}$ under the condition $\{y(t), t \in T, x_2(0) = \dot{x}_o x_1(0) = x_o\}$. On the other hand, it is easy to show that

$$-\frac{1}{2} \int_0^{t_f} |C(x)|^2_{R^2} \mathrm{d}t + \int_0^{t_f} C(x)' y\,\mathrm{d}t + \frac{1}{2} \int_0^{t_f} |C(\phi)|^2_{R^2} \mathrm{d}t - \int_0^{t_f} C(\phi)' y\,\mathrm{d}t$$

$$= -\frac{1}{2} \int_0^{t_f} (C(x) - C(\phi))'(C(x) + C(\phi))\,\mathrm{d}t + \int_0^{t_f} (C(x) - C(\phi))' y\,\mathrm{d}t$$

$$= I(\phi, y) \quad \text{(say)} \tag{22}$$

In the space $\{\omega; \sup_t |x_2(t) - \phi_2(t)|_{R^1} < \varepsilon\} \cap \{\omega : x_2(0) = \dot{x}_o x_1(0) = x_o\}$, we obtain

$$\begin{aligned} |I(\phi, y)| &\le C_1(\sup_t |x_2 - \phi_2|^2_{R^1} + \sup_t |x_2 - \phi_2|_{R^1}(1 + |y|^2_H)) \\ &\le C_2 \varepsilon \end{aligned}$$

Hence

$$\begin{aligned} &\exp(-C_2\varepsilon) E\{\chi_{\{\sup_t |x_2(t) - \phi_2(t)|_{R^1} < \varepsilon\} \cap \{\omega : x_2(0) = \dot{x}_o, x_1(0) = x_o\}}\} \\ &\quad \le \mathrm{APR}(\phi, y) \exp\left\{\frac{1}{2} \int_0^{t_f} |C(\phi)|^2_{R^2} \, \mathrm{d}t - \int_0^{t_f} C(\phi)' y \, \mathrm{d}t\right\} \\ &\quad \le \exp(C_2\varepsilon) E\{\chi_{\{\sup_t |x_2(t) - \phi_2(t)|_{R^1} < \varepsilon\} \cap \{\omega : x_2(0) = \dot{x}_o, x_1(0) = x_o\}}\} \end{aligned}$$

Consequently, the following inequality can be derived:

$$\begin{aligned} &\exp(-C_2\varepsilon) \frac{\mathcal{P}\{\sup_t |x_2(t) - \phi_2(t)|_{R_1} < \varepsilon \,|\, x_2(0) = \dot{x}_o x_1(0) = x_o\}}{\mathcal{P}\{\sup_t |w(t)|_{R^1} < \varepsilon\}} \\ &\quad \le \frac{\mathrm{APR}(\phi, y)}{\mathcal{P}\{\sup_t |w(t)|_{R^1} \le \varepsilon\}} \exp\left\{\frac{1}{2} \int_0^{t_f} |C(\phi)|^2_{R^2} \, \mathrm{d}t - \int_0^{t_f} C(\phi)' y \, \mathrm{d}t\right\} \\ &\quad \le \exp(C_2\varepsilon) \frac{\mathcal{P}\{\sup_t |x_2(t) - \phi_2(t)|_{R^1} < \varepsilon \,|\, x_2(0) = \dot{x}_o, x_1(0) = x_o\}}{\mathcal{P}\{\sup_t |w(t)|_{R^1} < \varepsilon\}} \end{aligned}$$

The proof has been completed.

III. HAMILTON–JACOBI EQUATION

In order to derive the recursive state estimator from the cost given by (13), we must take the optimization procedure twice. First, for a fixed t, the optimization problem with the cost given by (13) is rewritten as

$$\begin{cases} \dot{\phi}(s) = F(\phi(s)) + B' v(s), & 0 \le s < t \\ \phi(t) = x \end{cases} \tag{23}$$

with the cost

$$\begin{aligned} J(t, y, v) = l_0(\phi(0)) &+ \frac{1}{2\sigma^2} \int_0^t |v(s)|^2_{R^1} \, \mathrm{d}s + \frac{1}{2} \int_0^t B \frac{\partial F}{\partial \phi_2}(\phi) \, \mathrm{d}s \\ &+ \frac{1}{2} \int_0^t |C(\phi)|^2_{R^2} \, \mathrm{d}s - \int_0^t C(\phi) \cdot y \, \mathrm{d}s \end{aligned} \tag{24}$$

where $v \in L^2(0, t; R^1)$, we add the initial cost $l_0(\phi(0))$, and $a \cdot b = \Sigma_{i=1}^2 a_i b_i$.

Define

$$S(t,x) = \inf_v J(t,y,v) \tag{25}$$

where it should be noted that the filter (23) has a terminal condition $\phi(t) = x$ and we can use the usual dynamic programming argument to the value functional (25).

The second optimization step is to find the $\hat{\phi}^o(t)$ such that

$$S(t,\hat{\phi}^o(t)) \leq S(t,x), \quad \forall x \in R^2, \tag{26}$$

at each time t. So $\hat{\phi}^o(t)$ is an optimal state.

This is analogous to the minimum variance state estimation case that to compute the conditional mean $E\{x(t)|y(s);0 \leq s \leq t\}$, one can first average over all samples satisfying $x(t) = x$ and obtain the conditional density $p(t,x)$ and then one averages over $x \in R^2$.

Lemma 3.1 *In addition to all assumptions, we further set*

$$\begin{cases} |l_o(x)| \leq C(1+|x|^2_{R^2}) \\ |D_x l_o(x)| \leq C(1+|x|_{R^2}) \end{cases} \tag{27}$$

The value function $S(t,x)$ satisfies

$$|D_x S(t,x)|_{R^2} \leq C(1+|x|_{R^2}) \tag{28}$$

$$\int_0^{t_f} \left| \frac{\partial S(t,x)}{\partial t} \right| \mathrm{d}t \leq C\left(1+|x|^2_{R^2} + \int_0^{t_f} |y(t)|^2_{R^2}\,\mathrm{d}t\right) \tag{29}$$

$$|S(t,x)| \leq C(1+|x|^2_{R^2}) \tag{30}$$

Proof. Define

$$l(\phi,v,y) = \frac{1}{2\sigma^2}|v|^2_{R^1} + \frac{1}{2}\frac{\partial F}{\partial \phi_2}(\phi) + \frac{1}{2}|C(\phi)|^2_{R^2} - C(\phi)\cdot y \tag{31}$$

From (8), we have

$$l(\phi,v,y) \geq c_1|v|^2_{R^2} - c_0(1+|y|^2_{R^2}) \tag{32}$$

Hence from Bensoussan [1], Lemma 2.1, p. 18, we find that

$$\int_0^t |v(s)|^2_{R^1}\,\mathrm{d}s \leq c(1+|x|^2_{R^2}) \tag{33}$$

Furthermore, the estimates (28) and (30) are also derived in [1], p. 18. Here we shall prove (29), because the function l contains $y \in L^2(T;R^2)$. By using the optimality principle of dynamic programming, we get

$$S(t,x) = \inf\left\{ \int_{t-h}^{t} l(\phi,v,y)\,\mathrm{d}s + S(t-h,\phi^o_t(t-h)) \right\}, \tag{34}$$

where time moves in the backward direction, i.e., $\phi_t^o(t-h)$ is an optimal path at the time $t-h$ for the cost $J(t-h, y, v^o)$. In (34), we set $v(s) = \overline{v}$(constant) and get

$$S(t,x) \le \int_{t-h}^{t} l(\overline{\phi}_t, \overline{v}, y)\,ds + S(t-h, \overline{\phi}_t(t-h)) \tag{35}$$

where

$$\dot{\overline{\phi}}_t(s) = F(\overline{\phi}_t(s)) + B'\overline{v}, \overline{\phi}_t(t) = x \tag{36}$$

It follows from (36) that

$$|\overline{\phi}_t(t-h) - x|_{R^2} \le C|h|(1 + |x|_{R^2}) \tag{37}$$

From (28), we have

$$\begin{aligned} &|S(t-h, \overline{\phi}_t(t-h)) - S(t-h,x)| \\ &\quad \le D_x S(t-h,x)_{|x=\xi} \cdot (\overline{\phi}_t(t-h) - x), \quad \text{for } \min(\overline{\phi}_t, x) \le \xi \le \max(\overline{\phi}_t, x) \\ &\quad \le C|h|(1 + |x|_{R^2}^2) \end{aligned} \tag{38}$$

From (33), we obtain

$$\begin{aligned} S(t,x) &\le \int_{t-h}^{t} l(\overline{\phi}_t, \overline{v}, y)\,ds \\ &\qquad + S(t-h,x) + |S(t-h, \overline{\phi}_t(t-h)) - S(t-h,x)| \\ &\qquad \le C|h|(1 + |x|_{R^2}^2) + C\int_{t-h}^{t} |y(s)|_{R^2}^2\,ds + S(t-h,x) \end{aligned} \tag{39}$$

Hence,

$$\frac{\partial S(t,x)}{\partial t} \ge C(1 + |x|_{R^2}^2 + |y(t)|_{R^2}^2) \quad \text{a.e. } t. \tag{40}$$

On the other hand, for every control satisfying (33), we obtain

$$\begin{aligned} &|S(t-h, \phi_t(t-h)) - S(t-h,x)| \\ &\quad \le D_x S(t-h,x)_{|x=\xi} \cdot (\phi_t(t-h) - x) \\ &\quad \le C(1 + |x|_{R^2}) \cdot |\phi_t(t-h) - x|_{R^2} \\ &\quad \le C(1 + |x|_{R^2}^2)h + C(1 + |x|_{R^2}) \int_{t-h}^{t} |v(s)|_{R^1}\,ds \end{aligned} \tag{41}$$

It follows from (32) and (41) that

$$S(t,x) \le \inf\left\{ c_1 \int_{t-h}^{t} |v(s)|_{R^1}^2\,ds - c_0 h + S(t-h, \phi_t^o(t-h)) \right\}$$

$$\leq \inf\left\{c_1 \int_{t-h}^{t} |\nu(s)|^2_{R^1}\, ds - c_0 h + S(t-h,x)\right.$$

$$\left. - |S(t-h, \phi^o_t(t-h)) - S(t-h,x)|\right\}$$

$$\leq \inf\left\{c_1 \int_{t-h}^{t} |\nu(s)|^2_{R^1}\, ds - c_0 h\right.$$

$$- c(1 + |x|_{R^1}) \int_{t-h}^{t} |\nu(s)|_{R^1}\, ds$$

$$\left. - c(1 + |x|^2_{R^1})h + S(t-h,x)\right\}$$

$$\leq - \bar{C}(1 + |x|^2_{R^1})h + S(t-h,x) \tag{42}$$

This implies that

$$\frac{\partial S(t,x)}{\partial t} \leq - \bar{C}(1 + |x|^2_{R^2}) \tag{43}$$

Hence, (29) can be derived. The proof has been completed.

The dynamic programming equation becomes

$$\frac{\partial S(t,x)}{\partial t} - \inf\{l(x,\nu,y) - D_x S(t,x) \cdot \{F(x) + B'\nu\}\} = 0 \tag{44}$$

$$S(0,x) = l_o(x) \tag{45}$$

where l is given by

$$l(x,\nu,y) - D_x S(t,x) \cdot \{F(x) + B'\nu\} =$$

$$\frac{1}{2\sigma^2}|\nu - \sigma^2 D_{x_2} S(t,x)|^2_{R^1} - \frac{\sigma^2}{2}|D_{x_2} S(t,x)|^2_{R^2} - D_x S(t,x) \cdot f(x)$$

$$+ \frac{1}{2}|C(x)|^2_{R^2} - C(x) \cdot y + \frac{1}{2}\frac{\partial F}{\partial x_2}(x) - D_x S(t,x) \cdot F(x)$$

Hence, the optimal ν^o is given by

$$\nu^o = D_{x_2} S(t,x)$$

Consequently, we have the following Hamilton–Jacobi equation:

$$\frac{\partial S(t,x)}{\partial t}+\frac{\sigma^2}{2}|D_{x_2}S(t,x)|^2_{R^1}+D_xS(t,x)\cdot F(x)$$
$$-\frac{1}{2}|C(x)|^2_{R^d}+C(x)\cdot y(t)-\frac{1}{2}\frac{\partial F}{\partial x_2}(x)=0 \tag{46}$$

$$S(0,x)=l_o(x). \tag{47}$$

Theorem 3.1 *The function $S(t,x)$ defined by (25) satisfies (46). Moreover S is the maximum function satisfying (23–30) and (46).*

This proof has been found in Bensoussan [1], pp. 21–25.

IV. NUMERICAL PROCEDURE

From Lemma 3.1, (27) is characterized by

$$D_xS(t,x)_{|x=\hat{\mathscr{P}}(t)}=0 \tag{48}$$

Hence, for solving the Hamilton–Jacobi equation (46) numerically, we need only the local solution $S(t,x)$ around $D_xS=0$. This implies that we can set artificial boundary conditions. If we solve the Zakai equation instead of the Hamilton–Jacobi equation, we really need the global $p(t,x)$, because we shall calculate the average $\int_{-\infty}^{\infty}xp(t,x)\,\mathrm{d}x$. The property that from the local solution of $S(t,x)$ around $D_xS=0$ the optimal state estimate $\hat{\mathscr{P}}(t)$ can be derived is a main advantage for the formulation of the MAP state estimate. One more advantage is that the Hamilton–Jacobi equation can be numerically solved by using a finite difference scheme and its convergence property of the approximated solution to the original one has been proved by Kushner [4]. Here we shall present a numerical scheme for solving (46) and consider the one-dimensional case in order to simplify notation. First, we set the lattice

$$\Sigma_o^h=\{x=j\delta,j=0,\pm1,\pm2,\ldots,\pm(k-1)\} \tag{49}$$

with

$$\partial\Sigma_o^h=\{x=j\delta,j=\pm k\} \tag{50}$$

where δ is a spatial step, and the boundary $\partial\Sigma_o^h$ is sufficiently large. Before presenting the numerical scheme, we shall introduce the new notation: for any $\nu(t,x)$, let

$$\Delta_j^+ \nu = \frac{\nu(t,x+\delta) - \nu(t,x)}{\delta}, \quad x = j\delta$$

$$\Delta_j^- \nu = \frac{\nu(t,x) - \nu(t,x-\delta)}{\delta}, \quad x = j\delta$$

$$\nu^+ = \max(\nu, 0)$$

$$\nu^- = \max(-\nu, 0).$$

Here we shall present a numerical scheme for realizing the Hamilton–Jacobi equation [2]:

1. First, we find that $\hat{\phi}^o(0)$ which satisfies

$$D_x l_o(x)_{|x=\hat{\phi}^o(0)} = 0 \tag{51}$$

Here we assume that we can find a unique point $\hat{\phi}^o(0)$.

2. Reset Σ_o^h as

$$\Sigma_o^h = \{x = \hat{\phi}^o(0) + j\delta, j = 0, \pm 1, \pm 2, \ldots, \pm(k-1)\} \tag{52}$$

with

$$\partial\Sigma_o = \{X = \hat{\phi}^o(0) \pm k\delta\}$$

3. For the time stage i, construct an indicator function

$$\chi_{ij}^+ = \begin{cases} 1, & \text{for } \{x \in \sigma_{ij}^h | F(x) + \sigma^2 \Delta_j^+ S(t,x) > 0,\ t = \Sigma_{k=0}^i \delta t(k)\} \\ 0, & \text{for others} \end{cases}$$

$$\chi_{ij}^- = \begin{cases} 1, & \text{for } \{x \in \sigma_{ij}^h | F(x) + \sigma^2 \Delta_j^- S(t,x) < 0, t = \Sigma_{k=0}^i \delta t(k)\} \\ 0, & \text{for others} \end{cases}$$

where the time step $\delta t = \delta t(i)$ is chosen as

$$\max_j \left\{ \frac{\delta t(i)}{\delta} [\,|F(\hat{\phi}^o(0) + j\delta) + \sigma^2 \Delta_j^+ S|\chi_{ij}^+ + |F(\hat{\phi}^o(0) + j\delta) + \sigma^2 \Delta_j^- S|\chi_{ij}^-] \right\} \leq 1 \tag{53}$$

4. The observation data are modified by

$$y_i = \frac{1}{\delta t(i)} \int_{\Sigma_{k=0}^{(i-1)}\delta t(k)}^{i \cdot \delta t(i)} y(s)\, \mathrm{d}s \tag{54}$$

5. Denoting

$$S_{i,j} = S(t,x) \qquad \text{for } t = \Sigma_{k=0}^{i}\delta t(k), \quad x = \hat{\phi}^{o}(0) + j\delta \tag{55}$$

the following explicit finite difference equation can be solved:

$$\begin{aligned}\frac{S_{i+1,j} - S_{i,j}}{\delta t(i)} &+ (F_j + \sigma^2 \Delta_j^+ S_{i,j})\Delta_j^+ S_{ij} \cdot \chi_{ij}^{+} \\ &+ (F_j + \sigma^2 \Delta_j^- S_{i,j})\Delta_j^- S_{ij} \cdot \chi_{ij}^{-} \\ &- \frac{\sigma^2}{2}\{|\Delta_j^+ S_{ij}|^2 \chi_{ij}^{+} + |\Delta_j^- S_{ij}|^2 \chi_{ij}^{-}\} \\ -\frac{1}{2}|C_j|^2 + C_j y_i - \frac{1}{2}\left(\frac{\mathrm{d}F}{\mathrm{d}x}\right)_j &= 0, \quad \text{for } j \in \Sigma_o^h \\ &\begin{cases} S_{i+1,-k} = S_{i+1,-k+1} \\ S_{i+1,k} = S_{i+1,k-1} \end{cases}\end{aligned} \tag{56}$$

where

$$F_j = f(\hat{\phi}^{o}(0) + j\delta), \quad C_j = C(\hat{\phi}^{o}(0) + j\delta), \quad \left(\frac{\partial F}{\partial x}\right)_j = \frac{\mathrm{d}F}{\mathrm{d}x}_{|\hat{\phi}^{o}(0)+j\delta} \tag{57}$$

In the above numerical procedure, Fleming and Soner proposed that the spatial step δ is chosen as a function of δt which satisfies (53) in [2]. However, here the time step δt is adjusted, because of the easy implementation by using a digital computer.

6. In order to find an optimal state estimate $\hat{\phi}^{o}(i+1)$, first we find

   $\hat{j} = \operatorname{argmin}_{j \in \Sigma_o^h} S_{i+1,j}$.

7. Around the grid $\hat{j}$, we pick up

$$S_{i+1,\hat{j}-q}, S_{i+1,\hat{j}-q+1}, \ldots, S_{i+1,\hat{j}}, S_{i+1,\hat{j}+1}, \ldots, S_{i+1,\hat{j}+q} \tag{58}$$

   for some fixed $q < k$.
8. By using a cubic interpolation method, we construct a polynomial $\hat{S}_{j+1}(x)$ which satisfies (58).
9. Find a minimal point x of $\hat{S}_{j+1}(x)$. This is an optimal state estimate (see Fig. 1).

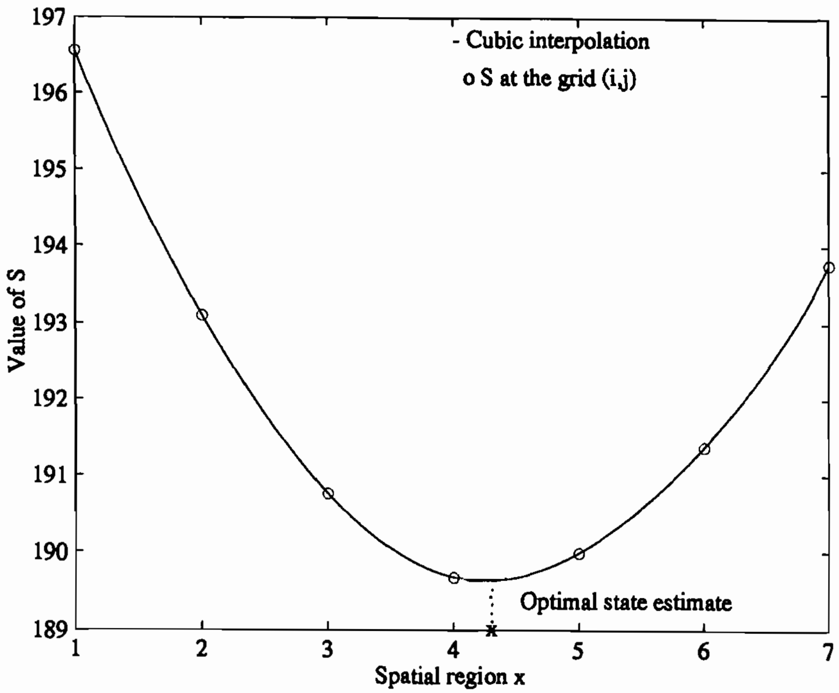


Figure 1 A drawing of how to seek an optimal estimate.

A. A Numerical Example

Since we were performing a minimization for fixed y, no change is effected by adding

$$\frac{1}{2}\int_0^t |y(s)|_{R^2}^2\,\mathrm{d}s$$

to the cost $J(t, y, v)$. So in the following simulations, S is the value function corresponding to the cost $J(t, y, v) + \frac{1}{2}\int_0^t |y(s)|_{R^2}^2\,\mathrm{d}s$.

We consider the following nonlinear function,

$$f\left(x, \frac{\mathrm{d}x}{\mathrm{d}t}\right) = -x^3, \quad \mu = 1.1 \tag{59}$$

$$C = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \tag{60}$$

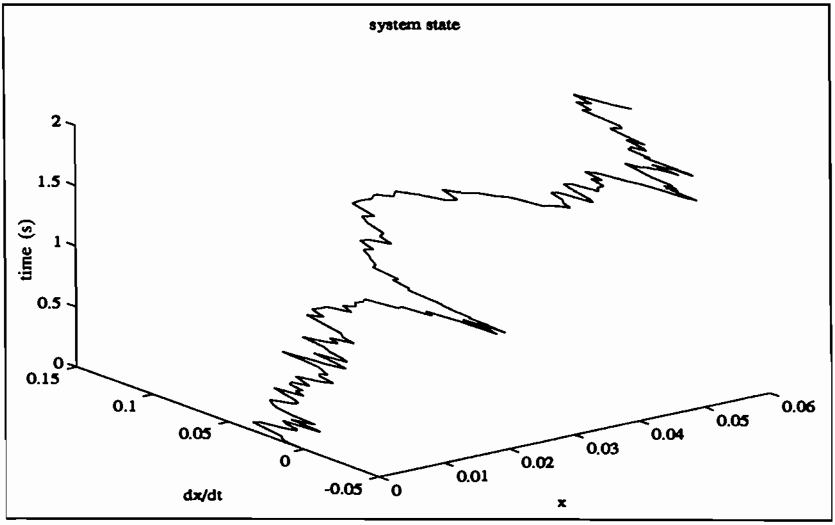


Figure 2 A sample run of true system states.

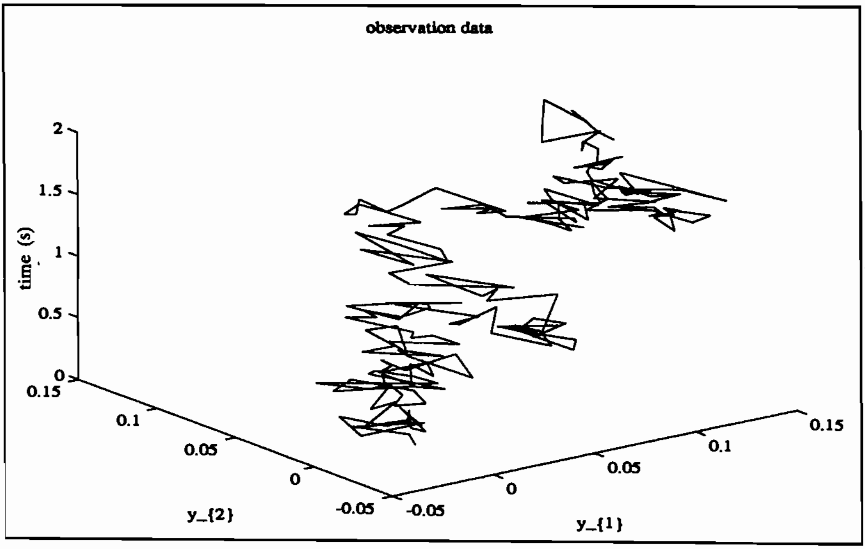


Figure 3 A sample run of the observation data.

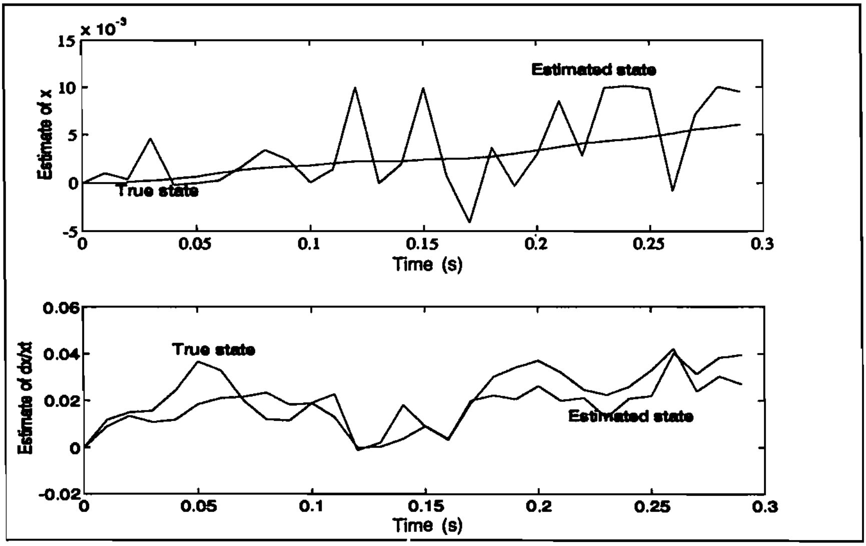


Figure 4 Sample runs of the optimal estimate and true state.

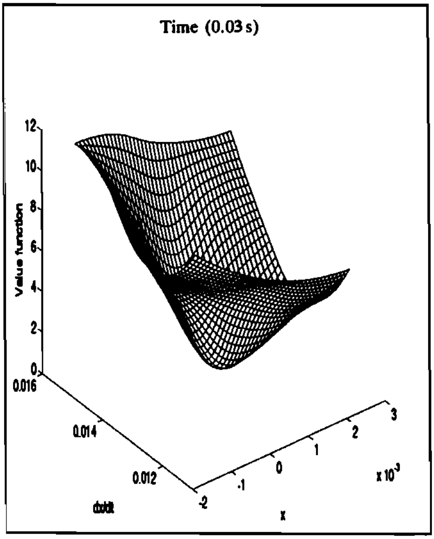


Figure 5 A sample run of the value function $S(t, x)$ at $t = 0.03$ sec.

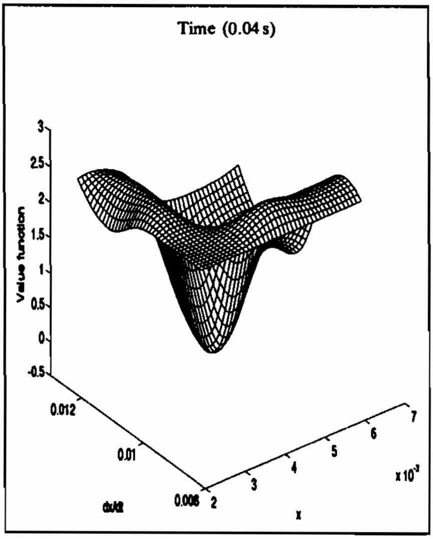


Figure 6 A sample run of the value function $S(t, x)$ at $t = 0.04$ sec.

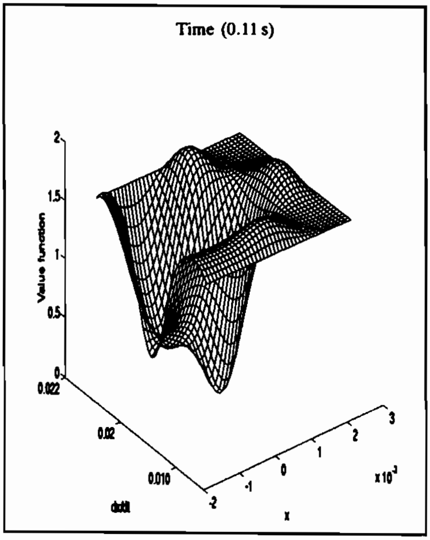


Figure 7 A sample run of the value function $S(t, x)$ at $t = 0.11$ sec.

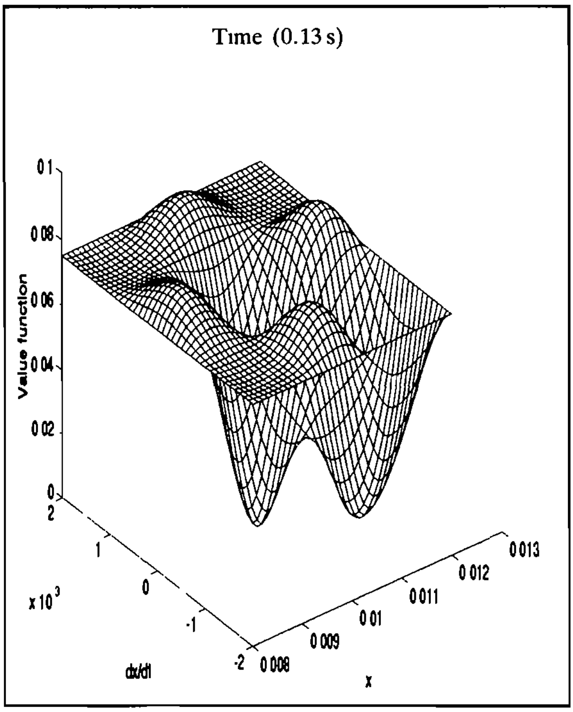


Figure 8 A sample run of the value function $S(t, x)$ at $t = 0.13$ sec.

This system equation is of so-called Duffing type. Theoretically speaking, we assume that for the outside of a sufficiently large ball $|x|_{R^2} \leq D$ the nonlinear function $f(x, \mathrm{d}x/\mathrm{d}t)$ is reset as a linear function. Furthermore, we replace e by $\sigma_o e(t)$. In the digital simulation studies, we set

$$\delta t(0) = 0.001, \qquad \delta = 0.01, \qquad \sigma = 0.1, \qquad \sigma_o = 0.001 \tag{61}$$

Furthermore, we set the initial condition $l_o(x)$ as

$$\frac{1}{2}x^2$$

In Figs 2 and 3, two sample runs of true system state $x(t)$ and observation data $y(t)$ are demonstrated.

From $S(t, x)$, we must find a point $x = \mathrm{argmin}_{x \in R}(S(t, x))$ and this is the optimal state estimate $\hat{\phi}(t)^o$. In this example, first we seek a grid point $\hat{j}$ and then we pick up 7 point around $\hat{j}$. Next by using a cubic interpolation method, we construct a polynominal and again find a minimal point. The optimal state estimate $\hat{\phi}(t)^o$ is shown in Fig. 4.

For the nonlinear system, the optimal value function $S(x, t)$ is no longer a quadratic form. We demonstrate the value function $S(x, t)$ at the time points $t = 0.03$, 0.04, 0.11, 0.13, 0.16, 0.2, 0.24, and 0.3 (sec), in Figs 5–12.

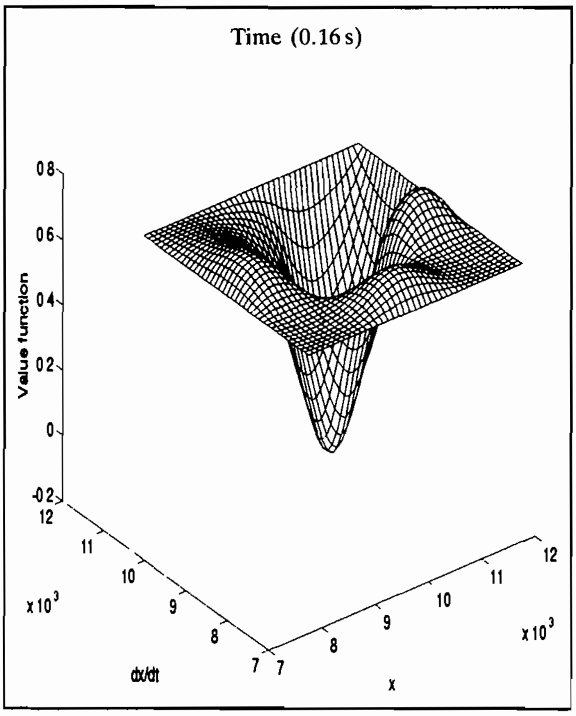


Figure 9 A sample run of the value function $S(t, x)$ at $t = 0.16$ sec.

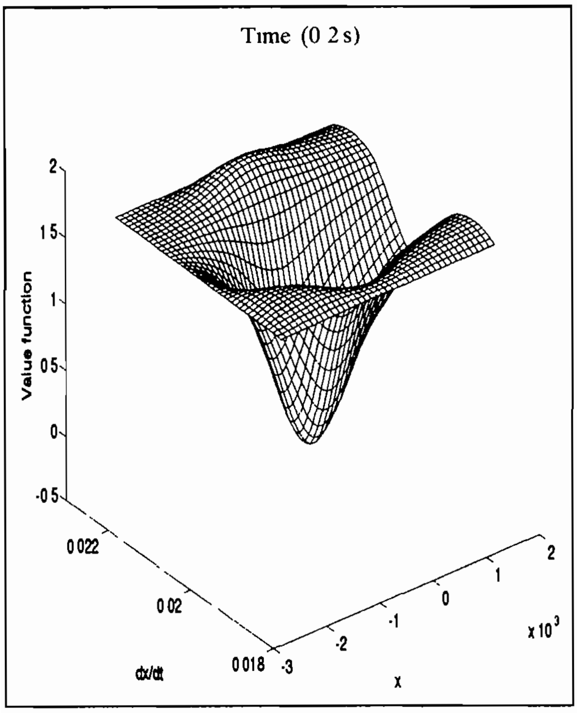


Figure 10 A sample run of the value function $S(t, x)$ at $t = 0.2$ sec.

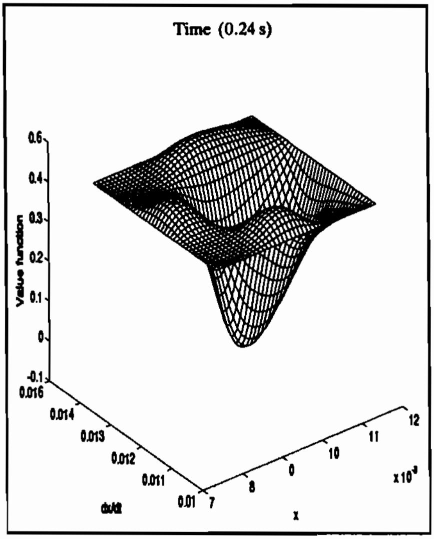


Figure 11 A sample run of the value function $S(t,x)$ at $t = 0.24$ sec.

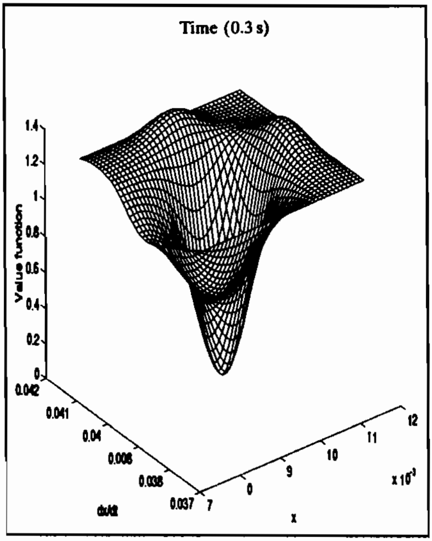


Figure 12 A sample run of the value function $S(t,x)$ at $t = 0.3$ sec.

REFERENCES

1. A Bensoussan. Perturbation Methods in Optimal Control. Chichester: John Wiley, 1988.
2. WH Fleming, HM Soner. Controlled Markov Processes and Viscosity Solutions. New York: Springer-Verlag, 1993.
3. G Kallianpur, RL Karandikar. White Noise Theory of Prediction, Filtering and Smoothing. New York: Gordon and Breach, 1988.
4. HJ Kushner, P Dupuis. Numerical Method for Stochastic Control Problems in Continuous Time. New York: Springer-Verlag, 1992.
5. LA Shepp, O Zeitouni. Ann Prob 20:652–654, 1992.
6. O Zeitouni, A Dembo. Stochastics 20:221–246, 1987.

This Page Intentionally Left Blank

8

Stochastic Properties of the H_∞ Filter

Kiyotsugu Takaba and Tohru Katayama

Kyoto University, Kyoto, Japan

I. INTRODUCTION

In recent years, the H_∞ filtering problem has received much attention as a new approach to robust filtering [1–7]. The H_∞ filtering problem is to find a state estimator such that the maximum energy in the estimation error over all possible noise disturbance trajectories is less than a prescribed bound γ. The filtering with H_∞ criterion is appropriate when there exists a significant uncertainty in the statistics of the noise disturbance.

The H_∞ filtering problem has been solved from various viewpoints [1–7]. An LQ optimization approach in the time domain to the H_∞ filtering and smoothing problems was given for the continuous-time case by Nagpal and Khargonekar [2]. For the discrete time case, the frequency domain approaches to the stationary filtering problem were given based on the bounded real lemma by Haddad *et al.* [3] and Yaesh and Shaked [4]. Yaesh and Shaked [5,6] gave game theoretic interpretations of the H_∞ filter for the finite-horizon problem. Fujita *et al.* [7] derived a solvability condition of the finite horizon H_∞ filtering problem based on completing the square and the conjugate point argument of a certain Riccati difference equation. They also showed the applicability of the H_∞ filter to active vision systems.

In the above works, the H_∞ filter was derived in the deterministic setting where the noise disturbances are deterministic L_2-signals. On the other hand, it is well known that the Kalman filter offers optimal estimates in the least-squares error sense for a stochastic system, and that the H_∞ filter converges to a Kalman filter as γ goes to infinity. This relationship between these two filters suggests the importance of the performance analysis of the H_∞ filter in the stochastic setting. Therefore, in this paper, we will consider the performance of the discrete-time H_∞ filter in the case where the underlying noise disturbances are zero mean Gaussian white noise processes.

This paper is organized as follows. In Section II, we give a brief review of the finite horizon H_∞ filtering problem for a discrete time system. In Section III, it is first shown that the H_∞ filter is optimal in the sense of the exponential quadratic error criterion for a stochastic system [9,10]. Secondly, we compare the performance of the H_∞ filter with that of the Kalman filter where the estimation error covariance matrix is employed as the performance index. Next, we show the monotonicity property of the H_∞ Riccati difference equation (RDE) with respect to the prescribed H_∞ error bound γ. Based on this property of the solution to the H_∞ RDE, we study the relationship between γ and the performance of the H_∞ filter. The results in this paper will provide an insight into the selection of the design parameter γ.

In this paper, we will use the standard notation: $(\cdot)^{\mathrm{T}}$ denotes the transpose, I the identity matrix with appropriate dimension, $E\{\cdot\}$ the expectation, and $\|\cdot\|$ denotes the Euclidean norm for a vector and the largest singular value for a matrix. We also define the weighted Euclidean norm $\|x\|_Q = (x^{\mathrm{T}}Qx)^{1/2}$ for a vector x and a positive definite matrix Q.

II. FINITE HORIZON H_∞ FILTERING PROBLEM

We consider a linear discrete time system described by

$$x_{k+1} = A_k x_k + B_k w_k \tag{1}$$

$$y_k = C_k x_k + D_k v_k \tag{2}$$

where $x_k \in \mathbf{R}^n, y_k \in \mathbf{R}^q$ are the state vector and the measurement output at time k, respectively. The exogenous signals $w_k \in \mathbf{R}^m$ and $v_k \in \mathbf{R}^q$ are the process disturbance and the measurement noise, respectively. Hereafter, we assume that $R_k := D_k D_k^{\mathrm{T}} > 0$ holds for any k. To add generality, we wish to estimate the vector $z_k \in \mathbf{R}^p$ defined by

$$z_k = L_k x_k, \quad L_k \neq 0 \tag{3}$$

The finite horizon H_∞ filtering problem is to find the estimates of z_k and x_k based on the measurement set $\{y_0, \ldots, y_k\}$ such that

$$\sup_{w,v,x_0} \frac{\sum_{k=0}^{N} \|z_k - \hat{z}_k\|^2}{\sum_{k=0}^{N} (\|w_k\|^2 + \|v_k\|^2) + \|x_0 - \bar{x}_0\|^2_{\Pi^{-1}}} < \gamma^2 \tag{4}$$

where $\hat{z}_k$ is the estimate of z_k, and $\bar{x}_0$ is the *a priori* estimate of the initial state x_0. Also, Π is a positive definite weighting matrix which represents the uncertainty of the initial state. The inequality (4) implies that the maximal energy ratio of the estimation error to the disturbance should be less than the prescribed bound γ.

We define the cost functional J by

$$\begin{aligned} J &= \sum_{k=0}^{N} \|z_k - \hat{z}_k\|^2 - \gamma^2\left(\sum_{k=0}^{N} \|w_k\|^2 + \sum_{k=0}^{N} \|v_k\|^2 + \|x_0 - \bar{x}_0\|^2_{\Pi^{-1}} \right) \\ &= \sum_{k=0}^{N} \|z_k - \hat{z}_k\|^2 - \gamma^2\left(\sum_{k=0}^{N} \|w_k\|^2 + \sum_{k=0}^{N} \|y_k - C_k x_k\|^2_{R_k^{-1}} + \right. \\ &\qquad \left. \|x_0 - \bar{x}_0\|^2_{\Pi^{-1}} \right) \end{aligned} \tag{5}$$

Then, it is easily seen that (4) is equivalent to

$$\begin{aligned} &J < 0 \quad \text{for all } (w_k, v_k, x_0) \text{ such that} \\ &\sum_{k=0}^{N} (\|w_k\|^2 + \|v_k\|^2) + (x_0 - \bar{x}_0)^{\mathrm{T}} \Pi^{-1} (x_0 - \bar{x}_0) \neq 0 \end{aligned} \tag{6}$$

A solution to the H_∞ filtering problem is given by the following theorem.

Theorem 1. *Suppose that A_k is nonsingular for all $k \in [0, N]$. Then, there exists a filter achieving the H_∞ error bound in (4) if there exists a positive definite solution P_k to the Riccati difference equation*

$$P_{k+1} = A_k P_k \Sigma_k^{-1} A_k^{\mathrm{T}} + B_k B_k^{\mathrm{T}}, \quad P_0 = \Pi \tag{7a}$$

$$\Sigma_k = I + (C_k^{\mathrm{T}} R_k^{-1} C_k - L_k^{\mathrm{T}} L_k) P_k \tag{7b}$$

and the following inequality holds.

$$V_k := \gamma^2 I - L_k P_k (I + C_k^{\mathrm{T}} R_k^{-1} C_k P_k)^{-1} L_k^{\mathrm{T}} > 0 \tag{8}$$

If such a solution P_k exists, then one of the filters achieving the H_∞ error bound (4) is given by

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K_k (y_k - C_k \hat{x}_{k/k-1}) \tag{9}$$

$$\hat{x}_{k+1/k} = A_k \hat{x}_{k/k}, \quad \hat{x}_{0/-1} = \bar{x}_0 \tag{10}$$

$$\hat{z}_k = L_k \hat{x}_{k/k} \tag{11}$$

$$K_k = P_k C_k^{\mathrm{T}} (R_k + C_k P_k C_k^{\mathrm{T}})^{-1} \tag{12}$$

where $\hat{x}_{k/t}$ denotes an estimate of x_k based on the measurement set $\{y_0, \ldots, y_t\}$.

Proof. Suppose that the RDE (7) has a positive definite solution P_k satisfying $V_k > 0$. It suffices to show that (6) holds for the filter in (9)–(12).

We now define $X_k = P_k^{-1}$. Since A_k is nonsingular, we see from the matrix inversion lemma that

$$X_k = A_k^{\mathrm{T}} X_{k+1} A_k + A_k^{\mathrm{T}} X_{k+1} B_k W_k^{-1} B_k^{\mathrm{T}} X_{k+1} A_k - C_k^{\mathrm{T}} R_k^{-1} C_k + \gamma^{-2} L_k^{\mathrm{T}} L_k \tag{13}$$

where $W_k := I - B_k^{\mathrm{T}} X_{k+1} B_k$. Also, the RDE (7) is rewritten as

$$P_{k+1} - B_k B_k^{\mathrm{T}} = A_k P_k (I + C_k^{\mathrm{T}} R_k^{-1} C_k P_k)^{-1} A_k^{\mathrm{T}} + A_k P_k L_k^{\mathrm{T}} V_k^{-1} L_k P_k A_k^{\mathrm{T}} > 0$$

This implies that $X_{k+1}^{-1} - B_k B_k^{\mathrm{T}} = X_{k+1}^{-\mathrm{T}/2} \{I - X_{k+1}^{\mathrm{T}/2} B_k B_k^{\mathrm{T}} X_{k+1}^{1/2}\} X_{k+1}^{-1/2} > 0$. Hence, we obtain $W_k = I - B_k^{\mathrm{T}} X_{k+1} B_k > 0$.

It is clear from (9)–(12) that

$$\tilde{x}_{k+1} = A_k \tilde{x}_k - A_k K_k \bar{y}_k + B_k w_k \tag{14}$$

where we define

$$\tilde{x}_k = x_k - \hat{x}_{k/k-1}, \qquad \bar{y}_k = y_k - C_k \hat{x}_{k/k-1}$$

Then, simple but tedious calculations using (7)–(14) yield

$$\tilde{x}_{k+1}^{\mathrm{T}} X_{k+1} \tilde{x}_{k+1} - \tilde{x}_k^{\mathrm{T}} X_k \tilde{x}_k = \|w_k\|^2 + \|v_k\|^2 - \gamma^{-2} \|z_k - \hat{z}_k\|^2 - \|w_k - w_k^*\|_{W_k}^2 - \|\bar{y}_k\|_{\Omega_k^{-1}}^2 \tag{15}$$

where

$$w_k^* = W_k^{-1} B_k^{\mathrm{T}} X_{k+1} A_k \tilde{x}_k, \quad \Omega_k = R_k + C_k P_k C_k^{\mathrm{T}}$$

By summing up (15) from $k = 0$ to $k = N$, we obtain

$$J = -\gamma^2 \left\{ \tilde{x}_{N+1}^{\mathrm{T}} X_{N+1} \tilde{x}_{N+1} + \sum_{k=0}^{N} (\|w_k - w_k^*\|_{W_k}^2 + \|\bar{y}_k\|_{\Omega_k^{-1}}^2) \right\} \leq 0 \tag{16}$$

Moreover, suppose that $J = 0$. Then, we get $\tilde{x}_{N+1} = 0$ and $w_k = w_k^*, \bar{y}_k = 0$ for all k. It thus follows from (14) that

$$\tilde{x}_{k+1} = (A_k + B_k W_k^{-1} B_k^{\mathrm{T}} X_{k+1} A_k) \tilde{x}_k, \quad \tilde{x}_{N+1} = 0$$

This implies $x_0 = \bar{x}_0$ and $w_k = 0$, $v_k = 0$ for all k. Therefore, the H_∞ error bound (6) is satisfied by the filter of (9)–(12). ■

Remark 1. The existence of a positive definite solution P_k to the H_∞ RDE (7) satisfying $V_k > 0$ is also necessary for the existence of an H_∞ filter. The proof is given based on the conjugate point argument of the solution to (7) by Fujita *et al.* [7].

Remark 2. The H_∞ filter in (9)–(12) is an optimal minimizing policy in the following minimax filtering problem [8].

$$\max_{v_0} \min_{\hat{z}_0} \max_{x_0, w_0} (\max_{v_1} \min_{\hat{z}_1} \max_{w_1} (\ldots (\max_{v_N} \min_{\hat{z}_N} \max_{w_N} (J)) \ldots)) \tag{17}$$

III. STOCHASTIC PROPERTIES OF H_∞ FILTER

In the previous section, we have derived an H_∞ filter in the deterministic setting. In this section, we will consider the performance of the H_∞ filter in the case when w_k and v_k are zero mean Gaussian white noise processes. In the following, we assume that w_k and v_k are Gaussian white noise processes such that

$$E\{w_k\} = 0, \qquad E\{v_k\} = 0 \tag{18}$$

$$E\left\{\begin{bmatrix} w_k \\ v_k \end{bmatrix}\begin{bmatrix} w_\tau \\ v_\tau \end{bmatrix}^{\mathrm{T}}\right\} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}\delta_{k\tau} \tag{19}$$

where $\delta_{k\tau}$ is the Kronekar delta. We also assume that the initial state x_0 is a Gaussian random vector with mean $\bar{x}_0$ and covariance π.

A. Exponential Quadratic Error Criterion

One of the well-known stochastic properties of the H_∞ filter is the optimality in the exponential quadratic error criterion [9,10].

Proposition 1 [9,10]. *Assume that w_k and v_k are zero mean Gaussian white noise processes with unit covariances and that x_0 is the Gaussian random vector with mean $\bar{x}_0$ and covariance Π. Then, H_∞ filter in (9)–(12) is optimal in the minimization of the following cost functional*

$$J_{\mathrm{EQ}} = E\left\{\exp\left(\gamma^{-2}\sum_{k=0}^{N}\|z_k - \hat{z}_k\|^2_{\Pi^{-1}}\right\}\right) \tag{20}$$

Proof. Since w_k and v_k are zero mean Gaussian white noise processes with unit covariance, and since x_0 is generated by the Gaussian distribution with mean $\bar{x}_0$ and covariance Π, the joint probability density function is given by

$$f(w, v, x_0) = \text{const.}\ \exp\left\{-\sum_{k=0}^{N}(\|w_k\|^2 + \|v_k\|^2) - \|x_0 - \bar{x}_0\|^2_{\Pi^{-1}}\right\} \tag{21}$$

where const. denotes an irrelevant positive constant. Then, by the definition of the expectation, we obtain

$$
\begin{aligned}
J_{\mathrm{EQ}} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\left(\gamma^{-2}\sum_{k=0}^{N}\|z_k-\hat{z}_k\|^2\right)f(w,v,x_0) \\
&\quad \mathrm{d}w_N\mathrm{d}v_N\ldots\mathrm{d}w_0\mathrm{d}v_0\mathrm{d}x_0 \\
&= \text{const.}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp(\gamma^{-2}J) \\
&\quad \mathrm{d}w_N\mathrm{d}v_N\ldots\mathrm{d}w_0\mathrm{d}v_0\mathrm{d}x_0
\end{aligned}
\tag{22}
$$

Recall that the measurement set $\{y_0,\ldots,y_k\}$ is available for the estimation at time k. It thus follows that the minimization of J_{EQ} with respect to $\hat{z}_k(k=0,\ldots,N)$ is expressed as

$$
\int_{-\infty}^{\infty}\min_{\hat{z}_0}\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\cdots\left[\int_{-\infty}^{\infty}\min_{\hat{z}_N}\left\{\int_{-\infty}^{\infty}\exp(\gamma^{-2}J)\,\mathrm{d}w_N\right\}\mathrm{d}v_N\right]\cdots\right]\right.
$$
$$
\left.\mathrm{d}w_0\mathrm{d}x_0\right\}\mathrm{d}v_0
$$

It is also shown that

$$
\int_{-\infty}^{\infty}\exp\{Z(x)\}\,\mathrm{d}x = \text{const.}\exp\{\max_{x} Z(x)\}
\tag{23}
$$

where $Z(x)$ is a quadratic function of x such that $\partial^2 Z/\partial x\,\partial x^{\mathrm{T}} < 0$ [10,11]. Since $\exp(\cdot)$ is a monotonically increasing function, the application of the above formula to (22) shows that the minimization of J_{EQ} with respect to $\hat{z}_k$ $(k=0,\ldots,n)$ is equivalent to the minimax filtering problem (17):

$$
\max_{v_0}\min_{\hat{z}_0}\max_{x_0,w_0}\left(\max_{v_1}\min_{\hat{z}_1}\max_{w_1}\left(\ldots\left(\max_{v_N}\min_{\hat{z}_N}\max_{w_N}(J)\right)\ldots\right)\right)
$$

As in Remark 2, this minimax optimization yields the H_∞ filter of (9)–(12). Therefore, the H_∞ filter minimizes the exponential quadratic cost J_{EQ}.

B. Estimation Error Covariance

As γ tends to infinity, the second term in the cost functional J of (5) becomes dominant, so that the minimax filtering problem (17) reduces to the minimization problem:

$$
\min_{w,x}\left\{\sum_{k=0}^{N}(\|w_k\|^2+\|y_k-C_kx_k\|^2_{R_k^{-1}})+\|x_0-\bar{x}_0\|^2_{\Pi^{-1}}\right\}
$$

As is well known, this minimization problem is equivalent to the minimum-variance estimation or least squares estimation problem in the case where w_k and v_k are zero mean Gaussian white noise processes with unit covariances, and where x_0 is the Gaussian random vector with mean $\bar{x}_0$ and covariance Π [12]. Among all causal state estimators, the following Kalman filter offers the optimal solution to the above minimization:

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + K'_k(y_k - C_k\hat{x}_{k/k-1}) \tag{24}$$

$$\hat{x}_{k+1/k} = A_k\hat{x}_{k/k}, \qquad \hat{x}_{0/-1} = \bar{x}_0 \tag{25}$$

$$K'_k = P'_kC_k^{\mathrm{T}}(R_k + C_kP'_kC_k^{\mathrm{T}})^{-1} \tag{26}$$

where the matrix P'_k is the optimal estimation error covariance matrix

$$P'_k = E\{(x_k - \hat{x}_{k/k-1})(x_k - \hat{x}_{k/k-1})^{\mathrm{T}}\},$$

which satisfies the following RDE.

$$P'_{k+1} = A_kP'_kA_k^{\mathrm{T}} - A_kP'_kC_k^{\mathrm{T}}(R_k + C_kP'_kC_k^{\mathrm{T}})^{-1}C_kP'_kA_k^{\mathrm{T}} + B_kB_k^{\mathrm{T}},$$
$$P'_0 = \Pi \tag{27}$$

It follows from the above observation that the H_∞ filter is a modified version of the Kalman filter by using the parameter γ. In fact, we easily see from (7), (9)–(12) and (24)–(27) that the H_∞ filter converges to the Kalman filter as γ goes to infinity. Hence, it is very important to compare the performances of these filters when w_k, v_k and x_0 are given by Gaussian white noise processes. As a measure to compare the performances of the H_∞ and Kalman filters, we adopt the estimation error covariance matrix which is a standard performance index in the stochastic setting. Hereafter, we assume that there exists a positive definite solution P_k to the RDE (7) satisfying $V_k > 0$ for all $k \in [0, N]$.

Theorem 2. *Suppose that w_k, v_k are zero mean Gaussian white noises with unit covariance matrices, and that x_0 is the Gaussian random vector with mean $\bar{x}_0$ and covariance Π. Define*

$$\Lambda_k = E\{(x_k - \hat{x}_{k/k-1})(x_k - \hat{x}_{k/k-1})^{\mathrm{T}}\}$$

for the H_∞ filter of (9)–(12). Then $P_k \geq \Lambda_k \geq P'_k$ holds for all $k \in [0, N]$.

Proof. We define

$$F_k = A_kK_k = A_kP_kC_k^{\mathrm{T}}(R_k + C_kP_kC_k^{\mathrm{T}})^{-1}$$
$$F'_k = A_kK'_k = A_kP'_kC_k^{\mathrm{T}}(R_k + C_kP'_kC_k^{\mathrm{T}})^{-1}$$

From (1)–(3), (9) and (10), the dynamics of the estimation error $\tilde{x}_k := x_k - \hat{x}_{k/k-1}$ is described by

$$\tilde{x}_{k+1} = (A_k - F_kC_k)\tilde{x}_k + B_kw_k - F_kD_kv_k, \qquad \tilde{x}_0 = x_0 - \bar{x}_0$$

It follows that

$$\Lambda_{k+1} = (A_k - F_k C_k)\Lambda_k (A_k - F_k C_k)^{\mathrm{T}} + F_k R_k F_k^{\mathrm{T}} + B_k B_k^{\mathrm{T}},$$
$$\Lambda_0 = \Pi \tag{28}$$

Also, after some simple calculations, the RDE (7) reduces to

$$P_{k+1} = (A_k - F_k C_k) P_k (A_k - F_k C_k)^{\mathrm{T}} + F_k R_k F_k^{\mathrm{T}}$$
$$+ B_k B_k^{\mathrm{T}} + A_k \bar{P}_k L_k^{\mathrm{T}} V_k^{-1} L_k \bar{P}_k A_k^{\mathrm{T}}, \qquad P_0 = \Pi \tag{29}$$

where $\bar{P}_k := P_k(I + C_k^{\mathrm{T}} R_k^{-1} C_k P_k)^{-1} \geq 0$. Subtracting (28) from (29) yields

$$P_{k+1} - \Lambda_{k+1} = (A_k - F_k C_k)(P_k - \Lambda_k)(A_k - F_k C_k)^{\mathrm{T}}$$
$$+ A_k \bar{P}_k L_k^{\mathrm{T}} (\gamma^2 I - L_k \bar{P}_k L_k^{\mathrm{T}})^{-1} L_k \bar{P}_k A_k^{\mathrm{T}},$$
$$P_0 - \Lambda_0 = 0$$

Since $V_k = \gamma^2 I - L_k \bar{P}_k L_k^{\mathrm{T}} > 0$ holds for all $k \in [0, N]$, we get $P_k - \Lambda_k \geq 0$ for all $k \in [0, N]$ by induction.

Next we prove $\Lambda_k \geq P_k'$. It is easily verified that

$$P_{k+1}' = (A_k - F_k' C_k) P_k' (A_k - F_k' C_k)^{\mathrm{T}} + F_k' R_k F_k'^{\mathrm{T}} + B_k B_k^{\mathrm{T}}$$
$$= (A_k - F_k C_k) P_k' (A_k - F_k C_k)^{\mathrm{T}} + B_k B_k^{\mathrm{T}} + F_k R_k F_k^{\mathrm{T}}$$
$$- (F_k - F_k') R_k (F_k - F_k')^{\mathrm{T}}, \qquad P_0' = \Pi \tag{30}$$

Subtracting (28) from this yields

$$P_{k+1}' - \Lambda_{k+1} = (A_k - F_k C_k)(P_k' - \Lambda_k)(A_k - F_k C_k)^{\mathrm{T}}$$
$$- (F_k - F_k') R_k (F_k - F_k')^{\mathrm{T}}$$

where $P_0' - \Lambda_0 = 0$. Since $R_k > 0$, we get $P_k' - \Lambda_k \leq 0$ for all $k \in [0, N]$ by induction. ■

Lemma 1. *We assume that $P, P' \in \mathbf{R}^{n \times n}$ and $R \in \mathbf{R}^{q \times q}$ are positive definite symmetric matrices. If $P \geq P' > 0$ holds, then we have*

$$P(I + C^{\mathrm{T}} R^{-1} C P)^{-1} \geq P'(I + C^{\mathrm{T}} R^{-1} C P')^{-1} \geq 0$$

Proof. We define $\bar{P} = P(I + C^{\mathrm{T}} R^{-1} C P)^{-1}$ and $\bar{P}' = P'(I + C^{\mathrm{T}} R^{-1} C P')^{-1}$. It is easily seen that

$$\bar{P} = (I - KC) P (I - KC)^{\mathrm{T}} + KRK^{\mathrm{T}} \geq 0 \tag{31}$$

$$\bar{P}' = (I - K'C) P' (I - K'C)^{\mathrm{T}} + K'RK'^{\mathrm{T}} \geq 0 \tag{32}$$

where $K = PC^{\mathrm{T}}(R + CPC^{\mathrm{T}})^{-1}$ and $K' = P'C^{\mathrm{T}}(R + CP'C^{\mathrm{T}})^{-1}$. We also rewrite (32) as

$$\begin{aligned}\bar{P}' &= (I - KC)P'(I - KC)^{\mathrm{T}} + KRK^{\mathrm{T}} \\ &\quad - (K' - K)(R + CP'C^{\mathrm{T}})(K' - K)^{\mathrm{T}} \end{aligned} \tag{33}$$

Subtracting this from (31) yields

$$\begin{aligned}\bar{P} - \bar{P}' &= (I - KC)(P - P')(I - KC)^{\mathrm{T}} \\ &\quad + (K' - K)(R + CP'C^{\mathrm{T}})(K' - K)^{\mathrm{T}} \end{aligned} \tag{34}$$

The right-hand side of the above equation is positive semidefinite since $P \geq P' > 0$ and $R > 0$. This completes the proof. ■

We now define

$$\bar{P}_k = P_k(I + C_k^{\mathrm{T}} R_k^{-1} C_k P_k)^{-1}$$
$$\bar{P}'_k = P'_k(I + C_k^{\mathrm{T}} R_k^{-1} C_k P'_k)^{-1}$$

where P_k and P'_k are the positive definite solutions to the RDEs (6) and (27), respectively. Then, the gain matrices K_k and K'_k can be expressed as

$$K_k = \bar{P}_k C_k^{\mathrm{T}} R_k^{-1}, \qquad K'_k = \bar{P}'_k C_k^{\mathrm{T}} R_k^{-1}$$

Thus, from Theorem 2 and Lemma 1, we get $\|K_k\|_F \geq \|K'_k\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm, i.e., $\|M\|_F = \sqrt{\mathrm{Tr}(M^{\mathrm{T}} M)}$. This implies that the H_∞ filter is more sensitive to $y_k - C_k \hat{x}_{k/k-1}$ than the Kalman filter. In the case where the measurement noise v_k is small, the estimate by the H_∞ filter will converge to the neighborhood of the actual state faster than the Kalman filter. On the other hand, when the measurement noise v_k is large, the estimate of the H_∞ filter is degraded more easily than that of the Kalman filter.

By defining $\omega_k = B_k w_k$, $\nu_k = D_k v_k$, and $Q_k = B_k B_k^{\mathrm{T}}$, we see that the least squares filtering problem for the system (1)–(3) where w_k and v_k are zero mean Gaussian white noise processes with unit covariances is equivalent to the least squares filtering problem for the system:

$$x_{k+1} = A_k x_k + \omega_k \tag{35}$$

$$y_k = C_k x_k + \nu_k \tag{36}$$

where ω_k and ν_k are zero mean Gaussian white noise processes satisfying

$$E\left\{\begin{bmatrix}\omega_k \\ \nu_k\end{bmatrix}\begin{bmatrix}\omega_\tau^{\mathrm{T}} & \nu_\tau^{\mathrm{T}}\end{bmatrix}\right\} = \begin{bmatrix} Q_k & 0 \\ 0 & R_k \end{bmatrix}\delta_{k\tau} \tag{37}$$

It is also straightforward to show that the H_∞ RDE (7) is expressed as

$$P_{k+1} = A_k P_k A_k^{\mathrm{T}} - A_k P_k C_k^{\mathrm{T}}(R_k + C_k P_k C_k^{\mathrm{T}})^{-1} C_k P_k A_k^{\mathrm{T}} + \bar{Q}_k \tag{38}$$
$$\bar{Q}_k = B_k B_k^{\mathrm{T}} + A_k \bar{P}_k L_k^{\mathrm{T}} V_k^{-1} L_k \bar{P}_k A_k^{\mathrm{T}} \tag{39}$$

Since V_k is positive definite for any k, comparison of the above equation with the Kalman filter RDE (27) shows that the H_∞ filter is optimal in the least squares sense for the system (35), (36) where ω_k and ν_k are mutually uncorrelated zero mean Gaussian white noise processes such that $E\{\omega_k \omega_\tau^{\mathrm{T}}\} = \bar{Q}_k \delta_{k\tau}$ and $E\{\nu_k \nu_\tau^{\mathrm{T}}\} = R_k \delta_{k\tau}$. It may also be noted that $Q_k \le \bar{Q}_k$ holds for all k, since $V_k > 0$. We see from the above discussion that the H_∞ filter is a robust filtering algorithm for the uncertainty in the covariance of the process disturbance ω_k. Let Q_k^{nom}, $\bar{Q}_k^{\mathrm{nom}}$, and Λ_k^{nom} be the nominal values of Q_k, $\bar{Q}_k$, and Λ_k, respectively. Then, we have

$$\Lambda_{k+1}^{\mathrm{nom}} = (A_k - F_k C_k)\Lambda_k^{\mathrm{nom}}(A_k - F_k C_k)^{\mathrm{T}} + F_k R_k F_k^{\mathrm{T}} + Q_k^{\mathrm{nom}},$$
$$\Lambda_0^{\mathrm{nom}} = \Pi \tag{40}$$
$$\Lambda_{k+1} = (A_k - F_k C_k)\Lambda_k(A_k - F_k C_k)^{\mathrm{T}} + F_k R_k F_k^{\mathrm{T}} + Q_k,$$
$$\Lambda_0 = \Pi \tag{41}$$

Subtracting (40) from (41) yields

$$\Lambda_{k+1}^{\mathrm{nom}} - \Lambda_{k+1} = (A_k - F_k C_k)(\Lambda_k^{\mathrm{nom}} - \Lambda_k)(A_k - F_k C_k)^{\mathrm{T}}$$
$$+ Q_k^{\mathrm{nom}} - Q_k, \qquad \Lambda_0^{\mathrm{nom}} - \Lambda_0 = 0$$

Thus, $\Lambda_k \le \Lambda_k^{\mathrm{nom}}$ holds if $Q_k \le Q_k^{\mathrm{nom}}$ for all k. This implies that the estimation error variance $\mathrm{Tr}\,\Lambda_k$ is not larger than the nominal performance $\mathrm{Tr}\,\Lambda_k^{\mathrm{nom}}$ when the actual covariance of ω_k is smaller than the nominal value. On the other hand, when Q_k is larger than Q_k^{nom}, we have $\Lambda_k \ge \Lambda_k^{\mathrm{nom}}$. However, similarly to the above discussion, we obtain $\Lambda_k \le P_k$ if $Q_k \le \bar{Q}_k$ for all k. Hence, $\mathrm{Tr}\,P_k$ provides an upper bound on the estimation error variance $\mathrm{Tr}\,\Lambda_k$ for the uncertain disturbance covariance $Q_k \in \{Q_k \mid Q_k \le \bar{Q}_k^{\mathrm{nom}}\}$. Further, if Q_k is sufficiently close to $\bar{Q}_k^{\mathrm{nom}}$, then the estimation error variance $\mathrm{Tr}\,\Lambda_k$ remains in the neighborhood of the optimum $\mathrm{Tr}\,P_k$.

C. Relationship Between γ and H_∞ RDE

In Section II, it was shown that the H_∞ filter is given in terms of the positive definite solution to the H_∞ RDE (7), which depends on the prescribed design parameter γ. Therefore, in order to study the performance of the H_∞ filter, it is important to investigate the behavior of the Riccati solution P_k as γ changes.

We first introduce the following lemma.

Lemma 2. *Define*

$$\psi(P,\gamma) = P + PL^{\mathrm{T}}(\gamma^2 I - LPL^{\mathrm{T}})^{-1}LP$$

for a given $p \times n$ matrix L. Assume that $P^{(1)} \geq P^{(2)} \geq 0$ and $\gamma^2 I - LP^{(1)}L^{\mathrm{T}} > 0$ hold for given $n \times n$ symmetric matrices $P^{(1)}$ and $P^{(2)}$. Then we have

$$\psi(P^{(1)},\gamma) \geq \psi(P^{(2)},\gamma) \tag{42}$$

Proof. We define

$$M^{(i)} = \gamma P^{(i)} L^{\mathrm{T}}(\gamma^2 I - LP^{(i)}L^{\mathrm{T}})^{-1}, \quad i = 1,2$$

$$\varphi(P,M) = \left(I + \frac{ML}{\gamma}\right) P \left(I + \frac{ML}{\gamma}\right)^{\mathrm{T}} - MM^{\mathrm{T}}$$

It is clear that $\psi(P^{(i)},\gamma) = \varphi(P^{(i)},M^{(i)})$.

For any matrix $M \in \mathbf{R}^{n \times p}$, we get

$$\begin{aligned}\varphi(P^{(i)},M^{(i)}) = {}& \frac{1}{\gamma^2}(M^{(i)} - M)(\gamma^2 I - LP^{(i)}L^{\mathrm{T}})(M^{(i)} - M)^{\mathrm{T}} \\ & + \varphi(P^{(i)},M), \quad i = 1,2\end{aligned}$$

It follows that

$$\begin{aligned}\psi(P^{(1)},\gamma) - \psi(P^{(2)},\gamma) &= \varphi(P^{(1)},M^{(1)}) - \varphi(P^{(2)},M^{(2)}) \\ &= \frac{1}{\gamma^2}(M^{(1)} - M^{(2)})(\gamma^2 I - LP^{(1)}L^{\mathrm{T}}) \\ &\quad \times (M^{(1)} - M^{(2)})^{\mathrm{T}} + \varphi(P^{(1)},M^{(2)}) \\ &\quad - \varphi(P^{(2)},M^{(2)}) \\ &= \left(I + \frac{M^{(2)}L}{\gamma}\right)(P^{(1)} - P^{(2)})\left(I + \frac{M^{(2)}L}{\gamma}\right)^{\mathrm{T}} \\ &\quad + \frac{1}{\gamma^2}(M^{(1)} - M^{(2)})(\gamma^2 I - LP^{(1)}L^{\mathrm{T}}) \\ &\quad \times (M^{(1)} - M^{(2)})^{\mathrm{T}}\end{aligned}$$

Since $P^{(1)} \geq P^{(2)}$ and $\gamma^2 I - LP^{(1)}L^{\mathrm{T}} > 0$, we obtain $\psi(P^{(1)},\gamma) \geq \psi(P^{(2)},\gamma)$. ∎

Let $P_k^{(i)}$ denote the solution to the RDE (7) for given γ_i $(i = 1,2)$. Then, $P_k^{(i)}$ satisfies

$$P_{k+1}^{(i)} = A_k P_k^{(i)} (\Sigma_k^{(i)})^{-1} A_k^{\mathrm{T}} + B_k B_k^{\mathrm{T}}, \qquad P_0^{(i)} = \Pi \tag{43a}$$

$$\Sigma_k^{(i)} = I + (C_k^{\mathrm{T}} R_k^{-1} C_k - \gamma^{-2} L_k^{\mathrm{T}} L_k) P_k^{(i)}, \quad i = 1,2 \tag{43b}$$

We also define

$$\bar{P}_k^{(i)} = P_k^{(i)}(I + C_k^{\mathrm{T}} R_k^{-1} C_k P_k^{(i)})^{-1}$$

Theorem 3. *Suppose that* $\gamma_1 \leq \gamma_2$, *and that the RDE of (7) has positive definite solutions* $P_k^{(i)}$ $(i = 1, 2)$ *such that* $\gamma_i^2 I - L_k \bar{P}_k^{(i)} L_k^T > 0$ *for all* $k \in [0, N]$. *Then,* $P_k^{(1)} \geq P_k^{(2)}$ *and* $\bar{P}_k^{(1)} \geq \bar{P}_k^{(2)}$ *holds for all* $k \in [0, N]$.

Proof. Since $P_k^{(1)} \geq P_k^{(2)}$ implies $\bar{P}_k^{(1)} \geq \bar{P}_k^{(2)}$ by Lemma 1, it suffices to prove $P_k^{(1)} \geq P_k^{(2)}$.

For $k = 0$, it is obvious that $P_0^{(1)} = P_0^{(2)} = \Pi$.

We assume that $P_k^{(1)} \geq P_k^{(2)}$ holds for $k = 0, 1, \ldots, t$. Since $P_k^{(i)} \geq 0$, $\bar{P}_k^{(i)}$ $(i = 1, 2)$ are well defined and positive definite. We see from Lemma 1 that $\bar{P}_k^{(1)} \geq \bar{P}_k^{(2)}$. It thus follows from Lemma 2 that

$$\psi_t(\bar{P}_t^{(1)}, \gamma_1) \geq \psi_1(\bar{P}_t^{(1)}, \gamma_2) \geq \psi_t(\bar{P}_t^{(2)}, \gamma_2) \tag{44}$$

where

$$\psi_k(P, \gamma) = P + PL_k^T(\gamma^2 I - L_k P L_k^T)^{-1} L_k P$$

By simple calculations, (43) reduces to

$$P_{t+1}^{(i)} = A_t \psi_t(\bar{P}_t^{(i)}, \gamma_i) A_t^T + B_t B_t^T \tag{45}$$

Thus, we get

$$P_{t+1}^{(1)} - P_{t+1}^{(2)} = A_t\{\psi_t(\bar{P}_t^{(1)}, \gamma_1) - \psi_t(\bar{P}_t^{(2)}, \gamma_2)\} A_t^T \geq 0 \tag{46}$$

As a result, we have shown by induction that $P_k^{(1)} \geq P_k^{(2)}$ holds for all $k \in [0, N]$. ■

The above theorem shows that the solution to the RDE (7) is monotonically nonincreasing with respect to the parameter γ. By a discussion similar to the previous subsection, as γ gets larger, P_k becomes smaller, while the H_∞ filter gets less sensitive to $\bar{y}_k := y_k - C_k \hat{x}_{k/k-1}$. As a result, the parameter γ represents a trade-off between the mean square error and the sensitivity to the estimation error $\bar{y}_k$.

IV. NUMERICAL EXAMPLE

Let us consider the system given by

$$x_{k+1} = \begin{bmatrix} 1 & 0.05 \\ 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w_k, \qquad x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{47}$$

$$y_k = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} x_k + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} v_k \tag{48}$$

$$z_k = \quad [1 \quad 0] x_k \tag{49}$$

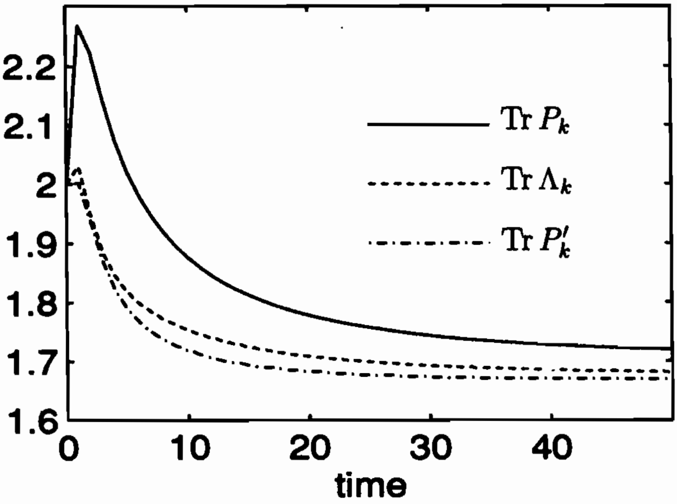


Figure 1 Relationship among P_k, Λ_k, and P'_k.

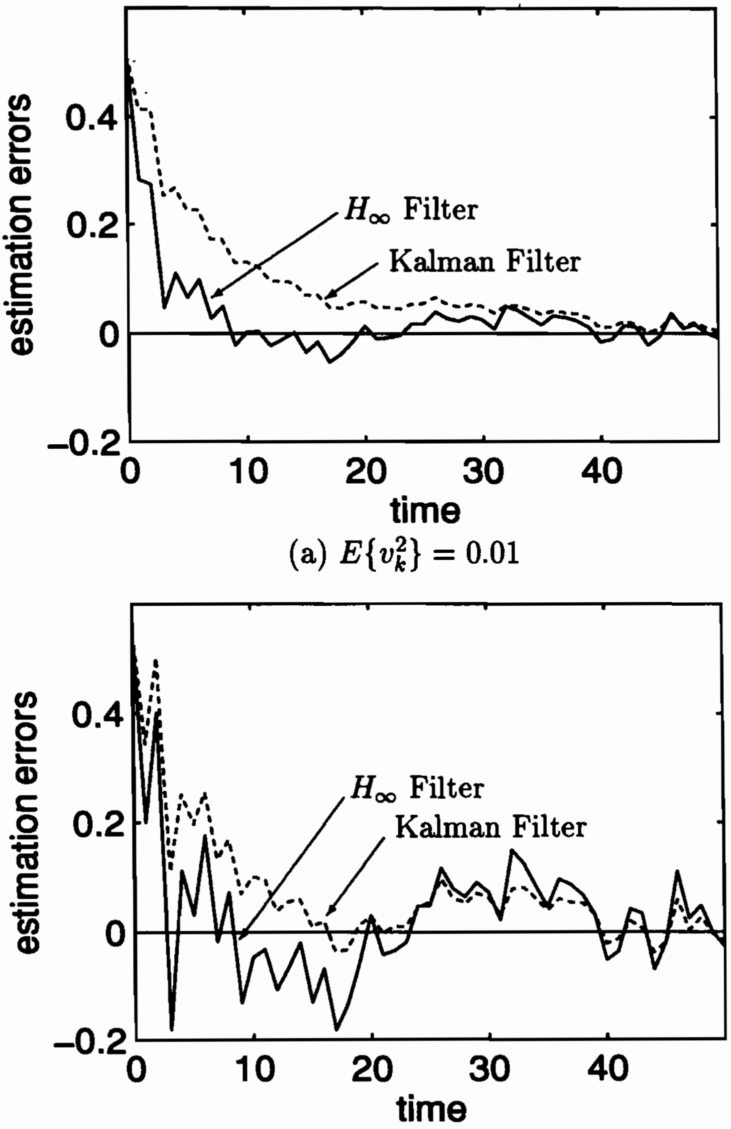
Hereafter, we assume $N = 50$ and $\Pi = I$. In this case, the optimal value of the parameter γ is $\gamma_{\text{opt}} = 1.000$. Figure 1 shows the relationship among P_k, P'_k, and Λ_k for $\gamma = 1.2$. We see from the figure that $\operatorname{Tr} P_k \geq \operatorname{Tr} \Lambda_k \geq \operatorname{Tr} P'_k$ holds for all k since $P_k \geq \Lambda_k \geq P'_k$ by Theorem 2.

Furthermore, in order to compare the performance of the H_∞ filter with that of the Kalman filter, we performed simulations (Fig. 2), in which we set $w_k \equiv 0$ to see the sensitivity of the filters to $\bar{y}_k = y_k - C_k \hat{x}_{k/k-1}$. The measurement noise v_k is zero mean Gaussian white noise. Figures 2a and b show the estimation errors $z_k - \hat{z}_k$ for $E\{v_k^2\} = 0.01$ and $E\{v_k^2\} = 0.1$, respectively. In the case when the measurement noise v_k is small (Fig. 2a), the H_∞ filter presents a much better transient response than the Kalman filter. On the other hand, when v_k is large (Fig. 2b), the H_∞ filter is more sensitive to the measurement noise than the Kalman filter.

Next, we consider the performance of the H_∞ and Kalman filters under uncertain process disturbance. We rewrite the state space equation (47) as

$$x_{k+1} = \begin{bmatrix} 1 & 0.05 \\ 0 & 1 \end{bmatrix} x_k + \omega_k, \qquad x_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{50}$$

where ω_k is a zero mean Gaussian white noise process with covariance



(a) $E\{v_k^2\} = 0.01$

(b) $E\{v_k^2\} = 0.1$

Figure 2 Comparison of H_∞ filter and Kalman filter.

Table 1 Error variances for various disturbance covariances.

		$E\{\omega_k \omega_k^T\} = Q_k$					
		$\frac{2Q_k^{\text{nom}}}{3}$	Q_k^{nom}	$\frac{2Q_k^{\text{nom}} + \bar{Q}_k^{\text{nom}}}{3}$	$\frac{Q_k^{\text{nom}} + 2\bar{Q}_k^{\text{nom}}}{3}$	$\bar{Q}_k^{\text{nom}}$	$\frac{4\bar{Q}_k^{\text{nom}}}{3}$
	nominal H_∞	1.291	1.682	1.694	1.707	1.720	1.732
$E\{\Vert\tilde{x}_k\Vert^2\}$	nominal Kalman	1.279	1.670	1.696	1.723	1.749	1.776
$(k = 50)$	optimal values	1.267	1.670	1.690	1.706	1.720	1.732

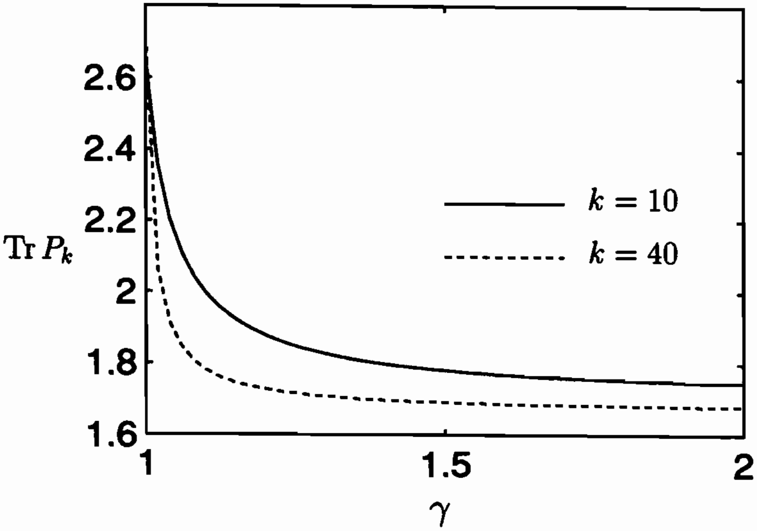


Figure 3 Relationship between γ and P_k.

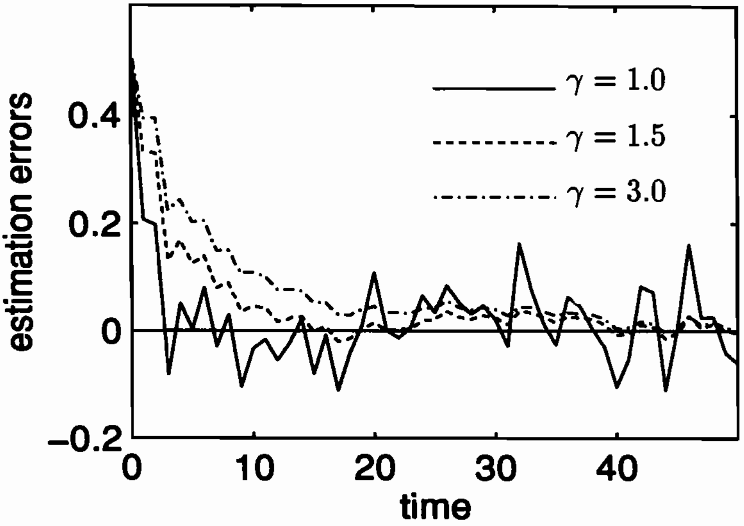


Figure 4 Estimation errors for various values of γ.

Q_k. Notice from (47) that the nominal value of Q_k is given by $Q_k^{\text{nom}} = \text{diag}[0 \quad 1]$. Moreover, we obtain $\min_{k\in[0,N]} \| \bar{Q}_k^{\text{nom}} - Q_k^{\text{nom}} \| = 0.007$ and $\max_{k\in[0,N]} \| \bar{Q}_k^{\text{nom}} - Q_k^{\text{nom}} \| = 0.266$, where $\bar{Q}_k^{\text{nom}}$ is the nominal value of $\bar{Q}_k$ in (39). Table 1 shows the estimation error variance $E\{\|\tilde{x}_k\|^2\}$ at $k = 50$ for a variety of Q_ks. When the disturbance covariance Q_k is larger than the nominal value, the estimation error variance of the (nominal) H_∞ filter is smaller than that of the nominal Kalman filter. Hence, we see from the table that the robustness of the H_∞ filter for large process disturbance is stronger than that of the Kalman filter.

The relationship between γ and $\text{Tr}\, P_k = 10, 40$ is illustrated in Fig. 3. In the figure, we see that P_{10}, P_{40} are monotonically nonincreasing with respect to γ. We also obtained similar results for other ks.

Simulation results for various γ are shown in Fig. 4, for which the measurement noise is the zero mean white noise with $E\{v_k^2\} = 0.01$. As shown in the figure, the H_∞ filter becomes less sensitive to $\tilde{y}_k$ as γ becomes large.

V. CONCLUSION

In this paper, we have derived some results on the performance of the H_∞ filter as applied to a stochastic system subject to zero mean Gaussian white noise disturbances. First, it is shown that the H_∞ filter is optimal in the sense of the exponential quadratic error criterion. Then, by using Riccati difference equations, we have compared the performances of the H_∞ filter and the Kalman filter in the above stochastic setting. The relationship between the prescribed H_∞ bound γ and the performance of the H_∞ filter is investigated based on the monotonicity property of the solution to the H_∞ RDE. The above results will provide a guideline for determining the value of the design parameter γ.

Although we restricted our discussions to the finite horizon problem in this paper, the results obtained here also apply to the infinite horizon H_∞ filtering problem for a time-invariant system by replacing the RDEs by the corresponding algebraic Riccati equations.

REFERENCES

1. U Shaked, Y Theodor. Proc 31st IEEE Conf Decision and Control, 1992, pp 2278–2286.
2. K Nagpal, PP Khargonekar. IEEE Trans Automatic Control AC-36:152–166, 1991.

3. WM Haddad, D Bernstein, M Mustafa. Systems and Control Letters 16:235–247, 1991.
4. I Yaesh, U Shaked. IEEE Trans Automatic Control AC-26:1264–1271, 1991.
5. I Yaesh, U Shaked. Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing (Proc MTNS '91), vol. I, 1991, pp 261–267.
6. I Yaesh, U Shaked. Int J Control 55:1443–1452, 1992.
7. M Fujita, A Maruyama, T Taniguchi, K Uchida. Proc 32nd IEEE Conf Decision and Control, 1993, pp 2194–2196.
8. K Takaba, T Katayama. Trans Inst Syst Contr Inform Eng 7:322–331, 1994 (in Japanese); also K Takaba. Studies on H_∞ Filtering Problems for Linear Discrete-Time Systems. Chapter 5, Dr Thesis, Kyoto University, 1996.
9. J Speyer, F Chih-hai, R Banavar. Proc 31st IEEE Conf Decision and Control, 1992, pp 2293–2298.
10. RS Mangoubi, BD Appleby, GC Verghese. Proc 33rd IEEE Conf Decision and Control, 1994, pp 3943–3948.
11. P Whittle. Risk-Sensitive Optimal Control. New York: Wiley, 1990.
12. AE Bryson, YC Ho. Applied Optimal Control. New York: Wiley, 1973.

9

Reduced Order Functional Estimator for Linear Stochastic Systems

Takayoshi Nakamizo

Ashikaga Institute of Technology, Tochigi, Japan

I. INTRODUCTION

It is well known from the Kalman filtering theory that the order of the optimal filter for the stochastic system is the same as the order of the system. In many practical situations, however, where we need not know all the state variables, but we may be interested in knowing only a few important state variables, a reduced-order filter would be preferable to the standard Kalman filter.

The standard problem of reduced-order filtering has been extensively reported for the case where some of the observations are assumed to be noise free [1–3]. In this chapter, unlike the previous studies, we consider the case where none of noise-free observation is assumed. This chapter deals with the design problem of a reduced-order filter which estimates a specific linear function of the state for the linear stochastic system [4–6]. It can be shown that this problem can be reformulated as a standard Kalman filtering problem for the reduced-order system obtained through an appropriate system transformation. Thus a method for designing a reduced-order filter offers the possibility of significant reduction in computational requirement and less complexity in physical implementation. However, the price to pay for these benefits is some loss of performance compared with the full-order Kalman filter. Finally the relation to the full-order Kalman filter is also carefully discussed in this chapter.

The structure of this chapter is as follows. The problem statement is given in Section II. The structure of a reduced-order filter is assumed. The unbiasness is first required, and then the remaining design freedom can be used to minimize the performance measure. An appropriate similar transformation of the system is proposed in Section III by means of which a very simple design procedure can be driven. It is shown that the estimation problem can be reformulated as a problem of a standard Kalman filter for the reduced-order stochastic system obtained through a system transformation. The proposed design procedure is completely parallel to the standard Kalman filter design. Furthermore the relationship between the full-order Kalman filter and the proposed reduced-order filter is discussed in Sections IV and V. This is another contribution of this article. In Section VI are presented simple examples to illustrate the design procedure of the reduced-order filter. The concluding remarks are given in Section VII.

II. PROBLEM STATEMENT

Consider the linear stochastic system described by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bw(t) \\ y(t) &= Cx(t) + v(t) \end{aligned} \tag{1}$$

where the state $x(t)$ is an n vector and the output $y(t)$ is an m vector; A, B, and C are constant matrices of appropriate dimensions. The plant disturbance $w(t)$ and the measurement noise $v(t)$ are zero mean white noise processes with

$$\begin{aligned} E\{w(t)w^T(\tau)\} &= W\delta(t-\tau) \\ E\{v(t)v^T(\tau)\} &= V\delta(t-\tau) \end{aligned} \tag{2}$$

Furthermore, for simplicity, $w(t)$ and $v(t)$ are assumed to be mutually independent. The problem of interest is to estimate a specific linear function of the state vector;

$$z_1(t) = Kx(t) \tag{3}$$

where $z_1(t)$ is of dimension $k \leqq n$. Define a reduced-order estimator as

$$\dot{\hat{z}}_1(t) = F\hat{z}_1(t) + Gy(t), \tag{4}$$

where $\hat{z}_1(t)$ is the estimate of $z_1(t)$. The matrices F and G are to be determined so that $\hat{z}_1(t)$ is unbiased, namely

$$E\{z_1(t) - \hat{z}_1(t)\} = 0 \tag{5}$$

and the estimation error criterion

$$J = E\{(z_1 - \hat{z}_1)^T(z_1 - \hat{z}_1)\} \tag{6}$$

is minimized.

III. REDUCED-ORDER FUNCTIONAL ESTIMATOR

A. System Transformation

It is assumed, without any great loss of generality, that the rank condition must hold;

$$\operatorname{rank}\begin{bmatrix} K \\ C \end{bmatrix} = n \tag{7}$$

By this is meant that $k + m \geqq n$ and the matrix C is assumed to have at least $n - k$ rows which are independent of the matrix K. The case where the above rank requirement does not hold will be briefly discussed at the end of this section.

First we pick any $\bar{m} = n - k$ independent rows among those of the matrix C, which are independent of the matrix K, and denote them $\bar{C}$. Introduce the nonsingular matrix defined by

$$T = \begin{bmatrix} K \\ \bar{C} \end{bmatrix} \tag{8}$$

by which the state vector x is transformed into $z = Tx$. Define also

$$T^{-1} = \begin{bmatrix} K \\ \bar{C} \end{bmatrix}^{-1} = (N, M)$$

Then

$$\begin{aligned} &KN = I_n, \qquad KM = 0 \\ &\bar{C}N = 0, \qquad \bar{C}M = I_m \\ &NK + M\bar{C} = I_{n+m} \end{aligned}$$

It follows that

$$TAT^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} KAN & KAM \\ CAN & CAM \end{bmatrix}$$

$$TB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} KB \\ CB \end{bmatrix}$$

The observation matrix C can be rearranged as

$$C = \begin{bmatrix} \tilde{C} \\ \bar{C} \end{bmatrix} \begin{matrix} \tilde{m} \\ \bar{m} \end{matrix}, \qquad m = \tilde{m} + \bar{m}$$

Thus we can write

$$y = \begin{bmatrix} \tilde{C} \\ \bar{C} \end{bmatrix} x + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

and the transformed matrix becomes

$$CT^{-1} = \begin{bmatrix} \tilde{C}N & \tilde{C}M \\ \bar{C}N & \bar{C}M \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ 0 & I_m \end{bmatrix}$$

Therefore the transformed system can be described by

$$\begin{cases} \dot{z}_1 = A_{11}z_1 + A_{12}z_2 + B_1 w \\ \dot{z}_2 = A_{21}z_1 + A_{22}z_2 + B_2 w \end{cases} \tag{9}$$

$$\begin{cases} y_1 = C_{11}z_1 + C_{12}z_2 + v_1 \\ y_2 = z_2 + v_2 \end{cases} \tag{10}$$

B. Problem Solution

The design procedure of the reduced-order filter is as follows:

1. Derive the transformed system described as (9) and (10).
2. Since, from $(10)_2$, $z_2 = y_2 - v_2$, we obtain

$$\begin{aligned} \dot{z}_1 &= A_{11}z_1 + A_{12}y_2 + B_1 w - A_{12}v_2 \\ y_1 &= C_{11}z_1 + C_{12}y_2 + v_1 - C_{12}v_2 \end{aligned} \tag{11}$$

3. To the system described by (11), the corresponding Kalman filter can be easily derived as

$$\begin{aligned} \dot{\hat{z}} &= A_{11}\hat{z}_1 + A_{12}y_2 + G_1(y_1 - C_{11}\hat{z}_1 - C_{12}y_1) \\ G_1 &= (PC_{11}^T - A_{12}R_{12}^T)R_{11}^{-1} \\ \dot{P} &= A_{11}P + PA_{11}^T - (PC_{11}^T - A_{12}R_{12}^T)R_{11}^{-1}(PC_{11} - A_{12}R_{12}^T) \\ &\quad + B_1 W B_1^T + A_{12}R_{22}A_{12}^T \end{aligned} \tag{12}$$

where

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix} = \begin{bmatrix} I & -C_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -C_{12}^T & I \end{bmatrix} \tag{13}$$

and

$$E\{v_i(t)v_j^T(\tau)\} = V_{ij}\delta(t - \tau), \quad i, j = 1, 2$$

In particular, if $C_{12} = 0$, then $R_{ij} = V_{ij}$.

The proof will be given in the next section. Eq. $(12)_3$ can also be rewritten as

$$\begin{aligned}\dot{P} = (A_{11} + A_{12}R_{12}^T R_{11}^{-1} C_{11})P + P(A_{11} + A_{12}R_{12}^T R_{11}^{-1} C_{11})^T \\ - PC_{11}^T R_{11}^{-1} C_{11} P + B_1 W B_1^T + A_{12}(R_{22} - R_{12}^T R_{11}^{-1} R_{12})A_{12}^T\end{aligned}$$

C. Derivation of Design Procedure

It will be shown in this section that the result given by (12) gives the optimal solution of the problem stated in Section II. Define the dynamics of a reduced-order filter by

$$\dot{\hat{z}}_1 = F\hat{z}_1 + G_1 y_1 + G_2 y_2 \tag{14}$$

where F, G_1, and G_2 are design parameters to be determined.

1. Error Equation

The error $e(t)$ is defined as

$$e(t) = z_1(t) - \hat{z}_1(t)$$

Then the estimation error equation is readily given by

$$\begin{aligned}\dot{e} = Fe + (A_{11} - F - G_1 C_{11})z_1 + (A_{12} - G_2 - G_1 C_{12})z_2 \\ + B_1 w - G_1 v_1 - G_2 v_2\end{aligned} \tag{15}$$

2. Requirement of Unbiasness

Assume that the matrix F is selected to be stable. The requirement of unbiasness of (5) is satisfied if

$$\begin{aligned}A_{11} - F - G_1 C_{11} = 0 \\ A_{12} - G_2 - G_1 C_{12} = 0\end{aligned} \tag{16}$$

from which we can obtain in terms of G_1

$$\begin{aligned}F = A_{11} - G_1 C_{11} \\ G_2 = A_{12} - G_1 C_{12}\end{aligned}$$

3. Error Covariance Equation

Substituting (16) into (15) yields

$$\dot{e} = (A_{11} - G_1 C_{11})e + B_1 w - A_{12} v_2 - G_1(v_1 - C_{12} v_2) \tag{17}$$

Define the error covariance matrix as

$$P(t) = E\{e(t)e^T(t)\}$$

From (17), the error covariance equation can be given by

$$\begin{aligned}\dot{P} &= (A_{11} - G_1 C_{11})P + P(A_{11} - G_1 C_{11})^T + B_1 W B_1^T \\ &\quad + G_1 R_{11} G_1^T + G_1 R_{12} A_{12}^T + A_{12} R_{12}^T G_1^T + A_{12} R_{22} A^{\mathrm{T}},\end{aligned} \tag{18}$$

where

$$\begin{aligned}R_{11} &= V_{11} - C_{12} V_{12}^T - V_{12} C_{12}^T + C_{12} V_{22} C_{12}^T \\ R_{12} &= V_{12} - C_{12} V_{22} \\ R_{22} &= V_{22}\end{aligned}$$

so that we have (13).

4. Minimization

For any time $\tau(\geqq t)$, the problem is to determine $G_1(t)$, $0 \leqq t \leqq \tau$ so that

$$J = E\{e^T(\tau) e(\tau)\} = \mathrm{tr}[P(\tau)] \tag{19}$$

is minimized by using the matrix minimum principle [7,8]. On the basis of (18), the Hamiltonian can be given by

$$\begin{aligned}H &= \mathrm{tr}[P(\tau)\Lambda^T(\tau)] \\ &= \mathrm{tr}[(A_{11}P + PA_{11}^T + B_1 W B_1^T + A_{12} R_{22} A_{12}^T)\Lambda^T] \\ &\quad + \mathrm{tr}[G_1 R_{11} G_1^T \Lambda^T] - \mathrm{tr}[G_1(C_{11}P - R_{12}A_{12}^T)\Lambda^T] \\ &\quad - \mathrm{tr}[(C_{11}P - PR_{12}A_{12}^T)^T G_1^T \Lambda^T]\end{aligned} \tag{20}$$

where Λ is the matrix of Lagrangian multipliers. According to the minimum principle,

$$\partial H / \partial G_1 = 0$$

becomes

$$\begin{aligned}&-\Lambda(C_{11}P - R_{12}A_{12}^T) - \Lambda^T(C_{11}P - R_{12}A_{12}^T)^T + \Lambda G_1 R_{11} \\ &\quad + \Lambda^T G_1 R_{11} = 0\end{aligned} \tag{21}$$

On the other hand, the Lagrangian multiplier is determined by

$$\dot{\Lambda} = -\partial H / \partial P = -\Lambda(A_{11} - G_1 C_{11}) - (A_{11} - G_1 C_{11})^T \Lambda \tag{22}$$

with the terminal condition

$$\Lambda(\tau) = \partial \mathrm{tr}[P(\tau)] / \partial P(\tau) = I$$

From (22), it can be seen that $\Lambda(t)$ is a positive symmetric matrix. Thus, from (21), the optimal value of G_1 must be given by

$$G_1 = (PC_{11}^T - A_{12} R_{12}^T) R_{11}^{-1} \tag{23}$$

It should be noted that $G_1(t)$ does not depend on time τ.

5. Solution

From the preceding derivation, the optimal parameters are obtained as

$$F = A_{11} - G_1 C_{11}$$
$$G_2 = A_{12} - G_1 C_{12}$$
$$G_1 = (PC_{11}^T - A_{12}R_{12}^T)R_{11}^{-1}$$

Therefore the reduced-order filter can be described as

$$\dot{\hat{z}}_1 = A_{11}\hat{z}_1 + A_{12}y_2 + G_1(y_1 - C_{11}\hat{z}_1 - C_{12}y_2)$$

where G_1 is given by $(12)_2$. Substituting (23) into (18) yields

$$\dot{P} = A_{11}P + PA_{11}^T - (PC_{11}^T A_{12}^T R_{12}^T)R_{11}^{-1}(PC_{11}^T - A_{12}R_{12}^{\mathrm{T}})^T$$
$$+ B_1 W B_1^T + A_{12}R_{22}A_{12}^T.$$

6. Remarks

1. The reduced-order filter obtained here may be said to be equivalent to the filter given by Nagpal, Helmick, and Sims [5]. In particular, they defined the reduced-order innovation process for the resultant filter and showed that the reduced-order innovation process possesses similar properties to that of the full-order Kalman filter. However, that must be self-evident only from the design procedure presented here.
2. When the rank requirement of (7) is not fulfilled, it is not possible to select $\bar{C}$ so that the matrix T becomes nonsingular. In this case, we could modify the linear function of the state to be estimated as

$$z_1 = \begin{bmatrix} K \\ \bar{K} \end{bmatrix} x = \tilde{K}x. \tag{24}$$

where the matrix K is to be selected such that

$$T = \begin{bmatrix} \tilde{K} \\ \bar{C} \end{bmatrix} \tag{25}$$

be nonsingular. The price to pay for this is the augmentation of the filter order.

3. Needless to say, the optimal estimate of the linear function $z_1 = Kx$ is given by $\hat{z}_1 = K\hat{x}$ in which $\hat{x}$ is generated by the full-order standard Kalman filter. The computational and implementational simplicity is achieved at the cost of somewhat poorer performance. In the following, we will investigate the relationship between the full-order optimal Kalman filter and the reduced-order filter proposed here. It will be also interesting to observe that the reduced-order filter

could be derived from the full-order Kalman filter as will be shown in Section V.

IV. REPRESENTATION OF THE FULL-ORDER KALMAN FILTER

From (9) and (10), the transformed system can be expressed as

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} w \tag{26}$$

and

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{27}$$

It follows from (26) and (27) that the standard Kalman filter can be readily given by

$$\begin{bmatrix} \dot{\hat{z}}_1 \\ \dot{\hat{z}}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \hat{z}_1 \\ \hat{z}_2 \end{bmatrix} + \Gamma \left\{ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} - \begin{bmatrix} C_{11} & C_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{z}_1 \\ \hat{z}_2 \end{bmatrix} \right\} \tag{28}$$

where the filter gain Γ is described by

$$\Gamma = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{bmatrix}^{-1}$$

A. Relationship Between R^{-1} and V^{-1}

From (13)

$$\begin{aligned} \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{bmatrix} &= \begin{bmatrix} V_{11} & V_{12} \\ V_{12}^T & V_{22} \end{bmatrix}^{-1} \\ &= \begin{bmatrix} I & 0 \\ -C_{12}^T & I \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix}^{-1} \begin{bmatrix} I & -C_{12} \\ 0 & I \end{bmatrix} \end{aligned} \tag{29}$$

Since the inversion of a partitioned matrix R can be expressed as

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{12}^T & R_{22} \end{bmatrix}^{-1} = \begin{bmatrix} R_{11}^{-1}(I + R_{12}S^{-1}R_{12}^T R_{11}^{-1}) & -R_{11}^{-1}R_{12}S^{-1} \\ -S^{-1}R_{12}^T R_{11}^{-1} & S^{-1} \end{bmatrix}$$

where

$$S = R_{22} - R_{12}^T R_{11}^{-1} R_{12},$$

we can obtain from (29)

$$\begin{aligned}
\Omega_{11} &= R_{11}^{-1}(I + R_{12}S^{-1}R_{12}^TR_{11}^{-1})\\
\Omega_{12} &= -R_{11}^{-1}(I + R_{12}S^{-1}R_{12}^TR_{11}^{-1})C_{12} - R_{11}^{-1}R_{12}S^{-1}\\
\Omega_{22} &= C_{12}^TR_{11}^{-1}(I + R_{12}S^{-1}R_{12}^TR_{11}^{-1})C_{12} + C_{12}^TR_{11}^{-1}R_{12}S^{-1}\\
&\quad + S^{-1}R_{12}^TR_{11}^{-1}C_{12} + S^{-1}
\end{aligned}$$

and also

$$\begin{aligned}
&C_{12}^T\Omega_{11} + \Omega_{12}^T = -S^{-1}R_{12}^TR_{11}^{-1}\\
&C_{12}^T\Omega_{12} + \Omega_{22} = S^{-1}(I + R_{12}^TR_{11}^{-1}C_{12})\\
&C_{12}^T\Omega_{11}C_{12} + \Omega_{12}^TC_{12} + C_{12}^T\Omega_{12} + \Omega_{22} = S^{-1}.
\end{aligned}\tag{30}$$

B. Full-Order Kalman Filter Representation

It can be easily shown from (28) that the Kalman filter can be expressed by

$$\begin{cases}
\dot{\hat{z}}_1 = A_{11}\hat{z}_1 + A_{12}\hat{z}_2 + \gamma_{11}(y_1 - C_{11}\hat{z}_1 - C_{12}\hat{z}_2) + \gamma_{12}(y_2 - \hat{z}_2)\\
\dot{\hat{z}}_2 = A_{21}\hat{z}_1 + A_{22}\hat{z}_2 + \gamma_{21}(y_1 - C_{11}\hat{z}_1 - C_{12}\hat{z}_2) + \gamma_{22}(y_2 - \hat{z}_2)
\end{cases}\tag{31}$$

where the filter gain can be given as

$$\begin{cases}
\gamma_{11} = P_{11}C_{11}^T\Omega_{11} + P_{12}(C_{12}^T\Omega_{11} + \Omega_{12}^T)\\
\gamma_{12} = P_{11}C_{11}^T\Omega_{12} + P_{12}(C_{12}^T\Omega_{12} + \Omega_{22})
\end{cases}\tag{32}$$

$$\begin{cases}
\gamma_{21} = P_{12}^TC_{11}^T\Omega_{11} + P_{22}(C_{12}^T\Omega_{11} + \Omega_{12}^T)\\
\gamma_{22} = P_{12}^TC_{11}^T\Omega_{12} + P_{22}(C_{12}^T\Omega_{12} + \Omega_{22})
\end{cases}\tag{33}$$

and the error covariance matrices P_{ij} are given by the solution of

$$\begin{aligned}
\dot{P}_{11} &= A_{11}P_{11} + P_{11}A_{11}^T + A_{12}P_{12}^T + P_{12}A_{12}^T + B_1WB_1^T - P_{12}\Omega_{22}P_{12}^T\\
&\quad - (P_{11}C_{11}^T + P_{12}C_{12}^T)\Omega_{11}(C_{11}P_{11} + C_{12}P_{12}^T)\\
&\quad - P_{12}\Omega_{12}^T(C_{11}P_{11} + C_{12}P_{12}^T) - (P_{11}C_{11}^T + P_{12}C_{12}^T)\Omega_{12}P_{12}^T
\end{aligned}$$

$$\begin{aligned}
\dot{P}_{12} &= A_{11}P_{12} + A_{12}P_{22} + P_{11}A_{21}^T + P_{12}A_{22}^T + B_1WB_2^T - P_{12}\Omega_{22}P_{22}\\
&\quad - (P_{11}C_{11}^T + P_{12}C_{12}^T)\Omega_{11}(C_{11}P_{12} + C_{12}P_{22})\\
&\quad - P_{12}\Omega_{12}^T(C_{11}P_{12} + C_{12}P_{22}) - (P_{11}C_{11}^T + P_{12}C_{12}^T)\Omega_{12}P_{22}
\end{aligned}$$

$$\begin{aligned}
\dot{P}_{22} &= A_{21}P_{12} + A_{22}P_{22} + P_{12}^TA_{21}^T + P_{22}A_{22}^T + B_2WB_2^T - P_{22}\Omega_{22}P_{22}\\
&\quad - (P_{12}^TC_{11}^T + P_{22}C_{12}^T)\Omega_{11}(C_{11}P_{12} + C_{12}P_{22})\\
&\quad - P_{22}\Omega_{12}^T(C_{11}P_{12} + C_{12}P_{22}) - (P_{12}^TC_{11}^T + P_{22} + P_{22}C_{12}^T)\Omega_{12}P_{22}
\end{aligned}$$

Making use of (30), the above equations can be written in terms of R_{ij}

$$\begin{aligned}
\dot{P}_{11} &= A_{11}P_{11} + P_{11}A_{11}^T - P_{11}C_{11}^T\Omega_{11}C_{11}P_{11} + B_1WB_1^T - P_{12}S^{-1}P_{12}^T \\
&\quad + P_{12}(A_{12}^T + S^{-1}R_{12}^TR_{11}^{-1}C_{11}P_{11}) \\
&\quad + (A_{12} + P_{11}C_{11}^TR_{11}^{-1}R_{12}S^{-1})P_{12}^T \\
\dot{P}_{12} &= A_{12}P_{12} + P_{11}A_{21}^T - P_{11}C_{11}^T\Omega_{11}C_{11}P_{12} + B_1WB_2^T - P_{12}S^{-1}P_{22} \\
&\quad + P_{12}(A_{22}^T + S^{-1}R_{12}^TR_{11}^{-1}C_{11}P_{12}) \\
&\quad + (A_{12} + P_{11}C_{11}^TR_{11}^{-1}R_{12}S^{-1})P_{22} \\
\dot{P}_{22} &= A_{21}P_{12} + P_{12}A_{21}^T - P_{12}^TC_{11}^T\Omega_{11}C_{11}P_{12} + B_2WB_2^T - P_{22}S^{-1}P_{22} \\
&\quad + P_{22}(A_{22}^T + S^{-1}R_{12}^TR_{11}^{-1}C_{11}P_{12}) \\
&\quad + (A_{22} + P_{12}^TC_{11}^TR_{11}^{-1}R_{12}S^{-1})P_{22} \qquad (34)
\end{aligned}$$

V. DERIVATION OF REDUCED-ORDER FILTER FROM FULL-ORDER KALMAN FILTER

A. Derivation 1

The dynamic equation of the full-order filter becomes from (31)

$$\dot{\hat{z}}_1 = (A_{11} - \gamma_{11}C_{11})\hat{z}_1 + (A_{12} - \gamma_{11}C_{12} - \gamma_{12})\hat{z}_2 + \gamma_{11}y_1 + \gamma_{12}y_2 \qquad (35)$$

In order to eliminate the term $\hat{z}_2$ from (35), we put

$$A_{12} - \gamma_{11}C_{12} - \gamma_{12} = 0 \qquad (36)$$

Then we obtain

$$\begin{aligned}
\dot{\hat{z}}_1 &= (A_{11} - \gamma_{11}C_{11})\hat{z}_1 + \gamma_{11}y_1 + \gamma_{12}y_2 \\
&= (A_{11} - \gamma_{11}C_{11})\hat{z}_1 + \gamma_{11}y_1 + (A_{12} - \gamma_{11}C_{12})y_2 \\
&= A_{11}\hat{z}_1 + A_{12}y_2 + \gamma_{11}(y_1 - C_{11}\hat{z}_1 - C_{12}y_2) \qquad (37)
\end{aligned}$$

which will lead us to $(12)_1$. Substituting γ_{11} and γ_{12} of (32) for (36) yields

$$\begin{aligned}
A_{12} &= \gamma_{11}C_{12} + \gamma_{12} \\
&= P_{11}C_{11}^T\Omega_{12} + P_{12}(C_{12}^T\Omega_{12} + \Omega_{22}) + P_{11}C_{11}^T\Omega_{11}C_{12} \\
&\quad + P_{12}(C_{12}^T\Omega_{11} + \Omega_{12}^T)C_{12} \\
&= P_{11}C_{11}^T(\Omega_{12} + \Omega_{11}C_{12}) \\
&\quad + P_{12}(C_{12}^T\Omega_{11}C_{12} + \Omega_{12}^TC_{12} + C_{12}^T\Omega_{12} + \Omega_{22})
\end{aligned}$$

Taking (30) into account, it follows that

$$A_{12} = (P_{12} - P_{11}C_{11}^TR_{11}^{-1}R_{12})S^{-1} \qquad (38)$$

or

$$P_{12} = A_{12}S + P_{11}C_{11}^T R_{11}^{-1} R_{12} \tag{39}$$

Substituting back into (33) yields

$$\begin{aligned}
\gamma_{11} &= P_{11}C_{11}^T\Omega_{11} + P_{12}(C_{12}^T\Omega_{11} + \Omega_{12}^T) \\
&= P_{11}C_{11}^T\Omega_{11} - P_{12}S^{-1}R_{12}^T R_{11}^{-1} \\
&\quad -P_{11}C_{11}^T\Omega_{11} - (A_{12}S + P_{11}C_{11}^T R_{11}^{-1}R_{12})S^{-1}R_{12}^T R_{11}^{-1} \\
&= P_{11}C_{11}^T(\Omega_{11} - R_{11}^{-1}R_{12}S^{-1}R_{12}^T R_{11}^{-1}) - A_{12}R_{12}^T R_{11}^{-1} \\
&= (P_{11}C_{11}^T - A_{12}R_{12}^T)R_{11}^{-1}
\end{aligned} \tag{40}$$

Similarly, applying (38) to (34), we can obtain

$$\begin{aligned}
\dot{P}_{11} &= A_{11}P_{11} + P_{11}A_{11}^T - P_{11}C_{11}^{\mathrm{T}}\Omega_{11}C_{11}P_{11} + B_1WB_1^T + P_{12}S^{-1}P_{12}^T \\
&= A_{11}P_{11} + P_{11}A_{11}^T - P_{11}C_{11}^{\mathrm{T}}\Omega_{11}C_{11}P_{11} + B_1WB_1^T \\
&\quad + (A_{12}S + P_{11}C_{11}^T R_{12})S^{-1}(SA_{12}^T + R_{12}^T R_{11}^{-1}C_{11}P_{11}) \\
&= (A_{11} + A_{12}R_{12}^T R_{11}^{-1}C_{11})P_{11} + P_{11}(A_{11}^T + C_{11}^T R_{11}^{-1}R_{12}^T A_{12}^T) \\
&\quad + B_1WB_1^T + A_{12}^T SA_{12}^T - P_{11}C_{11}^T(\Omega_{11} - R_{11}^{-1}R_{12}S^{-1}R_{12}R_{11}^{-1}) \\
&\quad \times C_{11}P_{11} \\
&= (A_{11} + A_{12}R_{12}^T R_{11}^{-1}C_{11})P_{11} + P_{11}(A_{11}^T + C_{11}^T R_{11}^{-1}R_{12}A_{12}^T) \\
&\quad - P_{11}C_{11}^T R_{11}^{-1}C_{11}P_{11} + B_1WB_1^T + A_{12}(R_{22} - R_{12}^T R_{11}^{-1}R_{12})A_{12}^T
\end{aligned} \tag{41}$$

This reveals that the reduced-order filter represented by (37), (40), and (41) is in coincidence with the reduced-order filter given by (12).

B. Derivation 2

The error covariance submatrix P_{11} of the full-order Kalman filter is given from (34) by

$$\begin{aligned}
\dot{P}_{11} = A_{11}P_{11} &+ P_{11}A_{11}^T - P_{11}C_{11}^T\Omega_{11}C_{11}P_{11} + B_1WB_1^T - P_{12}S^{-1}P_{12}^T \\
&+ P_{12}(A_{12}^T + S^{-1}R_{12}^T R_{11}^{-1}C_{11}P_{11}) \\
&+ (A_{12} + P_{11}C_{11}^T R_{11}^{-1}R_{12}S^{-1})P_{12}^T
\end{aligned}$$

Here we consider the artificial problem to determine the matrix P_{12} so that $\mathrm{tr}[P_{11}]$ is maximized. The Hamiltonian is defined as

$$H = \mathrm{tr}[P_{11}(t)\Lambda^T(t)]$$

From $\partial H/\partial P_{12} = 0$

$$\begin{aligned}
&\Lambda(A_{12}^T + S^{-1}R_{12}^T R_{11}^{-1}C_{11}P_{11})^T + \Lambda^T(A_{12} + P_{11}C_{11}^T R_{11}^{-1}S^{-1}) \\
&- \Lambda P_{12}S^{-1} - \Lambda^T P_{12}S^{-1} = 0
\end{aligned}$$

Therefore

$$P_{12} = A_{12}S + P_{11}C_{11}^{T}R_{11}^{-1}R_{12}$$

which is equal to (39).

VI. EXAMPLES

Example 1. A simple example will be given here to illustrate the design of the reduced-order filter. The system considered here is given by

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w \tag{42}$$

$$y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{43}$$

where the noise variances are given by

$$W = \sigma_w^2, \qquad V = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

A. Full-Order Optimal Kalman Filter

The dynamics of the full-order Kalman filter is expressed as

$$\begin{aligned} \dot{\hat{x}}_1 &= \hat{x}_2 + \gamma_{11}(y_1 - \hat{x}_1) + \gamma_{12}(y_2 - \hat{x}_2) \\ \dot{\hat{x}}_2 &= \gamma_{21}(y_1 - \hat{x}_1) + \gamma_{22}(y_2 - \hat{x}_2) \end{aligned} \tag{44}$$

where the filter gains are

$$\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^{T} & P_{22} \end{bmatrix} \begin{bmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{bmatrix}$$

Also the steady-state values P_{ij} of the error covariance are satisfied by

$$\begin{aligned} 2P_{12} - P_{11}^2/\sigma_1^2 - P_{12}^2/\sigma_2^2 &= 0 \\ P_{22} - P_{11}P_{12}/\sigma_1^2 - P_{12}P_{22}/\sigma_2^2 &= 0 \\ P_{12}^2/\sigma_1^2 - P_{22}^2/\sigma_2^2 + \sigma_w^2 &= 0 \end{aligned} \tag{45}$$

the solution of which is given by

$$\begin{aligned} P_{11} &= \sigma_1\sigma_2(2rs + r^2s^2)^{1/2}/(1 + rs) \\ P_{12} &= \sigma_1\sigma_w/(1 + rs) \\ P_{22} &= \sigma_2\sigma_w(2rs + r^2s^2)^{1/2}/(1 + rs) \end{aligned} \tag{46}$$

where

$$r = \sigma_w/\sigma_2, \qquad s = \sigma_1/\sigma_2$$

B. Design of the Reduced-Order Filter for $z = (1, 0)x$

The system can be described from (42) and (43) as

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = w \end{cases} \tag{47}$$

$$\begin{cases} y_1 = x_1 + v_1 \\ y_2 = x_2 + v_2 \end{cases} \tag{48}$$

Substituting $x_2 = y_2 - v_2$ into (47) yields

$$\dot{x}_1 = y_2 - v_2$$
$$y_1 = x_1 + v_1$$

Thus, the reduced-order filter can be readily obtained as

$$\begin{aligned} \dot{\hat{x}}_1 &= y_2 + G_1(y_1 - \hat{x}_1) \\ G_1 &= P/\sigma_1^2 \\ P &= \sigma_1 \sigma_2 \end{aligned} \tag{49}$$

It should be noted that the filter is independent of σ_w^2.

1. Alternative Derivation from the Full-Order Kalman Filter

From $(44)_1$

$$\begin{aligned} \dot{\hat{x}}_1 &= \gamma_{12} y_2 + \gamma_{11}(y_1 + \hat{x}_1) + (1 - \gamma_{12})\hat{x}_2 \\ \gamma_{11} &= P_{11}/\sigma_1^2 \\ \gamma_{12} &= P_{12}/\sigma_2^2 \end{aligned} \tag{50}$$

To eliminate $\hat{x}_2$ from (50), putting $\gamma_{12} = 1$ yields

$$\dot{\hat{x}}_1 = y_2 + \gamma_{11}(y_1 - \hat{x}_1)$$

and from $(45)_1$, taking $\gamma_{12} = 1$ and $P_{12} = \gamma_{22}$ into account,

$$\sigma_2^2 - P_{11}^2/\sigma_1^2 = 0$$

Hence, we can obtain $P_{11} = \sigma_1, \sigma_2$ and $\gamma_{11} = \sigma_2/\sigma_1$. Thus we have the reduced-order filter given by (49).

Another observation gives the same result. In fact, from (45)

$$P_{11}^2 = 2\sigma_1^2 P_{12} - \sigma_1^2 P_{12}^2/\sigma_2^2$$

Maximization of P_{11} with respect to P_{12} gives

$$P_{12} = \sigma_2^2$$
$$P_{11}^2 = \sigma_1^2 \sigma_2^2$$

which leads us to the same reduced-order filter as (49).

It is interesting here to note that

$$P = \lim_{\sigma_w \to \infty} P_{11}$$

where P_{11} is given by $(46)_1$: namely the error variance of the reduced-order filter is equal to the limiting value of the error variance P_{11} of the full-order Kalman filter as σ_w tends to infinity.

C. Derivation of the Reduced-Order Filter for $z = (0, 1)x$

From (42) and (43), the reduced-order filter can be designed on the basis of the dynamics:

$$\begin{aligned} \dot{x}_2 &= w \\ y_2 &= x_2 + v_2 \end{aligned}$$

and is readily given by

$$\begin{aligned} \dot{\hat{x}}_2 &= G_1(y_2 - \hat{x}_2) \\ G_1 &= P/\sigma_{22} \\ P &= \sigma_2 \sigma_w \end{aligned} \tag{51}$$

It should be noted that the filter does not depend on σ_1^2.

Another observation also gives the same results. From $(45)_3$,

$$P_{22}^2 = \sigma_2^2 \sigma_w^2 - \sigma_2^2 P_{12}^2/\sigma_1^2$$

The maximum value of P_{22} is achieved when $P_{12} = 0$, i.e.,

$$\begin{aligned} P_{12} &= 0 \\ P_{22}^2 &= \sigma_2^2 \sigma_w^2 \end{aligned}$$

which easily reduces to (51). Also we can observe that

$$P = \lim_{\sigma_1 \to \infty} P_{22}$$

where P_{22} is expressed by $(46)_3$.

Example 2. An example will be presented here to illustrate the design of the reduced-order filter where the rank condition does not hold. Suppose that the dynamic system is described as

$$\dot{x} = \begin{bmatrix} -1 & 4 & 5 \\ 4 & -2 & 2 \\ 0 & 0 & -3 \end{bmatrix} x + \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} w$$

$$y = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix} x + v$$

with noise variances

$$W = \sigma_w^2, \qquad V = \sigma_v^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Consider the case where the desired linear function to be estimated is given as

$$z_1 = (1\ 2\ 3)x \tag{52}$$

Since the rank condition of (7) does not hold, the first-order filter cannot be obtained. As shown in (24), we can consider the second-order filter by modifying (52) as

$$z_1 = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 0 \end{bmatrix} x$$

An appropriate transformation matrix T of (25) can be defined as

$$T = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

The system transformation gives

$$\dot{z} = \begin{bmatrix} 1 & -1 & -1 \\ 3 & -3 & -1 \\ 3 & 0 & 4 \end{bmatrix} z + \begin{bmatrix} 11 \\ 3 \\ 5 \end{bmatrix} w$$

$$y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} z + v$$

Based on the above system, we can readily obtain the second-order filter for $\sigma_w^2 = 0.1$ and $\sigma_v^2 = 1.0$ as

$$\dot{\hat{z}}_1 = \begin{bmatrix} -34.4804 & -1 \\ -8.5517 & -3 \end{bmatrix} \hat{z}_1 + \begin{bmatrix} 35.4804 \\ 11.5517 \end{bmatrix} y_1 - \begin{bmatrix} 1 \\ 1 \end{bmatrix} y_2$$

with the steady-state value of the estimation error variance $P_{11} = 3.548$. It is interesting to note that the optimal value of error variance achieved by the full-order Kalman filter can be calculated as $P = 3.240$.

VII. CONCLUDING REMARKS

In this chapter, we have described a computationally feasible design technique for the reduced-order unbiased filter which estimates a specified linear function of the state for the stochastic system. The solution of the present problem is simply to derive the Kalman filter by a standard method for the reduced-order stochastic system obtained from an appropriate system transformation. Due to the reduction of filter order, the result is of course suboptimal. The price to pay for the computational simplicity is some deterioration of performance. Finally the relation to the full-order Kalman filter has also been discussed.

REFERENCES

1. BJ Uttam, WF O'Halloran. On observer and reduced order optimal filter for linear stochastic systems, JACC paper 29-4, 1971.
2. E Tse. Inform Control 22:405–434, 1973.
3. Y Halevy. IEEE Trans Autom Control 34:777–781, 1981.
4. CS Sims. Advances in Control and Dynamic Systems, Vol. 18 (CT Leondes, ed.). New York: Academic Press, pp 55–103, 1982.
5. KG Bernstein, DC Hyland. IEEE Trans Autom Control 30:583–588, 1985.
6. KM Nagpal, RE Helmick, CS Sims. Int J Control 45:1867–1888, 1987.
7. M Athans, E Tse. IEEE Trans Autom Control 12:690–698, 1967.
8. M Athans. Informat Control 11:592–606, 1968.

10

Shares in Emergent Markets: Dynamics and Statistical Properties of Equilibrium Classification of Agents in Evolutionary Models

Masanao Aoki

University of California, Los Angeles, Los Angeles, California

I. INTRODUCTION

Much attention has recently been paid to the construction of macroeconomic models which incorporate nonpairwise externalities among agents, such as fads or bandwagon effects. These effects are called field effects in [1], [2], and [3], and referred to as social influences in [4] and [5]. See also [6], [7], [8], or [9]. An example in sociology is [10]. By modeling these externalities we hope to deduce an "emergent property" of a system of agents, i.e., aggregate properties or characteristics of situations involving a large number of interacting economic agents. These interactions are not pairwise in nature, but are between individual agents and aggregate environments, which are in turn determined by or dependent on the behavior or decisions of the collection of agents.

To address this need, we start with a more disaggregate description of a collection of agents than commonly used in traditional macroeconomics, and investigate the nature of patterns of interactions which may emerge as the number of agents increases, i.e., we look for conditions under which statistically significant stable patterns of interactions emerge.

To begin, we classify agents by their types or characteristics, and use their joint statistical descriptions or dynamics of transitions of agents from one type to another to derive their stationary distributions. See [1] and [2], and [11] and [12] for modeling approaches in a similar spirit. By "types," we mean any scheme to classify or categorize agents, such as behavioral or decisions rules used by agents, stores they shop, or brands of goods they buy, and so on, depending on the context of models we wish to construct. We thus employ Markov processes with sets of random partitions of agents into separate types, classes, or categories as the state spaces.

For example, in a simulation study of a fish market which involves interaction among a large number of buyers and a small number of sellers of fresh fish, Weibusch, Kirman, and Herreiner discovered in [13] that a stable pattern of market shares for the sellers emerges under certain conditions.* In this example, the customers are partitioned into types by the stores they buy fish. When the number of sellers is small, say 2 or 3, the situations can be modeled using the master equations exactly as in [1] and [2], and we can examine its equilibrium distributions.† Aldous has an example in [15] which he attributes to J. Pitman, called the Chinese restaurant process, in which people arrive sequentially at an initially empty restaurant with a large number of tables and seat themselves. In this example, customers are classified by the tables they sit at. This perspective can be applied to noneconomic situations as well. For example, relative sizes of basins of attraction of random maps of a finite sets of points can be so examined; compare [15] with [16]. The former adopts this viewpoint, while the latter does brute force calculations. Here the points are classified into types by the basins in which it will eventually be attracted. We introduce the notion of frequency spectrum in Section IV by which these seemingly unrelated topics can be unified.

When the number of types is large, it is analytically messy and not practical to use the master equations directly. Instead of dealing with approximations to the master equations, we adopt an alternative approach in this chapter. As a first step, we derive the stationary distributions of random partitions of agents into types. In the case of the fish market with four or more sellers, we look for stable distributions of (the order statistics of) market shares of sellers by regarding patterns of buyers among sellers as random partitions of the set of buyers (and the process of changing patterns as Markov processes on the set of random partitions).‡

*See also [14]. Slightly generalizing their formulation, customers on a given day may be regarded as a sample from the population of the city, hence the sample size may vary from day to day. The number of sellers may also vary from day to day.

†This alternative approach is more informative than their study which relies heavily on simulation experiments since the existence and local stability properties may be analyzed to supplement and augment the simulation studies.

‡As we discuss later we focus on the order statistics of the composition, i.e., the fractions of agents per type rather than their absolute numbers in all applications. It is known that this gives a tractable problem formulation, when the number of agents becomes very large. Over the course of time, some buyers change their minds and shop at a different store. This is captured by transition rates of the Markov process over partitions.

In this chapter we primarily discuss stationary distributions of the compositions of agents by focusing on the order statistics of the compositions of a population of economic agents by some characteristics or traits, such as types, or choices, and examining the fractions of agents of these characteristics. Thus we treat agents as exchangeable in the technical sense known in probability theory. We focus on asymptotic results for the stationary distributions of the shares (types) of customers in markets, i.e., shares of the sellers when the number of buyers becomes large and when the number of sellers also becomes large in the fish market, for example.*

The usual approach of modeling such interaction phenomena in the economic literature would involve random matching with anonymous partners or some other game theoretic devices. The game theoretic approach does not seem to be useful when the number of participants is large, since game theoretic algorithms are probably nonimplementable for the case with a large number of participants.† As will become apparent, the number of types or categories need not be fixed. The approach we propose here can accommodate new types appearing randomly, such as new consumer goods, decision or behavioral rules and so forth, as results of innovations or learning.

We use results on random partitions and random maps, as summarized in [15] for example, and introduce into the economic literature the equilibrium distributon called Ewens' sampling formula, developed in the context of population biology, see [17], to describe the order statistics of the market shares. This formula arises naturally as the random average of multinomial distributions with the parameter vector governed by Dirichlet distributions. This approach can deal with large numbers of types of participants at least approximately due to a result of Kingman who in [18] and [19] demonstrated that any model which has the Ewens sampling property must necessarily have a Poisson–Dirichlet distribution for the decreasing order statistics of the proportion or market shares.‡ There is

*The numbers of agents of any type all approach zero as the number of types becomes unbounded. This is why we focus on the order statistics.

†For example, [23] or [24] deals with two or three types of agents at the most. This is due to the sheer combinatorial computational burden which overwhelms the computational capability at our disposal. These kinds of difficulties also favor statistical descriptions of the stationary compositions of populations of agents by types adopted in this chapter.

‡The Poisson–Dirichlet distribution is a generalization of the Dirichlet distribution, as the number of agents and types go to infinity in a certain way. See [25] or [26] on the Dirichlet distribution. See [15] or [27] for a useful interpretation of this distribution in terms of a Pólya-like urn model.

a key parameter in the Ewens distribution, which is denoted by θ in this chapter. We can interpret it in terms of the probability of two agents belonging to the same subset in the partition of the set of total number of agents. Here, we use the notion of the frequency spectrum or intensity which is introduced in Section IV.

There is a related literature on Zipf's law and size distribution of market shares or firm sizes as discussed in [20], [21], [22], and others, and Bose–Einstein statistics. We comment on the relation of these with Dirichlet distributions and the Ewens' sampling formula in Section V. The connection here is shown to be the logarithmic series distribution, which appears as the mean of Poisson random variables which would be independent except for the constraint imposed on them by the total numbers of agents in the model. Whittle has developed similar results in [28]. Keener, Rothman, and Starr have developed a two-parameter family of distributions on partitions which cover these distributions in [29]. See also [30], [31], and [32].

II. DIRICHLET DISTRIBUTIONS AND EWENS' SAMPLING FORMULA

A probability distribution of a population composition is defined on a finite dimensional simplex. It is a collection of nonnegative numbers, x_i, $i = 1, 2, \ldots, n$, satisfying the sum constraint, $\Sigma_i x_i = 1$, by regarding x_i as the coordinates of a point $x = (x_1, x_2, \ldots, x_n)$ in an n-dimensional Euclidean space. We think of x_i as the proportion or fraction of agents of type i in the population. Since the labeling of types as type 1, 2, and so on, is arbitrary in the sense that there is no intrinsic meaning, economic or otherwise, to the orders of labels, x_i are exchangeable and the joint distributions are symmetric in these variables.*

An exchangeable Dirichlet distribution has the density

$$\frac{\Gamma(n\alpha)}{\Gamma^n(\alpha)}(x_1 x_2 \ldots x_n)^{\alpha-1}$$

where $x_n = 1 - x_1 - \ldots - x_{n-1}$. It is regarded as a density for x_i, $i = 1, \ldots n-1$. We denote this by $D(\alpha, n)$ for short. Here, α is a parameter of the distribution. We interpret it later.

Several ways of obtaining Dirichlet distributions are known [25]. One is as follows: suppose Y_i, $i = 1, 2, \ldots, n$, are independently and identically

*See [15] or [18] on exchangeability and exchangeable partitions.

distributed (i.i.d.) random variables with density $y^{\alpha-1}e^{-y}/\Gamma(\alpha)$, for positive y. Define $S = Y_1 + \ldots + Y_n$, and a new set of random variables by $V_i = Y_i/S$, $i = 1, 2, \ldots, n-1$. Change the variables from Y_i to V_i, $i = 1, \ldots, n-1$, and S. The Jacobian is S^{n-1}, and we deduce the joint probability density function

$$g(v_1, v_2, \ldots, v_{n-1}) = \frac{\Gamma(n\alpha)}{\Gamma^n(\alpha)} \prod_{i=1}^{n} v_i^{\alpha-1}$$

where $v_n = 1 - v_1 - v_2 - \ldots - v_{n-1}$. We also see that S is independent of V_i, $i = 1, 2, \ldots, n-1$.

Dirichlet distributions possess important properties of neutrality introduced in [33], and size-biased permutation or rescaling mentioned in [34]. See Hoppe in [35] and [36] for a discussion in the more general setting of Poisson–Dirichlet distributions.

Consider a random sample of size n drawn from a population with K types of agents with relative composition or fractions, $x_1, x_2, \ldots, x_K$. Let n_i be the number of agents of type i. Suppose that they have a multinomial distribution

$$P(n_1, n_2, \ldots, n_K; K) = \frac{n!}{n_1! \ldots n_K!} x_1^{n_1} \ldots x_K^{n_K}$$

where $\Sigma_i n_i = n$. For a simpler explanation assume that n_i is positive for all i.

When we average this expression with respect to a Dirichlet distribution $D(\varepsilon, n)$* we obtain the probability

$$P(n_1, \ldots, n_K; K) = \frac{n!}{n_1! \ldots n_K!} \frac{\Gamma(K\varepsilon)}{\Gamma^K(\varepsilon)} \frac{\Gamma(\varepsilon + n_1) \ldots \Gamma(\varepsilon + n_K)}{\Gamma(K\varepsilon + n)}$$

Now arrange n_i in decreasing order $n_{(1)} \geq n_{(2)} \geq \ldots n_{(k)}$ for some $k \leq K$. The probability of the first k order statistics is given by

$$P(n_{(1)}, \ldots, n_{(k)}; k) = \frac{n!}{n_{(1)}! \ldots n_{(k)}!} \frac{\Gamma(K\varepsilon)}{\Gamma^k(\varepsilon)} \frac{\Gamma(\varepsilon + n_{(1)}) \ldots \Gamma(\varepsilon + n_{(k)})}{\Gamma(K\varepsilon + r)} \times M$$

*The index ε is the same for all fractions because we assume them to be exchangeable.

where $r = n_{(1)} + \ldots + n_{(k)}$, and M is the number of ways that the order statistics can be distributed among K types, i.e.,

$$M = \frac{K!}{a_1!a_2!\ldots a_r!(K-k)!}$$

and where a_j is the number of n_i equal to j. See [37] and [38].

If we let K go to infinity, and ε to zero, in such a way that $K\varepsilon$ approaches a constant, θ, then

$$\frac{K!}{(K-k)!\Gamma(\varepsilon)^k} \to \theta^k$$

and we obtain the limiting expression

$$P(n_{(1)}, \ldots, n_{(k)}; k) = \frac{n!\,\theta^k \Gamma(\theta)}{n_{(1)} \ldots n_{(k)} a_1! \ldots a_r! \Gamma(\theta + r)}$$

This is the equilibrium distribution of the order statistics of the fractions of the ways agents distribute themselves among different types, and is the same as the Ewens sampling formula [19] to which we return shortly. The joint probability when a total of n agents are involved is obtained by replacing r by n and k by K. An alternative expression for the probability distribution is obtained by noting that $\Pi_{j=1}^n n_j$ is expressible as $1^{a_1}2^{a_2}\ldots j^{a_j}\ldots n^{a_n}$, see [39].

Some special cases shed light on this distribution. When n agents are all of the same type, then we have $a_i = 0$, $i = 1, \ldots, n-1$, and $a_n = 1$, and

$$P(a_1 = \ldots a_{n-1} = 0, a_n = 1; n) = \prod_{j=1}^{n-1} \frac{j}{\theta + j}$$

When $n = 2$, this yields the probability

$$P(a_1 = 0, a_2 = 1; 2) = \frac{1}{\theta + 1}$$

This is the probability that two agents chosen randomly in succession are of the same type. This expression may be used to interpret the parameter θ.

The probability that all n agents are of a different type is obtained by setting a_1 to n and all other as to zero:

$$P(a_1 = n, a_2 = \ldots = a_n = 0; n) = \prod_{j=1}^{n-1} \frac{\theta}{\theta + j}$$

A. An Urn Model Representation

All these expressions strongly suggest a connection with a representation of these probabilities by a Pólya-like urn model. This is indeed true, and has been shown in [27] and [36]. Ewens' development in [17] is also very much in the urn model framework. Consider an urn which contains only θ black balls initially.* Each time a black ball is drawn, it is returned to the urn together with one ball of a new type (color). When a nonblack ball is drawn, then it is returned together with a ball of the same type (color). Therefore, after j draws, the probability that the $(j+1)$th draw is a black ball is $\theta/(\theta+j)$, and the probability that the $(j+1)$th draw is not a black ball is $j/(\theta+j)$. The probability that the first j draws are all of the same type agrees with the expression obtained above with $a_i=0$, $i=1,\dots,j-1$, and $a_j=1$,

$$P(a_1=\dots=a_{j-1}=0, a_j=1)=\frac{1}{\theta+1}\frac{2}{\theta+2}\cdots\frac{(j-1)}{\theta+j-1}=\frac{\theta(j-1)!\Gamma(\theta)}{\Gamma(\theta+j)}$$

Let $q_{j,i}$ be the probability that the first j draws yield i types of balls, i.e., j balls are of i different colors. For example, $q_{j,1}$ is the expression we just derived, and

$$q_{j,j}=\frac{\theta^j}{(\theta+1)\dots(\theta+j-1)}$$

We have a recursion relation

$$q_{j+1,i}=q_{j,i}\frac{j}{\theta+j}+q_{j,i-1}\frac{\theta}{\theta+j}$$

Suppose we write these $q_{j,i}$ as

$$q_{j,i}=\frac{c(j,i)\,\theta^i}{[\theta]^j}$$

where we define

$$[\theta]^j=\theta(\theta+1)\dots(\theta+j-1)$$

Then we easily see that $c(j,i)$ satisfies the recursion relation

$$c(n,i)=c(n-1,i-1)+(n-1)c(n-1,i)$$

where we used n rather than j.

*Allow for fractional numbers. Otherwise think of one black ball with weight θ and make the probability of drawing a ball proportional to the weight.

This recursion is the same as the (unsigned) Stirling number of the first kind.* By noting that $\Sigma_j a_j = K$, we can rewrite the probability as

$$P(a_1, \ldots, a_n; K) = \frac{1}{{}_{\theta+n-1}C_{\theta-1}} \prod_{j=1}^{n} \left\{\frac{\theta}{j}\right\}^{a_j} \frac{1}{a_j!}$$

Rewritten this way, we see a connection with the logarithmic series expression for the probability in [11] to which we return later.

B. Neutrality

One might question the appropriateness of the Dirichlet distributions for the vector of random fractions of economic agents. We describe next why this distribution is the most natural one to use in our context.

First, Fabius shows in [40] that if none of the fractions x_i nor $1 - \Sigma x_i$ vanishes almost surely in a random vector of fractions $(x_1, \ldots, x_K)$, then these fractions are distributed as a Dirichlet distribution or the limit of such distributions. Second, we refer to the notion of neutrality in [33] and that of a size-biased permutation (re-sampling) or rescaling in [34]. They both refer to a population with fractions which have the representation

$$x_1 = u_1$$

and

$$x_m = u_m \prod_{j=1}^{m-1} (1 - u_j), \qquad m \geq 2$$

where u_j are independent random variables, called residual fractions, and the model, called (infinite) residual fraction model. This is the same as the random alms model of Halmos in [41], although he did not discuss these notions.

In a random vector of fractions $(x_1, \ldots, x_n)$, x_1 is said to be neutral if x_1 is independent of $\{x_2/(1 - x_1), x_3/(1 - x_1), \ldots, x_n/(1 - x_1)\}$. This means that x_1 does not influence the manner in which the remaining fractions proportionally divide the remainder of the unit interval. These are rescaled fractions. If we introduce another set of random variables by

$$z_i = \frac{x_i}{\sum_{j=i}^{n} x_j}$$

*See [42] or [43] on Stirling numbers.

for $i = 1, 2, \ldots, n-1$, and $z_n = 1$, then z_1 is independent of $(z_2, \ldots, z_{n-1})$. If $(x_1, x_2, \ldots, x_j)$ is independent of $(x_{j+1}, \ldots, x_n)/(1 - x_1 - x_2 - \ldots - x_j)$ for all j, we call the vector x completely neutral. Thus, if x is completely neutral, then $z_1, z_2, \ldots, z_n$ are independent. The converse is also true. See [35].

Let the density of z_i be given by

$$B(\alpha, b_i)^{-1} z_i^{\alpha-1}(1 - z_i)^{b_i-1}$$

$i = 1, \ldots, n-1$, where $B(\cdot,\cdot)$ is the beta function. Then the density function of the x_i is $D(\alpha, n)$ when $b_i = (n-i)\alpha$. See [33].

Consider an infinite sequence, i.e., let K go to infinity. If $u_1, u_2, \ldots$ are sequences of independent beta variates with the same density $a(1-u)^{a-1}$, then $E(u_i) = \theta$ with $\theta = (1+\beta)^{-1}$, for all i, and

$$E(x_i) = \theta(1-\theta)^{i-1}$$

The random variables x_i are such that

$$\lim_{j\to\infty} E(x_1 + x_2 + \ldots + x_j) = 1$$

and $x_1 + x_2 + \ldots + x_j$ converges to 1 in probability. Furthermore,

$$\sum_{i=1}^{\infty} E(x_i^{k+1}) = \frac{\Gamma(a+1)\Gamma(k+1)}{\Gamma(a+k+1)}$$

Next, compare this with

$$E(x^k) = E\{E(x^k | x_1, x_2, \ldots, x_n)\} = E(\Sigma x_i^{k+1})$$

where x_i are governed by $D(\varepsilon, n)$. The moment is given by

$$\frac{\Gamma(K\varepsilon+1)\Gamma(K+k+1)}{\Gamma(K+1)\Gamma(K\varepsilon+k+1)} \to \frac{\Gamma(\theta+1)\Gamma(k+1)}{\Gamma(\theta+k+1)}$$

for all nonnegative k as K goes to infinity.

These two expressions are equal when we set $a = \theta$. The sequence of random fractions given by the Dirichlet distribution tends in distribution to that specified by the residual fraction representation, because their moments become the same. (See [26] for a sufficient condition for the convergence of distributions of bounded random variables.) Moreover, it is known that a population with infinitely many types which are completely neutral or an infinite residual allocation model with independent and identically distributed residual fractions with density $\theta(1-u)^{\theta-1}$ is the only infinite residual allocation model which is invariant in distribution under size-biased permutation, as shown in [35].

The limit of order statistics distributions as n goes to infinity is named the Poisson–Dirichlet distribution by Kingman in [18]. Patil and Taillie point out in [34] that Kingman's and the model sketched above are permutations of each other, and are merely alternative description of the same underlying population structure.

For later comparison, we note that when a sample $x_1, \ldots, x_K$ is arranged in nonincreasing order from the largest to the smallest, $X_1^* \geq X_2^* \geq \ldots$, its order statistics are such that

$$f(X_1^* = x) = \theta x^{-1}(1-x)^{\theta-1}$$

for $1/2 \leq x \leq 1$, and

$$f(X_1^* = x, X_2^* = y) = \theta^2 x^{-1} y^{-1}(1-x-y)^{\theta-1}$$

for $0 \leq y \leq x \leq 1$, and $0 \geq 1-x \geq 2y$. These expressions are derived in [37]. See also [38].

III. RANDOM PARTITIONS AND EWENS' SAMPLING FORMULA

We think of statistical distributions of buyers among stores, or more generally, of economic agents among various types as randomly partitioning a set of agents, i.e., a set of n points into K subsets for some K. Alternatively we may think of n agents and group them into K subsets. We can also relate these to partitioning permutations into cycles where agents in a cycle may be regarded as of the same type as we do in Section IV.A.

As time progresses agents randomly switch stores to buy or randomly change their types. We treat these processes as Markov processes on the set of partitions. Is there going to be some useful statistical characterization of market structure which emerges from such random visits? The answer is yes under some technical conditions. The Ewens sampling formula introduced in Section II is one example of such distributions.

Suppose there are K stores and call agent type i if he buys at store i at some given point in time. Let a_j be the number of stores which have exactly j customers, i.e.,

$$\sum_j a_j = K$$

and

$$\sum_j j a_j = n$$

where n is the number of agents in this sample. If it helps, one may visualize K nonempty boxes with a_j of them containing j unmarked, i.e., nondistinguishable, balls. Then the Ewens sampling formula states that the distribution of balls in K boxes is

$$P(a_1, a_{2,} \ldots, a_n; K) = \frac{n!}{\prod_{j=1}^{n} (j)^{a_j} a_j!} \frac{\theta^K}{[\theta]^n}$$

Kingman in [18] and [19], and Watterson in [44] and [37] have derivations of this formula. We comment on the Kingman derivation based on the consistency requirements on the partition, since this approach has a direct implication on the sampling of market shares on a particular day. The basic idea is that probability distributions on the set of all possible partitions of n agents must be consistent in the following sense: think of n agents as a sample from N agents, i.e., on a given day only n of the total number, N, of agents come to market. Then the partition of n into various stores must be such that a further sampling of m agents out of n should be the same as the sampling of m from the initial N agents in the distributions. Kingman shows that the Ewens formula satisfies this property and he proves the characterization theorem that the partition structure with consistency implies that the distribution is the Ewens form for some θ. See [45] for a useful summary.

IV. FREQUENCY SPECTRUM

Here we follow [36] to describe the concept of frequency spectrum or intensity which is the probability of a fraction being in some range. This notion was introduced in [17]. We follow Hoppe in [36].

Let f be a bounded measurable function on $[0, 1]$. Let Q_1 be the first component in the size-biased permutation of the composition vector, $x = (x_1, x_2, \ldots, x_K)$, i.e., the fraction of the agents of the same type as the agent first drawn. Thus

$$\sum_i f(x_i) x_i = E\{f(Q_1) \,|\, x\}$$

In the Poisson–Dirichlet population, Q_1 has beta $(1, \theta)$ density. Taking the expectation of the above

$$E\left\{ \sum_i f(x_i) x_i \right\} = E\{f(Q_1)\} = \int_0^1 f(u)\, \theta (1 - u)^{\theta - 1} \mathrm{d}u$$

Now choose a particular f which is equal to u^{-1} if $u \geq t$ and zero otherwise for t in the unit interval. Then, the expected number of types whose frequency exceeds t is given by

$$\int_t^1 \theta u^{-1}(1-u)^{\theta-1}\mathrm{d}u$$

namely,

$$\phi(x) = \theta x^{-1}(1-x)^{\theta-1}$$

is the frequency spectrum.

Having obtained the frequency spectrum many interesting results can be calculated. For example, the probability that two randomly drawn agents are of the same type is given by

$$E(x^2) = \int_0^1 x^2 ax^{-1}(1-x)^{a-1}\mathrm{d}x = \frac{1}{1+a}$$

Hence, the parameter $a(=\theta)$ which characterizes the intensity may be interpreted accordingly. Recall our earlier discussion in Section II.A.

A. Attractors in Discrete Random Dynamics

Given a set of n points in a space S, a random map assigns a point to each point of S. For simplicity we assume that each element of the set of all transformations of S into S is equally likely.*

Because n is finite and a map stays fixed once chosen, attractors of a given random map consist of cycles and paths leading into the cycles. We associate weight w_s with the sth attractor which is the fraction of points which fall in this attractor out of the total of n points. Katz shows in [47] how to decompose or partition random maps into components. See [16] and [46] for background material.

To calculate the probability distribution or the frequency spectrum of the weights, consider the probability, $Q(T_1, T_2, \ldots, T_m)$, that a randomly selected configuration S_1 at time zero visits T_1 different points in S (at time $0, 1, \ldots, T_1 - 1$) before it falls on a point already visited, that a randomly selected initial state S_2 visits T_2 different point of S before it falls on the point already visited, and that a randomly selected initial state S_m visits T_m different points before it falls on the points previously visited by all

*There are n^n possible random maps. If we consider only the set of one-to-one maps there are $n!$ such maps. There are other possibilities. See [46].

the initial states. The expression is in Derrida and Flyvbjerg [16]. The probability Y_m that the m initial states fall on the same attractor is

$$Y_m = \sum_{T_1=1}^{\infty} \sum_{T_2=0}^{\infty} \cdots \sum_{T_m=0}^{\infty} Q(\dot{T}_1, T_2, \ldots, T_m)$$

For a given random map, the probability Y_m that m initial states fall on the same attractor is

$$Y_m = \sum_s w_s^m$$

Therefore, on the average

$$Y_m = \int_0^1 f(W)\, W^m \mathrm{d}W$$

where $f(W)\mathrm{d}W$ is the average number of attractors which have weights between W and $W + \mathrm{d}W$. The expression $f(W)$ is the frequency spectrum. Calculating Y_m, we identify

$$f(W) = \frac{1}{2} W^{-1}(1 - W)^{-1/2}$$

This corresponds to the case with $\theta = \frac{1}{2}$ of the frequency spectrum, and follows from the fact that for exchangeable random partitions the probability that agents 1 and 2 belong to the same set in the partition is $\frac{2}{3}$. See [15] for the calculations involved.

We next sketch the direct calculation used in [16]. Let us illustrate the case of $m = 2$, when there are N points. With the first initial point, the probability of choosing T_1 distinct points is

$$1 \times \left(1 - \frac{1}{N}\right) \times \left(1 - \frac{2}{N}\right) \cdots \times \left(1 - \frac{T_1}{N}\right) \times \frac{T_1}{N},$$

because the $(T_1 + 1)$th point must fall on one of the T_1 points. With the second initial condition, the probability of choosing T_2 points and that of the $(T_2 + 1)$th point falling on one of the first T_1 points (otherwise these two initial points do not belong to the same basin) is obtained by multiplying the above by

$$\left(1 - \frac{T_1}{N}\right)\left(1 - \frac{T_1 + 1}{N}\right) \cdots \left(1 - \frac{T_1 + T_2 - 1}{N}\right) \times \frac{T_1}{N}$$

Change variables to $T_1 = \sqrt{N} t_1$ and $7'_1 + T_2 = \sqrt{N} t_2$ and carry out the integral to conclude that θ is $\frac{1}{2}$.

The probability that a randomly chosen initial state belongs to a basin of attraction of weight W is given by

$$g(W) = Wf(W)$$

When $f(x) = \theta x^{-1}(1-x)^{\theta-1}$ more generally, we obtain $g(x) = \theta(1-x)^{\theta-1}$. This is a well-known density in the literature of ecology according to [45].

B. Markov Processes on the Set of Exchangeable Partitions

Evolutionary dynamics of a large number of agents of a (large) number of types may be modeled as Markov processes on partitions of the set of agents who are exchangeable. Here we give an example of such dynamics due to Kingman, [48] and [18]. See also [49]. Poisson–Dirichlet distributions arise in many ways. One way is via the notion of empirical distributions of frequencies of different types of agents.* Consider a set of N agents, each of whom is one of S types. Fix N for simpler explanation. Consider a discrete time Markov chain $\{G_t\}$ for convenience of discussion. Its state space is a finite set of all possible types of agents. Let $p(i,j)$ be the transition probability of a type i agent at time t changing into a type j agent at time $t+1$. An example is†

$$p(i,j) = (1-u)\,\delta_{i,j} + uq(i,j)$$

where $Q = \{q(i,j)\}$ is a stochastic matrix, and u is a parameter of small value. Here it is exogenously given. It may depend on some economic variables such as price vectors in economic applications. This implies that customers mostly revisit the same store but occasionally go to another store with the probability shown in Q.

At time t label agents in random order $X_r(t)$, $r = 1,2,\ldots,N$. Think of this as the rth element to be drawn at time t in random sampling without replacement. These random variables are exchangeable by construction and hence

$$\Pr\{X_r(t+1) = j \mid G_t\} = \frac{1}{N}\sum_{a=1}^{N} p(X_a(t),j)$$

Let the probability of being in state j at time t be denoted by

$$\pi_t(j) = \Pr(X_1(t) = j)$$

*Types are to be interpreted broadly: classes, choices of stores to buy, and so on.

†Other forms of transition rates or transition probabilities are discussed in [2], Chapter 5, [11], and [12].

It satisfies

$$\pi_{t+1}(j) = \frac{1}{N}\sum E\{p(X_a(t), j)\} = \sum_i \pi_t(i)p(i,j)$$

where the summation is over all states.

Suppose that $P = \{p(i,j)\}$ is irreducible, aperiodic, and positive recurrent, with stationary probability measure $\{\pi(i)\}$. Then, regardless of the initial state of the Markov chain, we have

$$\pi_t(j) \to \pi(j)$$

as time goes to infinity for all j.

Let

$$\pi_t(i,j) = \Pr\{X_1(t) = i, X_2(t) = j\}$$

Then

$$\pi_{t+1}(i,j) = \frac{1}{N^2}\sum_{a,b} E\{p(X_a(t), i)p(X_b(t), j)\}$$

By exchangeability, if $a \neq b$, the right-hand side becomes

$$E\{p(X_1(t), i)p(X_2(t), j)\} = \sum_{k,l} \pi_t(k,l)p(k,i)p(l,j)$$

and if $a = b$, then it is

$$E\{p(X_1(t), i)p(X_2(t), j)\} = \sum_k \pi_t(k)p(k,i)p(k,j)$$

We also have

$$\pi_t(i,j) \to \pi(i,j)$$

In general,

$$\pi_t(j_1, j_2, \ldots, j_n) = \Pr\{X_1(t) = j_1, \ldots, X_n(t) = j_n\} \to \pi(j_1, j_2, \ldots, j_n)$$

Note that $\pi(j_1, \ldots, j_n)$ is a symmetric function of its n arguments, and $\pi(j_1, \ldots, j_n)$ forms a consistent family in the sense that

$$\sum_{j_n} \pi(j_1, \ldots, j_n) = \pi(j_1, \ldots, j_{n-1})$$

For the example of $p(i,j)$ mentioned at the beginning of this subsection we can write this transition as

$$n(n+\theta-1)\pi(j_1,j_2,\ldots,j_n) = \theta\sum_a\sum_i \pi(j_1,\ldots,j_{a-1},i,j_{a+1},\ldots,j_n) \times q(i,j_a) + \sum_a \pi(j_1,\ldots,j_{a-1},j_{a+1},\ldots j_n)\nu_a$$

where $\theta = 2Nu$ and $\nu_a = |\{a \neq b, j_b = j_a\}|$.

By de Finetti's theorem, these properties are enough for the existence of a family of random variables $p(i)$ which are nonnegative and sum to one, such that

$$\pi(j_1,j_2,\ldots,j_n) = E\{p(j_1)p(j_2)\ldots p(j_n)\}$$

which has a natural interpretation as the limiting distribution as n goes to infinity of the empirical distribution of n random variables X_r.

More specifically, $p(i)$ are the frequencies of the random fractions of types in the population, i.e., the frequency spectrum. Let f be any continuous function on $[0,1]$. Then

$$E\left[\sum p(j)f\{p(j)\}\right] = \int_0^1 f(x)\mu(\mathrm{d}x) = \int_0^1 f(x)x\phi(x)\,\mathrm{d}x$$

where we write μ as $x\phi(x)$. Here $\phi(x)$ is the frequency spectrum, i.e., the probability of relative frequency lying between x and $x + \mathrm{d}x$.

Now set $j_1 = j_2 = \ldots = j_n = j$ in the arguments of the πs. This equation becomes

$$(n-1+\theta)\int_0^1 x^n\phi(x)\,\mathrm{d}x = \theta\sum_{i,j}\pi(i,j,\ldots,j)q(i,j) + (n-1)\sum_j \pi(j,\ldots,j)$$

Note that

$$\sum_j \pi(j,j,\ldots,j) = E\left\{\sum_j p(j)^n\right\} = \int_0^1 x^n\phi(x)\,\mathrm{d}x$$

In the first term on the right-hand side of this equation

$$\sum_i \pi(i,j,\ldots,j) = E\left\{\sum_i p(i)q(i,j)p(j)^{n-1}\right\} = E\{\psi_j[p(j)]p(j)^{n-1}\}$$

where $\psi_j(p(j)) = \sum_i p(i)q(i,j)$. Assume that ψ_j is independent of j. This assumption leads to the expression

$$(n-1+\theta)\int_0^1 x^n \phi(x)\,\mathrm{d}x = \theta\int_0^1 \psi(x)x^{n-1}\phi(x)\,\mathrm{d}x + (n-1) \\ \times \int_0^1 x^{n-1}\phi(x)\,\mathrm{d}x$$

which can be made into a first-order difference equation for

$$h_n = \int_0^1 x^n \phi(x)\,\mathrm{d}x$$

with the forcing term

$$k_{n-1} = \int_0^1 \psi(x)x^{n-1}\phi(x)\,\mathrm{d}x$$

$$(n-1+\theta)h_n = \theta k_{n-1} + (n-1)h_{n-1}$$

On solving this

$$h_n = \theta\sum_{r=0}^{n-1}\frac{(n-1)(n-2)\ldots(r+1)}{(n-1+\theta)\ldots(r+\theta)}k_r$$

where we set $h_1 = 1$ and $k_0 = 1$.

We can rewrite this expression as

$$h_n = \theta\sum_{r=0}^{n-1} {}_{n-1}C_r \int_0^1 u^{r+\theta-1}(1-u)^{n-r-1}\mathrm{d}u \int_0^1 \psi(v)v^r\phi(v)\,\mathrm{d}v$$

The right-hand side can be written compactly as

$$\theta\int_0^1\int_0^1 (1-u+uv)^{n-1}u^{\theta-1}\psi(v)\phi(v)\,\mathrm{d}u\,\mathrm{d}v$$

Changing variables from (u, v) to (x, v), with $x = 1-u+uv$, we convert the integral expression into

$$\theta\int_0^1 x^{n-1}(1-x)^{\theta-1}\int_0^x (1-v)^{-\theta}\psi(v)\phi(v)\,\mathrm{d}v\,\mathrm{d}x$$

Since this equals h_n for all $n \geq 1$ we identify

$$\phi(x) = \theta x^{-1}(1-x)^{\theta-1}\int_0^x (1-v)^{-\theta}\psi(v)\phi(v)\,\mathrm{d}v$$

We convert this into a differential equation for ϕ and solve it as

$$\phi(x) = Cx^{-1}(1-x)^{\theta-1}\exp\left\{-\theta\int_x^1 y^{-1}(1-y)^{-1}\psi(y)\,\mathrm{d}y\right\}$$

Kingman ([48], p. 34) shows that this frequency spectrum is approximately given by

$$\phi(x) \approx \theta x^{-1}(1-x)^{\theta-1}$$

Another example of a transient frequency spectrum is in [49].

V. LOGARITHMIC SERIES DISTRIBUTION

In this section we connect the (ordered) market share distributions with Zipf's distributions, and the like, used in the economic literature on size distributions. In [11] Kelly used a birth-and-death process with immigration to study family sizes and clustering of agents in social environments such as at cocktail parties. His models may be reinterpreted in terms of agents of different types in the spirit of this chapter. The state space of a continuous time Markov process is $n = (n_1, n_2, \ldots)$, where only a finite number of components are nonzero.

A birth in a collection of categories with j agents may be interpreted as one agent changing his type, i.e., moving from a category which contains j agents to form a new category with $(j+1)$ agents, i.e., n changes into

$$(n_1, \ldots, n_{j-1}, n_j - 1, n_{j+1} + 1, \ldots)$$

A death in a category with j agents mean that n changes into

$$(n_1, \ldots, n_{j-1} + 1, n_j - 1, n_{j+1}, \ldots)$$

The boundary condition is n changing into $(n_1 + 1, n_2, \ldots)$ when an agent of a new type enters, and a single agent departs.

In the simple case of constant birth rate λ and death rate μ with immigration rate ν Kelly has established that the equilibrium distribution given by

$$p(n) = \text{const} \times \prod_{j=1}^{\infty} \frac{\beta_j^{n_j}}{n_j!}$$

where

$$\beta_j = \nu\frac{x^j}{j}, \qquad x = \lambda/\mu$$

and where x is assumed to be less than one, is a stationary distribution for $n = (n_1, n_2, \ldots)$, where n_j is either the number of types containing exactly j agents, or the number of agents of type j, depending on the interpretation. Note that

$$E(n_j) = \beta_j$$

The random variable n_j is Poisson with mean which is inversely related to j and is called a logarithmic series distribution in [37]. See [50] for additional information of this distribution.

To extend this model to market share models, we follow [37] in part, and assume that the as are independent Poisson random variables with mean

$$\Pr(a(i) = j) = \theta \frac{x^j}{j}$$

where $\theta^{-1} = -\ln(1-x)$, with x less than one, and where $a(j)$ is as defined in Section III, i.e., it denotes the number of stores with exactly j customers.

From the assumption, we have

$$P\{a(1), a(2), \ldots, a(N)\} = \exp\left\{-\theta \sum_{i=1}^{N} \frac{x^i}{i}\right\} \prod_{j=1}^{N} \left(\frac{\theta x^j}{j}\right)^{a(i)} \frac{1}{a(i)!}$$

subject to the constraints

$$\sum_{j=1}^{N} a(j) = K$$

and

$$\sum_{j=1}^{N} ja(j) = N$$

Using the dummy variable s_j, the joint probability generating function of these random variables is

$$E\left\{\prod_{j=1}^{N} s_j^{a(j)}\right\} = \prod_{j=1}^{N} \exp\left\{\frac{\theta x^j}{j}(s_j - 1)\right\}$$

Set $s_j = s^j$. Then

$$E(s^N) = \exp\left(-\theta \sum \frac{x^j}{j}\right) \exp\left(\theta \sum \frac{(xs)^j}{j}\right)$$

since $\Sigma_j ja(j) = N$. We can use $(1-xs)^{-\theta}$ for the second exponential expression since the powers in s match up to the Nth in our generating function calculations. The coefficient of s^N yields

$$P\left(\sum ja(j) = N \right) = \exp\left(-\theta \sum \frac{x^j}{j} \right) {}_{N+\theta-1}C_N x^N$$

where we use the relation between the negative binomial and binomial coefficients

$${}_{-\theta}C_N(-x)^N = {}_{N+\theta-1}C_N x^N$$

We thus recover Ewens' formula by dividing the joint probability by the above:

$$P\{a(1), a(2), \ldots, a(N) | N\} = \frac{1}{{}_{N+\theta-1}C_N} \prod \left(\frac{\theta x^j}{j} \right)^{a(j)} \frac{1}{a(j)!}$$

By setting s_j to s, we obtain

$$E(s^K) = \exp\left(-\theta \sum \frac{x^j}{j} \right)(1-x)^{-\theta_s}$$

After replacing s_j by $s\phi^j$ and proceeding analogously, we use

$$E(s^K \phi^N) = \exp\left(-\theta \sum \frac{x^j}{j} \right)(1-x\phi)^{-\theta_s}$$

to obtain

$$E(s^K | N) = \frac{\Gamma(N+\theta s)}{\Gamma(\theta s)} \frac{\Gamma(\theta)}{\Gamma(N+\theta)}$$

By replacing s_j with $s_j\phi^j$ in the probability generating function of the as and differentiating $E(\Pi(s_j)^{a(j)}\phi^{ja(j)}|N)\, k_j$ times with respect to s_j and setting it to 1, we have

$$E\{(a(j))_{k_j} | N\} = \phi^M \prod \left(\frac{\theta}{j} \right)^{k_j} \exp\left(\theta \sum \frac{\phi^j}{j} \right) \Big/ {}_{N+\theta-1}C_N$$

where $M = \Sigma jk_j$, and where $(m)_k = m(m-1)\ldots(m-k+1)$. Up to terms of degree N in ϕ, the exponential term agrees with $(1-\phi)^{-\theta}$, in which ϕ^{N-M} has the coefficient ${}_{N-M+\theta-1}C_{N-M}$. Note that $k_1 + 2k_2 + \ldots + Nk_N = M \le N$.

A special case is

$$E(a(j) | N) = \frac{\theta}{j} {}_{N-j+\theta-1}C_{N-j} / {}_{N+\theta-1}C_N$$

which is the case with $k_j = 1$, all other ks being zero. When we evaluate the right-hand side, by approximating the factorials by Stirling's formula it is approximately equal to

$$\frac{\theta}{j}\left(\frac{1-\dfrac{j}{N+\theta-1}}{1-\dfrac{j}{N}}\right)^{N-j}\quad\left(\frac{N}{N-\theta+1}\right)^{j}\left(1-\frac{j}{N+\theta-1}\right)^{\theta-1}$$

The second and third factors converge to 1 as N become large. The last factor converges to $(1-j/N)^{\theta-1}$. Therefore we recover an approximate expression for the frequency spectrum:

$$E(a(j)\,|\,N)\approx\theta x^{-1}(1-x)^{\theta-1}\mathrm{d}x$$

where x is the relative frequency j/N and $\mathrm{d}x$ is approximated by $1/N$. Recall that we have earlier derived this frequency spectrum directly.

VI. ZIPF'S LAW AND BOSE–EINSTEIN STATISTICS

Hill establishes a connection to Zipf's law with the logarithmic series distribution in [22] which is the starting-point of Watterson's analysis described above.*

Let K/N converge to θ in probability. Hill shows that $a(j)/K$ converges in probability to $\theta(1-\theta)^{j}$ in probability. Now, if θ is distributed as a logarithmic series

$$f(\theta)=\frac{-1}{\theta\ln(1-x)}$$

for $\theta\geq(1-x)$, and zero otherwise, then

$$E\left\{\frac{a(j)}{K}\right\}=\gamma x^{j}/j$$

VII. DISCUSSIONS

This chapter proposes the use of Ewens' sampling formula to describe stationary distributions of compositions of agents into a large number of types, and describes two notions, neutrality and consistency, as sufficient conditions for the formula to be useful in economic applications. Whether these conditions are satisfied, as least asymptotically, in economic

*An equivalent description is the following: after ensuring each type is represented by at least one agent, the remaining number of agents is distributed in a Bose–Einstein way.

applications has not been tested empirically. What kind of economic phenomena or market structures possess this neutrality property? This question requires empirical testing.

We have alluded to the econometric aspects of estimating various parameters or treating the number of agents in markets as samples from a potentially much larger number of agents so that only a small number of them actually come to markets. Estimating the biomass of different species of fish from data on catches would be an interesting application.

We can go further in the directions indicated. For example, we can treat situations where sellers drop out or exit the market when their profit over some period drops below some critical level, i.e., we can introduce a pure death process into the modeling process. We do not discuss this here for space limitation. See [51] for discussions on transient distributions, and [52] for related developments on asymptotic theory of extreme order statistics.

VIII. APPENDIX

A. Order Statistics

Let $X_i, i = 1, \ldots, K$, be nonnegative exchangeable random variables which sum to one. Let $\phi(x_1, \ldots, x_K)$ be its density function.

The decreasing order statistics, denoted by $X_{(1)} \geq X_{(2)} \geq \ldots$, are such that their joint proability density function is given by

$$f(X_{(1)}, \ldots, X_{(r)}) = K(K-1)(K-2)\ldots(K-r+1)\int\ldots\int \phi(x_1, \ldots, x_K) \\ \times \mathrm{d}x_1, \ldots, \mathrm{d}x_{K-1}$$

where the integration is over the appropriate region in $(K-1)$-dimensional space, and which is shown to be given by

$$\theta^r \Gamma(\theta)\, \mathrm{e}^{\gamma\theta} g\left(\frac{1 - x_1 - \ldots - x_r}{x_r}\right) \prod_{j=1}^{r-1} x_j^{-1}$$

See [37] and [38]. In this expression γ is Euler's constant, and $g(\cdot)$ is the density of a random variable Z which has Laplace transform

$$E(\mathrm{e}^{-tZ}) = \exp\left\{\theta \int_0^1 (\mathrm{e}^{-ty} - 1) y^{-1} \mathrm{d}y\right\}$$

The product $\Gamma(\theta)e^{\gamma\theta}g(z)$ equals $z^{\theta-1}$ in the range $0 \leq z \leq 1$. Therefore the joint density for the` first r nondecreasing order statistics is given by

$$f(x_1, x_2, \ldots, x_r) = \theta^r \left(\prod_{j=1}^{r} x_j^{-1} (1 - x_1 - \ldots x_r)^{\theta-1} \right)$$

B. Exchangeable Random Partitions

This section collects some relevant results and definitions from [15] for easy reference.

A function f from a set $\{1, 2, \ldots, N\}$ onto itself induces a partition of this set: points i and j belong to the same component if and only if some k-fold iteration of f evaluated at i is the same as some m-fold iteration evaluated at j. Here we take f to be uniform over the set of all N^N possible functions, then induced partitions are exchangeable random partitions.

The main result we use in this paper is Theorem 11.14, and Proposition 11.9 in [15]. Proposition 11.9 defines the Poisson–Dirichlet distribution on the set

$$\Delta = \left\{ (p_1, p_2, \ldots : p_1 \geq p_2 \geq \ldots, \sum_j p_j \leq 1 \right\}$$

Let L_N be a map from the set of all partitions of $\{1, 2, \ldots, N\}$, denoted by $A = \{A_i\}$, into partitions of N into sums of nonincreasing integers

$$I_N = \left\{ (n_i) : n_1 \geq n_2 \geq \ldots, \sum_i n_i = N \right\}$$

where n_i is the number of elements in A_i, denoted by $|A_i|$.

As N goes to infinity, $N^{-1} I_N(R^N) \rightarrow D = (D_1 D_2, \ldots)$ almost surely, where R^N denotes the set of random partitions of positive integers up to N, and where D is an element of Δ. Its distribution is called the Poisson–Dirichlet distribution by Kingman in [19].

Define

$$Q_N(a_1, a_2, \ldots, a_N) = P\{a(R^N) = (a_1, \ldots, a_N)\}$$

Then, this expression is known as Ewens' sampling formula in the literature of population genetics,

$$Q_N(a_1, \ldots, a_N) = \frac{N!}{\theta(\theta + 1) \ldots (\theta + N - 1)} \prod_{j=1}^{N} \frac{\theta^j}{j^{a_j} a_j!}$$

In this expression, the factor,

$$\frac{N!}{\prod_j j^{a_j} a_j!}$$

is the number of permutations of $\{1, 2, \ldots, N\}$ with exactly a_j cycles of length $j, j \geq 1$. Note that $\sum_j j a_j = N$.

The parameter θ in it is related or interpreted by the fact

$$Q_2(0, 1) = \frac{1}{1+\theta}$$

where the left-hand side is the probability $P(R^2\{1, 2\})$, i.e., 1 and 2 belongs to the same set in the partition of $\{1, 2\}$. In the uniform random partition $\theta = \frac{1}{2}$ as shown by Aldous. This number has been obtained by more elementary and direct calculation by Derrida and Flyvbjerg in [16].

Let $D = (D_1, D_2, \ldots)$ be as above, governed by the Poisson–Dirichlet distribution. As we have discussed in the main part of this chapter, we should think of D_i as the fraction of agents of type i. Define

$$\phi(y)\,\mathrm{d}y = \Pr(\text{some } D_i \in (y, y + \mathrm{d}y))\}$$

This function is called the intensity or frequency spectrum function. Aldous also has a short derivation of this function given by

$$\phi(y) = \theta y^{-1}(1-y)^{\theta-1}$$

Using this function we can calculate many interesting functions. For example

$$\Pr(1, 2, \ldots, N \text{ in distinct components}) = \prod_{i=1}^{N} \frac{\theta}{\theta + i - 1}$$

In view of this and other related relations we can interpret the remainder of Ewens' sampling formula. See [17] for details.

REFERENCES

1. M Aoki. Jap Econ Rev 46:148–165, 1995.
2. M Aoki. A New Approach to Macroeconomic Modeling: Evolutionary Stochastic Dynamics, Multiple Equilibria, and Externalities as Field Effects. Cambridge University Press, 1996.
3. M Aoki, K. Miyahara. Stochastic aggregation and dynamic field effects. Univ California, Department of Economics, Working Paper No 3, 1993.
4. GS Becker. J Pol Econ 82:1063–1093, 1974.
5. GS Becker. J Pol Econ 99:1109–1116, 1990.

6. G Akerlof. Quart J Econ 94:749–775, 1980.
7. B Arthur. Self-reinforcing mechanism in economics. In PW Anderson, KJ Arrow, D Pines eds, Economy as an Evolving Complex System. Redwood City: Addison-Wesley, 1988.
8. RJ Caballero. Amer Econ Rev 82:1279–1292, 1992.
9. M Aoki. J Econ Dyn Control 18:865–877, 1994.
10. N Glance, BA Huberman. J Math Sociol 17:281–302, 1993.
11. FP Kelly. Reversibility and Stochastic Networks, New York: Wiley, 1979.
12. PK Pollett. J Appl Prob 23:880–900, 1986.
13. G Weisbuch, A Kirman, D Herreiner. Market Organization. Santa Fe Institute Working Paper 95-11-102, 1995.
14. A Kirman. Economies with interacting agents. Mimeograph. The European University Institute, San Domenico, Italy, 1995.
15. DJ Aldous. Exchangeability and Related Topics. Lecture Notes in Mathematics 117, Berlin: Springer, 1985.
16. B Derrida, H Flyvbjerg. J Physique 48:971–978, 1987.
17. WJ Ewens. Theor Pop Biol 3:87–112, 1972.
18. JFC Kingman. Ann Prob 6:183–197, 1978.
19. JFC Kingman. Proc R Soc Lond A 361:1–20, 1978.
20. GK Zipf. Human Behavior and the Principle of Least Effort. New York: Addison-Wesley, 1949.
21. H Simon, CP Bonini. Amer Econ Rev 48:607–17, 1958.
22. BM Hill. J Amer Statis Assoc 65:1220–1232, 1970.
23. KN Matsuyama, N Kiyotaki, A Matsui. Rev Econ Studies 60:283–307, 1993.
24. R Wright. J Ecol Dyn Control 19:181–206, 1995.
25. JFC Kingman. J Roy Stat Soc B-37:1–22, 1975.
26. SS Wilks. Mathematical Statistics. New York: Wiley, 1962.
27. FM Hoppe. J Math Biology 20:91–94, 1984.
28. P Whittle. Stochastic Equilibrium. New York: Wiley, 1989.
29. R Keener, E Rothman, N. Starr. Ann Stat 15:1466–1481, 1988.
30. W Chen. Appl Probab 17:611–622, 1980.
31. W Chen. J Appl Probab 18:139–147, 1981.
32. W Chen. Ann Inst Statist Math 33:405–447, 1981.
33. RJ Connor, JE Mosimann. J Am Statist Assoc 64:194–206, 1969.
34. GP Patil, SW Joshi, C Taillie. Bull Inst Inter Stat 47:497–515, 1977.
35. FM Hoppe. J Appl Prob 23:1008–1012, 1986.
36. FM Hoppe. J Math Biol 25:123–159, 1987.
37. GA Watterson. J Appl Prob 13:639–651, 1976.
38. GA Watterson, HA Guess. Theor Pop Biol 11:141–160, 1977.
39. GE Andrews. The Theory of Partitions. Reading: Addison-Wesley, 1976.
40. J Fabius. Ann Stat 3:583–587, 1973.
41. PR Halmos. Ann Math Stat 15:182–189, 1944.
42. M Abramovitz, IA Stegun. Handbook of Mathematical Functions. New York: Dover, 1968.

43. R Stanley. Enumerative Combinatorics, Vol 1. Monterey: Wadsworth and Brooks, 1986.
44. GA Watterson. Adv Appl Prob 6:217–250, 1974.
45. W J Ewens. In S Lessard, ed., Mathematical and Statistical Developments of Evolutionary Theory. Dordrecht: Kluwer, 1990.
46. B Harris. Ann Math Stat 31:1045–1062, 1960.
47. L Katz. Ann Math Stat 20:512–517, 1955.
48. JFC Kingman. Mathematics of Genetic Diversity. Philadelphia: SIAM, 1980.
49. RC Griffiths. Adv Appl Prob 11:326–354, 1979.
50. GP Patil, SW Joshi. A Dictionary and Bibliography of Discrete Distributions. Edinburgh: Oliver and Boyd, 1968.
51. SN Ethier, TG Kurtz. Adv Appl Prob 13:429–452, 1981.
52. J Galambos. The Asymptotic Theory of Extreme Order Statistics. 2nd ed. Malabar, Florida: R Krieger Pub Co, 1987.

11

Fuzzy Random Data Obtained as Vague Perceptions of Random Phenomena

Tokuo Fukuda

Otemon Gakuin University, Osaka, Japan

I. INTRODUCTION

The purpose of this chapter is to explore fuzzy random vectors (FRVs) or equivalently multidimensional fuzzy random variables which will be a theoretical basis for investigating multidimensional fuzzy stochastic processes given as an extension of scalar ones proposed by the author and his colleagues [1–4].

Suppose that a person feels the atmospheric temperature and humidity, and that he expresses the degree of comfort by using a word like "sultry," "comfortable," "very sultry," "very comfortable," etc. The linguistic data obtained from him as mentioned above have two outstanding features. One is fuzziness due to the intrinsic vagueness of the words like "sultry" and "comfortable", and another is the randomness caused by the random fluctuations of the temperature and humidity. Furthermore, it should be noted that a person feels the temperature and the humidity simultaneously and he expresses the degree of comfort synthetically. Hereafter, we call these data fuzzy random data (FRD). Then, a natural question arising with respect to FRD is evidently how to describe them mathematically and what properties they have.

There have been many publications concerned with fuzzy random variables with various definitions. See, e.g., Kwakernaak [5,6], Miyakoshi and Simbo [7], Kruse and Meyer [8], Boswell and Taylor [9], Inoue [10], Klement, Puri and Ralescu [11], Puri and Ralescu [12–14], and Uemura [15]. However, it may be fair to say that little effort has been made to investigate FRVs from the viewpoint of the vague perceptions of non-fuzzy random vectors.

In this chapter, using the set representation method of fuzzy sets, the concept of FRVs is introduced, and some of their statistical properties are investigated theoretically.

Section II is devoted to describing the basic results concerned with general fuzzy sets, and the functions of multidimensional fuzzy sets (fuzzy vectors) are also given in the same section. FRVs are introduced in Section III, where the methodology proposed by Kwakernaak [5] is adopted with the help of correspondence theory. Furthermore, in Section IV, applying multivalued logic, the statistical moments of FRVs up to second moments are proposed and some of their properties are examined theoretically. Proofs of propositions and theorems are deferred to the appendix.

II. SET REPRESENTATIONS OF FUZZY VECTORS AND MATRICES

In this chapter, a fuzzy set $\tilde{U}$ given as a vague perception of $x_o \in \mathrm{X}$ is defined by the triple

$$\tilde{U} = (\mathrm{X}, \{\tilde{U}_\alpha | \alpha \in (0,1)\}, s_{\tilde{U}}) \tag{1}$$

where X is called the basic space; $s_{\tilde{U}}$ is the predicate, i.e., $s_{\tilde{U}}: \mathrm{X} \rightarrow \mathcal{P}$ with $\mathcal{P}$ the "universe of discourse" defined by a set of statements, and assigns a statement

$$s_{\tilde{U}}(x) = \{x \text{ coincides with } x_o\} \tag{2}$$

to each element $x \in \mathrm{X}$; and $\{\tilde{U}_\alpha | \alpha \in (0,1)\}$ is the family of subsets of X satisfying

$$L_\alpha \tilde{U} \subseteq \tilde{U}_\alpha \subseteq L_{\overline{\alpha}} \tilde{U} \quad \text{for } \forall \alpha \in (0,1) \tag{3}$$

In (3), $L_\alpha \tilde{U}$ and $L_{\overline{\alpha}} \tilde{U}$ are the strong cut and the level set of $\tilde{U}$ at α defined respectively by

$$L_\alpha \tilde{U} = \{x | (x \in \mathrm{X}) \wedge ((\tilde{U})(x) > \alpha)\} \quad \text{for } \alpha \in [0,1) \tag{4}$$

$$L_{\overline{\alpha}} \tilde{U} = \{x | (x \in \mathrm{X}) \wedge ((\tilde{U})(x) \geq \alpha)\} \quad \text{for } \alpha \in (0,1] \tag{5}$$

where $(\tilde{U})(x)$ is the membership function of $\tilde{U}$ given by

$$(\tilde{U})(x) = t(s_{\tilde{U}}(x)) \tag{6}$$

and $t(*)$ in (6) is the truth function of $*$ in the sense of multivalued logic. The crisp point x_o in X, which gives the vague perception, i.e., the fuzzy set $\tilde{U}$, is called the original point of $\tilde{U}$. The definition of fuzzy sets adopted in this chapter is a modified one originally proposed by Kwakernaak [5]. Furthermore, we call the family of subsets $\{\tilde{U}_\alpha | \alpha \in (0,1)\}$ defined by (3) the set representation of a fuzzy set $\tilde{U}$. This view is slightly but not essentially different from its usual definition (see, e.g., [8] or [16]).

As seen from the definition of fuzzy sets given by (1), there may be many fuzzy sets which have the same strong cuts and level sets at every level $\alpha \in (0,1)$. This leads to the following definition of the equivalence of fuzzy sets:

Definition 2.1 Two fuzzy sets $\tilde{U}$ and $\tilde{V}$ are called equivalent, and denoted by $\tilde{U} = \tilde{V}$, if

$$L_\alpha \tilde{U} = L_\alpha \tilde{V} \quad \text{and} \quad L_{\bar{\alpha}} \tilde{U} = L_{\bar{\alpha}} \tilde{V} \quad \text{for } \forall \alpha \in (0,1) \tag{7}$$

It can be shown that if the following relation

$$L_\alpha \tilde{U} \subseteq \tilde{V}_\alpha \subseteq L_{\bar{\alpha}} \tilde{U} \quad \text{for } \forall \alpha \in (0,1) \tag{8}$$

holds, then $\tilde{U} = \tilde{V}$, where $\{\tilde{V}_\alpha | \alpha \in (0,1)\}$ is the set representations of $\tilde{V}$. Furthermore, let $\{\alpha_r^+ | r = 1, 2, \ldots\}$ and $\{\alpha_r^- | r = 1, 2, \ldots\}$ be a strictly increasing and decreasing sequence, respectively, such that

$$\alpha = \lim_{r \to \infty} \alpha_r^- \quad \text{for } \alpha \in [0,1), \qquad \text{and} \qquad \alpha = \lim_{r \to \infty} \alpha_r^+ \quad \text{for } \alpha \in (0,1] \tag{9}$$

as well as $\alpha_r^- \in (\alpha, 1)$ and $\alpha_r^+ \in (0, \alpha)$. Then,

$$L_\alpha \tilde{U} = \bigcup_{r=1}^{\infty} \tilde{U}_{\alpha_r^-} \quad \text{for } \alpha \in [0,1) \qquad \text{and}$$

$$L_{\bar{\alpha}} \tilde{U} = \bigcap_{r=1}^{\infty} \tilde{U}_{\alpha_r^+} \quad \text{for } \alpha \in (0,1] \tag{10}$$

are valid [8].

Let the basic space X of a fuzzy set $\tilde{U}$ be a metric space, and let $\{\tilde{U}_\alpha | \alpha \in (0,1)\}$ be the set representation of $\tilde{U}$. Then, the collection of all fuzzy sets in X is denoted by $\mathcal{F}(\mathrm{X})$ when it satisfies the following conditions:

1. $\displaystyle\bigcap_{\alpha \in (0,1)} \tilde{U}_\alpha = L_{\bar{1}} \tilde{U} \neq \varnothing$
2. $\tilde{U}_\alpha$ is compact for each $\alpha \in (0,1)$
3. $\operatorname{supp} \tilde{U}$ defined by

$$\operatorname{supp} \tilde{U} = \operatorname{cl} \bigcup_{\alpha \in (0,1)} \tilde{U}_\alpha = \operatorname{cl} L_0 \tilde{U} \tag{11}$$

is also compact.

Proposition 2.1 *Let $\tilde{U} = (\mathrm{X}, \{\tilde{U}_\alpha | \alpha \in (0,1)\}, s_{\tilde{U}})$ be a fuzzy set in $\mathcal{G}(\mathrm{X})$, where* X *is a metric space, and $f: \mathrm{X} \to \mathrm{Y}$ be a mapping from* X *to another metric space* Y. *Then,*

$$f(\tilde{U}) = (Y, \{f(\tilde{U}_\alpha) | \alpha \in (0,1)\}, s_{f(\tilde{U})}) \tag{12}$$

is the fuzzy set satisfying

$$f(L_\alpha \tilde{U}) = L_\alpha f(\tilde{U}) \quad \text{for } \forall \alpha \in [0,1) \tag{13}$$

and

$$f(L_{\overline{\alpha}} \tilde{U}) \subseteq L_{\overline{\alpha}} f(\tilde{U}) \quad \text{for } \forall \alpha \in (0,1] \tag{14}$$

where

$$s_{f(\tilde{U})}(v) = \{v \text{ coincides with } f(u_o)\} \quad (u_o\text{: the original point of } \tilde{U}) \tag{15}$$

Furthermore, when f is a continuous mapping, the equality in (14) holds, i.e.,

$$f(L_{\overline{\alpha}} \tilde{U}) = L_{\overline{\alpha}} f(\tilde{U}) \quad \text{for } \forall \alpha \in (0,1] \tag{16}$$

and $\tilde{U}$ is a member of $\mathcal{G}(\mathrm{X})$.

The family of fuzzy sets is denoted by $\mathcal{G}_c(\mathrm{X})$ when each element $\tilde{U}$ in $\mathcal{G}_c(\mathrm{X})$ is a member of $\mathcal{G}(\mathrm{X})$ and its set representation is composed of compact and convex sets. When X is the n-dimensional Euclidean space R^n, the fuzzy sets in $\mathcal{G}_c(\mathrm{R}^n)$ are called (n-dimensional) fuzzy vectors in this chapter.

Let us consider the fuzzy set obtained as a vague perception of a crisp $m \times n$ matrix A_o (original matrix),

$$\tilde{\mathbf{A}} = (\mathrm{R}^{m \times n}, \{\tilde{\mathbf{A}}_\alpha | \alpha \in (0,1)\}, s_{\tilde{\mathbf{A}}}) \tag{17}$$

where the basic space is given by $\mathrm{R}^{m \times n}$, and its set representation $\{\tilde{\mathbf{A}}_\alpha | \alpha \in (0,1)\}$ is given by the family of sets of $m \times n$ matrices satisfying

$$L_\alpha \tilde{\mathbf{A}} \subseteq \tilde{\mathbf{A}}_\alpha \subseteq L_{\overline{\alpha}} \tilde{\mathbf{A}} \quad \text{for } \forall \alpha \in (0,1) \tag{18}$$

with

$$L_\alpha \tilde{\mathbf{A}} = \{A | (A \in \mathrm{R}^{m \times n}) \wedge ((\tilde{\mathbf{A}})(A) > \alpha)\} \tag{19}$$

$$L_{\overline{\alpha}} \tilde{\mathbf{A}} = \{A | (A \in \mathrm{R}^{m \times n}) \wedge ((\tilde{\mathbf{A}})(A) \geq \alpha)\} \tag{20}$$

$$(\tilde{\mathbf{A}})(A) = t(s_{\tilde{\mathbf{A}}}(A)) \tag{21}$$

and

$$s_{\tilde{\mathbf{A}}}(A) = \{A \text{ coincides with } A_o\} \tag{22}$$

As given in (17), the fuzzy sets on the basic space $\mathrm{R}^{m\times n}$ will be represented by using bold-faced capital letters with a tilde. The families of fuzzy sets $\mathcal{G}(\mathrm{R}^{m\times n})$ is defined in the same manner as $\mathcal{G}(\mathrm{R}^n)$. The fuzzy set in $\mathcal{G}(\mathrm{R}^{m\times n})$ is called a fuzzy matrix in this chapter.

III. FUZZY RANDOM VECTORS AND MATRICES

A. Fuzzy Random Vectors

Let $(\Omega, \mathcal{A}, P)$ be a complete probability space, where $\mathcal{A}$ is the σ-algebra generated by the subsets of Ω, and P is a non-atomic probability measure. Let $x_o \in \mathrm{R}^n$ be a random vector defined on this probability space. Assume now that we perceive the value of x_o through a set of rectangles $\{W_i | i \in J\}$ with J a finite or countable set, each representing a rectangle in R^n, such that

$$W_i \bigcap W_j = \varnothing \quad \text{for } i \neq j \quad \text{and} \quad \bigcup_{i\in J} W_i = \mathrm{R}^n$$

Let us define the special type of fuzzy set given by

$$\tilde{I}_i = (\mathrm{R}^n, \{\tilde{I}_{i,\alpha} | \alpha \in (0,1)\}, s_{\tilde{I}_i}) \tag{23}$$

where

$$s_{\tilde{I}_i}(x) = \{x \text{ coincides with } x_o \text{ in } W_i\}$$

with $\tilde{I}_{i,\alpha} = W_i$ for $\forall \alpha \in (0,1)$, and

$$t(s_{\tilde{I}_i}(x)) = (\tilde{I}_i)(x) = \begin{cases} 1 & \text{if } x \in W_i \\ 0 & \text{otherwise} \end{cases}$$

Then, we can consider that knowing in what rectangle the point x_o exists is equivalent to getting the fuzzy set defined by (23). Therefore, the mapping

$$\tilde{x} : \Omega \rightarrow \{\tilde{I}_i | i \in J\}$$

characterizes a special type of FRVs, and this implies we can define FRVs by replacing the set representations $\{\tilde{I}_{i,\alpha} | \alpha \in (0,1)\} (i \in J)$ by the collection of general set representations.

In order to define FRVs mathematically, first define the reduced probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ generated by $\mathcal{F} \subset \mathcal{A}$, where $\tilde{\Omega}$ is the quotient space of Ω, i.e.,

$$\tilde{\Omega} = \Omega/\sim \tag{24}$$

with the equivalence relation

$$\omega_1 \sim \omega_2 \Leftrightarrow [\forall \mathrm{A} \in \mathcal{F}, \omega_1 \in \mathrm{A} \Leftrightarrow \omega_2 \in \mathrm{A}] \tag{25}$$

$\bar{\mathcal{A}}$ is the reduction of $\mathcal{A}$ defined by

$$\bar{\mathcal{A}} = \{\bar{\mathrm{A}} \mid c^{-1}(\bar{\mathrm{A}}) \in \mathcal{A}\} \tag{26}$$

where $c(\cdot)$ is the canonical projection, i.e.,

$$c(\omega_1) = c(\omega_2) = \bar{\omega} \in \bar{\Omega} \quad \text{if } \omega_1 \sim \omega_2 \quad (\omega_1, \omega_2 \in \Omega) \tag{27}$$

and $\bar{P}$ is the probability measure on $\bar{\mathrm{A}}$ given by

$$\bar{P}(\bar{\mathrm{A}}) = P(c^{-1}(\bar{\mathrm{A}})) \quad \text{for } \forall \bar{\mathrm{A}} \in \bar{\mathcal{A}} \tag{28}$$

Let $\mathcal{K}_c^n$ be the collection of all the nonempty compact convex subsets of R^n, and $\mathcal{B}(\mathcal{K}_c^n)$ stand for the Borel family of $\mathcal{K}_c^n$, i.e., the minimum σ-algebra generated by the open subsets of $\mathcal{K}_c^n$, where the topology of $\mathcal{K}_c^n$ is generated by the Hausdorff metric d_H given by

$$d_H(\mathrm{A}, \mathrm{B}) = \max\left\{ \sup_{\mathrm{u} \in \mathrm{A}} \inf_{\mathrm{v} \in \mathrm{B}} \|\mathrm{u} - \mathrm{v}\|, \sup_{\mathrm{u} \in \mathrm{B}} \inf_{\mathrm{v} \in \mathrm{A}} \|\mathrm{u} - \mathrm{v}\| \right\} \tag{29}$$

for arbitrary elements A and B in $\mathcal{K}_c^n$. Then, we have the following definition.

Definition 3.1 The FRV $\tilde{x}$ obtained as the vague perception of an ordinary random vector x_o is defined by the mapping $\tilde{x} : \bar{\Omega} \rightarrow \mathcal{G}_c(\mathrm{R}^n)$ with

$$\tilde{x}(\bar{\omega}) = (\mathrm{R}^n, \{\tilde{x}_\alpha(\bar{\omega}) \mid \bar{\omega} \in \bar{\Omega}, \alpha \in (0, 1)\}, s_{\tilde{x}}) \tag{30}$$

where the set representation $\{\tilde{x}_\alpha(\bar{\omega}) \mid \bar{\omega} \in \bar{\Omega}, \alpha \in (0, 1)\}$ is composed of $\bar{\mathcal{A}}$-measurable correspondences from $\bar{\Omega}$ to R^n [17]; and $s_{\tilde{x}}$ is the predicate associated with the statement

$$s_{\tilde{x}}(x) = \{x(\omega) = x_o(\omega) \text{ a.s. in } \Omega\} \tag{31}$$

The measurability of $\tilde{x}$ is given through that of the function $\tilde{x}_\alpha$ from $\bar{\Omega}$ to $\mathcal{B}(\mathcal{K}_c^n)$, i.e.,

$$\tilde{x}_\alpha^{-1}(\mathrm{E}) \in \bar{\mathcal{A}} \quad \text{for every } \mathrm{E} \in \mathcal{B}(\mathcal{K}_c^n) \tag{32}$$

A correspondence $\tilde{x}_\alpha$ between $\bar{\Omega}$ and R^n is a relation between them whose domain is the whole of $\bar{\Omega}$. Thus, $\tilde{x}_\alpha$ can be considered as a function from $\bar{\Omega}$ to the family of the subsets of R^n with the property that $\tilde{x}_\alpha(\bar{\omega}) \neq \emptyset$ for all $\bar{\omega} \in \bar{\Omega}$.

Definition 3.2 The random vector x_o, of which the FRV $\tilde{x}$ is a fuzzy perception, is called the original random vector of the FRV. We denote the admissible set of all possible original random vectors by χ.

Throughout this chapter, we assume that the following conditions hold:

(A-1) $x_{\mathrm{o}} \in \chi$ and $x_{\mathrm{o}} \in \operatorname{supp} \tilde{x}$ (x_{o}: the original random vector)

(A-2) $\left\{ \xi \,\middle|\, (\xi \in \chi) \wedge \left(\xi \in \bigcap_{\alpha \in (0,1)} \tilde{x}_\alpha \text{ a.s. in } \Omega \right) \right\} \neq \emptyset$

The condition **(A-1)** is surely natural since FRVs are obtained as the vague perception of random phenomena. **(A-2)** implies that the correspondence $\tilde{x}_\alpha$ has at least one almost everywhere selection. For more detailed aspects of almost everywhere selections of correspondences see, e.g., [18] and [17].

Proposition 3.1 *The following four measurabilities are equivalent*:

(a) $\tilde{x}_\alpha^{-1}(\mathrm{E}) \in \bar{\mathcal{A}}$ for every $\mathrm{E} \in \mathcal{B}(\mathcal{K}_c^n)$
(b) $\tilde{x}_\alpha^w(\mathrm{G}) = \{\tilde{\omega} \in \tilde{\Omega} \mid \tilde{x}_\alpha(\tilde{\omega}) \cap \mathrm{G} \neq \emptyset\} \in \bar{\mathcal{A}}$ for every $\mathrm{G} \in \mathcal{B}(\mathrm{R}^n)$
(c) $\tilde{x}_\alpha^s(\mathrm{G}) = \{\tilde{\omega} \in \tilde{\Omega} \mid \tilde{x}_\alpha(\tilde{\omega}) \subset \mathrm{G}\} \in \bar{\mathcal{A}}$ for every $\mathrm{G} \in \mathcal{B}(\mathrm{R}^n)$
(d) $\mathrm{Gr}(\tilde{x}_\alpha) = \{(\tilde{\omega}, x) \in \tilde{\Omega} \times \mathrm{R}^n \mid x \in \tilde{x}_\alpha(\tilde{\omega})\} \in \bar{\mathcal{A}} \times \mathcal{B}(\mathrm{R}^n)$

where $\bar{\mathcal{A}} \times \mathcal{B}(\mathrm{R}^n)$ *is the product σ-algebra on* $\tilde{\Omega} \times \mathrm{R}^n$, *i.e., the smallest σ-algebra containing all the products of the form* $\bar{\mathrm{A}} \times \mathrm{B}$ *with* $\bar{\mathrm{A}} \in \bar{\mathcal{A}}$ *and* $\mathrm{B} \in \mathcal{B}(\mathrm{R}^n)$.

Since $\tilde{x}_\alpha$ is a correspondence from $\tilde{\Omega}$ to R^n, the composition $\tilde{x}_\alpha \circ c$ is a correspondence from Ω to R^n. Therefore, the following corollary also holds:

Corollary 3.1 *Let* x_α *be a correspondence defined by the composition* $\tilde{x}_\alpha \circ c$. *Then, the following four measurabilities are equivalent*:

(a) $x_\alpha^{-1}(\mathrm{E}) \in \mathcal{A}$ for every $\mathrm{E} \in \mathcal{B}(\mathcal{K}_c^n)$
(b) $x_\alpha^w(\mathrm{G}) = \{\omega \in \Omega \mid x_\alpha(\omega) \cap \mathrm{G} \neq \emptyset\} \in \mathcal{A}$ for every $\mathrm{G} \in \mathcal{B}(\mathrm{R}^n)$
(c) $x_\alpha^s(\mathrm{G}) = \{\omega \in \Omega \mid x_\alpha(\omega) \subset \mathrm{G}\} \in \mathcal{A}$ for every $\mathrm{G} \in \mathcal{B}(\mathrm{R}^n)$
(d) $\mathrm{Gr}(x_\alpha) = \{(\omega, x) \in \Omega \times \mathrm{R}^n \mid x \in x_\alpha(\omega)\} \in \mathcal{A} \times \mathcal{B}(\mathrm{R}^n)$

where $\mathcal{A} \times \mathcal{B}(\mathrm{R}^n)$ *is the product σ-algebra on* $\Omega \times \mathrm{R}^n$.

Theorem 3.1 *Let* $\tilde{x}$ *be an FRV and* $\{\tilde{x}_\alpha \mid \alpha \in (0,1)\}$ *its set representation. Then, the following properties hold*:

(i) $L_\alpha \tilde{x}(\tilde{\omega})$ *is an* $\bar{\mathcal{A}}$*-measurable correspondence for* $\forall \alpha \in (0,1)$.
(ii) $\{L_{\bar{\alpha}} \tilde{x}(\tilde{\omega}) \mid \tilde{\omega} \in \tilde{\Omega}, \alpha \in (0,1)\}$ *is a set representation of the FRV equivalent to* $\tilde{x}$.
(iii) $\operatorname{supp} \tilde{x}(\tilde{\omega})$ *is an* $\bar{\mathcal{A}}$*-measurable correspondence.*
(iv) $L_{\mathrm{T}} \tilde{x}(\tilde{\omega})$ *is an* $\bar{\mathcal{A}}$*-measurable correspondence.*

Theorem 3.2 *Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Then, for each $\alpha \in (0,1)$ and $i \in \{1, 2, \ldots, n\}$,* $\inf \tilde{x}_{\alpha,i}(\tilde{\omega})$ *and* $\sup \tilde{x}_{\alpha,i}(\tilde{\omega})$ *are $\tilde{\mathcal{A}}$-measurable random variables, where*

$$\tilde{x}_{\alpha,i}(\tilde{\omega}) = \{x_i | x \in \tilde{x}_\alpha(\tilde{\omega}), x = (x_1, \ldots, x_n)'\} \tag{33}$$

Let $\tilde{x}$ be a scalar FRV, i.e., a fuzzy random variable. Then, using Theorem 3.1(iii) and the same procedure for the proof of Theorem 3.2, we can confirm $\tilde{\mathcal{A}}$- measurability of $\inf(\operatorname{supp} \tilde{x})$ and $\sup(\operatorname{supp} \tilde{x})$. Then, the following corollary is obtained.

Corollary 3.2 *Let $\tilde{x}$ be a fuzzy random variable. Then,*

(i) $\inf L_\alpha \tilde{x}$ *and* $\sup L_\alpha \tilde{x}$ *are $\tilde{\mathcal{A}}$-measurable for each $\alpha \in (0,1)$.*
(ii) $\inf L_{\bar{\alpha}} \tilde{x}$ *and* $\sup L_{\bar{\alpha}} \tilde{x}$ *and $\tilde{\mathcal{A}}$-measurable for each $\alpha \in (0,1]$.*
(iii) $\inf(\operatorname{supp} \tilde{x})$ *and* $\sup(\operatorname{supp} \tilde{x})$ *are $\tilde{\mathcal{A}}$-measurable.*

B. Fuzzy Random Matrices

Let $\mathcal{K}^{m\times n}$ be a collection of all the non-empty compact subsets of $\mathrm{R}^{m\times n}$, and $\mathcal{B}(\mathcal{K}^{m\times n})$ stand for the Borel family of $\mathcal{K}^{m\times n}$, where the topology of $\mathcal{K}^{m\times n}$ is generated by the Hausdorff metric d_H given by (29) with the matrix norm

$$\|A\| = \sqrt{\lambda_{\max}(A'A)} \tag{34}$$

instead of the vector one, where $\lambda_{\max}$ is the maximum eigenvalue of $A'A$. Then, fuzzy random matrices (FRMs) are defined in the same manner as that for FRVs:

Definition 3.3 The FRM $\tilde{\mathrm{M}}$ obtained as the vague perception of an ordinary random matrix M_o is defined by the mapping $\tilde{\mathrm{M}}: \tilde{\Omega} \to \mathcal{G}(\mathrm{R}^{m\times n})$ with

$$\tilde{\mathrm{M}}(\tilde{\omega}) = (\mathrm{R}^{m\times n}, \{\tilde{\mathrm{M}}_\alpha | \alpha \in (0,1)\}, s_{\tilde{\mathrm{M}}}) \tag{35}$$

where $\{\tilde{\mathrm{M}}_\alpha(\tilde{\omega}) | \alpha \in (0,1)\}$ is composed of $\tilde{\mathcal{A}}$-measurable correspondences from $\tilde{\Omega}$ to $\mathcal{B}(\mathcal{K}^{m\times n})$; and

$$s_{\tilde{M}}(M) = \{M(\omega) = M_o(\omega) \text{ a.s. in } \Omega\} \tag{36}$$

Let $\tilde{x}$ be an n-dimensional FRV. Then, as a kind of fuzzy version of the matrix xx' generated by the ordinary random vector x, the fuzzy set $\tilde{x} \boxdot \tilde{x}'$ generated by $\tilde{x}$ is given as follows:

$$\tilde{x} \boxdot \tilde{x}' = (\mathrm{R}^{n\times n}, \{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha | \alpha \in (0,1)\}, s_{\tilde{x}\boxdot\tilde{x}'}) \tag{37}$$

where $\tilde{x}_\alpha$ is the element of the set representation of $\tilde{x}$,

$$\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha = \{W \mid W = xx', x \in \tilde{x}_\alpha\} \tag{38}$$

and

$$s_{\tilde{x} \boxdot \tilde{x}'}(W) = \{W \text{ coincides with } x_\mathrm{o} x'_\mathrm{o}\} \tag{39}$$

Proposition 3.2 *Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha \mid \alpha \in (0,1)\}$ its set representation. Then, the fuzzy set given by (37) is a fuzzy random matrix, i.e., $\tilde{x} \boxdot \tilde{x}'$ is the element of $\mathcal{G}(\mathrm{R}^{n \times n})$ and*

$$(\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha)^{-1}(\mathrm{E}) \in \bar{\mathcal{A}} \quad \textit{for every } \mathrm{ER} \in \mathcal{B}(\mathcal{K}^{n \times n}) \tag{40}$$

Furthermore, $\tilde{x} \boldsymbol{\boxdot} \tilde{x}'$ satisfies

$$L_\alpha(\tilde{x} \boxdot \tilde{x}') = (L_\alpha \tilde{x}) \boxdot (L_\alpha \tilde{x})' \quad \text{for } \forall \alpha \in [0,1) \tag{41}$$

and

$$L_{\overline{\alpha}}(\tilde{x} \boldsymbol{\boxdot} \tilde{x}') = (L_{\overline{\alpha}} \tilde{x}) \boldsymbol{\boxdot} (L_{\overline{\alpha}} \tilde{x})' \quad \text{for } \forall \alpha \in (0,1] \tag{42}$$

IV. STATISTICAL PROPERTIES OF FRVs

A. Expectation of FRVs

Let ξ be a possible original random variable of the FRV $\tilde{x}$, i.e., a random vector in χ. In order to investigate the statistical properties of $\tilde{x}$, consider first the following statement:

$$s_{\tilde{x}}(\xi) = \{\xi \text{ coincides with the original random vector } x_\mathrm{o} \text{ a.s. in } \Omega\} \tag{43}$$

Then, the above statement $s_{\tilde{x}}(\xi)$ is rewritten as

$$\begin{aligned} s_{\tilde{x}}(\xi) = \wedge \{\xi(\omega) = x_\mathrm{o}(\omega) \mid \omega \in \Omega \text{ except for } \forall \omega \in \mathrm{A} \in \mathcal{A} \\ \text{such that } \xi(\omega) \neq x_\mathrm{o}(\omega) \text{ and } P(\mathrm{A}) = 0\} \end{aligned} \tag{44}$$

Using the results given by Bellman and Giertz [19], it can be found that the truth value of $s_{\tilde{x}}(\xi)$ is evaluated as

$$t(s_{\tilde{x}}(\xi)) = \operatorname*{essinf}_{\omega \in \Omega} \{(\tilde{x}(\bar{\omega}))(\xi(\omega))\} \tag{45}$$

where

$$\begin{aligned} \operatorname*{essinf}_{\omega \in \Omega} \{(\tilde{x}(\bar{\omega}))(\xi(\omega))\} = \inf \{(\tilde{x}(\bar{\omega}))(\xi(\omega)) \mid \omega \in \Omega, \bar{\omega} = c(\omega) \text{ except for} \\ \forall \omega \in \mathrm{A} \in \mathcal{A} \text{ such that } \xi(\omega) \neq x_\mathrm{o}(\omega) \\ \text{and } P(\mathrm{A}) = 0\} \end{aligned} \tag{46}$$

Incidentally, the proposition

$$s_{E\{\tilde{x}\}}(x) = \{x \text{ coincides with the expectation of } x_o\} \tag{47}$$

is given by

$$s_{E\{\tilde{x}\}}(x) = \bigvee_{\xi \in \chi} \{s_{\tilde{x}}(\xi) \wedge (E\{\xi\} = x)\} \tag{48}$$

Then, the truth value of the composite proposition $s_{E\{\tilde{x}\}}(x)$ is given by

$$t(s_{E\{\tilde{x}\}}(x)) = \sup_{\xi \in \chi} \left\{ \left(\operatorname*{essinf}_{\omega \in \Omega} (\tilde{x}(\bar{\omega}))(\xi(\omega)) \right) \wedge (E\{\xi\} = x) \right\} \tag{49}$$

with $\sup \varnothing = 0$.* We introduce here the Aumann-like integral (abbreviated as AL-integral) given by

$$(\mathrm{AL}) \int \tilde{x}_\alpha \, \mathrm{d}P = \left\{ x \,\middle|\, x = \int \xi \, \mathrm{d}P, \xi \in S(\tilde{x}_\alpha) \right\} \tag{50}$$

where $\tilde{x}_\alpha$ is an arbitrary element of the set representation of $\tilde{x}$, and $S(\tilde{x}_\alpha)$ is the selection set of $\tilde{x}_\alpha$, i.e.,

$$S(\tilde{x}_\alpha) = \{\xi \mid \xi \in \chi; \xi(\omega) \in \tilde{x}_\alpha(\bar{\omega}) \text{ a.s. in } \Omega; \bar{\omega} = c(\omega)\} \tag{51}$$

Using the AL-integral given above, the expectation of $\tilde{x}_\alpha$ is defined by

$$E\{\tilde{x}_\alpha\} = (\mathrm{AL}) \int \tilde{x}_\alpha \, \mathrm{d}P \tag{52}$$

When the admissible set of possible original vectors χ is composed of integrable random vectors satisfying the condition **(A-2)**, it is easy to show that

$$E\{\tilde{x}_\alpha\} \neq \varnothing \quad \text{for each } \alpha \in (0,1) \tag{53}$$

Then, we can confirm the following property holds.

Proposition 4.1 *Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Assume that every element of χ is integrable. Then,*

$$\{x \,|\, t(s_{E\{\tilde{x}\}}(s)) > \alpha\} \subseteq E\{\tilde{x}_\alpha\} \subseteq \{x \,|\, t(s_{E\{\tilde{x}\}}(x)) \geq \alpha\} \quad \text{for } \forall \alpha \in (0,1) \tag{54}$$

Therefore, the following definition of the expectation of a FRV $\tilde{x}$ should be reasonable.

Definition 4.1 Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation.

*Hereafter, we always consider $\sup \varnothing = 0$.

Assume that every element of χ is integrable. Then the expectation of the FRV $\tilde{x}$ is defined by

$$E\{\tilde{x}\} = (\mathrm{R}^n, \{E\{\tilde{x}_\alpha\} \,|\, \alpha \in (0,1)\}, s_{E\{\tilde{x}\}}), \tag{55}$$

where $s_{E\{\tilde{x}\}}$ is the predicate associated with the statement given by (47), and $\{E\{\tilde{x}_\alpha\} \,|\, \alpha \in (0,1)\}$ is the set representation of $E\{\tilde{x}\}$ given through (52).

It is clear from (54) that

$$L_\alpha E\{\tilde{x}\} \subseteq E\{\tilde{x}_\alpha\} \subseteq L_{\overline{\alpha}} E\{\tilde{x}\} \quad \text{for } \forall \alpha \in (0,1) \tag{56}$$

with

$$L_\alpha E\{\tilde{x}\} = \{x \,|\, (E\{\tilde{x}\})(x) > \alpha\} \tag{57}$$

$$L_{\overline{\alpha}} E\{\tilde{x}\} = \{x \,|\, (E\{\tilde{x}\})(x) \geq \alpha\} \tag{58}$$

and

$$(E\{\tilde{x}\})(x) = t(s_{E\{\tilde{x}\}}(x)) \tag{59}$$

which means that $\{E\{\tilde{x}_\alpha\} \,|\, \alpha \in (0,1)\}$ surely satisfies the condition imposed on set representations of fuzzy sets.

The following corollary is obtained immediately from the proof of Proposition 4.1.

Corollary 4.1 *Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Assume that every element of χ is integrable. Then $\{E\{L_\alpha \tilde{x}\} \,|\, \alpha \in (0,1)\}$ and $\{E\{L_{\overline{\alpha}} \tilde{x}\} \,|\, \alpha \in (0,1)\}$ are the set representations of the fuzzy set equivalent to $E\{\tilde{x}\}$, and they satisfy*

$$L_\alpha E\{\tilde{x}\} \subseteq E\{L_\alpha \tilde{x}\} \subseteq E\{L_{\overline{\alpha}} \tilde{x}\} \subseteq L_{\overline{\alpha}} E\{\tilde{x}\} \tag{60}$$

By using Lyapunov's convexity theorem (see, e.g., [17]), we have the following result.

Proposition 4.2 *If the admissible set of possible original random vectors χ consists of all integrable ones, then $E\{\tilde{x}\}$ is convex.*

For the convexity of $E\{\tilde{x}\}$, we have another result:

Proposition 4.3 *Let $\tilde{x}$ be a FRV and χ also convex, i.e.,* if *$\xi_1 \in \chi$ and $\xi_2 \in \chi$, then $\xi = \lambda \xi_1 + (1-\lambda)\xi_2 \in \chi$ for $\lambda \in [0,1]$. Then, $E\{\tilde{x}\}$ is convex, if $E\{\tilde{x}\}$ exists.*

In order to investigate further the properties of the expected FRVs, we have to introduce the concept of integrably boundedness of χ.

Definition 4.2 The admissible set of possible original random vectors

χ is L^n-integrably bounded if χ consists of all the random vectors satisfying

$$\|\xi(\omega)\|^n \leq h(\omega) \quad \text{a.s. in } \Omega \tag{61}$$

where $h(\omega)$ is the random variable such that

$$\int h \, \mathrm{d}P < +\infty \tag{62}$$

Theorem 4.1 *Let $\tilde{x}$ be a fuzzy random vector, i.e., $\tilde{x} \in \mathcal{G}_c(\mathrm{R}^n)$, and assume that χ satisfies the following conditions:*

(i) *χ is convex*

(ii) *χ is composed of the random vectors satisfying (61) and (62) with $n = 1$*

(iii) *χ is closed with respect to convergent sequences, i.e., if $u_j \rightarrow u$ a.s. in $\Omega(u_j \in \chi)$, then u is the element of χ.*

Then, $E\{\tilde{x}\}$ is a fuzzy vector, $E\{\tilde{x}\} \in \mathcal{G}_c(\mathrm{R}^n)$.

Theorem 4.2 *Let $\tilde{x}$ be a fuzzy random vector. Assume that χ is L^1-integrably bounded. Then, $E\{\tilde{x}\}$ is a fuzzy vector, and it satisfies*

$$\mathrm{cl}\, L_\alpha E\{\tilde{x}\} = E\{\mathrm{cl}\, L_\alpha \tilde{x}\} \tag{63}$$

and

$$L_{\bar{\alpha}} E\{\tilde{x}\} = E\{L_{\bar{\alpha}} \tilde{x}\} \tag{64}$$

for every $\alpha \in (0, 1)$.

B. Second Moments of FRVs

Since $\tilde{x} \boxdot \tilde{x}'$ is a fuzzy random matrix when $\tilde{x}$ is a FRV as shown in Proposition 3.2, it is possible to consider the second moment of $\tilde{x}$. Let x_o be the original random vector of a FRV $\tilde{x}$. Then, the statement "M coincides with $E\{x_o x_o'\}$" is given by

$$s_{E\{\tilde{x}\boxdot\tilde{x}'\}}(M) = \bigvee_{\xi \in \chi} \{s_{\tilde{x}}(\xi) \wedge (E\{\xi\xi'\} = M)\} \tag{65}$$

where $s_{\tilde{x}}(\xi)$ is the proposition given by (43). Using the same procedure as for $s_{E\{\tilde{x}\}}$ given by (48), the truth value of the composite statement $s_{E\{\tilde{x}\boxdot\tilde{x}'\}}(M)$ is given by

$$t(s_{E\{\tilde{x}\boxdot\tilde{x}'\}}(M)) = \sup_{\xi \in \chi} \left\{ \left(\operatorname*{essinf}_{\omega \in \Omega} (\tilde{x}(\bar{\omega}))(\xi(\omega)) \right) \wedge (E\{\xi\xi'\} = M) \right\} \tag{66}$$

Therefore, using the second moment $E\{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha\}$ of the correspondence $\tilde{x}_\alpha$, i.e.,

$$E\{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha\} = (\mathrm{AL}) \int \tilde{x}_\alpha \boxdot \tilde{x}'_\alpha \mathrm{d}P \tag{67}$$

the following relation is confirmed.

Proposition 4.4 *Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Assume that each element ξ of χ is square integrable, i.e.,*

$$\int \|\xi\|^2 \mathrm{d}P < +\infty \quad \text{for } \forall \xi \in \chi \tag{68}$$

Then,

$$\{M | t(s_{E\{\tilde{x}\boxdot\tilde{x}'\}}(M)) > \alpha\} \subseteq E\{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha\} \subseteq \{M | t(s_{E\{\tilde{x}\boxdot\tilde{x}'\}}(M)) \geq \alpha\}$$
$$\text{for } \forall \alpha \in (0,1) \tag{69}$$

Definition 4.3 Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Assume that every element in χ is square integrable. Then, the second moment $E\{\tilde{x} \boxdot \tilde{x}'\}$ of $\tilde{x}$ is defined by

$$E\{\tilde{x} \boxdot \tilde{x}'\} = (\mathrm{R}^{n\times n}, \{E\{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha\} | \alpha \in (0,1)\}, s_{E\{\tilde{x}\boxdot\tilde{x}'\}}) \tag{70}$$

where $s_{E\{\tilde{x}\boxdot\tilde{x}'\}}$ is the predicate associated with the statement given by (65), and $\{E\{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha\} | \alpha \in (0,1)\}$ is the set representation of $E\{\tilde{x} \boxdot \tilde{x}'\}$ given through (67).

Furthermore, the proposition "M coincides with the variance of the original random vector x_o of $\tilde{x}$" should be given by

$$s_{\mathrm{var}\tilde{x}}(M) = \bigvee_{\xi \in \chi} \{s_{\tilde{x}}(\xi) \wedge (\mathrm{var}\, \xi = M)\} \tag{71}$$

where

$$\mathrm{var}\, \xi = E\{(\xi - E\{\xi\})(\xi - E\{\xi\})'\} \tag{72}$$

Then, using the same procedure as for $s_{E\{\tilde{x}\boxdot\tilde{x}'\}}$ defined by (65), the truth value of the composite proposition $s_{\mathrm{var}\tilde{x}}(M)$ is given by

$$t(s_{\mathrm{var}\tilde{x}}(M)) = \sup_{\xi \in \chi} \left\{ \left(\operatorname*{essinf}_{\omega \in \Omega} (\tilde{x}(\tilde{\omega}))(\xi(\omega)) \right) \wedge (\mathrm{var}\, \xi = M) \right\} \tag{73}$$

The following proposition is proved by using the same procedure as that for Proposition 4.4.

Proposition 4.5 *Let $\tilde{x}$ be a FRV and $\{\bar{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Assume that each element of χ is square integrable. Then, we have*

$$\{M | t(s_{\mathrm{var}}\bar{x}(M)) > \alpha\} \subseteq \operatorname{var}\tilde{x}_\alpha \subseteq \{M | t(s_{\mathrm{var}\,\bar{x}}(M)) \geq \alpha\} \quad \forall \alpha \in (0,1) \tag{74}$$

where

$$\operatorname{var}\tilde{x}_\alpha = E\{(\tilde{x}_\alpha \boxminus E\{\tilde{x}_\alpha\}) \boxdot (\bar{x}_\alpha \boxminus E\{\tilde{x}_\alpha\})'\} \tag{75}$$

and

$$\tilde{x}_\alpha \boxminus E\{\tilde{x}_\alpha\} = \{x | x = \xi - E\{\xi\}, \xi \in \tilde{x}_\alpha\} \tag{76}$$

Definition 4.4 Let $\tilde{x}$ be a FRV and $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ its set representation. Assume that every element of χ is square integrable. Then, the variance of $\tilde{x}$ is defined by

$$\operatorname{var}\tilde{x} = (\mathrm{R}^{n\times n}, \{\operatorname{var}\tilde{x}_\alpha | \alpha \in (0,1)\}, s_{\mathrm{var}\,\tilde{x}}) \tag{77}$$

where $s_{\mathrm{var}\,\tilde{x}}$ is the predicate associated with the statement given by (71), and $\{\operatorname{var}\tilde{x}_\alpha | \alpha \in (0,1)\}$ is the set representation of $\operatorname{var}\tilde{x}$ given through (75).

Proposition 4.6 *Let $\tilde{x}$ be a FRV and assume that each element of χ is square integrable. Then $\{E\{L_\alpha(\tilde{x} \boxdot \tilde{x}')\} | \alpha \in (0,1)\}$ and $\{E\{L_{\overline{\alpha}}(\tilde{x} \boxdot \tilde{x}')\} | \alpha \in (0,1)\}$ are the set representations of the fuzzy sets equivalent to $E\{\tilde{x} \boxdot \tilde{x}'\}$, and they satisfy*

$$\begin{aligned} L_\alpha E\{\tilde{x} \boxdot \tilde{x}'\} \subseteq E\{L_\alpha(\tilde{x} \boxdot \tilde{x}')\} \subseteq E\{\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha\} \subseteq E\{L_{\overline{\alpha}}(\tilde{x} \boxdot \tilde{x}')\} \\ \subseteq L_{\overline{\alpha}} E\{\tilde{x} \boxdot \tilde{x}'\} \end{aligned} \tag{78}$$

for each $\alpha \in (0,1)$. Furthermore, $\operatorname{var} L_\alpha\tilde{x}$ *and* $\operatorname{var} L_{\overline{\alpha}}\tilde{x}$ *are also the set representations of the fuzzy sets equivalent to* $\operatorname{var}\tilde{x}$ *and they satisfy*

$$L_\alpha(\operatorname{var}\tilde{x}) \subseteq \operatorname{var} L_\alpha\tilde{x} \subseteq \operatorname{var}\tilde{x}_\alpha \subseteq \operatorname{var} L_{\overline{\alpha}}\tilde{x} \subseteq L_{\overline{\alpha}}(\operatorname{var}\tilde{x}) \tag{79}$$

for each $\alpha \in (0,1)$, where $\operatorname{var} L_\alpha\tilde{x}$ *and* $\operatorname{var} L_{\overline{\alpha}}\tilde{x}$ *are given by (75) with $\tilde{x}_\alpha = L_\alpha\tilde{x}$ and $\tilde{x}_\alpha = L_{\overline{\alpha}}\tilde{x}$, respectively.*

Theorem 4.3 *Let $\tilde{x}$ be a FRV, and assume that the admissible set of possible original random vectors χ satisfies the following conditions:*

(i) *χ is convex*
(ii) *χ is composed of the random vectors satisfying (61) and (62) with $n = 2$.*
(iii) *χ is closed with respect to convergent sequences, i.e., if $u_j \rightarrow u$* a.s. *in $\Omega(u_j \in \chi)$, then u is the element of χ.*

Then, $E\{\tilde{x} \boxdot \tilde{x}'\}$ *and* $\operatorname{var}\tilde{x}$ *are fuzzy matrices, i.e.., they are the elements of* $\mathcal{G}(\mathrm{R}^{n\times n})$.

Theorem 4.4 *Let* $\tilde{x}$ *be a FRV, and assume that* χ *is* L^2*-integrably bounded. Then,* $E\{\tilde{x} \boxdot \tilde{x}'\}$ *and* $\operatorname{var}\tilde{x}$ *are the elements of* $\mathcal{G}(\mathrm{R}^{n\times n})$ *and they satisfy*

$$\mathrm{cl}\, L_\alpha E\{\tilde{x} \boxdot \tilde{x}'\} = E\{\mathrm{cl}((L_\alpha\tilde{x}) \boxdot (L_\alpha\tilde{x})')\} \tag{80}$$

$$L_{\overline{\alpha}} E\{\tilde{x} \boxdot \tilde{x}'\} = E\{(L_{\overline{\alpha}}\tilde{x}) \boxdot (L_{\overline{\alpha}}\tilde{x})'\} \tag{81}$$

$$\mathrm{cl}\, L_\alpha(\operatorname{var}\tilde{x}) = \operatorname{var}(\mathrm{cl}\, L_\alpha\tilde{x}) \tag{82}$$

$$L_{\overline{\alpha}}(\operatorname{var}\tilde{x}) = \operatorname{var}(L_{\overline{\alpha}}\tilde{x}) \tag{83}$$

for every $\alpha \in (0,1)$.

With the help of Proposition 3.2, the propositions and theorems in this subsection are proved by using the same procedures as those for $E\{\tilde{x}\}$ in Subsection IV.A, and hence, they are omitted here.

V. CONCLUSIONS

In this chapter, a class of FRVs and their statistical moments up to second moments have been introduced from the consistent viewpoint of fuzzy perceptions of ordinary nonvague random vectors, and some properties of FRVs and their statistical moment have been examined theoretically, where set representations of fuzzy sets are thoroughly adopted because of the feasibility for describing operations between FRVs.

It may be fair to say that by adopting the family of compact convex correspondences as the set representation of an FRV, we are able to introduce a reasonable definition of FRV, which is basically compatible with the (scalar) fuzzy random variable proposed by Kwakernaak [5] and is never obtained from a simple extension of Kwakernaak's. Furthermore, the proposed concept of FRVs as vague perceptions of random vectors may be, in some sense, a bridge between two major types of fuzzy random variables: one proposed by Kwakernaak and the other by Puri and Ralescu [12].

It should be noted that since the set representations of proposed FRVs are constructed by collections of correspondences, the rich theory of correspondences is expected to be applicable for further investigation of our FRVs.

VI. APPENDIX: PROOFS OF PROPOSITIONS AND THEOREMS

Proof of Proposition 2.1

(i) Since we can rewrite the propositions $s_{f(\tilde{U})}$ by

$$s_{f(\tilde{U})}(v) = \bigvee_{u \in X} \{s_{\tilde{U}} \wedge (v = f(u))\} \tag{A.1}$$

it follows that

$$t(s_{f(\tilde{U})}(v)) = \sup_{u \in X} \{(\tilde{U})(u) \,|\, v = f(u)\} \tag{A.2}$$

Assume here that

$$y \in \{v \,|\, t(s_{f(\tilde{U})}(v)) > \alpha\} = \left\{ v \,\middle|\, \sup_{u \in X} \{(\tilde{U})(u) \,|\, v = f(u)\} > \alpha \right\} \tag{A.3}$$

Then, it can be found that there is some element $x \in X$ such that $y = f(x)$ and $x \in L_\alpha \tilde{U} \subseteq \tilde{U}_\alpha$. Hence, it is clear that

$$y = f(x) \subseteq \{v \,|\, v = f(u), \quad u \in \tilde{U}_\alpha\} \tag{A.4}$$

or equivalently

$$\{v \,|\, t(s_{f(\tilde{U})}(v)) > \alpha\} \subseteq f(L_\alpha \tilde{U}) \subseteq f(\tilde{U}_\alpha) \tag{A.5}$$

On the other hand, when there is a y such that

$$y \in f(\tilde{U}_\alpha) = \{v \,|\, v = f(u), \quad u \in \tilde{U}_\alpha\} \tag{A.6}$$

some element x in X satisfies $y = f(x)$ and $x \in \tilde{U}_\alpha \subseteq L_{\bar{\alpha}} \tilde{U}$, which means

$$y \subseteq \{v \,|\, v = f(u), \quad u \in L_{\bar{\alpha}} \tilde{U}\} \subseteq \{v \,|\, t(s_{f(\tilde{U})}(v)) \geq \alpha\} \tag{A.7}$$

or equivalently

$$f(\tilde{U}_\alpha) \subseteq f(L_{\bar{\alpha}} \tilde{U}\} \subseteq \{v \,|\, t(s_{f(\tilde{U})}(v)) \geq \alpha\} \tag{A.8}$$

Combining (A.5) with (A.8), we have

$$(v \,|\, t(s_{f(\tilde{U})}(v)) > \alpha\} \subseteq f(\tilde{U}_\alpha) \subseteq \{v \,|\, t(s_{f(\tilde{U})}(v)) \geq \alpha\} \tag{A.9}$$

for $\forall \alpha \in (0,1)$, which shows that $\{f(\tilde{U}_\alpha) \,|\, \alpha \in (0,1)\}$ is a set representation of the fuzzy set $f(\tilde{U})$.

(ii) From (A.5) and the definition of the membership function of $f(\tilde{U})$, it is clear that

$$L_\alpha f(\tilde{U}) \subseteq f(L_\alpha \tilde{U}) \quad \text{for } \alpha \in [0,1) \tag{A.10}$$

Assume here that

$$y \in f(L_\alpha \tilde{U}) = \{v \,|\, v = f(u), \quad u \in L_\alpha \tilde{U}\} \tag{A.11}$$

Then, we can find that there exists at least one element x of $L_\alpha \tilde{U}$ satisfying $y = f(x)$, which means

$$\begin{aligned} y = f(x) &\subseteq \left\{ v \,\middle|\, \sup_{u \in X} \{(\bar{U})(u) \,|\, v = f(u)\} > \alpha \right\} \\ &= L_\alpha f(\bar{U}), \end{aligned} \tag{A.12}$$

or equivalently

$$f(L_\alpha \bar{U}) \subseteq L_\alpha f(\bar{U}) \quad \text{for } \alpha \in [0,1) \tag{A.13}$$

Therefore, combining (A.10) with (A.13), it follows that

$$f(L_\alpha \bar{U}) = L_\alpha f(\bar{U}) \quad \text{for } \forall \alpha \in [0,1) \tag{A.14}$$

(iii) It is obvious from (A.8) that (14) holds. Hence, in order to show (16), it is sufficient to prove

$$f(L_{\overline{\alpha}} \bar{U}) \supseteq L_{\overline{\alpha}} f(\bar{U}) \quad \text{for } \alpha \in (0,1] \tag{A.15}$$

Define $x_\alpha(v)$ and $m(v)$ by

$$x_\alpha(v) = \{u \in \mathrm{X} \,|\, (v = f(u)) \wedge (u \in L_{\overline{\alpha}} \bar{U})\} \tag{A.16}$$

and

$$m(v) = \{(\bar{U})(u) \,|\, (u \in \mathrm{X}) \wedge (v = f(u))\} \tag{A.17}$$

respectively. When $x_\alpha(v)$ is not empty, it is easy to verify that

$$L_{\overline{\alpha}} f(\bar{U}) \subseteq f(L_{\overline{\alpha}} \bar{U}) \quad \text{for } \alpha \in (0,1] \tag{A.18}$$

If $x_\alpha(v)$ is empty, $L_{\overline{\alpha}} f(\tilde{U})$ is rewritten as

$$L_{\overline{\alpha}} f(\bar{U}) = \left\{ y \in \mathrm{Y} \,\middle|\, \sup_{u \in \mathrm{X}} m(y) \geq \alpha \right\} \tag{A.19}$$

Hence, if $v \in L_{\overline{\alpha}} f(\tilde{U})$ holds, we obtain

$$\sup_{u \in \mathrm{X}} m(v) \geq \alpha \tag{A.20}$$

Since we have assumed $x_\alpha(v) = \varnothing$, it follows that

$$\beta < \alpha \quad \text{for all } \beta \in m(v) \tag{A.21}$$

Therefore, α is an accumulation point of $m(v)$, i.e., there is a strictly increasing sequence $\{\alpha_k^+ \,|\, k = 1, 2, \ldots\}$ of elements $\alpha_k^+ \in m(v)$ with

$$\lim_{k \to \infty} \alpha_k^+ = \alpha \tag{A.22}$$

The property $\alpha_k^+ \in m(v)$ is connected with the existence of $x_k \in \mathrm{X}$, which satisfies the condition $v = f(x_k)$ and $(\bar{U})(x_k) = \alpha_k^+$, and hence $x_k \in L_{\overline{\alpha_k^+}}\bar{U}$.

Since $\bar{U} \in \mathcal{G}(\mathrm{X})$, it can be shown that $L_{\overline{\alpha_k^+}}\bar{U}$ is a compact subset of X for each $k \in N$. Therefore, we can conclude that the sequence $\{x_k | k \in N\}$ has an accumulation point x and especially that it contains a partial sequence $\{x_{i_k} | k \in N\}$ convergent to x. Furthermore, from the compactness of $L_{\bar{\alpha}}\bar{U}$ for $\alpha \in (0,1]$, it follows

$$x \in \bigcap_{k=1}^{\infty} L_{\overline{\alpha_k^+}}\bar{U} = L_{\bar{\alpha}}\bar{U} \tag{A.23}$$

Since f is a continuous mapping and $L_{\overline{\alpha_k^+}}\bar{U} \supset L_{\overline{\alpha_{k+1}^+}}\bar{U}$, $\forall k \in N$, we obtain

$$\lim_{k \to \infty} x_{i_k} = x \Rightarrow \lim_{k \to \infty} f(x_{i_k}) = f(x) \tag{A.24}$$

For all $k \in N$, $f(x_k) = v$, and hence $f(x) = v$ with $x \in L_{\bar{\alpha}}\bar{U}$ as shown in (A.23). Therefore,

$$\begin{aligned} x &\in \{u \in \mathrm{X} | (v = f(u)) \wedge (u \in L_{\bar{\alpha}}\bar{U})\} \\ &= x_\alpha(v) \end{aligned} \tag{A.25}$$

which is a contradiction to the assumption $x_\alpha(v) = \varnothing$. Therefore, only the case '$x_\alpha(v) \neq \varnothing$' occurs and (16) is proved.

(iv) Assume that f is a continuous mapping. Then, it is obvious that $f(\bar{U})$ is compact. Furthermore, using the equalities given by (13) and (16), it follows that

$$\bigcap_{\alpha \in (0,1)} f(\bar{U}_\alpha) = L_{\bar{1}}f(\bar{U}) = f(L_{\bar{1}}\bar{U}) = f\left(\bigcap_{\alpha \in (0,1)} \bar{U}_\alpha \right) \neq \varnothing \tag{A.26}$$

and

$$\begin{aligned} \operatorname{supp} f(\bar{U}) &= \mathrm{cl} \bigcup_{\alpha \in (0,1)} f(\bar{U}) = \mathrm{cl}\, L_0 f(\bar{U}) = \mathrm{cl} f(L_0\bar{U}) \\ &\subseteq \mathrm{cl} f(\mathrm{cl}\, L_0\bar{U}) = f(\mathrm{cl}\, L_0\bar{U}) = f(\operatorname{supp} \bar{U}) \end{aligned} \tag{A.27}$$

where the compactness of $\operatorname{supp} \tilde{U}$ has been used. Therefore, we can conclude $f(\bar{U}) \in \mathcal{G}(\mathrm{X})$.

Proof of Proposition 3.1

In order to prove Proposition 3.1, we need the following Lemmas [17].

Lemma A.1 *The following statements are equivalent:*

(a) $\tilde{x}_\alpha^{-1}(\mathrm{E}) \in \tilde{\mathcal{A}}$ *for every* $\mathrm{E} \in \mathcal{B}(\mathcal{K}_0)$.
(b) $\tilde{x}_\alpha^{w}(\mathrm{B}) \in \tilde{\mathcal{A}}$ *for every open subset* $B \in \mathrm{R}^n$.

Lemma A.2 *Let* $\tilde{x}_\alpha^{w}(\mathrm{B})$ *be an element of* $\tilde{\mathcal{A}}$ *for every open subset* $\mathrm{B} \in \mathrm{R}^n$. *Then*,

(d) $\mathrm{Gr}(\tilde{x}_\alpha) \in \tilde{\mathcal{A}} \times \mathcal{B}(\mathrm{R}^n)$.

Lemma A.3 *Let* $\mathrm{Gr}(\tilde{x}_\alpha)$ *be* $\tilde{\mathcal{A}} \times \mathcal{B}(\mathrm{R}^n)$*-measurable. Then*,

(c) $\tilde{x}_\alpha^{s}(\mathrm{G}) \in \tilde{\mathcal{A}}$ *for every* $\mathrm{G} \in \mathcal{B}(\mathrm{R}^n)$.

(Proof of Proposition 3.1)

(a)→(d) and (d)→(c): These are easily obtained from Lemmas A.1, A.2, and A.3.

(c)→(b): Let B be an arbitrary element of $\mathcal{B}(\mathrm{R}^n)$. Then, the complement B^c is also an element of $\mathcal{B}(\mathrm{R}^n)$ and hence $\tilde{\Omega}\backslash\tilde{x}_\alpha^s(\mathrm{B}^c) \in \tilde{\mathcal{A}}$. Furthermore, we have

$$\begin{aligned}\tilde{\Omega}\backslash\tilde{x}_\alpha^s(\mathrm{B}^c) &= \tilde{\Omega}\backslash\{\tilde{\omega} \in \tilde{\Omega} \mid \tilde{x}_\alpha(\tilde{\omega}) \subset \mathrm{R}^n\backslash\mathrm{B}\} \\ &= \{\tilde{\omega} \in \tilde{\Omega} \mid \tilde{x}_\alpha(\tilde{\omega}) \cap \mathrm{B} \neq \varnothing\} \\ &= \tilde{x}_\alpha^w(\mathrm{B}) \end{aligned} \tag{A.28}$$

which implies $\tilde{x}_\alpha^w(\mathrm{B}) \in \tilde{\mathcal{A}}$.

(b)→(a): An arbitrary open subset G of R^n is clearly an element of $\mathcal{B}(\mathrm{R}^n)$ and this means that (b)→(b)′. Then, using Lemma A.1, it follows that (b)→(b)′→(a).

Proof of Theorem 3.1

(i) Let $\{\alpha_r^- \mid r = 1, 2, \ldots\}$ and $\{\alpha_r^+ \mid r = 1, 2, \ldots\}$ be respectively strictly decreasing and increasing sequences satisfying (10). Then, it follows that

$$L_\alpha \tilde{x} = \bigcup_{r=1}^{\infty} \tilde{x}_{\alpha_r^-}(\tilde{\omega}), \qquad L_{\overline{\alpha}} \tilde{x} = \bigcap_{r=1}^{\infty} \tilde{x}_{\alpha_r^+}(\tilde{\omega}), \tag{A.29}$$

and $L_{\overline{\alpha}}\tilde{x}$ is compact and convex because so is $\tilde{x}_{\alpha_r}^+$. Since $\tilde{x}_{\alpha_r^-}(\tilde{\omega})$ and $\tilde{x}_{\alpha_r^+}(\tilde{\omega})$ are $\tilde{\mathcal{A}}$-measurable and $\tilde{x}_{\alpha_1^+}(\tilde{\omega}) \supseteq \tilde{x}_{\alpha_2^+}(\tilde{\omega}) \supseteq \ldots$ for each $\tilde{\omega} \in \tilde{\Omega}$, it follows that

$$(L_\alpha \tilde{x})^s(\mathrm{A}) = \left\{ \tilde{\omega} \in \tilde{\Omega} \,\middle|\, \bigcup_{r=1}^{\infty} \tilde{x}_{\alpha_r^-}(\tilde{\omega}) \subset \mathrm{A} \right\}$$

$$= \bigcap_{r=1}^{\infty} \{\tilde{\omega} \in \tilde{\Omega} \,|\, \tilde{x}_{\alpha_r^-}(\tilde{\omega}) \subset \mathrm{A}\} \in \tilde{\mathcal{A}} \tag{A.30}$$

and

$$(L_{\bar{\alpha}} \tilde{x})^w(\mathrm{A}) = \left\{ \tilde{\omega} \in \tilde{\Omega} \,\middle|\, \bigcap_{r=1}^{\infty} (\tilde{x}_{\alpha_r^+}(\tilde{\omega}) \cap \mathrm{A}) \neq \varnothing \right\}$$

$$= \bigcap_{r=1}^{\infty} (\tilde{\omega} \in \tilde{\Omega} \,|\, \tilde{x}_{\alpha_r^+}(\tilde{\omega}) \cap \mathrm{A} \neq \varnothing\} \in \tilde{\mathcal{A}} \tag{A.31}$$

where A is an element of the Borel family of R^n, i.e., $\mathrm{A} \in \mathcal{B}(\mathrm{R}^n)$. Then, using Proposition 3.1, we know that $L_\alpha \tilde{x}$ is $\tilde{\mathcal{A}}$-measurable, and $\{L_{\bar{\alpha}} \tilde{x} \,|\, \alpha \in (0,1)\}$ is the set representation of FRV equivalent to $\tilde{x}$.

(ii) For every closed subset A of R^n, we have $(\mathrm{supp}\,\tilde{x})^s(\mathrm{A}) = (L_0 \tilde{x})^s(\mathrm{A})$. Then, it can be shown by using the same procedure as given in (i) that $(L_0 \tilde{x})^s(\mathrm{A}) \in \tilde{\mathcal{A}}$, and hence $(\mathrm{supp}\,\tilde{x})^s(\mathrm{A}) \in \tilde{\mathcal{A}}$. Furthermore, it can be shown that the following relation is valid:

$$\tilde{\Omega} \backslash (\mathrm{supp}\,\tilde{x})^s(\mathrm{A}) = (\mathrm{supp}\,\tilde{x})^w(\mathrm{A}^c) \tag{A.32}$$

and this implies $(\mathrm{supp}\,\tilde{x})^w(\mathrm{B}) \in \tilde{\mathcal{A}}$ for every open subset B in R^n. Therefore, using Lemma A.1, we know that $\mathrm{supp}\,\tilde{x}$ is an $\tilde{\mathcal{A}}$-measurable correspondence.

(iii) Since the right equation of (A.29) is valid for $\alpha = 1$, the same procedure as for (i) is applicable to show the $\tilde{\mathcal{A}}$-measurability of $L_{\mathrm{T}} \tilde{x}(\tilde{\omega})$.

Proof of Theorem 3.2

Let us define an open subset A in R^n by

$$\mathrm{A} = \{x \in \mathrm{R}^n \,|\, x_i < c, x_i; i\text{th element of } x = (x_1, \ldots, x_n)'\} \tag{A.33}$$

Since $\tilde{x}_\alpha(\tilde{\omega})$ is compact, we know for each element x_i of $x = (x_1, \ldots, x_n)'$ $(i = 1, \ldots, n)$ that

$$\inf \tilde{x}_{\alpha,i}(\tilde{\omega}) = \min\{x_i \,|\, x \in \tilde{x}_\alpha(\tilde{\omega})\} \tag{A.34}$$

and

$$\sup \tilde{x}_{\alpha,i}(\tilde{\omega}) = \max\{x_i \,|\, x \in \tilde{x}_\alpha(\tilde{\omega})\} \tag{A.35}$$

hold. Then, one has

$$\begin{aligned}\{\tilde{\omega}\in\tilde{\Omega} \mid \inf \tilde{x}_{\alpha,i}(\tilde{\omega}) < c\} &= \{\tilde{\omega}\in\tilde{\Omega} \mid \tilde{x}_{\alpha}(\tilde{\omega}) \cap \mathrm{A} \neq \varnothing\} \\ &= \tilde{x}_{\alpha}^{w}(\mathrm{A}) \in \tilde{\mathcal{A}}\end{aligned} \tag{A.36}$$

and

$$\begin{aligned}\{\tilde{\omega}\in\tilde{\Omega} \mid \sup \tilde{x}_{\alpha,i}(\tilde{\omega}) < c\} &= \{\tilde{\omega}\in\tilde{\Omega} \mid \tilde{x}_{\alpha}(\tilde{\omega}) \subset \mathrm{A}\} \\ &= \tilde{x}_{\alpha}^{s}(\mathrm{A}) \in \tilde{\mathcal{A}}\end{aligned} \tag{A.37}$$

since $\tilde{x}_\alpha$ is $\tilde{\mathcal{A}}$-measurable and the equivalence property among the four measurabilities in Proposition 3.1 holds.

Proof of Proposition 3.2

(i) First, we shall show that $\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha$ is a continuous function from $\mathcal{K}_c^n$ to $\mathcal{K}^{n\times n}$. For two arbitrary points $\tilde{x}_\alpha$ and $\tilde{y}_\alpha$ in $\mathcal{K}_c^n$, the distance between $\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha$ and $\tilde{y} \boxdot \tilde{y}'_\alpha$ is given by

$$\begin{aligned}&d_H(\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha, \tilde{y}_\alpha \boxdot \tilde{y}'_\alpha) = \\ &\max\left\{\sup_{a\in\tilde{x}_\alpha}\ \inf_{b\in\tilde{y}_\alpha} \|aa' - bb'\|,\ \sup_{b\in\tilde{y}_\alpha}\ \inf_{a\in\tilde{x}_\alpha} \|aa' - bb'\|\right\}\end{aligned} \tag{A.38}$$

because of

$$A \in \tilde{x}_\alpha \boxdot \tilde{x}'_\alpha = \{W \mid W = xx', x \in \tilde{x}_\alpha\} \Rightarrow A = aa', \quad a \in \tilde{x}_\alpha \tag{A.39}$$

and

$$B \in \tilde{y}_\alpha \boxdot \tilde{y}'_\alpha = \{W \mid W = yy', y \in \tilde{y}_\alpha\} \Rightarrow B = bb', \quad b \in \tilde{y}_\alpha \tag{A.40}$$

Since $\tilde{x}_\alpha$ and $\tilde{y}_\alpha$ are nonempty compact subsets of R^n, there is some constant M such that

$$\sup_{a\in\tilde{x}_\alpha} \|a\| \le \frac{M}{2} \quad \text{and} \quad \sup_{b\in\tilde{y}_\alpha} \|b\| \le \frac{M}{2} \tag{A.41}$$

Furthermore, since $\|aa' - bb'\| < (\|a\| + \|b\|)\|a - b\|$, it follows from (A.41) that

$$\sup_{a\in\tilde{x}_\alpha}\ \inf_{b\in\tilde{y}_\alpha} \|aa' - bb'\| \le M \sup_{a\in\tilde{x}_\alpha}\ \inf_{b\in\tilde{y}_\alpha} \|a - b\| \tag{A.42}$$

and

$$\sup_{b\in\tilde{y}_\alpha}\ \inf_{s\in\tilde{x}_\alpha} \|aa' - bb'\| \le M \sup_{b\in\tilde{y}_\alpha}\ \inf_{a\in\tilde{x}_\alpha} \|a - b\| \tag{A.43}$$

Hence, by substituting (A.42) and (A.43) into (A.38), we have

$$d_H(\tilde{x}_\alpha \boxdot \tilde{x}_\alpha, \tilde{y}_\alpha \boxdot \tilde{y}'_\alpha) \le M d_H(\tilde{x}_\alpha, \tilde{y}_\alpha) \tag{A.44}$$

Therefore, setting $\delta = \varepsilon/M$ for $\forall\varepsilon > 0$, we can confirm that, if

$d_H(\tilde{x}_\alpha, \tilde{y}_\alpha) < \delta$, then

$$d_H(\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha, \tilde{y}_\alpha \boxdot \tilde{y}'_\alpha) < \varepsilon \tag{A.45}$$

which means $\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha$ is continuous for any point in $\mathcal{K}^n_c$. Then, applying Proposition 2.1 to the function $f(\tilde{x}) = \tilde{x} \boxdot \tilde{x}'$, we know that $\tilde{x} \boxdot \tilde{x}'$ is an element of $\mathcal{G}(\mathrm{R}^{n \times n})$ and the equalities (41) and (42) hold.

(ii) Let E be an arbitrary open subset of $\mathcal{K}^{n \times n}$ in the measurable space $(\mathcal{K}^{n \times n}, \mathcal{B}(\mathcal{K}^{n \times n}))$. Then, since $\tilde{x}_\alpha \boxdot \tilde{x}'_\alpha$ is a continuous function of $\tilde{x}_\alpha$,

$$(\tilde{x}_\alpha \boxdot \tilde{x}_\alpha)^{-1}(\mathrm{E}) = \{\tilde{x}_\alpha | \tilde{x}_\alpha \in \mathcal{K}^n_c, \tilde{x}_\alpha \boxdot \tilde{x}'_\alpha \in \mathrm{E}\} \tag{A.46}$$

is an open subset of $\mathcal{K}^n_c$. Therefore, we have (40).

Proof of Proposition 4.1

From (10), we know that

$$L_{\mathrm{T}}\tilde{x} = \bigcap_{r=1}^{\infty} (L_{\alpha_r^+}\tilde{x}) \tag{A.47}$$

holds, where $\{\alpha_r^+ | r = 1, 2, \ldots\}$ is the sequence defined by (9). Then, from (A.47), (3), and the definition of the selection set given in (51), it is clear that

$$L_{\mathrm{T}}\tilde{x} \subseteq L_\alpha \tilde{x} \subseteq \tilde{x}_\alpha \subseteq L_{\overline{\alpha}}\tilde{x} \quad \text{for each } \alpha \in (0, 1) \tag{A.48}$$

Under the assumption **(A-2)** we have that

$$\varnothing \neq S(L_\alpha \tilde{x}) \subseteq S(\tilde{x}_\alpha) \subseteq S(L_{\overline{\alpha}}\tilde{x}) \tag{A.49}$$

from which follows

$$(\mathrm{AL})\int L_\alpha \tilde{x} \mathrm{d}P \subseteq (\mathrm{AL})\int \tilde{x}_\alpha \mathrm{d}P \subseteq (\mathrm{AL})\int L_{\overline{\alpha}}\tilde{x} \mathrm{d}P \tag{A.50}$$

where the existence of each integral is guaranteed because the admissible set of possible original random vectors χ is assumed to be composed of integrable ones.

(i) If there exists a vector ζ such that

$$\begin{aligned} \zeta &\in \{x | t(s_{E\{\tilde{x}\}}(x)) > \alpha\} \\ &= \left\{ x \,\middle|\, \sup_{\xi \in \chi} \left\{ \left(\operatorname*{essinf}_{\omega \in \Omega} (\tilde{x}(\tilde{\omega}))(\xi(\omega)) \right) \wedge (E\{\xi\} = x) \right\} > \alpha \right\} \end{aligned} \tag{A.51}$$

there is some element ξ of χ satisfying $E\{\xi) = \zeta$ and

$$\xi(\omega) \in L_\alpha \tilde{x}(\tilde{\omega}) \quad \text{a.s. in } \Omega \tag{A.52}$$

It is clear that ξ in (A.52) is an element of $S(L_\alpha \tilde{x})$ and hence we have

$$\zeta = E\{\xi\} = \int \xi \mathrm{d}P \in (\mathrm{AL}) \int L_\alpha \tilde{x} \mathrm{d}P \tag{A.53}$$

or equivalently we can conclude that

$$\{x \mid t(s_{E\{\tilde{x}\}}(x)) > \alpha\} \subseteq E\{L_\alpha \tilde{x}\} \tag{A.54}$$

(ii) If there exists ζ such that

$$\zeta \in (\mathrm{AL}) \int L_{\bar{\alpha}} \tilde{x} \, \mathrm{d}P = \left\{ x \,\middle|\, x = \int \xi \mathrm{d}P, \xi \in S(L_{\bar{\alpha}} \tilde{x}) \right\} \tag{A.55}$$

there is some element ξ of χ satisfying $\zeta = \int \xi \mathrm{d}P$ and

$$(\tilde{x}(\tilde{\omega}))(\xi(\omega)) \geq \alpha \quad \text{a.s. in } \Omega \tag{A.56}$$

(A.56) implies that there exists $\xi \in \chi$ satisfying $\zeta = E\{\xi\}$ and

$$\operatorname*{essinf}_{\omega \in \Omega} (\tilde{x}(\tilde{\omega}))(\xi(\omega)) \geq \alpha \tag{A.57}$$

which means that

$$\begin{aligned} \zeta \in & \left\{ x \,\middle|\, \sup_{\xi \in \chi} \left\{ \left(\operatorname*{essinf}_{\omega \in \Omega} (\tilde{x}(\tilde{\omega}))(\xi) \right) \wedge (E\{\xi\} = x) \right\} \geq \alpha \right\} \\ & = \{x \mid t(s_{E\{\tilde{x}\}}(x)) \geq \alpha\} \end{aligned} \tag{A.58}$$

or equivalently

$$E\{L_{\bar{\alpha}} \tilde{x}\} \subseteq \{x \mid t(s_{E\{\tilde{x}\}}(x)) \geq \alpha\} \tag{A.59}$$

(iii) From (A.50), (i) and (ii), we have

$$\begin{aligned} & \{x \mid t(s_{E\{\tilde{x}\}}(x)) > \alpha\} \subseteq E\{L_\alpha \tilde{x}\} \subseteq E\{\tilde{x}_\alpha\} \subseteq E\{L_{\bar{\alpha}} \tilde{x}\} \\ & \subseteq \{x \mid t(s_{E\{\tilde{x}\}}(x)) \geq \alpha\} \end{aligned} \tag{A.60}$$

for each $\alpha \in (0, 1)$.

Proofs of Theorems 4.1 and 4.2

First, we have to invoke the definitions of the inferior limit and the superior limit concerned with sequences of subsets of R^n in the sense of Kuratowski [20]. They are defined respectively as follows:

$$\begin{aligned} & \liminf_{r \to \infty} \mathrm{A}_r = \left\{ a \in \mathrm{R}^n \,\middle|\, a = \lim_{r \to \infty} a_r; a_r \in \mathrm{A}_r \right\} \\ & \limsup_{r \to \infty} \mathrm{A}_r = \{ a \in \mathrm{R}^n \mid a \text{ is a cluster point of } \{a_r\}, a_r \in \mathrm{A}_r \} \end{aligned} \tag{A.61}$$

$$= \bigcap_{e=1}^{\infty} \mathrm{cl}\left(\bigcup_{m=r}^{\infty} \mathrm{A}_m\right) \tag{A.62}$$

where $\{\mathrm{A}_r | r = 1, 2, \ldots\}$ is a sequence of subsets of R^n. Then, the limit of $\{\mathrm{A}_r\}$ is defined by

$$\lim_{r\to\infty} \mathrm{A}_r = \liminf_{r\to\infty} \mathrm{A}_r = \limsup_{r\to\infty} \mathrm{A}_r \tag{A.63}$$

if $\liminf_{r\to\infty} \mathrm{A}_r = \limsup_{r\to\infty} \mathrm{A}_r$.

In order to prove Theorems 4.1 and 4.2, the following lemmas are required [17].

Lemma A.4 *Let* $\{\mathrm{A}_r | r = 1, 2, \ldots\}$ *be a constant sequence of subsets of a topological space, i.e.,*

$$\mathrm{A}_r = \mathrm{A}_{\mathrm{O}} \quad \text{for all } r = 1, 2, \ldots \tag{A.64}$$

Then, $\lim_{r\to\infty} \mathrm{A}_r$ *exists and*

$$\lim_{r\to\infty} \mathrm{A}_r = \mathrm{cl}\,\mathrm{A}_{\mathrm{O}} \tag{A.65}$$

Lemma A.5 *Let* $\{f_r | r = 1, 2, \ldots\}$ *be a sequence of* ($\mathcal{A}$*-measurable*) *integrable functions such that*

$$\|f_r(\omega))\| \leq h(\omega) \quad \text{a.s. in } \Omega \quad \text{for } r = 1, 2, \ldots \tag{A.66}$$

and

$$\int f_r \mathrm{d}P \to \xi \tag{A.67}$$

Then, there is an integrable function f satisfying

$$\xi = \int f\,\mathrm{d}P = \lim_{r\to\infty} \int f_r \mathrm{d}P \tag{A.68}$$

with

$$f(\omega) \in \limsup_{r\to\infty} \{f_r(\omega))\} \quad \text{a.s. in } \Omega \tag{A.69}$$

Lemma A.6. *Let* $\{u_r | r = 1, 2, \ldots\}$ *be a sequence of* $\tilde{\mathcal{A}}$*-measurable correspondences, and let the admissible set of possible original random vectors* χ *be* L^1*-integrably bounded. Then*

$$(\mathrm{AL}) \int \liminf_{r\to\infty} u_r \mathrm{d}P \subseteq \liminf_{r\to\infty} \left\{ (\mathrm{AL}) \int u_r \mathrm{d}P \right\} \tag{A.70}$$

Throughout the proofs of Theorems 4.1 and 4.2, let $\{\tilde{x}_\alpha | \alpha \in (0,1)\}$ be the set representation of $\tilde{x}$.

(Proof of Theorem 4.1)

(i) Let $\{u_r(\tilde{\omega}) | r = 1, 2, \ldots : \tilde{\omega} \in \tilde{\Omega}\}$ be a sequence of correspondences from $\tilde{\Omega}$ to R^n with the property that there is some integrable function $h(\omega)$ such that

$$\|x(\omega)\| \leq h(\omega) \quad \text{a.s. in } \Omega \tag{A.71}$$

for each $x \in S(u_r)$, i.e.,

$$x \in S(u_r) = \{\eta | \eta \in \chi, \eta(\omega) \in u_r(\tilde{\omega}) \quad \text{a.s. in } \Omega\} \tag{A.72}$$

Assume that

$$\zeta = \limsup_{r \to \infty} \{\zeta_r\} \tag{A.73}$$

where

$$\zeta_r \in (\mathrm{AL}) \int u_r \mathrm{d}P \quad (r = 1, 2, \ldots) \tag{A.74}$$

Then, there is a subsequence $\{\zeta_{r_j} | j = 1, 2, \ldots\}$ with

$$\zeta = \lim_{j \to \infty} \zeta_{r_j} = \lim_{j \to \infty} \int \varphi_j \mathrm{d}P \quad (\varphi_j \in S(u_{r_j})) \tag{A.75}$$

because that ζ is a cluster point of $\{\zeta_r | r = 1, 2, \ldots\}$. Applying Lemma A.5 to $\{\varphi_j\}$ given by (A.75), we can confirm that there is an integrable vector φ with

$$\zeta = \int \varphi \mathrm{d}P = \lim_{j \to \infty} \int \varphi_j \mathrm{d}P \tag{A.76}$$

and

$$\varphi \in \limsup_{j \to \infty} \{\varphi_j\} \subseteq S(\limsup_{r \to \infty} u_r) \quad \text{a.s. in } \Omega \tag{A.77}$$

because of condition (iii). Hence, we have

$$\limsup_{r \to \infty} \left\{(\mathrm{AL}) \int u_r \mathrm{d}P\right\} \subseteq (\mathrm{AL}) \int \limsup_{r \to \infty} u_r \mathrm{d}P \tag{A.78}$$

(ii) For $r = 1, 2, \ldots$, set

$$\tilde{x}_{\alpha.r}(\tilde{\omega}) = \tilde{x}_\alpha(\tilde{\omega}) \quad \text{a.s. in } \tilde{\Omega} \tag{A.79}$$

Since $\tilde{x}_\alpha(\tilde{\omega})$ is compact and all elements of χ are integrable, it follows from

Lemma A.4 that

$$\lim_{r\to\infty} \bar{x}_{\alpha,r}(\bar{\omega}) = \mathrm{cl}\,\bar{x}_\alpha(\bar{\omega}) = \bar{x}_\alpha(\bar{\omega}) \quad \text{a.s. in } \tilde{\Omega} \tag{A.80}$$

and

$$\lim_{r\to\infty} E\{\bar{x}_{\alpha,r}\} = \mathrm{cl}\,E\{\bar{x}_\alpha\} \tag{A.81}$$

Then, using (A.78), it can be found that

$$\begin{aligned} \mathrm{cl}\,E\{\bar{x}_\alpha\} &= \limsup_{r\to\infty}\left\{(\mathrm{AL})\int \bar{x}_{\alpha,r}\mathrm{d}P\right\} \\ &\subseteq (\mathrm{AL})\int \limsup_{r\to\infty} \bar{x}_{\alpha,r}\mathrm{d}P \\ &= (\mathrm{AL})\int \mathrm{cl}\,\bar{x}_\alpha \mathrm{d}P \\ &= E\{\bar{x}_\alpha\} \end{aligned} \tag{A.82}$$

which means that $E\{\bar{x}_\alpha\}$ is closed. On the other hand, since each element of χ is bounded by an integrable real-valued function because of condition (ii), it follows that each element of $E\{\bar{x}_\alpha\}$ is also bounded. Therefore, $E\{\bar{x}_\alpha\}$ is compact. Using the same procedure as for $E\{\bar{x}_\alpha\}$, the compactness of $E\{\mathrm{supp}\,\bar{x}\}$ is also proved.

(iii) It is clear from (**A-2**) that

$$E\{\bar{x}_\alpha\} \neq \varnothing \quad \text{for each } \alpha \in (0,1) \tag{A.83}$$

which means

$$\bigcap_{\alpha\in(0,1)} E\{\bar{x}_\alpha\} \neq \varnothing \tag{A.84}$$

Furthermore, from Proposition 4.3, we know that $E\{\bar{x}_\alpha\}$ is convex with condition (i). Therefore, employing the results of (ii), we can conclude that $E\{\bar{x}\}$ satisfies all the conditions for $\mathcal{G}_c(\mathrm{R}^n)$.

(Proof of Theorem 4.2)

(i) When χ is L^1-integrably bounded, it is clear that conditions (i), (ii), and (iii) in Theorem 4.1 are satisfied. Hence, it remains to prove (63) and (64).

(ii) From (10), we have

$$L_\alpha \bar{x}(\bar{\omega}) = \bigcup_{r=1}^{\infty} \bar{x}_{\alpha_r^-}(\bar{\omega}) \quad \text{a.s. in } \tilde{\Omega}, \qquad L_\alpha E\{\bar{x}\} = \bigcup_{r=1}^{\infty} E\{\bar{x}_{\alpha_r^-}\} \tag{A.85}$$

and

$$L_{\bar{\alpha}}\tilde{x}(\tilde{\omega}) = \bigcap_{r=1}^{\infty} \tilde{x}_{\alpha_r^+}(\tilde{\omega}) \quad \text{a.s. in } \tilde{\Omega}, \qquad L_{\bar{\alpha}}E\{\tilde{x}\} = \bigcap_{r=1}^{\infty} E\{\tilde{x}_{\alpha_r^+}\} \tag{A.86}$$

where $\{\alpha_r^-\}$ and $\{\alpha_r^+\}$ are the sequences defined in (9). Since $\{\tilde{x}_{\alpha_r^-} | r = 1, 2, \ldots\}$ and $\{\tilde{x}_{\alpha_r^+} | r = 1, 2, \ldots\}$ are respectively increasing and decreasing sequences of correspondences, it can be found that

$$\lim_{r \to \infty} \tilde{x}_{\alpha_r^-} = \mathrm{cl}\left(\bigcup_{r=1}^{\infty} \tilde{x}_{\alpha_r^-}(\tilde{\omega}) \right) = \mathrm{cl}\, L_{\alpha}\tilde{x}(\tilde{\omega}) \quad \text{a.s. in } \tilde{\Omega} \tag{A.87}$$

$$\lim_{r \to \infty} E\{\tilde{x}_{\alpha_r^-}\} = \mathrm{cl}\left(\bigcup_{r=1}^{\infty} E\{\tilde{x}_{\alpha_r^-}\} \right) = \mathrm{cl}\, L_{\alpha}E\{\tilde{x}\} \tag{A.88}$$

$$\lim_{r \to \infty} \tilde{x}_{\alpha_r^+} = \bigcap_{r=1}^{\infty} \tilde{x}_{\alpha_r^+}(\tilde{\omega}) = L_{\bar{\alpha}}\tilde{x}(\tilde{\omega}) \quad \text{a.s. in } \tilde{\Omega} \tag{A.89}$$

$$\lim_{r \to \infty} E\{\tilde{x}_{\alpha_r^+}\} = \bigcap_{r=1}^{\infty} E\{\tilde{x}_{\alpha_r^+}\} = L_{\bar{\alpha}}E\{\tilde{x}\} \tag{A.90}$$

Then, applying Lemma A.6 and (A.78)–(A.87), (A.88), (A.89) and (A.90), it follows that

$$\begin{aligned} E\{\mathrm{cl}\, L_{\alpha}\tilde{x}\} &= E\left\{ \lim_{r \to \infty} \tilde{x}_{\alpha_r^-} \right\} \subseteq \liminf_{r \to \infty} E\{\tilde{x}_{\alpha_r^-}\} \\ &\subseteq E\left\{ \limsup_{r \to \infty} \tilde{x}_{\alpha_r^-} \right\} = E\{\mathrm{cl}\, L_{\alpha}\tilde{x}\} \end{aligned} \tag{A.91}$$

$$\begin{aligned} E\{ L_{\bar{\alpha}}\tilde{x}\} &= E\left\{ \lim_{r \to \infty} \tilde{x}_{\alpha_r^+} \right\} \subseteq \liminf_{r \to \infty} E\{\tilde{x}_{\alpha_r^+}\} \\ &\subseteq E\left\{ \limsup_{r \to \infty} \tilde{x}_{\alpha_r^+} \right\} = E\{ L_{\bar{\alpha}}\tilde{x}\} \end{aligned} \tag{A.92}$$

which means (63) and (64) hold for every $\alpha \in (0, 1)$.

REFERENCES

1. Y Sunahara, T Fukuda, H Matsuo. On a method of parameter identification from fuzzy random data. In SICE '87, 1987, pp 1085–1088.
2. T Fukuda, Y Sunahara. Fuzzy stochastic processes and their applications to parameter identification. In 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation, 1991.

3. T Fukuda. In F Kozin, T Ono, eds, Systems and Control-Topics in Theory and Applications. Mita Press, Tokyo, Japan, 1991, pp 61–72.
4. T Fukuda, Y Sunahara. In 23rd ISCIE Symposium on Stochastic Systems Theory and its Applications. 1991, pp 79–82.
5. H Kwakernaak. Inform Sci 15:1–29, 1978.
6. H Kwakernaak. Inform Sci 17:253–278, 1979.
7. M Miyakoshi, M Simbo. Fuzzy Sets Syst 12:133–142, 1984.
8. R Kruse, KD Meyer. Statistics with Vague Data. Dordecht: D Reidel Publishing Company, 1987.
9. SB Boswell, MS Taylor. Fuzzy Sets Syst 24:331–344, 1987.
10. H Inoue. Fuzzy Sets Syst 41:285–291, 1991.
11. EP Klement, ML Puri, DA Ralescu. Proc R Soc Lond A407:171–182, 1986.
12. ML Puri, DA Ralescu. J Math Anal Appl 114:409–422, 1986.
13. ML Puri, DA Ralescu. J Math Anal Appl 91:552–558, 1983.
14. ML Puri, DA Ralescu. Ann Prob 13(4):1373–1379, 1985.
15. T Uemura. Fuzzy Sets Syst 59:181–188, 1993.
16. R Kruse, J Gebhardt, F Klawonn. Foundations of Fuzzy Systems. Chichester: John Wiley, 1994.
17. E Klein, AC Thompson. Theory of Correspondences. Netherlands: John Wiley, 1988.
18. W Hildenbrand. Core and Equilibria of a Large Economy. New Jersey: Princeton Unversity Press, 1974.
19. RE Bellman, M Giertz. Inform Sci 5:149–156, 1973.
20. G Matheron. Random Sets and Integral Geometry. New York: John Wiley, 1975.

12

Theory of Cyclostationary Processes and Its Applications

Hideaki Sakai

Kyoto University, Kyoto, Japan

Shuichi Ohno

Shimane University, Shimane, Japan

I. INTRODUCTION

Recently, there has been growing interest in cyclostationary processes in the signal processing community. This is partly due to the fact that signals appearing in many communication systems exhibit cyclostationarity.

A stochastic process whose mean and second-order covariance are periodically time varying was originally called a periodically correlated process by Gladyshev [1] and later was called a periodic nonstationary process by Ogura [2] and a cyclostationary process by Gardner and Franks [3] and Gardner [4], respectively. The fundamental spectral theory of discrete-time cyclostationary processes was first developed by Gladyshev [1]. The continuous-time case was treated by Ogura [2], Gardner and Franks [3], and Gardner [4].

In the first part of this article, we present the spectral theory of discrete-time cyclostationary (DTCS) processes including the meaning of the cyclic spectral density matrix from an engineering point of view, based on the ideal bandpass filtering in [2,3]. Then, two multichannel *stationary* processes are constructed from a DTCS process and we show that the spectral density matrices of these two stationary processes are connected by Gladyshev's formula in [1]. To estimate the cyclic spectral density matrix, Ogura and Yoshida [5] proposed a periodogram-based nonparametric method. They also applied the multichannel entropy (equivalent to autoregresive (AR) model fitting) method to a certain vector process constructed from the original DTCS process. Using their approach we state a parametric formula for the cyclic spectral density matrix of a periodic AR process in [6]. We also mention a method of generating the process using a multichannel time-invariant filter.

In the second part of this article we apply the above spectral theory of a DTCS process to design an optimal filter bank matched to input signal statistics.

Recently there has been a great deal of work concerning the multirate subband filtering method [7,8]. In this method, in the analysis part a signal is first low-pass/high-pass filtered and then downsampled by decimator. These signals are transmitted to the synthesis part. In the synthesis part these two signals are first upsampled by expander and then low-pass/high-pass filtered and finally added. If the input is a stationary stochastic signal, then the output is no longer stationary. In fact the output signal is cyclostationary with period M, where M is the rate of decimation and interpolation. To fully characterize the output, it is necessary to know the cyclic spectral density matrix of the cyclostationary output signal.

Here we first derive this density matrix by using Gladyshev's formula. Then we show the fact that the output of the alias-free filter banks is stationary for any stationary input and see the perfect reconstruction (PR) condition from the spectral point of view.

Next the result is used to derive the averaged mean squared reconstruction error when the high-pass band signal in the synthesis part is dropped in the 2-band filter bank. This is used as a criterion to optimize the low-pass filter in the analysis part under the PR condition. The criterion is a generalization of the one by Vandendorpe [9] for CQF (conjugate quadrature filter) banks. Some numerical results are presented for CQF and general PR filter banks.

II. SPECTRAL THEORY OF CYCLOSTATIONARY PROCESSES

In this section, for simplicity, we assume that the mean $m(t)$ of a cyclostationary process $x(t)$ with period M is zero. Let the autocovariance be defined by

$$R(s,t) = E[x(s)\overline{x(t)}] \tag{1}$$

where "$^-$" denotes the complex conjugate. From the definition of cyclostationarity we have

$$R(s+M,t+M) = R(s,t) \tag{2}$$

For fixed τ, $R(t+\tau,t)$ is a periodic sequence of period M so that the following discrete Fourier expansion is obtained:

$$R(t+\tau,t)=\sum_{k=0}^{M-1} c_k(\tau)\,W^{-kt}\quad (W=\mathrm{e}^{-\mathrm{j}(2\pi/M)}) \tag{3}$$

On the other hand, from the orthogonality property

$$\frac{1}{M}\sum_{t=0}^{M-1} W^{t(m-n)}=\begin{cases}1 & (m\equiv n \ (\mathrm{mod}\ M))\\ 0 & (\text{otherwise})\end{cases} \tag{4}$$

conversely, the kth "modulated autocovariance" is expressed as

$$c_k(\tau)=\frac{1}{M}\sum_{t=0}^{M-1} R(t+\tau,t)\,W^{kt} \tag{5}$$

It is stated in [1] that $c_k(\tau)$ has the following Fourier representation:

$$c_k(\tau)=\int_0^{2\pi} \mathrm{e}^{\mathrm{j}\omega\tau}F_k(\omega)\,\mathrm{d}\omega \tag{6}$$

$$F_k(\omega)=\frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} c_k(\tau)\,\mathrm{e}^{-\mathrm{j}\omega\tau}$$

where we assume that the densities exist and that $c_{k+M}(\tau)=c_k(\tau)$, $F_{k+M}(\omega)=F_k(\omega)$ for all $k\in Z$. These $F_k(\omega)$ are in a sense 'spectra' of this cyclostationary process, although only $F_0(\omega)$ is real and nonnegative. We note that the time averaged variance over one period is given by

$$c_0(0)=\frac{1}{M}\sum_{t=0}^{M-1} R(t,t) \tag{7}$$

and is also expressed as

$$c_0(0)=\int_0^{2\pi} F_0(\omega)\,\mathrm{d}\omega \tag{8}$$

Putting $s=t+\tau$ and substituting (6) into (3) we have

$$R(s,t)=\int_0^{2\pi}\sum_{k=0}^{M-1} F_k(\omega)\,\mathrm{e}^{\mathrm{j}\omega(s-t)}\mathrm{e}^{\mathrm{j}\omega_0 kt}\mathrm{d}\omega \tag{9}$$

$$=\int_0^{2\pi}\int_0^{2\pi} \tilde{F}(\omega,\nu)\,\mathrm{e}^{\mathrm{j}(\omega s-\nu t)}\,\mathrm{d}\omega\,\mathrm{d}\nu \tag{10}$$

Hence, $\tilde{F}(\omega,\nu)$ must be expressed as

$$\tilde{F}(\omega,\nu)=\sum_{k=-M+1}^{M-1} F_k(\omega)\,\delta(\omega-\nu-k\omega_0)\quad \left(\omega_0=\frac{2\pi}{M}\right) \tag{11}$$

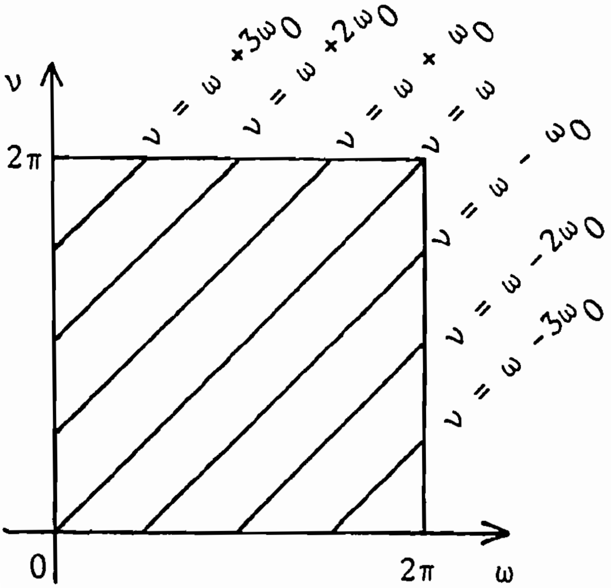


Figure 1 The support of the spectral measure $\bar{F}(\omega, \nu)$.

where $\delta(\cdot)$ is the Dirac delta function. That is, the support of $\bar{F}(\omega, \nu)$ is concentrated on the $(2M-1)$ lines $\omega = \nu + k\omega_0$ in Fig. 1.

Now we change the whole band from $[0, 2\pi)$ to $[-\omega_0/2, 2\pi - \omega_0/2)$ and divide it into M subbands whose center frequencies and bandwidth are $k\omega_0$ $(k = 0, \ldots, M-1)$ and ω_0, respectively. Then the original signal $x(t)$ is decomposed as

$$x(t) = \sum_{k=0}^{M-1} \{e^{j\omega_0 kt} h(t)\} \otimes x(t) \tag{12}$$

where $\otimes$ denotes the convolution operation and $\{h(t)\}$ is the impulse response of the ideal low-pass filter whose center frequency and bandwidth are 0 and ω_0, respectively, with the frequency response

$$H(e^{j\omega}) = \sum_{\tau=-\infty}^{\infty} h(\tau) e^{-j\omega\tau} = \begin{cases} 1 & (|\omega| < \omega_0/2) \\ 0 & (\text{otherwise}) \end{cases} \tag{13}$$

So we have

$$x(t) = \sum_{k=0}^{M-1} e^{j\omega_0 kt} x_k(t) \tag{14}$$

where the amplitude process of the kth harmonics $x_k(t)$ is given by

$$x_k(t) = \sum_{\tau=-\infty}^{\infty} h(\tau)\,\mathrm{e}^{-\mathrm{j}\omega_0 k(t-\tau)} x(t-\tau) \tag{15}$$

Next from these amplitude processes we construct the following multichannel process:

$$\boldsymbol{x}(t) = (x_0(t), \ldots, x_{M-1}(t))^T \tag{16}$$

This is an M-channel weakly stationary process. To show this the covariance between $x_k(t)$ and $x_l(t)$ is calculated. Using (10) we have

$$\begin{aligned} E[x_k(s)\overline{x_l(t)}] &= \sum_{\tau_1}\sum_{\tau_2} h(\tau_1)\overline{h(\tau_2)}\,\mathrm{e}^{-\mathrm{j}\omega_0 k(s-\tau_1)}\mathrm{e}^{\mathrm{j}\omega_0 l(t-\tau_2)} \\ &\quad \times \int_0^{2\pi}\int_0^{2\pi} \tilde{F}(\omega,\nu)\,\mathrm{e}^{\mathrm{j}\{\omega(s-\tau_1)-\nu(t-\tau_2)\}}\,\mathrm{d}\omega\,\mathrm{d}\nu \\ &= \int_0^{2\pi}\int_0^{2\pi} \tilde{F}(\omega,\nu)\,H(\mathrm{e}^{\mathrm{j}(\omega-\omega_0 k)})\overline{H(\mathrm{e}^{\mathrm{j}(\nu-\omega_0 l)})} \\ &\quad \times \mathrm{e}^{\mathrm{j}\{\omega s-\nu t-\omega_0(ks-lt)\}}\,\mathrm{d}\omega\,\mathrm{d}\nu \end{aligned} \tag{17}$$

Since from (11) the support of $\tilde{F}(\omega,\nu)$ is concentrated on $\nu=\omega-m\omega_0$ $(m=-M+1,\ldots,M-1)$, the right-hand side of (17) becomes

$$\begin{aligned} &\int_0^{2\pi}\sum_m F_m(\omega)\,H(\mathrm{e}^{\mathrm{j}(\omega-\omega_0 k)})\overline{H(\mathrm{e}^{\mathrm{j}(\omega-\omega_0(l+m))})} \\ &\times \mathrm{e}^{\mathrm{j}\{\omega(s-t)-\omega_0(ks-lt-mt)\}}\,\mathrm{d}\omega \end{aligned} \tag{18}$$

But from the characteristic of $H(\mathrm{e}^{\mathrm{j}\omega})$ in (13) we note that $H(\mathrm{e}^{\mathrm{j}(\omega-k\omega_0)})=1$, $(|\omega-k\omega_0|<\omega_0/2)$, 0 (otherwise) so that in the summation in (18) only terms for $k\equiv l+m\ (\mathrm{mod}\ M)$ remain. Hence, the covariance between $x_k(t)$ and $x_l(t)$ is given by

$$\begin{aligned} E[x_k(s)\overline{x_l(t)}] &= \int_0^{2\pi} F_{k-l}(\omega)\,H(\mathrm{e}^{\mathrm{j}(\omega-k\omega_0)})\,\mathrm{e}^{\mathrm{j}(\omega-k\omega_0)(s-t)}\,\mathrm{d}\omega \\ &= \int_{-\omega_0/2}^{\omega_0/2} F_{k-l}(\omega+k\omega_0)\,\mathrm{e}^{\mathrm{j}\omega(s-t)}\,\mathrm{d}\omega \end{aligned} \tag{19}$$

This depends on the time difference $s-t$, and $\boldsymbol{x}(t)$ in (16) is weakly stationary. Since the bandwidth of $x_k(t)$ is ω_0, for $|\omega| < \omega_0/2$ the corresponding spectral density $F_{kl}(\omega)$ is given by

$$F_{kl}(\omega) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\left(\int_{-\omega_0/2}^{\omega_0/2} F_{k-1}(\nu + k\omega_0)\,\mathrm{e}^{\mathrm{j}\nu\tau}\mathrm{d}\nu\right)\mathrm{e}^{-\mathrm{j}\omega\tau}$$

$$= \int_{-\omega_0/2}^{\omega_0/2}\left\{\frac{1}{2\pi}\sum_{\tau}\mathrm{e}^{\mathrm{j}(\nu-\omega)\tau}\right\}F_{k-l}(\nu + k\omega_0)\,\mathrm{d}\nu \tag{20}$$

Using the identity

$$\frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\mathrm{e}^{\mathrm{j}\omega\tau} = \sum_{m=-\infty}^{\infty}\delta(\omega + 2\pi m) \tag{21}$$

and noting that $\nu - \omega + 2\pi m$ becomes 0 only for $m = 0$, we have

$$F_{kl}(\omega) = F_{k-l}(\omega + k\omega_0) \quad \left(|\omega| < \frac{\omega_0}{2}\right) \tag{22}$$

This means that the spectral density matrix of $x(t)$ in (16) is completely determined by the functions $F_i(\omega)$ $(i = 0, \ldots, M-1)$ in (6). This result is due to Gladyshev [1].

Next we form another M-channel process $\{y(t)\}$ by concatenating M consecutive samples from $x(t)$ as

$$y(t) := \begin{pmatrix} y_0(t) \\ y_1(t) \\ \vdots \\ y_{M-1}(t) \end{pmatrix} = \begin{pmatrix} x(M(t-1)) \\ x(M(t-1)+1) \\ \vdots \\ x(M(t-1)+M-1) \end{pmatrix} \tag{23}$$

Then this $y(t)$ is also an M-channel stationary process. To see this let us define

$$r_{kl}(s,t) = E[y_k(s)\overline{y_l(t)}] \tag{24}$$

From (23), (2) we have

$$r_{kl}(s,t) = E[x(M(s-1)+k)\overline{x(M(t-1)+l)}]$$
$$= R(M(s-1)+k, M(t-1)+l) = R(M(s-t)+k, l)$$

This depends on the time difference $s - t$. Conversely, if $y(t)$ is an M-channel stationary process and $x(t)$ is generated by the relation $x(M(t-1)+i) = y_i(t)$ in (23), then $x(t)$ is cyclostationary with period M. This can be seen by reversing the above argument. Now let the spectral density matrix of $y(t)$ be denoted by $S(\omega) = (S_{kl}(\omega))$ $(k, l = 0, \ldots, M-1)$. That is,

$$S_{kl}(\omega) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} r_{kl}(\tau, 0)\, \mathrm{e}^{-\mathrm{j}\omega\tau}$$

We examine the relation between $S(\omega)$ and the cyclic spectral density matrix $F(\omega) = (F_{kl}(\omega))$ $(k, l = 0, \ldots, M-1)$ of $x(t)$ in (16). In (14) putting $s = t + \tau$ and substituting into (24), and using (19), the covariance of $y(t)$ in (24) is written as

$$\begin{aligned} r_{kl}(\tau, 0) &= E\Bigg[\sum_{m=0}^{M-1} \mathrm{e}^{\mathrm{j}\omega_0 m(M(t+\tau-1)+k)} x_m(M(t+\tau-1)+k) \\ &\qquad \times \sum_{n=0}^{M-1} \mathrm{e}^{-\mathrm{j}\omega_0 n(M(t-1)+l)} \overline{x_n(M(t-1)+l)}\Bigg] \\ &= \sum_m \sum_n \mathrm{e}^{\mathrm{j}\omega_0(mk-nl)} \int_{-\omega_0/2}^{\omega_0/2} \mathrm{e}^{\mathrm{j}\omega(\tau M+k-l)} F_{m-n}(\omega + m\omega_0)\, \mathrm{d}\omega \qquad (25) \end{aligned}$$

Hence, for $|\omega| < \omega_0/2 = \pi/M$, the corresponding spectral density is given by

$$\begin{aligned} \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} r_{kl}(\tau, 0)\, \mathrm{e}^{-\mathrm{j}\omega M\tau} &= \sum_m \sum_n \mathrm{e}^{\mathrm{j}\omega_0(mk-nl)} \\ &\quad \times \int_{-\omega_0/2}^{\omega_0/2} \left\{ \frac{1}{2\pi} \sum_\tau \mathrm{e}^{\mathrm{j}(\nu-\omega)M\tau} \right\} \\ &\quad \times \mathrm{e}^{\mathrm{j}\nu(k-l)} F_{m-n}(\nu + m\omega_0)\, \mathrm{d}\nu \\ &= \sum_m \sum_n \mathrm{e}^{\mathrm{j}\omega_0(mk-nl)} \\ &\quad \times \int_{-\omega_0/2}^{\omega_0/2} \sum_{i=-\infty}^{\infty} \delta((\nu-\omega)M + 2\pi i) \\ &\quad \times \mathrm{e}^{\mathrm{j}\nu(k-l)} F_{m-n}(\nu + m\omega_0)\, \mathrm{d}\nu \\ &= \frac{1}{M} \sum_m \sum_n \mathrm{e}^{\mathrm{j}\omega_0(mk-nl)} \mathrm{e}^{\mathrm{j}\omega(k-l)} \\ &\quad \times F_{m-n}(\omega + m\omega_0) \qquad (26) \end{aligned}$$

where we use the identity (21). Next, we define the following unitary matrix:

$$U(\omega) = (U_{kl}(\omega)) = \left\{ \frac{1}{\sqrt{M}} \mathrm{e}^{-\mathrm{j}(\omega_0 kl + \omega l)} \right\} \tag{27}$$

Then from (26), (27) the following Gladyshev's formula in [1] is obtained as

$$F(\omega) = U(\omega) S(\omega M) U^*(\omega) \quad \left(|\omega| < \frac{\omega_0}{2} \right) \tag{28}$$

where * denotes the complex conjugate transpose. So we can estimate the cyclic spectral density matrix $F(\omega)$ of a DTCS process by first estimating the spectral density matrix $S(\omega)$ of a stationary process $y(t)$ in (23) by some method and transforming it by (28).

Also, we note that using $F(\omega)$, (7) can be expressed as

$$c_0(0) = \int_{-\pi/M}^{\pi/M} \mathrm{tr} F(\omega) \mathrm{d}\omega \tag{29}$$

On the other hand, Ogura and Yoshida [5] used the following method. First, we define an M-channel process as

$$\tilde{y}(t) := \begin{pmatrix} 1 \\ W^t \\ \vdots \\ W^{(M-1)t} \end{pmatrix} x(t) \equiv g(t) x(t) \quad (W = \mathrm{e}^{-\mathrm{j}(2\pi/M)}) \tag{30}$$

This is not a stationary process. But by the ergodicity of DTCS processes, the estimate of the autocovariance matrix of (30) converges, i.e.,

$$\begin{aligned} \left(\frac{1}{N} \sum_{t=1}^{N} \tilde{y}(t+\tau) \tilde{y}(t)^* \right)_{kl} &= \frac{1}{N} \sum_{t=1}^{N} X(t+\tau) \overline{X(t)} W^{t(k-l)} W^{k\tau} \\ &\rightarrow \tilde{c}_{kl}(\tau) \\ &:= \frac{1}{M} \sum_{t=0}^{M-1} R(t+\tau, t) W^{t(k-l)} W^{k\tau} \end{aligned} \tag{31}$$

as $N \rightarrow \infty$. Using (5) and taking the discrete Fourier transform of $\{\tilde{c}_{kl}(\tau)\}$ gives

$$\begin{aligned} \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \tilde{c}_{kl}(\tau) \mathrm{e}^{-\mathrm{j}\omega\tau} &= \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} c_{k-l}(\tau) W^{k\tau} \mathrm{e}^{-\mathrm{j}\omega\tau} \\ &= F_{k-l}(\omega + k\omega_0) \end{aligned} \tag{32}$$

Hence, comparing (32) with (22) we see that the cyclic spectral density matrix $\boldsymbol{F}(\omega)$ can be estimated by applying any spectral estimation method for multichannel stationary processes to the data $\{\tilde{y}(1),\ldots,\tilde{y}(N)\}$. However, this method is redundant as it estimates $(M^2+M)/2$ components of the cyclic spectral density matrix, whereas the density is determined by only $F_0(\omega),\ldots,F_{M-1}(\omega)$. Ogura and Yoshida [5] used the multichannel maximum entropy (or AR) method, but if the AR model with order p is fitted, the number of parameters required is about pM^2. In the following we introduce a periodic AR process and derive a parametric formula for $\boldsymbol{F}(\omega)$ containing about pM parameters.

II. PERIODIC AR PROCESSES

A periodic AR process $x(t)$ is defined by

$$x(t)+\sum_{i=1}^{p_t} a_t(i)x(t-i)=\varepsilon(t) \tag{33}$$

where $\{\varepsilon(t)\}$ is an uncorrelated (white noise) cyclostationary process with mean 0 and variance σ_t^2 and for simplicity we assume that $x(t)$ takes real values. Also, all the parameters are periodically time varying with period M. That is,

$$p_t=p_{t+M}, \qquad a_t(i)=a_{t+M}(i) \quad (i=1,\ldots,p_t), \qquad \sigma_t^2=\sigma_{t+M}^2 \tag{34}$$

Thus, the process (33) is characterized by M orders $(p_0,p_1,\ldots,p_{M-1})$ and M sets of AR parameters $\{\boldsymbol{a}_k(p_k):=(a_k(1),\ldots,a_k(p_k))^T,\sigma_k^2\}$ $(k=0,\ldots,M-1)$.

Replacing t with $M(t-1)+k$, (33) is written as

$$x(M(t-1)+k)+\sum_{i=1}^{p_k} a_k(i)x(M(t-1)+k-i)=\varepsilon(M(t-1)+k) \tag{35}$$

Multiplying by $x(M(t-1)+k-v)$ on both sides and taking the expectation, we have

$$R(k,k-v)+\sum_{i=1}^{p_k} a_k(i)R(k-i,k-v)=\delta_{v,0}\sigma_k^2 \quad (v\geq 0) \tag{36}$$

Combining (36) for $v=0,\ldots,p_k$, the following Yule–Walker-type equation is obtained as

$$\boldsymbol{R}_{k-1}(p_k-1)\boldsymbol{a}_k(p_k)=-(R(k,k-1),\ldots,R(k,k-p_k))^T \tag{37}$$

where

$$R_k(j) = \begin{pmatrix} R(k,k) & \dots & R(k,k-j) \\ \vdots & & \vdots \\ R(k-j,k) & \dots & R(k-j,k-j) \end{pmatrix} \tag{38}$$

Thus, if the autocovariances $R(k,v) = E[x(k)x(v)]$ are given, the AR parameters $a_k(p_k)$, σ_k^2 $(k = 0,\dots,M-1)$ are obtained by solving (36). Given a set of data $\{X(0), X(1),\dots,X(N-1)\}$, we estimate $R(k,v)$ by

$$\hat{R}(k,v) = \frac{M}{N}\sum_{i=0}^{T} x(k+Mi)x(v+Mi)$$
$$(T = [(N-1-\max(k,v))/M]), \tag{39}$$

substitute it into (36), and obtain the estimate of the AR parameters $\hat{a}_k$, $\hat{\sigma}_k^2$. The circular Levinson algorithm has been derived in [10] for solving the Yule–Walker-type equation in (37) recursively for increasing orders $p_k = 1, 2, \dots$.

IV. A SPECTRAL FORMULA OF PERIODIC AR PROCESSES

For notational simplicity let us define $p = \max(p_0, p_1,\dots,p_{M-1})$ and put $a_t(i) = 0$ $(p_t \leq i \leq p)$. Then we can replace p_t in (33) by p. Since $a_t(i), \sigma_t^2$ are periodic sequences with period M, as in (3), it follows that

$$a_t(i) = \sum_{k=0}^{M-1} a_{ik} W^{kt} \quad (W = e^{-j(2\pi/M)}) \tag{40}$$

$$\sigma_t^2 = \sum_{k=0}^{M-1} s_k W^{kt} \tag{41}$$

From (33), (30), and (40) we have

$$x(t) + (a_{10} W a_{11} \dots W^{M-1} a_{1,M-1})\tilde{y}(t-1)$$
$$+ \dots + (a_{p0} W a_{p1} \dots W^{p(M-1)} a_{p,M-1})\tilde{y}(t-p) = \varepsilon(t)$$

Multiplying by W^{it} $(i = 1,\dots,M-1)$ on both sides, and noting $W^{nM} = 1$ (n; integer) and using $\tilde{y}(t)$ in (30), it follows that

$$\tilde{y}(t) = \begin{pmatrix} a_{10} & a_{11} & a_{1,M-1} \\ a_{1,M-1} & a_{10} & a_{1,M-2} \\ \vdots & \vdots & \vdots \\ a_{11} & a_{12} & a_{10} \end{pmatrix} D\tilde{y}(t-1) + \dots$$

$$+\begin{pmatrix} a_{p0} & a_{p1} & \cdots & a_{p,M-1} \\ a_{p,M-1} & a_{p0} & \cdots & a_{p,M-2} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{p0} \end{pmatrix} D^p \tilde{y}(t-p) = g(t)\,\varepsilon(t) \tag{42}$$

where D is the diagonal matrix

$$D = \mathrm{diag}(1, W, \ldots, W^{M-1}) \tag{43}$$

Since the matrices on the left-hand side of (42) are constants, $\tilde{y}(t)$ is an output of an M-channel time-invariant filter

$$H(z) = I + H_1 z^{-1} + H_2 z^{-2} + \ldots$$

(I: $M \times M$ identity matrix) with an input signal $e(t) = g(t)\,\varepsilon(t)$ whose statistics are periodically time varying. Hence, by the whiteness and the ergodicity of $e(t)$, the limit of the sample autocovariance of $\tilde{y}(t)$ is given by

$$\left(\frac{1}{N}\sum_{t}\sum_{i,j} H_i e(t+u-i) e^*(t-j) H_j^*\right)_{mn}$$

$$\to \tilde{R}_{mn}(u) := \left(\sum_{j} H_{j+u} Q H_j^*\right)_{mn} \tag{44}$$

where we put

$$Q = \frac{1}{M}\sum_{t=0}^{M-1} g(t) g^*(t)\,\sigma_t^2 \tag{45}$$

From (41), (30) we have

$$(Q)_{mn} = \frac{1}{M}\sum_{t=0}^{M-1} W^{t(m-n)} \sum_{k=0}^{M-1} s_k W^{kt} = s_{n-m} \tag{46}$$

where $s_k = s_{k+M}(-M < k < 0)$. This means that $\tilde{R}_{mn}(u)$ in (44) is identical to the autocovariance matrix of an M-channel stationary AR process with the same coefficient matrices on the left-hand side of (42) and the variance–covariance matrix Q in (45). Thus, $F(\omega)$ in (22) can be obtained as the spectral density matrix of this AR process.

But, since the nondiagonal matrices on the left-hand side of (42) are circulant [11], these are decomposed as

$$\begin{pmatrix} a_{i0} & a_{i1} & \cdots & a_{i,M-1} \\ a_{i,M-1} & a_{i0} & \cdots & a_{i,M-2} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{i0} \end{pmatrix} = V\Gamma_i V^* \tag{47}$$

where $\boldsymbol{V}$ is the DFT matrix with its (m, n) element

$$(\boldsymbol{V})_{mn} = \frac{1}{\sqrt{M}} W^{mn} \quad (m, n = 0, \ldots, M-1) \tag{48}$$

and $\boldsymbol{\Gamma}_i$ is diagonal with its mth element

$$(\boldsymbol{\Gamma}_i)_m = \sum_{k=0}^{M-1} a_{ik} W^{km} = a_m(i) \tag{49}$$

Thus, the polynomial matrix corresponding to the transfer function in the left-hand side of (42) is written as

$$\boldsymbol{A}(z) = \boldsymbol{I} + \sum_{i=1}^{p} \boldsymbol{V}\boldsymbol{\Gamma}_i \boldsymbol{V}^* \boldsymbol{D}^i z^{-i} \tag{50}$$

Here we note the following relation:

$$\boldsymbol{V}^* \boldsymbol{D}^i \boldsymbol{V} = \boldsymbol{\Pi}^i \tag{51}$$

where $\boldsymbol{\Pi}$ is a permutation matrix defined by

$$\boldsymbol{\Pi} = \begin{pmatrix} 0 & \ldots & 0 & 1 \\ 1 & \ldots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \ldots & 1 & 0 \end{pmatrix} \tag{52}$$

Then using the original AR parameters, $A(z)$ in (50) is expressed as

$$\begin{aligned} \boldsymbol{V}^* \boldsymbol{A}(z) \boldsymbol{v} = \boldsymbol{I} &+ \begin{pmatrix} 0 & \ldots & 0 & a_0(1) \\ a_1(1) & \ldots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \ldots & a_{M-1}(1) & 0 \end{pmatrix} z^{-1} \\ &+ \begin{pmatrix} 0 & \ldots & 0 & a_0(2) & 0 \\ 0 & \ldots & 0 & 0 & a_1(2) \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \ldots & a_{M-1}(2) & 0 & 0 \end{pmatrix} z^{-2} \\ &+ \ldots \\ &:= \bar{A}(z) \end{aligned} \tag{53}$$

Therefore, $F(\omega)$ for $|\omega| \leq \omega_0/2$ in (22) is given by

$$\begin{aligned} 2\pi F(\omega) &= A^{-1}(z)Q(A(z))^{*-1}|_{z=e^{j\omega}} \\ &= V(\tilde{A}^{-1}(z)\tilde{Q}\tilde{A}^{-T}(z^{-1})|_{z=e^{j\omega}})V^* \\ &:= 2\pi V\tilde{F}(\omega)V^* \end{aligned} \tag{54}$$

where

$$\tilde{Q} = V^*QV = (\sigma_m^2\delta_{mn}) \tag{55}$$

and δ_{mn} is the Kronecker delta. From (54) $F(\omega)$ is derived from $\tilde{F}(\omega)$ by the similarity transformation with the DFT matrix, and $\tilde{F}(\omega)$ is the spectral density matrix of an AR process generated by passing white noise with diagonal variance–covariance matrix $\tilde{Q}$ through the filter whose transfer function is $\tilde{A}^{-1}(z)$. From (53) and (55) $\tilde{F}(\omega)$ contains at most $(p+1)M$ parameters. Ogura and Yoshida [5] did not utilize the special structure in the matrices on the left-hand side of (42) and applied the M-channel AR model fitting directly with $pM^2 + (M^2 + M)/2$ parameters. So their method is redundant compared with ours.

Alternatively, $F(\omega)$ can be calculated by substituting the spectral density matrix $S(\omega)$ of $y(t)$ in (23) into Gladyshev's formula (28). When $x(t)$ is a periodic AR process in (33), it is shown by Pagano [12] that $y(t)$ is an M-channel stationary AR process. Actually, from (23), (35) we see that

$$\begin{pmatrix} 1 & & \\ a_1(1) & 1 & 0 \\ \vdots & & \ddots \\ a_{M-1}(M-1) & \ldots & 1 \end{pmatrix} \begin{pmatrix} x(M(t-1)) \\ x(M(t-1)+1) \\ \vdots \\ x(M(t-1)+M-1) \end{pmatrix} +$$

$$\begin{pmatrix} a_0(M) & \ldots & a_0(1) \\ a_1(M+1) & \ldots & a_1(2) \\ \vdots & & \vdots \\ a_{M-1}(2M-1) & \ldots & a_{M-1}(M) \end{pmatrix} \begin{pmatrix} x(M(t-2)) \\ x(M(t-2)+1) \\ \vdots \\ x(M(t-2)+M-1) \end{pmatrix} +$$

$$\ldots = \begin{pmatrix} \varepsilon(M(t-1)) \\ \varepsilon(M(t-1)+1) \\ \vdots \\ \varepsilon(M(t-1)+M-1) \end{pmatrix} := \boldsymbol{\varepsilon}(t) \tag{56}$$

We write this as

$$\mathcal{A}_0 y(t) + \mathcal{A}_1 y(t-1) + \ldots + \mathcal{A}_q y(t-p') = \varepsilon(t) \tag{57}$$

where $p' = \max_k[(p_k - k - 1)/M] + 1$. The elements of the coefficient matrices $\mathcal{A}_m$ are given by

$$(\mathcal{A}_m)_{kl} = a_k(mM + k - l) \quad (k, l = 0, \ldots, M-1) \tag{58}$$

where $a_k(0) = 1$, $a_k(i) = 0\,(i < 0)$. Since the variance–covariance matrix of the white noise $\varepsilon(t)$ is $\tilde{Q}$ in (55), from (56) we have

$$S(\omega) = \frac{1}{2\pi}\mathcal{A}^{-1}(z)\tilde{Q}\mathcal{A}^{-T}(z^{-1})|_{z=\mathrm{e}^{\mathrm{j}\omega}} \tag{59}$$

where

$$\mathcal{A}(z) = \mathcal{A}_0 + \mathcal{A}_1 z^{-1} + \ldots + \mathcal{A}_{p'} z^{-p'} \tag{60}$$

From (27), (53), and (58) we have

$$\begin{aligned} U(\omega)\mathcal{A}(\mathrm{e}^{\mathrm{j}\omega M})U^*(\omega) = \frac{1}{M}\sum_{m,n} \mathrm{e}^{\mathrm{j}\omega_0(km-ln)} \sum_i a_m(iM + m - n) \\ \times \mathrm{e}^{-\mathrm{j}\omega(iM+m-n)} \end{aligned} \tag{61}$$

But from (48), (53) it can be seen that

$$U(\omega)\mathcal{A}(\mathrm{e}^{\mathrm{j}\omega M})U^*(\omega) = V\tilde{A}(\mathrm{e}^{\mathrm{j}\omega})V^* \tag{62}$$

Substituting (59) into (28) together with (62) and noting that $U(\omega)$ is unitary and $U(\omega)\tilde{Q}U^*(\omega) = Q$, we see that $F(\omega)$ is again given by (54). This shows that $F(\omega)$ based on $\tilde{y}(t)$ and that based on Gladyshev's formula coincide, as they should.

However, by the former approach we can show an almost time-invariant filter structure generating $x(t)$ in (33). Actually, from (4), (30), and (48) we have

$$x(t) = \frac{1}{\sqrt{M}}(1 \quad 1 \cdots 1)\, V^*\tilde{y}(t) \tag{63}$$

Also, we note

$$(V^* g(t)\,\varepsilon(t))_m = \begin{cases} \sqrt{M}\varepsilon(t) & (t \equiv m \pmod{M}) \\ 0 & (\text{otherwise}) \end{cases} \tag{64}$$

This means that after multiplying by V^* on both sides of (42) the input term $V^* g(t)\,\varepsilon(t)$ is created by a switching operation with period M. Since reducing the input by $1/\sqrt{M}$ and amplifying the output by $\sqrt{M}$ do not

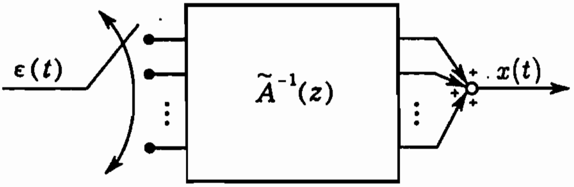


Figure 2 An "almost" time-invariant filter generating a periodic AR process.

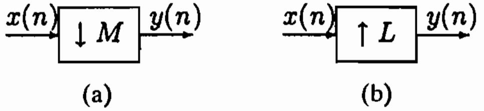


Figure 3 Decimator and expander.

change the linear input–output relationship, we finally obtain the whole block diagram in Fig. 2 for the generation of the process, where every part except the switching operation is time invariant. This result can be easily extended to the case where $x(t)$ is a periodic ARMA process [13]. Also, a similar lattice filter structure for generating a periodic AR process is shown in [6].

V. MULTIRATE SYSTEMS

In this section, we briefly review the fundamentals of multirate systems. Here we treat filters with real coefficients and follow the notation in [7].

The M-fold decimation in Fig. 3a is to produce the output sequence

$$y(n) = x(Mn) \tag{65}$$

for an input sequence $x(n)$ where M is an integer. Only those samples of $x(n)$ which occur at time instants equal to multiples of M are retained. In the transform domain, it can be expressed by

$$Y(z) = \frac{1}{M}\sum_{k=0}^{M-1} X(z^{1/M}W^k) := X(z)|_{\downarrow M} \tag{66}$$

where $W = e^{-j(2\pi/M)}$ and $X(z), Y(z)$ are the z-transform of $x(n), y(n)$, respectively.

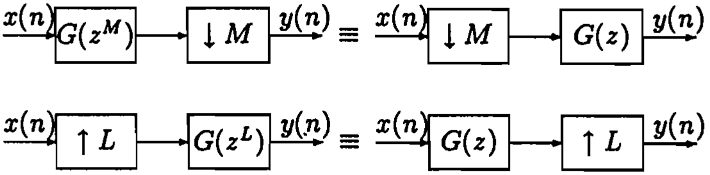


Figure 4 Noble identities.

The L-fold expander in Fig. 3b is to produce the output sequence

$$y(n) = \begin{cases} x(n/L) & n\text{: multiple of } L \\ 0 & \text{otherwise} \end{cases} \tag{67}$$

with an integer L. The output is obtained by inserting $L-1$ zeros between the input signals. In the transform domain, it can be expressed by

$$Y(z) = X(z^L) := X(z)|_{\uparrow L} \tag{68}$$

Here we have

$$\begin{aligned} (G(z^M)X(z))|_{\downarrow M} &= G(z)(X(z)|_{\downarrow M}) \\ G(z^L)(X(z)|_{\uparrow L}) &= (G(z)X(z))|_{\uparrow L} \end{aligned} \tag{69}$$

These are known as the noble identities [7]. Therefore, we can redraw the left part in Fig. 4 as the right part.

Separating the coefficients $h(n)$ of $H(z)$ in terms of n modulo M, we can write

$$H(z) = \sum_{l=0}^{M-1} z^{-l} E_l(z^M) \tag{70}$$

where

$$E_l(z) = \sum_{n=-\infty}^{\infty} h(Mn+l) z^{-n}, \quad 0 \le l \le M-1$$

Similarly $H(z)$ can be written as

$$H(z) = \sum_{l=0}^{M-1} z^{-(M-1-l)} R_l(z^M) \tag{71}$$

with

$$R_l(z) = E_{M-1-l}(z) \quad 0 \le l \le M-1 \tag{72}$$

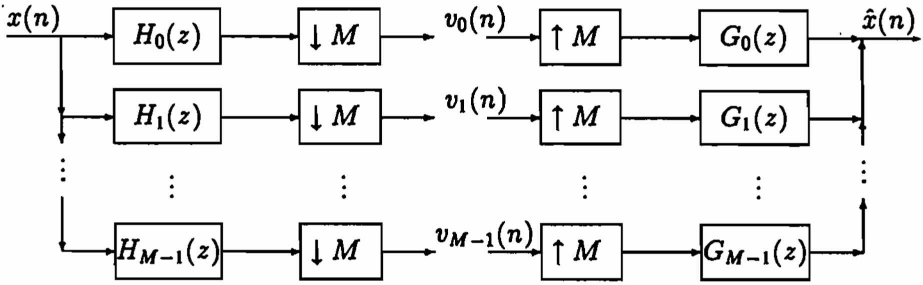


Figure 5 A block diagram of an M-channel filter bank.

The expressions in (70) and (71) are called Type 1 and Type 2 polyphase representation of $H(z)$, respectively [7].

Defining analysis filters and synthesis filters in the M-band filter bank in Fig. 5 as

$$\boldsymbol{H}(z) = (H_0(z), H_1(z), \ldots, H_{M-1}(z))^T \tag{73}$$

$$\boldsymbol{G}(z) = (G_0(z), G_1(z), \ldots, G_{M-1}(z))^T \tag{74}$$

respectively, using the Type 1 polyphase representation,

$$H_k(z) = \sum_{l=0}^{M-1} z^{-l} E^h_{kl}(z^M), \qquad G_k(z) = \sum_{l=0}^{M-1} z^{-l} E^g_{kl}(z^M) \tag{75}$$

we can write

$$\boldsymbol{H}(z) = \boldsymbol{E}_h(z^M)\boldsymbol{e}(z), \qquad \boldsymbol{G}(z) = \boldsymbol{E}_g(z^M)\boldsymbol{e}(z) \tag{76}$$

where

$$\boldsymbol{e}(z) = (1, z^{-1}, \ldots, z^{-M+1})^T \tag{77}$$

and

$$(\boldsymbol{E}_h(z))_{k,l} = E^h_{kl}(z), \qquad (\boldsymbol{E}_g(z))_{k,l} = E^g_{kl}(z) \tag{78}$$

Using the noble identities (69), we interchange the filters by the decimator and the expander respectively, so that Fig. 6 can be obtained. It should be noted that in [7], p. 231, the Type 2 polyphase representation of $G(z)$ is used as

$$\boldsymbol{G}(z) = \boldsymbol{R}_g^T(z^M)(z^{-M+1}, \ldots, z^{-1}, 1)^T \tag{79}$$

where

$$\boldsymbol{R}_g^T(z) = \boldsymbol{E}_g(z)\boldsymbol{J}_M \tag{80}$$

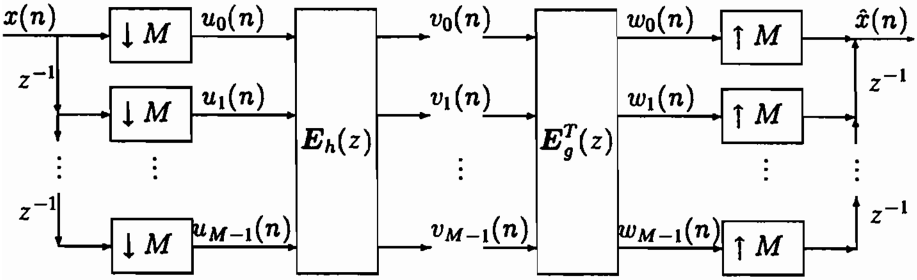


Figure 6 Filter bank using polyphase matrix.

and

$$J_M = \begin{pmatrix} 0 & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & 0 \end{pmatrix} \tag{81}$$

The alias component (AC) matrix of a filter bank $\boldsymbol{H}(z)$ is given by

$$\boldsymbol{H}_{\mathrm{AC}}(z) = \begin{pmatrix} H_0(z) & H_1(z) & \dots & H_{M-1}(z) \\ H_0(zW) & H_1(zW) & \dots & H_{M-1}(zW) \\ \vdots & \vdots & \ddots & \vdots \\ H_0(zW^{M-1}) & H_1(zW^{M-1}) & \dots & H_{M-1}(zW^{M-1}) \end{pmatrix} \tag{82}$$

which can be rewritten as

$$\begin{aligned} \boldsymbol{H}_{\mathrm{AC}}^T(z) &= (\boldsymbol{H}(z), \boldsymbol{H}(zW), \dots, \boldsymbol{H}(zW^{M-1})) \\ &= \boldsymbol{E}_h(z^M)(\boldsymbol{e}(z), \boldsymbol{e}(zW), \dots, \boldsymbol{e}(zW^{M-1})) \\ &= \boldsymbol{E}_h(z^M)\,\boldsymbol{\Lambda}(z)\,\boldsymbol{W}^* \end{aligned} \tag{83}$$

where $\boldsymbol{W}$ is the “unnormalized” DFT matrix such that

$$(\boldsymbol{W})_{i,j} = W^{ij} \tag{84}$$

and

$$\boldsymbol{\Lambda}(z) = \mathrm{diag}(1, z^{-1}, \dots, z^{-M+1}) \tag{85}$$

VI. THE OUTPUT OF FILTER BANKS

In this section, we derive the cyclic spectral density matrix $\boldsymbol{F}(\omega)$ of the output $\hat{x}(n)$ of the M-band filter bank in Fig. 6 when the input $x(n)$ is a stationary process with zero mean and the covariance $R_x(n)$ and the spectral density is $S_x(\omega)$.

Now we consider the block diagram in Fig. 6 and define M-channel processes $\boldsymbol{u}(n)$, $\boldsymbol{v}(n)$, and $\boldsymbol{w}(n)$ as

$$\boldsymbol{u}(n) = (u_0(n), u_1(n), \ldots, u_{M-1}(n))^T \tag{86}$$

$$\boldsymbol{v}(n) = (v_0(n), v_1(n), \ldots, v_{\mathrm{M}-1}(n))^T \tag{87}$$

$$\boldsymbol{w}(n) = (w_0(n), w_1(n), \ldots, w_{M-1}(n))^T \tag{88}$$

respectively.

First, we consider the analysis filter bank where $\boldsymbol{u}(n)$ can be expressed by

$$\boldsymbol{u}(n) = (x(Mn), x(Mn-1), \ldots, x(Mn-M+1))^T \tag{89}$$

Since $x(n)$ is stationary, $\boldsymbol{u}(n)$ is an M-channel stationary process. Define the (i,k)th element of the spectral density matrix as

$$(\boldsymbol{S}_u(\omega))_{i,k} = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}(E[\boldsymbol{u}(t+\tau)\boldsymbol{u}^T(t)])_{i,k}\mathrm{e}^{-\mathrm{j}\omega\tau} \quad \text{for } |\omega| \leq \pi \tag{90}$$

Now we show the relation between $S_x(\omega)$ and $\boldsymbol{S}_u(\omega)$. From (89), we have

$$(\boldsymbol{S}_u(\omega))_{i,k} = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} R_x(M\tau+k-i)\,\mathrm{e}^{-\mathrm{j}\omega\tau} \tag{91}$$

Putting $l = k - i$ and using the identity (4) we have

$$\begin{aligned}(\boldsymbol{S}_u(\omega))_{i,k} &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} R_x(\tau)\frac{1}{M}\sum_{m=0}^{M-1} W^{(\tau-l)m}\mathrm{e}^{-\mathrm{j}\omega((\tau-l)/M)} \\ &= \frac{1}{M}\sum_{m=0}^{M-1} W^{-lm}\mathrm{e}^{\mathrm{j}(l/M)\omega}\frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} R_x(\tau)\,W^{\tau m}\mathrm{e}^{-\mathrm{j}(\tau/M)\omega} \\ &= \frac{1}{M}\sum_{m=0}^{M-1} W^{-lm}\mathrm{e}^{\mathrm{j}(l/M)\omega}\, S_x\!\left(\frac{\omega+2\pi m}{M}\right)\end{aligned}$$

Using $\boldsymbol{W}$ in (84) and $\boldsymbol{\Lambda}(z)$ in (85), we have

$$\boldsymbol{S}_u(\omega) = \frac{1}{M}\boldsymbol{\Lambda}(\mathrm{e}^{\mathrm{j}(\omega/M)})\,\boldsymbol{W}\boldsymbol{S}_x(\omega)\,\boldsymbol{W}^*\boldsymbol{\Lambda}(\mathrm{e}^{-\mathrm{j}(\omega/M)}) \tag{92}$$

with

$$\boldsymbol{S}_x(\omega) = \mathrm{diag}\left(S_x\left(\frac{\omega}{M}\right), S_x\left(\frac{\omega+2\pi}{M}\right), \ldots, S_x\left(\frac{\omega+2\pi(M-1)}{M}\right)\right) \tag{93}$$

Since $\boldsymbol{v}(n)$ is filtered by $\boldsymbol{E}_h(z)$, it is also an M-channel stationary process with the spectral density matrix

$$\begin{aligned} \boldsymbol{S}_{\boldsymbol{v}}(\omega) &= \frac{1}{M}\boldsymbol{E}_h(\mathrm{e}^{\mathrm{j}\omega})\boldsymbol{\Lambda}(\mathrm{e}^{\mathrm{j}(\omega/M)})\boldsymbol{W}\boldsymbol{S}_x(\omega)\boldsymbol{W}^*\boldsymbol{\Lambda}(\mathrm{e}^{-\mathrm{j}(\omega/M)})\boldsymbol{E}_h^T(\mathrm{e}^{-\mathrm{j}\omega}) \\ &= \frac{1}{M}\boldsymbol{H}_{\mathrm{AC}}^*(\mathrm{e}^{-\mathrm{j}(\omega/M)})\boldsymbol{S}_x(\omega)\boldsymbol{H}_{\mathrm{AC}}(\mathrm{e}^{-\mathrm{j}(\omega/M)}) \end{aligned} \tag{94}$$

Secondly, in the synthesis filter bank, since $\boldsymbol{v}(n)$ is filtered by $\boldsymbol{E}_g^T(z)$ to produce the output $\boldsymbol{w}(n)$, $\boldsymbol{w}(n)$ is also an M-channel stationary process with the spectral density matrix

$$\boldsymbol{S}_w(\omega) = \boldsymbol{E}_g^T(\mathrm{e}^{\mathrm{j}\omega})\boldsymbol{S}_{\boldsymbol{v}}(\omega)\boldsymbol{E}_g(\mathrm{e}^{-\mathrm{j}\omega}) \tag{95}$$

The relation between $\hat{x}(n)$ and $w_i(n)$ in Fig. 6 is given by

$$\hat{x}(Mn+i) = w_i(n) \quad \text{for } i = 0,\ldots,M-1 \tag{96}$$

By the argument concerning (23), $\hat{x}(n)$ is a cyclostationary process with period M.

From Gladyshev's formula (28), its cyclic spectral density matrix is given by

$$\boldsymbol{F}(\omega) = \frac{1}{M}\boldsymbol{W}\boldsymbol{\Lambda}(\mathrm{e}^{\mathrm{j}\omega})\boldsymbol{S}_w(M\omega)\boldsymbol{\Lambda}(\mathrm{e}^{-\mathrm{j}\omega})\boldsymbol{W}^* \quad \text{for } |\omega| \le \pi/M \tag{97}$$

where we use that $\boldsymbol{U}(\omega)$ in (27) is rewritten as $\boldsymbol{W}\boldsymbol{\Lambda}(\mathrm{e}^{\mathrm{j}\omega})/\sqrt{M}$. Substituting (94) and (95) into (97), we finally have

$$\boldsymbol{F}(\omega) = \frac{1}{M^2}(\boldsymbol{H}_{\mathrm{AC}}(z)\boldsymbol{G}_{\mathrm{AC}}^T(z))^*\boldsymbol{S}_x(M\omega)\boldsymbol{H}_{\mathrm{AC}}(z)\boldsymbol{G}_{\mathrm{AC}}^T(z)\big|_{z=\mathrm{e}^{-\mathrm{j}\omega}} \quad \text{for } |\omega| \le \pi/M, \tag{98}$$

where

$$\boldsymbol{G}_{\mathrm{AC}}(z) = \boldsymbol{W}^*\boldsymbol{\Lambda}(z)\boldsymbol{E}_g^T(z^M) \tag{99}$$

Now we show the following theorem.

Theorem. *If a filter bank is alias free, then its output for any stationary input is stationary.*

Proof. $\hat{x}(n)$ is stationary if $\boldsymbol{F}(\omega)$ is diagonal. From (98) it is sufficient to show that

$$\boldsymbol{H}_{\mathrm{AC}}(z)\boldsymbol{G}_{\mathrm{AC}}^T(z) = \boldsymbol{W}^*\boldsymbol{\Lambda}(z)\boldsymbol{E}_h^T(z^M)\boldsymbol{E}_g(z^M)\boldsymbol{\Lambda}(z)\boldsymbol{W}^* \tag{100}$$

is diagonal since $\boldsymbol{S}_x(M\omega)$ is diagonal.

A filter bank is said to be alias free iff

$$\begin{aligned} \boldsymbol{P}(z) &:= \boldsymbol{R}_g(z)\boldsymbol{E}_h(z) \\ &= \begin{pmatrix} p_0(z) & p_1(z) & \dots & p_{M-2}(z) & p_{M-1}(z) \\ z^{-1}p_{M-1}(z) & p_0(z) & \dots & p_{M-3}(z) & p_{M-2}(z) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ z^{-1}p_1(z) & z^{-1}p_2(z) & \dots & z^{-1}p_{M-1}(z) & p_0(z) \end{pmatrix} \end{aligned} \tag{101}$$

holds [7], pp. 249–253. Using the Type 1 polyphase representation of $\boldsymbol{G}(z)$, from (80) and (101), we have

$$\begin{aligned} \boldsymbol{E}_h^T(z)\boldsymbol{E}_g(z) &= \boldsymbol{E}_h^T(z)\boldsymbol{R}_g^T(z)\boldsymbol{J}_M = \boldsymbol{P}^T(z)\boldsymbol{J}_M \\ &= \begin{pmatrix} z^{-1}p_1(z) & z^{-1}p_2(z) & \dots & z^{-1}p_{\mathrm{M}-1}(z) & p_0(z) \\ z^{-1}p_2(z) & p_3(z) & \dots & p_0(z) & p_1(z) \\ \vdots & \vdots & \iddots & \vdots & \vdots \\ p_0(z) & p_1(z) & \dots & p_{M-2}(z) & p_{M-1}(z) \end{pmatrix} \end{aligned} \tag{102}$$

By defining

$$\boldsymbol{Q}(z) = \boldsymbol{\Lambda}(z)\boldsymbol{E}_h^T(z^M)\boldsymbol{E}_g(z^M)\boldsymbol{\Lambda}(z) \tag{103}$$

it can be easily shown that $\boldsymbol{Q}(z)$ is a left circulant matrix

$$\boldsymbol{Q}(z) = \begin{pmatrix} q_1(z) & \dots & q_{M-1}(z) & q_0(z) \\ q_2(z) & \dots & q_0(z) & q_1(z) \\ \vdots & \iddots & \vdots & \vdots \\ q_0(z) & \dots & q_{M-2}(z) & q_{M-1}(z) \end{pmatrix} \tag{104}$$

where

$$q_k(z) = z^{-M+1-k}p_k(z^M) \tag{105}$$

From [11] this can be expressed as

$$\boldsymbol{Q}(z) = \sum_{k=0}^{M-1} q_k(z)\boldsymbol{J}_M\boldsymbol{\Pi}'^k \tag{106}$$

with a permutation matrix

$$\boldsymbol{\Pi}' = \begin{pmatrix} 0 & \boldsymbol{I}_{M-1} \\ 1 & 0 \end{pmatrix} \tag{107}$$

Next we show that

$$\boldsymbol{W}^*\boldsymbol{J}_M\boldsymbol{\Pi}'^k\boldsymbol{W}^* = M\boldsymbol{\Lambda}(W^{k-1}) \tag{108}$$

Postmultiplying the mth row of W^* by $J_M\Pi'^k$, we have

$$\begin{aligned}&(1, W^{-m}, W^{-2m}, \ldots, W^{-(M-1)m})J_M\Pi'^k\\ &= (W^{-(M-1)m}, W^{-(M-2)m}, \ldots, W^{-km}, \ldots, 1)\Pi'^k\\ &= (W^{-(k-1)m}, W^{-(k-2)m}, \ldots, 1, W^{-(M-1)m}, \ldots, W^{-km})\end{aligned}$$

Then from (4),

$$\begin{aligned}(W^* J_M \Pi'^k W^*)_{mn} &= \sum_{l=0}^{M-1} W^{-(k-1-l)m} W^{-ln} = W^{-(k-1)m} \sum_{l=0}^{M-1} W^{l(m-n)}\\ &= MW^{-(k-1)m}\delta_{mn}\end{aligned}$$

By (85), (108) is obtained. Therefore from (103), (105), (106), and (108) we have

$$\begin{aligned}\frac{1}{M}H_{\mathrm{AC}}(z)G_{\mathrm{AC}}^T(z) &= \frac{1}{M}W^* Q(z) W^*\\ &= \frac{1}{M}\sum_{k=0}^{M-1} q_k(z) W^* J_M \Pi'^k W^*\\ &= z^{-M}\sum_{k=0}^{M-1} p_k(z^M) z^{-(k-1)} \Lambda(W^{k-1}) \qquad (109)\end{aligned}$$

Since $\Lambda(W^{k-1})$ is diagonal, $H_{\mathrm{AC}}(z)G_{\mathrm{AC}}^T(z)$ is also diagonal. This shows that the output is stationary. □

Let us denote the 0th diagonal element of (109) as

$$T(z) = z^{-M}\sum_{k=0}^{M-1} p_k(z^M) z^{-(k-1)} \qquad (110)$$

Then it can be easily shown that

$$T(zW^i) = \left(\frac{1}{M}H_{\mathrm{AC}}(z)G_{\mathrm{AC}}^T(z)\right)_{i,i} \qquad (111)$$

Then, from (32) and (98), for $i = 0, \ldots, M-1$,

$$F_0(\omega + i2\pi/M) = |T(e^{j(\omega + i2(\pi/M))})|^2 S_x(\omega + i2\pi/M) \quad \text{for } |\omega| \leq \pi/M, \qquad (112)$$

that is,

$$F_0(\omega) = |T(e^{j\omega})|^2 S_x(\omega) \quad \text{for } |\omega| \leq \pi \qquad (113)$$

This shows that the alias-free filter bank characterized by (101) is equivalent to a scalar linear time-invariant system given by (110).

Moreover, a filter bank is said to be a perfect reconstruction (PR) filter bank iff

$$\hat{x}(n) = cx(n - n_0), \quad n_0\text{: integer} \tag{114}$$

for some integer n_0 and some constant $c \neq 0$. The necessary and sufficient condition for this is

$$\boldsymbol{P}(z) = \boldsymbol{R}_g(z)\boldsymbol{E}_h(z) = cz^{-m_0}\begin{pmatrix} 0 & \boldsymbol{I}_{M-r} \\ z^{-1}\boldsymbol{I}_r & 0 \end{pmatrix} \tag{115}$$

for some integer r with $0 \leq r \leq M - 1$ and $Mm_0 = n_0 - r - M + 1$ [7], p. 253. That is,

$$p_k(z) = \begin{cases} cz^{-m_0} & \text{for } k = r \\ 0 & \text{otherwise} \end{cases} \tag{116}$$

and

$$T(z) = cz^{-M}z^{-(r-1)}z^{-Mm_0} = cz^{-n_0} \tag{117}$$

Thus, for a PR filter bank, the output is stationary and $F_0(\omega) = c^2 S_x(\omega)$.

VII. OPTIMIZATION OF FILTER BANKS

We optimize a filter bank with a stationary input signal $x(n)$ by minimizing the averaged variance of the reconstruction error when some subband signals are dropped under the PR condition. We assume that high-pass bands from K to M are dropped and denote the output of the filter bank as $y(n)$ in Fig. 7. Also we define an M-channel process $\hat{\boldsymbol{w}}(n)$ as

$$\hat{\boldsymbol{w}}(n) = (\hat{w}_0(n), \hat{w}_1(n), \ldots, \hat{w}_{M-1}(n))^{\mathrm{T}} \tag{118}$$

The difference $\boldsymbol{e}(n)$ between $\boldsymbol{w}(n)$ and $\hat{\boldsymbol{w}}(n)$ is given by

$$\begin{aligned} \boldsymbol{e}(n) = \boldsymbol{w}(n) - \hat{\boldsymbol{w}}(n) &= \boldsymbol{E}_g^T(q)\,\boldsymbol{v}(n) - \boldsymbol{E}_g^T(q)\begin{pmatrix} \boldsymbol{I}_{M-K} & 0 \\ 0 & 0 \end{pmatrix}\boldsymbol{v}(n) \\ &= \boldsymbol{E}_g^T(q)\boldsymbol{I}'\,\boldsymbol{v}(n) \end{aligned} \tag{119}$$

where

$$\boldsymbol{I}' = \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{I}_K \end{pmatrix} \tag{120}$$

and q^{-1} is the delay operator $q^{-1}x(n) = x(n-1)$. Since $\boldsymbol{v}(n)$ is an M-channel stationary process, so is $\boldsymbol{e}(n)$. Then, as in (96),

$$y(Mn + i) = \hat{w}_i(n) \tag{121}$$

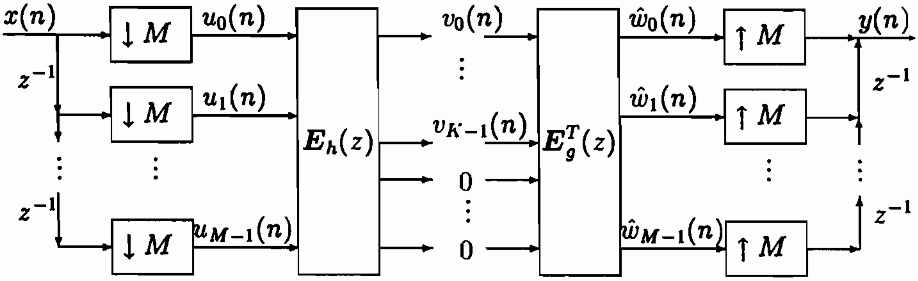


Figure 7 Filter bank when from K to $M-1$ band signals are dropped.

so that by defining the reconstruction error as $e(n) = \hat{x}(n) - y(n)$

$$(e(n))_i = \hat{x}(Mn+i) - y(Mn+i) = e(Mn+i), \tag{122}$$

that is,

$$e(n) = (e(Mn), e(Mn+1), \ldots, e(Mn+M-1))^T \tag{123}$$

Since the spectral density matrix of $e(n)$ is given by

$$S_e(\omega) = \boldsymbol{E}_g(\mathrm{e}^{\mathrm{j}\omega}) \boldsymbol{I}' S_v(\omega) \boldsymbol{I}' \boldsymbol{E}_g^T(\mathrm{e}^{-\mathrm{j}\omega}) \tag{124}$$

and from (94) and Gladyshev's formula (28), by using $\boldsymbol{H}_{\mathrm{AC}}(z)$ and $\boldsymbol{G}_{\mathrm{AC}}(z)$, the cyclic spectral density matrix of $e(n)$ is given by

$$\boldsymbol{F}_e(\omega) = \frac{1}{M^2}(\boldsymbol{H}_{\mathrm{AC}}(z)\boldsymbol{I}'\boldsymbol{G}_{\mathrm{AC}}^T(z))^* S_x(M\omega)\boldsymbol{H}_{\mathrm{AC}}(z)\boldsymbol{I}'\boldsymbol{G}_{\mathrm{AC}}^T(z)\big|_{\mathrm{e}^{-\mathrm{j}\omega}}$$
$$\text{for } |\omega| \leq \pi/M \tag{125}$$

From (29), the averaged variance of the reconstruction error is given by

$$\sigma_e^2 = \int_{-\pi/M}^{\pi/M} \mathrm{tr}\boldsymbol{F}_e(\omega)\,\mathrm{d}\omega \tag{126}$$

When $M=2$ and the high-pass band signal is dropped, we have

$$\boldsymbol{H}_{\mathrm{AC}}(z)\boldsymbol{I}'\boldsymbol{G}_{\mathrm{AC}}^T(z) = \begin{pmatrix} H_0(z) & H_1(z) \\ H_0(-z) & H_1(-z) \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} G_0(z) & G_0(-z) \\ G_1(z) & G_1(-z) \end{pmatrix}$$

$$= \begin{pmatrix} H_1(z) \\ H_1(-z) \end{pmatrix} (G_1(z) \quad G_1(-z))$$

So, (126) reduces to

$$\sigma_e^2 = \frac{1}{4}\int_{-\pi/2}^{\pi/2} \mathrm{tr}\left\{\begin{pmatrix} G_1(\mathrm{e}^{\mathrm{j}\omega}) \\ G_1(-\mathrm{e}^{\mathrm{j}\omega}) \end{pmatrix} (H_1(\mathrm{e}^{\mathrm{j}\omega}) \quad H_1(-\mathrm{e}^{\mathrm{j}\omega})) \right.$$

$$\left. + \begin{pmatrix} S_x(\omega) & 0 \\ 0 & S_x(\omega+\pi) \end{pmatrix} \begin{pmatrix} H_1(\mathrm{e}^{-\mathrm{j}\omega}) \\ H_1(-\mathrm{e}^{-\mathrm{j}\omega}) \end{pmatrix} (G_1(\mathrm{e}^{-\mathrm{j}\omega}) \quad G_1(-\mathrm{e}^{-\mathrm{j}\omega})) \right\} \mathrm{d}\omega$$

$$= \frac{1}{4}\int_{-\pi/2}^{\pi/2} (|G_1(\mathrm{e}^{\mathrm{j}\omega})|^2 + |G_1(\mathrm{e}^{\mathrm{j}(\omega+\pi)}|^2)|H_1(\mathrm{e}^{\mathrm{j}\omega})|^2 S_x(\omega)\,\mathrm{d}\omega$$

$$+ \ \frac{1}{4}\int_{-\pi/2}^{\pi/2} (|G_1(\mathrm{e}^{\mathrm{j}\omega})|^2 + |G_1(\mathrm{e}^{\mathrm{j}(\omega+\pi)})|^2)|H_1(\mathrm{e}^{\mathrm{j}(\omega+\pi)})|^2$$

$$\times S_x(\omega+\pi)\,\mathrm{d}\omega$$

$$= \frac{1}{4}\int_{0}^{2\pi} (|G_1(\mathrm{e}^{\mathrm{j}\omega})|^2 + |G_1(\mathrm{e}^{\mathrm{j}(\omega+\pi)})|^2)|H_1(\mathrm{e}^{\mathrm{j}\omega})|^2 S_x(\omega)\,\mathrm{d}\omega \tag{127}$$

where we use that $|G_1(\mathrm{e}^{\mathrm{j}\omega})|^2 + |G_1(\mathrm{e}^{\mathrm{j}(\omega+\pi)})^2|, |H_1(\mathrm{e}^{\mathrm{j}\omega})|^2$ and $S_x(\omega)$ are periodic functions of ω with period 2π. Now let us consider the PR filter bank. If $H_0(z)$ and $H_1(z)$ are causal finite impulse response (FIR) filters with order N and we put

$$G_0(z) = H_1(-z), \qquad G_1(z) = -H_0(-z) \tag{128}$$

then, the PR condition (117) reduces to

$$T(z) = \frac{1}{2}(H_1(z)\,G_1(z) - H_1(-z)\,G_1(-z)) = cz^{-n_0}$$

$$n_0\text{: odd integer, } c \neq 0 \tag{129}$$

This filter bank is also called biorthogonal [14]. We put $c = 1$ without loss of generality. Then the problem reduces to designing one pair of filters among $H_0(z)$, $H_1(z)$, $G_0(z)$, and $G_1(z)$ to minimize (127) under the PR condition (129).

VIII. SOME EXAMPLES

In this section, we give some numerical results for the conjugate quadrature filter (CQF) bank and the PR filter bank. The properties of the CQF bank are shown in [7].

By choosing

$$H_1(z) = (-z)^{-N}\tilde{H}_0(-z) \tag{130}$$

and imposing the power complementary condition

$$\tilde{H}_0(z)H_0(z)+\tilde{H}_0(-z)H_0(-z)=2 \tag{131}$$

(129) becomes

$$T(z)=z^{-N} \tag{132}$$

where $\tilde{H}(z)=\sum_{n=0}^{N}h^*(n)z^n$ and N is odd. This shows that the PR condition is satisfied. So, the criterion in (127) reduces to

$$\frac{1}{2}\int_0^{\pi}|H_1(e^{j\omega})|^2 S_x(\omega)\,d\omega \tag{133}$$

This is originally derived in [9]. The filters satisfying (130) and (131) are said to be conjugate quadrature filters (CQF) or orthogonal. It is shown in [7] that this CQF bank is parametrized by $(N+1)/2$ free parameters.

We use the Newton–Raphson method for optimization. It is very difficult to obtain a global solution of a nonlinear optimization problem with many variables. For the PR filter bank with N free parameters we use the coefficients of the optimal CQF bank as the initial values for optimization.

For the input signal with zero mean and the covariance $R_x(n)=0.9^{|n|}$, the frequency responses of optimized $H_0(z)$ and $H_1(z)$ with order $N=7$ for CQF and PR filter banks are shown in Figs 8 and 9, respectively.

The averaged variance of the reconstruction error σ_e^2 of the CQF bank are 0.038916, 0.036127, and 0.035031 for $N=3$, 5, 7, respectively, while those of the PR bank are 0.038846, 0.036045, and 0.034997 for $N=3$, 5, 7, respectively. These are much smaller than the one-step prediction error variance 0.19. Although the difference between the variances of both optimal filter banks for $N=7$ is very small, the corresponding frequency responses are fairly different. The numerical result for the PR linear phase filter bank has been presented in [15].

IX. CONCLUSION

In this article, first the spectral theory of cyclostationary processes has been reviewed from an engineering viewpoint. Two multichannel stationary processes are constructed and the relation between the two spectral density matrices is presented as Gladyshev's formula. Then, a parametric expression for a periodic AR process has been given.

Second, since the output of the filter bank for the stationary input is a cyclostationary process, its cyclic spectral density matrix has been

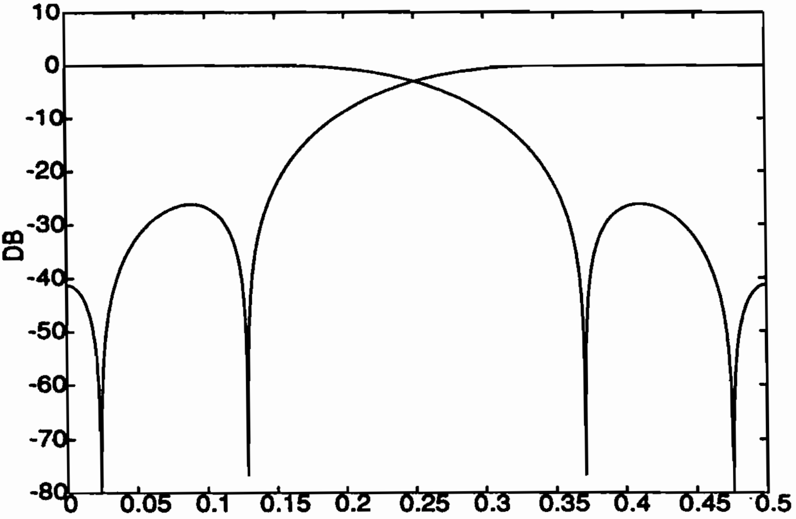


Figure 8 Frequency responses $H_0(e^{j\omega})$ and $H_1(e^{j\omega})$ of an 8-tap optimal CQF bank ($N = 7$).

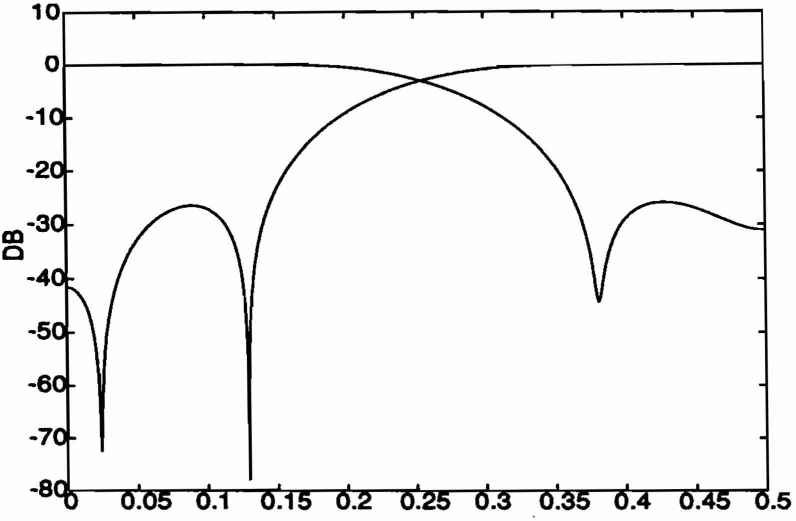


Figure 9 Frequency responses $H_0(e^{j\omega})$ and $H_1(e^{j\omega})$ of an 8-tap optimal PR bank ($N = 7$).

derived. Using this result, we have derived the general criterion to obtain optimal 2-band PR filter banks that minimize the averaged variance of the reconstruction error when the high-pass band signal is dropped. Some numerical results have been presented.

REFERENCES

1. EG Gladyshev. Sov Math 2:385–388, 1961.
2. H Ogura. IEEE Trans Inform Theory IT-17:143–149, 1971.
3. WA Gardner, LE Franks. IEEE Trans Inform Theory IT-21:4–14, 1975.
4. WA Gardner. Introduction to Random Processes with Applications to Signal and Systems. New York: McGraw Hill, 1989.
5. H Ogura, Y Yoshida. Trans Inst Electr Comm Japan J65-A:22–29, 1982 (in Japanese).
6. H Sakai. Electr Comm Japan 73:9–15. Scripta Technica, Inc, 1990.
7. PP Vaidyanathan. Multirate Systems and Filter Banks. Englewood Cliffs: Prentice Hall, 1993.
8. AN Akansu, RA Haddad. Multiresolution Signal Decomposition. New York: Academic Press, 1992.
9. L Vandendorpe. Signal Processing 29:237–249, 1992.
10. H Sakai. IEEE Trans Acoust Speech Signal Processing ASSP-30:279–287, 1982.
11. PJ Davis. Circulant Matrices. New York: John Wiley, 1979.
12. M Pagano. Ann Statist 6:1310–1317, 1978.
13. H Sakai. J Time Ser Anal 12:73–82, 1991.
14. M Vetterli, C Herley. IEEE Trans Signal Processing 40:2207–2232, 1992.
15. S Ohno, H Sakai. Proc ICASSP, vol. 2, 1995, pp 1292–1295, Detroit; IEEE Trans Signal Processing 44:2718–2725, 1996.

13

Stochastic System Identification Using Polyspectra

Jitendra K. Tugnait

Auburn University, Auburn, Alabama

I. INTRODUCTION

It is common practice in the system identification literature to assume that measurements of the system output are noisy but the measurements of the input to the system are *perfect*. This assumption is not necessarily true in system and control applications where the input is not under the analyst's (complete) control; rather, it can only be measured. Clearly, it may not always be possible to neglect the noise introduced by the sensor itself, or by the ambient environment.

In this chapter we consider multivariable system identification under "symmetric modeling" [1,2] of stochastic systems by allowing the input measurements also to be noise contaminated, in addition to having noisy output measurements. Such models are called *errors-in-variables* models in the econometrics literature. Past work on dynamic system identification with noisy input has concentrated overwhelmingly on exploitation of second-order statistics [1–4]. It is known that, in general, there does not exist a unique solution when only second-order statistics are used. Use of higher-order statistics can alleviate this problem at the cost of higher-variance estimates [5–9]. Our recent work [9] dealing with single-input/single-output models has utilized frequency-domain approaches coupled with a novel entity, the integrated polyspectrum, which leads to computationally simpler and statistically more accurate parameter estimates than heretofore possible with higher-order statistics, including the approaches of [5–8]. One of the objectives of this chapter is to extend the approach of [9] to multivariable models.

In this chapter we consider a specific class of systems where the input process is non-Gaussian and the measurement noise at the input as well as the output is Gaussian if the input process has nonvanishing trispectrum. The noise processes are allowed to be non-Gaussian with vanishing bispectrum but nonvanishing trispectrum if the input process has nonvanishing bispectrum. Clearly, this model may not always be appropriate but there are several situations of practical interest where such assumptions are valid. For instance, a pseudo-random binary sequence is often used to probe a control system for identification purposes [3]; such sequences are clearly non-Gaussian with nonvanishing trispectrum. On the other hand, the proposed approaches will fail if the system input process is Gaussian.

There exists a large body of literature ([12–16] and references therein) where system identification is carried out using frequency response data (i.e., transfer function measurements at a set of frequencies). In contrast to these approaches, our starting-point is time-domain input–output data. The class of inputs in (most of) these papers ([12–16]) is restricted to multisines (periodic inputs) whereas we use "arbitrary" (non-Gaussian stochastic; pseudo-random binary sequences for instance) inputs. Once we transform the time-domain data into frequency-domain using integrated polyspectra (see Section III), some of the existing techniques become applicable for parameter estimation, but not necessarily for analysis. For instance, the parameter estimation formulation of Section III.A is the same as in [13] (see also [12]) and that of Section III.C is quite similar to that of [14]. Other approaches following [15] and [16] (and others) are possible for parameter estimation given the system multivariable transfer function. Note, however, that [15–16] require measurements of the transfer function to be available on certain frequencies that are equidistant on $[0, \pi]$; in our formulation frequencies can be located arbitrarily on $(0, \pi)$.

Consistency results proven in [15] are based on letting N_f (number of frequencies) $\rightarrow \infty$. Our consistency results are based upon letting T (number of time samples) $\rightarrow \infty$. Implicit in the formulation of [15] is that $T \rightarrow \infty$. This follows from the fact that the noise spectrum in [15] is taken to be complex Gaussian and (mutually) independent over the grid of frequencies $2\pi k/N_f$, $(k = 0, 1, \ldots, N_f - 1)$, where N_f is the number of time samples in a "batch" (cf. [10] Section 4.4). Independence holds only for frequencies on the above-mentioned grid and this assumption is crucial for the consistency results of [15] to hold true. Clearly the above number of frequencies cannot tend to ∞ without the batch size also tending to ∞, else the independence assumption will not hold true.

Past approaches to the problem of stochastic linear system identification may be divided into two classes: those that exploit only the second-order statistics, and those that use higher (higher than second)-order cumulant statistics. A good survey of the work done prior to about 1980 is given in [19]. For later work, see [1–9] and [20–29]. Higher-order statistics have been exploited in [1], [5–9], [22], [23], [25], [28], and [29].

Söderström [19] allows only white additive noise at the input, and furthermore, the input and the output noises are assumed to be mutually uncorrelated. Most of the early work in this area has been done in econometrics where such models have been called errors-in-variables models. When only second-order statistics are exploited, it is known that, in general, there does not exist a unique solution [20–24]. Therefore, attention has been focused on characterization of the class of transfer functions which fit the data. Fairly complete results are available for the scalar dynamic case [20–23] but the multivariable case remains incomplete [24].

The use of higher-order cumulant statistics [30] can, in principle, yield consistent parameter estimates. Deistler [22] (see also [28]) has shown how to estimate the transfer function of an SISO system in the frequency domain by use of the higher-order cumulant spectrum of the output and the higher-order cumulant cross-spectrum of the input–output record. That is, [22] presents a nonparametric approach.

Instrumental variable-type approaches have been presented in [6] using third-order cumulants and in [8] using fourth-order cumulants of the input–output record. The consistency results of [6] and [8] have been proven only for i.i.d. (independent and identically distributed) inputs. Otherwise one needs to check the invertibility of a matrix which depends upon the unknown parameters to be estimated; therefore, the invertibility condition check is not practical. The consistency results of the approach proposed in this chapter hold under far more general conditions.

In [7] a novel cost function involving the third-order cumulants of the input–output data has been proposed and it has been shown to be proportional to a conventional mean-square error criterion based upon noiseless data. Its extension to the fourth-order cumulant case has not been provided in [7]. Note that when the systen input has a symmetric PDF, its third-order statistics will vanish, rendering the approaches based upon third-order statistics useless. One example is that of a pseudo-random binary sequence as the system input. Consistency of the approach of [7] has been established under several restrictive conditions such as the system input $u(t)$ is a linear process with nonzero power spectrum at zero frequency. It should be noted that, unlike the second-order statistics case, one cannot, in general, model a stationary random process with a given higher-order cumulant spectrum as having been generated by driving a linear system with an i.i.d. sequence [31]. In this chapter we do not require any such constraint. Moreover, our approaches also apply to fourth-order statistics case unlike [7]. Also [7] requires that the noise processes, if non-Gaussian with symmetric PDF, should be linear processes. We do not need such an assumption in this chapter. In [29] several linear/iterative approaches using the auto- and/or cross-third-order cumulants of the input-output processes have been presented. Conditions under which the proposed approaches will yield consistent parameter estimators have not been provided in [29]. The simulation results presented in [29] show that the cross-cumulants-based approach of [29] outperforms the approach of [7] by a wide margin for the presented example.

In [5] the square root of the magnitude of the fourth cumulant of a generalized error signal is proposed as a performance criterion for parameter estimation. Both single-input single-output and multiple-input multiple-output models have been considered in [5]. Strong consistency of the proposed parameter estimator has been established for linear inputs in [5] for Gaussian noise processes. The approach of [5] results in a nonlinear estimator that requires a good initial guess for convergence; unfortunately, no method for reliable initialization was provided in [5].

In this chapter attention is focused on frequency-domain approaches where the integrated polyspectrum (bispectrum or trispectrum) of the input and the integrated cross-polyspectrum, respectively, of the input–output are exploited. Two new classes of parametric, frequency-domain approaches are proposed and analyzed. An integrated polyspectrum-based persistence of excitation condition on system input is defined and related to parameter identifiability of the multivariable system. Both classes of the parameter estimators are shown to be consistent in any measurement noise sequences with vanishing bispectra when integrated bispectrum-based approaches are used. The proposed parameter estimators are shown to be consistent in Gaussian measurement noise when integrated trispectrum-based approaches are used. The input to the system need not be a linear process but must have nonvanishing bispectrum or trispectrum. Performance analysis of the parameter estimators is also presented.

The chapter is organized as follows. In Section II a more precise statement of the parameter estimation problem under consideration is provided along with a definition and some analysis of the integrated polyspectrum of interest in this chapter. The integrated polyspectrum (bispectrum and trispectrum)-based approaches are described in Section III which includes estimation of the integrated polyspectra, their large sample properties, a linear parameter estimator, and a nonlinear parameter estimator the cost function of which is asymptotically equivalent to some negative log-likelihood function. Consistency of the proposed parameter estimators is discussed in Section IV under some mild sufficiency conditions. Performance analysis of the proposed estimators is presented in Section V.

II. MODEL ASSUMPTIONS AND INTEGRATED POLYSPECTRUM

A. Model Assumptions

The system model is given by

$$\mathbf{s}(t) = -\sum_{i=1}^{n_a} A_i\mathbf{s}(t-i) + \sum_{i=1}^{n_b} B_i\mathbf{u}(t-i) \tag{1}$$

where the p-column vector $\mathbf{s}(t)$ is the system output and the m-column vector $\mathbf{u}(t)$ is the system input. Noisy measurements of the system input and output are available as

$$\mathbf{x}(t) = \mathbf{u}(t) + \mathbf{v}_i(t), \qquad \mathbf{y}(t) = \mathbf{s}(t) + \mathbf{v}_o(t) \tag{2}$$

Define the unknown parameter matrix

$$\theta = \{\text{vec}\{A_i\},\quad 1 \leq i \leq n_a, \qquad \text{vec}\{B_j\},\quad 1 \leq j \leq n_b\} \tag{3}$$

where vec denotes the column stacking operator. We wish to estimate θ given certain statistics of the input–output data record $\{\mathbf{x}(t), \mathbf{y}(t), t = 1, 2, \ldots\}$.

The following conditions are assumed to be true.

(AS1) $\det(A(z)) \neq 0$ for $|z| \geq 1$ where $A(z) = I + \Sigma_{i=1}^{n_a} A_i z^{-i}$ and z is a complex variable. Moreover, $A(z)$ and $B(z) = \Sigma_{i=1}^{n_b} B_i z^{-i}$ are coprime and rank $[A_{n_a} \vdots B_{n_b}] = p$.

(AS2) All the processes involved (i.e., $\mathbf{x}(t)$, $\mathbf{y}(t)$, $\mathbf{v}_i(t)$, and $\mathbf{v}_o(t)$) are zero-mean and jointly stationary. Furthermore, the noise sequences $\{\mathbf{v}_i(t)\}$ and $\{\mathbf{v}_o(t)\}$ are independent of $\{\mathbf{u}(t)\}$, hence of $\{\mathbf{s}(t)\}$.

(AS3) The cumulant/cross-cumulant sequences of the various processes involved satisfy the following summability conditions:

$$\sum_{\tau_1,\ldots,\tau_{k-1}=-\infty}^{\infty} [1 + |\tau_j|]\,|C_{z_1z_2\cdots z_k}(\tau_1, \ldots, \tau_{k-1})| < \infty$$

for each $j = 1, 2, \ldots, k-1$ and each $k = 2, 3, \ldots$ where $z_i(t)$ denotes a component of any of the involved processes such as $\mathbf{u}(t)$, $\mathbf{s}(t)$, $\mathbf{v}(t)$, etc., and $C_{z_1z_2\cdots z_k}(\tau_1, \ldots, \tau_{k-1})$ denotes the kth joint cumulant of the random variables $\{z_1(t+\tau_1), \ldots, z_{k-1}(t+\tau_{k-1}), z_k(t)\}$.

(AS4) The noise processes are jointly Gaussian if we exploit the integrated trispectrum of the data. The noise processes are assumed to have vanishing bispectrum if we exploit the integrated bispectrum of the data.

Finally define the set of unknown parameters

$$\Theta = \{\theta | \det(A(z;\theta)) \neq 0 \text{ for } |z| \geq 1\}$$

where $A(z;\theta)$ denotes $A(z)$ explicitly parametrized by θ.

Remarks. Condition (AS1) ensures that the full polynomial canonical form of the multivariable ARMA model transfer function is unique [3], Lemma C6.1.1. Condition (AS3) is needed to invoke certain asymptotic results from [10], Chap. 7 concerning the various polyspectral estimators discussed in Section III.A.

B. Polyspectrum and Integrated Polyspectrum

Consider a zero-mean, scalar, real-valued stationary stochastic process $\{s(t)\}$. Its third-order cumulant function $C_{sss}(i,k)$ is given by

$$C_{sss}(i,k) := E\{s(t+i)s(t+k)s(t)\} \tag{4}$$

Its bispectrum $B_{sss}(\omega_1,\omega_2)$ is defined as

$$B_{sss}(\omega_1,\omega_2) = \sum_{i=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} C_{sss}(i,k) \exp\{-j(\omega_1 i + \omega_2 k)\} \tag{5}$$

From the above definitions it is easy to see that

$$C_{sss}(i,k) = \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} B_{sss}(\omega_1,\omega_2) \exp\{j(\omega_1 i + \omega_2 k)\}\, d\omega_1\, d\omega_2 \tag{6}$$

Define

$$w(t) := s^2(t) - E\{s^2(t)\} \qquad \text{and} \qquad \tilde{w}(t) := s^2(t) \tag{7}$$

Then both the cross-spectrum between the process $\{\tilde{w}(t)\}$ and $\{s(t)\}$ and the cross-spectrum between the process $\{w(t)\}$ and $\{s(t)\}$ are given by

$$\begin{aligned} S_{ws}(\omega) &:= \sum_{k=-\infty}^{\infty} E\{w(t+k)s(t)\} \exp\{-j\omega k\} \\ &= \sum_{k=-\infty}^{\infty} C_{sss}(k,k) \exp\{-j\omega k\} = S_{\tilde{w}s}(\omega) \end{aligned} \tag{8}$$

It then follows that

$$C_{sss}(k,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\bar{w}s}(\omega)\exp\{j\omega k\}\,\mathrm{d}\omega = C_{sss}(0,-k)$$
$$= C^*_{sss}(0,-k) \tag{9}$$

where * denotes complex conjugation. Compare (6) with (9) to deduce that

$$S^*_{\bar{w}s}(\omega) = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_{sss}(\omega,\omega_2)\,\mathrm{d}\omega_2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} B_{sss}(\omega_1,\omega)\,\mathrm{d}\omega_1 = S^*_{ws}(\omega) \tag{10}$$

Notice that the cross-spectrum between the signal $s(t)$ and its square can be interpreted as an integrated bispectrum of $s(t)$. This integrated bispectrum will form a basis (along with the integrated trispectrum, to be defined later) for unknown parameter estimation. It is easy to see that since the bispectrum of a Gaussian process is identically zero, so is its integrated bispectrum. In the sequel it will be easier to work with the centered (zero-mean) $s^2(t)$, i.e., $w(t)$.

Turning to the trispectrum, it is defined as

$$T_{ssss}(\omega_1,\omega_2,\omega_3) := \sum_{i=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty} C_{ssss}(i,k,l)$$
$$\exp\{-j(\omega_1 i + \omega_2 k + \omega_3 l)\} \tag{11}$$

where

$$\begin{aligned} C_{ssss}(i,k,l) := {} & E\{s(t)s(t+i)s(t+k)s(t+l)\} \\ & - E\{s(t)s(t+i)\}\,E\{s(t+k)s(t+l)\} \\ & - E\{s(t)s(t+k)\}\,E\{s(t+l)s(t+i)\} \\ & - E\{s(t)s(t+l)\}\,E\{s(t+k)s(t+i)\} \end{aligned} \tag{12}$$

is the fourth-order cumulant function of the process $\{s(t)\}$. It then follows that

$$C_{ssss}(i,k,l) = \frac{1}{(2\pi)^3}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} T_{ssss}(\omega_1,\omega_2,\omega_3)\,e^{j(\omega_1 i + \omega_2 k + \omega_3 l)}$$
$$\times\,\mathrm{d}\omega_1\,\mathrm{d}\omega_2\,\mathrm{d}\omega_3 \tag{13}$$

Define

$$\bar{r}(t) := s^3(t) - 3s(t)\,E\{s^2(t)\} \qquad \text{and} \qquad r(t) = \bar{r}(t) - E\{\bar{r}(t)\} \tag{14}$$

Then both the cross-spectrum between the process $\{\bar{r}(t)\}$ and $\{s(t)\}$ and the cross-spectrum between the process $\{r(t)\}$ and $\{s(t)\}$ are given by

$$\begin{aligned} S_{rs}(\omega) &:= \sum_{k=-\infty}^{\infty} E\{r(t+k)s(t)\}\exp\{-j\omega k\} \\ &= \sum_{k=-\infty}^{\infty} C_{ssss}(k,k,k)\exp\{-j\omega k\} = S_{\bar{r}s}(\omega) \end{aligned} \tag{15}$$

It then follows that

$$\begin{aligned} C_{ssss}(k,k,k) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\bar{r}s}(\omega)\exp\{j\omega k\}\,\mathrm{d}\omega \\ &= C_{ssss}(0,0,-k) = C^{*}_{ssss}(0,0,-k) \end{aligned} \tag{16}$$

Compare (13) with (16) to deduce that

$$S^{*}_{\bar{r}s}(\omega) = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} T_{ssss}(\omega,\omega_2,\omega_3)\,\mathrm{d}\omega_2\mathrm{d}\omega_3 = S^{*}_{rs}(\omega) \tag{17}$$

Notice that the cross-spectrum between the signal $s(t)$ and a function of its cube can be interpreted as an integrated trispectrum of $s(t)$.

We now turn to vector processes. Define the m-column vectors $\mathbf{r}_{3u}(t)$ and $\mathbf{r}_{2u}(t)$ whose ith components $r_{3ui}(t)$ and $r_{2ui}(t)$, respectively, are given by

$$r_{3ui}(t) = u_i^3(t) - 3u_i(t)E\{u_i^2(t)\} - E\{u_i^3(t)\} \tag{18}$$

$$r_{2ui}(t) = u_i^2(t) - E\{u_i^2(t)\} \tag{19}$$

where $u_i(t)$ is the ith component of $\mathbf{u}(t)$. We will use the notation $\mathbf{r}_{3x}(t)$ and $\mathbf{r}_{2x}(t)$ to denote the above vectors when $\mathbf{u}(t)$ is replaced with $\mathbf{x}(t)$. Mimicking the univariate formulation of (4)–(17), we now define the "component-by-component" integrated bispectrum of $\mathbf{u}(t)$ as

$$S_{r_{2u}u}(\omega) = \sum_{k=-\infty}^{\infty} E\{\mathbf{r}_{2u}(t+k)\mathbf{u}^T(t)\}\,\mathrm{e}^{-\mathrm{j}\omega k} \tag{20}$$

Similarly we have the integrated trispectrum

$$S_{r_{3u}u}(\omega) = \sum_{k=-\infty}^{\infty} E\{\mathbf{r}_{3u}(t+k)\mathbf{u}^T(t)\}\,\mathrm{e}^{-\mathrm{j}\omega k} \tag{21}$$

III. TWO PARAMETER ESTIMATORS

Consider the cross-spectrum ($l = 2, 3$)

$$S_{r_{lx}y}(\omega) = \sum_{k=-\infty}^{\infty} E\{\mathbf{r}_{lx}(t+k)\mathbf{y}^T(t)\}\,\mathrm{e}^{-\mathrm{j}\omega k} \tag{22}$$

It then follows from (1)–(3), assumption (AS2), (20), (21), and (22) that

$$\begin{aligned} S_{r_{lx}y}(\omega) &= S_{r_{lu}s}(\omega) = S_{r_{lu}u}(\omega)\, G^{\mathcal{H}}(\mathrm{e}^{\mathrm{j}\omega}) \\ &= S_{r_{lx}x}(\omega)\, G^{\mathcal{H}}(\mathrm{e}^{\mathrm{j}\omega}) \quad (l = 2, 3) \end{aligned} \tag{23}$$

where $G(z) = A^{-1}(z)\,B(z)$ is the transfer function of (1), and $\mathcal{H}$ denotes the complex conjugate transpose (hermitian) operation. Therefore, if $S^{-1}_{r_{lu}xu}(\omega)$ exists, then we have the transfer function matrix expression

$$G(\mathrm{e}^{\mathrm{j}\omega}) = [S^{-1}_{r_{lx}x}(\omega)\, S_{r_{lx}y}(\omega)]^{\mathcal{H}} \quad (l = 2, 3) \tag{24}$$

The basic approach to model parameter estimation consists of two steps. First obtain a consistent estimator $\hat{G}(\mathrm{e}^{\mathrm{j}\omega})$ of $G(\mathrm{e}^{\mathrm{j}\omega})$ $(=G(\mathrm{e}^{\mathrm{j}\omega}; \theta_0))$ via the consistent estimators $\hat{S}^{-1}_{r_{lx}x}(\omega)$ and $\hat{S}_{r_{lx}y}(\omega)$ of $S^{-1}_{r_{lx}x}(\omega)$ and $S_{r_{lx}y}(\omega)$, respectively, based upon the input–output record $\{\mathbf{x}(t), \mathbf{y}(t), t = 1, 2, \ldots, T\}$. Next estimate the system parameters using the estimated transfer function matrix as "data".

A. Estimation of Integrated Polyspectrum

This is little more than estimation of the cross-spectrum between $\{\mathbf{x}(t)\}$ (or $\{\mathbf{y}(t)\}$) and $\{\mathbf{r}_{lx}(t)\}$ $(l = 2, 3)$. Numerous techniques are available for this purpose; see [10] and references therein. We will follow the approach of smoothing in the frequency domain [10], Section 7.4. Given a record of length T, let $\mathbf{X}(\omega)$ denote the DFT of $\{\mathbf{x}(t),\ 1 \le t \le T\}$ given by

$$\mathbf{X}(\omega_k) = \sum_{t=0}^{T-1} \mathbf{x}(t+1) \exp(-\mathrm{j}\omega_k t) \tag{25}$$

where

$$\omega_k = \frac{2\pi}{T} k, \quad k = 0, 1, \ldots, T-1 \tag{26}$$

Similarly define $\mathbf{Y}(\omega_k)$. Also let $\mathbf{R}_{lx}(\omega)$ denote the DFT of $\{\mathbf{r}_{lx}(t)\}$ obtained by using the relations (recall (18) and (19))

$$r_{3xj}(t) = x_j^3(t) - 3x_j(t)\,\hat{\mu}_{2xy} \tag{27}$$

$$\hat{\mu}_{2xj} = \frac{1}{N} \sum_{t=1}^{N} x_j^2(t) \tag{28}$$

$$r_{2xj}(t) = x_j^2(t) \tag{29}$$

$(j = 1, 2, \ldots, m)$. [Note that the mean of $\mathbf{r}_{lx}(t)$ has not been subtracted since it influences only the d.c. (frequency ω_0 in (26)) which is not used for parameter estimation purposes in this chapter.]

Given the above DFTs, following [10], Section 7.4 we define the cross-spectrum estimators as

$$\hat{S}_{r_{lx}x}(k) = \frac{2\pi}{T^2} \sum_{s=1}^{T-1} W^{(T)}(\omega_{k-s}) \mathbf{R}_{lx}(\omega_{k-s}) [\mathbf{X}(\omega_{k-s})]^{\mathcal{H}} \tag{30}$$

$$\hat{S}_{r_{lx}y}(k) = \frac{2\pi}{T^2} \sum_{s=1}^{T-1} W^{(T)}(\omega_{k-s}) \mathbf{R}_{lx}(\omega_{k-s}) [\mathbf{Y}(\omega_{k-s})]^{\mathcal{H}} \tag{31}$$

for $1 \leq k \leq T-1$, where the scalar weighting function $W^{(T)}(\alpha)$ is given by

$$W^{(T)}(\alpha) = B_T^{-1} \sum_{i=-\infty}^{\infty} W(B_T^{-1}[\alpha + 2\pi i]) \tag{32}$$

such that $W(\beta), -\infty < \beta < \infty$, is real-valued, even, of bounded variation satisfying $\int_{-\infty}^{\infty} W(\beta)\, d\beta = 1$ and $\int_{-\infty}^{\infty} |W(\beta)|\, d\beta < \infty$ [10], Sections 5.6 and 7.4. It is convenient to take $W(\beta) = 0$ for $|\beta| > 2\pi$ and $W(\beta) = (4\pi)^{-1}$ for $|\beta| \leq 2\pi$. In this case (30) involves uniform weighting of the $2B_T T + 1$ cross-periodogram ordinates whose frequencies fall in the interval $(\omega_k - 2\pi B_T, \omega_k + 2\pi B_T)$. Thus (30) reduces to

$$\hat{S}_{r_{lx}x}(k) = \frac{1}{T(2m_T + 1)} \sum_{i=-m_T}^{m_T} \mathbf{R}_{lx}(\omega_{k-i}) [\mathbf{X}(\omega_{k-i})]^{\mathcal{H}} \tag{33}$$

where $m_T = B_T T$. Similar modification holds for (31).

Let us choose B_T to be such that as $T \rightarrow \infty$, we have $B_T \rightarrow 0$ and $B_T T \rightarrow \infty$. Let $k_l(T)$ with $T = 1, 2, \ldots$ be a sequence of integers such that $\lim_{T\rightarrow\infty} k_l(T)/T = f_l$, a fixed frequency (in Hz). Under the cumulant summability conditions stated in assumption (AS3), it follows from [10], Thm. 7.4.1 that as $T \rightarrow \infty$, we have

$$\begin{aligned} \lim_{T\rightarrow\infty} E\{\hat{S}_{r_{lx}x}(k_l(T))\} &= S_{r_{lx}x}(2\pi f_l), \\ \lim_{T\rightarrow\infty} E\{\hat{S}_{r_{lx}y}(k_l(T))\} &= S_{r_{lx}y}(2\pi f_l) \end{aligned} \tag{34}$$

Moreover, it follows from [10], Thm. 7.4.3 that for $0 \leq f_l, f_m \leq 0.5$,

$$\begin{aligned} &\lim_{T\rightarrow\infty} \Delta_T \operatorname{cov}\{[\hat{S}_{r_{lx}x}(k_l(T))]_{i_1,j_1}, [\hat{S}_{r_{lx}y}(k_m(T))]_{i_2,j_2}\} \\ &\qquad = \delta(l-m)[S_{r_{lx}r_{lx}}(2\pi f_l)]_{i_1,i_2}[S_{yx}(2\pi f_l)]_{j_2,j_1} \end{aligned} \tag{35}$$

$$\begin{aligned} &\lim_{T\rightarrow\infty} \Delta_T \operatorname{cov}\{[\hat{S}_{r_{lx}x}(k_l(T))]_{i_1,j_1}, [\hat{S}^*_{r_{lx}y}(k_m(T))]_{i_2,j_2}\} \\ &\qquad = \delta(l-m)[S_{r_{lx}y}(2\pi f_l)]_{i_1,i_2}[S_{r_{lx}x}(2\pi f_l)]_{j_2,j_1} \end{aligned} \tag{36}$$

where

$$\Delta_T = \frac{B_T T}{2\pi \int_{-\infty}^{\infty} W^2(\alpha)\, \mathrm{d}\alpha} = 2B_T T \quad \text{(if (33) is used)}, \tag{37}$$

$\operatorname{cov}\{X, Y\} = E\{XY^{\mathscr{H}}\} - E\{X\}E\{Y^{\mathscr{H}}\}, [A]_{ij}$ denotes the ijth element of matrix A, and the other covariances may be deduced in a similar manner. Convergence in (34)–(36) is uniform in f.

In light of (33) define a coarser frequency grid:

$$\omega_l = \frac{2\pi l}{L_T} = \frac{2\pi l(2m_T+1)}{T} = \frac{2\pi l(2B_T T+1)}{T}$$
$$\text{with } l = 0, 1, \ldots, L_{T-1} \tag{38}$$

where $L_T = \lfloor T/(2m_T+1)\rfloor$. Then by Theorems 7.2.4, 7.3.6 and 7.4.4 of [10] (see also discussion on p. 252 of [10]), it follows that for large T (as $B_T \to 0$ and $B_T T \to \infty$), $\hat{S}_{r_{lx}x}(k)$ and $\tilde{S}_{r_{lx}y}(k)$ for $k = l(2m_T+1)$, $l = 0, 1, \ldots, (L_T/2) - 1$, are (asymptotically) jointly Gaussian with covariance structure specified by (35) and (36). It follows therefore that $\hat{S}_{r_{lx}x}(k)$ and $\hat{S}_{r_{lx}y}(k)$ for $k = l(2m_T+1)$, $l = 0, 1, \ldots, (L_T/2) - 1$, are asymptotically independent. Results (34)–(36) imply weak consistency of the estimators (30) and (31). However, by [10], Thm. 7.7.5 the estimators (30) and (31) are also *strongly* consistent.

Using the estimated integrated polyspectra we have an estimator of the system transfer function matrix at frequency ω_k

$$\hat{G}(e^{j\omega_k}) = [\hat{S}^{-1}_{r_{lx}x}(k)\hat{S}_{r_{lx}y}(k)]^{\mathscr{H}} \quad (l = 2, 3), \tag{39}$$

provided that $\hat{S}^{-1}_{r_{lx}x}(k)$ exists. If $S^{-1}_{r_{lx}x}(\omega_k)$ exists, then it follows from [11] Prop. 1.2.16 (see also proof of [10] Thm. 8.11.1) that

$$\begin{aligned}\lim_{T\to\infty}\hat{G}(e^{j2\pi f}) &= \lim_{T\to\infty}[\hat{S}^{-1}_{r_{lx}x}(k(T))\hat{S}_{r_{lx}y}(k(T))]^{\mathscr{H}} \\ &= G(e^{j2\pi f}) \quad \text{w.p.1}\end{aligned} \tag{40}$$

where $\lim_{T\to\infty}k(T)/T = f$. Convergence in (40) is uniform in f. Finally, by the asymptotic independence results stated earlier, it follows that $\hat{G}(e^{j\omega_k})$ for $k = l(2m_T+1)$, $l = 0, 1, \ldots, (L_T/2) - 1$, are (asymptotically) independent. It remains to specify its correlation structure and asymptotic distribution, which is what we do next.

A perturbation expansion of $\hat{G}^{\mathscr{H}}(e^{j\omega_k})$ (see also proofs of Thm. 8.7.1 and Lemma P8.1 in [10]) yields

$$\begin{aligned}\hat{G}^{\mathscr{H}}(e^{j\omega_k}) = {}&(E\{\hat{S}_{r_{lx}x}(k)\})^{-1}E\{\hat{S}_{r_{lx}y}(k)\} + E\{\hat{S}_{r_{lx}x}(k)\})^{-1}[\hat{S}_{r_{lx}y}(k) \\ &- E\{\hat{S}_{r_{lx}y}(k)\}](E\{\hat{S}_{r_{lx}x}(k)\})^{-1}[\hat{S}_{r_{lx}x}(k) - E\{\hat{S}_{r_{lx}x}(k)\}] \\ &\times (E\{\hat{S}_{r_{lx}x}(k)\})^{-1}E\{\hat{S}_{r_{lx}y}(k)\} + o_p(\Delta_T^{-0.5}),\end{aligned} \tag{41}$$

where $o_p(a_n)$ implies a random sequence $\{z_n\}$ such that $\lim_{n\to\infty}z_n a_n^{-1} = 0$ in probability (i.p.). Using the asymptotic distribution of the estimators (30) and (31) on the coarse grid (38), [10], Thm. P5.2, and (41), it follows (as in [10], Thm. 8.8.1) that $\hat{G}(e^{j\omega_k})$ for $k = l(2m_T+1)$,

$l = 0, 1, \ldots, (L_T/2) - 1$, are (asymptotically) jointly Gaussian. Using (35)–(37), (41), and the Kronecker product notation, it can be shown that asymptotically

$$\begin{aligned}
&\lim_{T\to\infty}\Delta_T \mathrm{cov}(\mathrm{vec}(\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k})), \mathrm{vec}(\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k})))\\
&= (S_{xr_{lx}}^{-T}(\omega_k)\, S_{r_{lx}r_{lx}}(\omega_k)[S_{xr_{lx}}^{*}(\omega_k)]^{-1})\\
&\quad \otimes [S_{yy}(\omega_k) + G(\mathrm{e}^{\mathrm{j}\omega_k})\, S_{xx}(\omega_k)\, G^{\mathcal{H}}(\mathrm{e}^{\mathrm{j}\omega_k}) - S_{yx}(\omega_k) G^{\mathcal{H}}(\mathrm{e}^{\mathrm{j}\omega_k})\\
&\quad - G(\mathrm{e}^{\mathrm{j}\omega_k})\, S_{xy}(\omega_k)]
\end{aligned} \tag{42}$$

$$\lim_{T\to\infty}\Delta_T \mathrm{cov}(\mathrm{vec}(\hat{G}^{\mathcal{H}}(\mathrm{e}^{\mathrm{j}\omega_k})), \mathrm{vec}(\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k}))) = 0 \tag{43}$$

where we have also used the fact that $\lim_{T\to\infty} E\{\hat{S}_{r_{lx}x}(k)\} = S_{r_{lx}x}(\omega_k)$, etc. (cf. (34) and (40)). It follows from (43) that

$$\lim_{T\to\infty}\Delta_T \mathrm{cov}(\mathrm{vec}(\hat{G}^{*}(\mathrm{e}^{\mathrm{j}\omega_k})), \mathrm{vec}(\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k}))) = 0 \tag{44}$$

since $\mathrm{vec}(\hat{G}^{\mathcal{H}})$ and $\mathrm{vec}(\hat{G}^{*})$ are linearly related via a (full-rank) permutation matrix. Thus, $\mathrm{vec}(\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k}))$ on the coarse grid (38) is asymptotically a complex Gaussian (in the sense of [10] Section 4.2) random vector, independent at distinct frequencies on the coarse grid over $(0, \pi)$, with the covariance structure (42).

B. An Equation Error Formulation

It follows from the definition of $G(\mathrm{e}^{\mathrm{j}\omega})$ that

$$-\sum_{i=1}^{n_a} G^T(\mathrm{e}^{\mathrm{j}\omega_k}) A_i^T \mathrm{e}^{-\mathrm{j}\omega_k i} + \sum_{i=1}^{n_b} B_i^T \mathrm{e}^{-\mathrm{j}\omega_k i} = G^T(\mathrm{e}^{\mathrm{j}\omega_k}) \tag{45}$$

for any ω_k. Noting that the A_is and B_is are real and $G(\mathrm{e}^{\mathrm{j}\omega})$ is, in general, complex-valued, we rewrite (45) after replacing $G(\mathrm{e}^{\mathrm{j}\omega_k})$ with its estimate $\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k})$ (see (39)), as

$$-\sum_{i=1}^{n_a} \mathrm{Re}\{\hat{G}^T(\mathrm{e}^{\mathrm{j}\omega_k})\mathrm{e}^{-\mathrm{j}\omega_k i}\} A_i^T + \sum_{i=1}^{n_b} \mathrm{Re}\{\mathrm{e}^{-\mathrm{j}\omega_k i}\} B_i^T = \mathrm{Re}\{\hat{G}^T(\mathrm{e}^{\mathrm{j}\omega_k})\} \tag{46}$$

and

$$-\sum_{i=1}^{n_a} \mathrm{Im}\{\hat{G}^T(\mathrm{e}^{\mathrm{j}\omega_k})\mathrm{e}^{-\mathrm{j}\omega_k i}\} A_i^T + \sum_{i=1}^{n_b} \mathrm{Im}\{\mathrm{e}^{-\mathrm{j}\omega_k i}\} B_i^T = \mathrm{Im}\{\hat{G}^T(\mathrm{e}^{\mathrm{j}\omega_k})\} \tag{47}$$

Using frequencies $\omega_{\mathrm{k}} = 2\pi(k-1)/L_T$ for $1 \le k \le L = (L_T/2)$, (46) and (47) may be rewritten in a matrix-equation form as

$$\mathbf{F}_T \vartheta = \mathbf{f}_T \tag{48}$$

where

$$\vartheta = [A_1 \vdots A_2 \vdots \cdots \vdots A_{n_a} \vdots B_1 \vdots B_2 \vdots \cdots \vdots B_{n_b}]^T, \tag{49}$$

$$\begin{aligned}\mathbf{f}_T = [&\mathrm{Re}\{\hat{G}(\mathrm{e}^{\mathrm{j}\omega_1})\} \vdots \mathrm{Im}\{\hat{G}(\mathrm{e}^{\mathrm{j}\omega_1})\} \vdots \cdots \vdots \mathrm{Re}\{\hat{G}(\mathrm{e}^{\mathrm{j}\omega_L})\} \\ &\vdots \mathrm{Im}\{\hat{G}(\mathrm{e}^{\mathrm{j}\omega_L})\}]^T,\end{aligned} \tag{50}$$

and $\mathbf{F}_T$ is a $(2Lm) \times (n_a p + n_b m)$ matrix composed of appropriate elements from the left side of (46) and (47). An ordinary least-squares solution to (48) is given by

$$\hat{\vartheta}_T = (\mathbf{F}_T^T \mathbf{F}_T)^{-1} \mathbf{F}_T^T \mathbf{f}_T \tag{51}$$

a numerically well-conditioned solution is obtained via a singular value decomposition formulation.

The above approach also occurs in [13] (see also [12]); however, they do not have any counterparts to our Theorems 1 and 2 (see Section IV).

C. A Fitting Error Formulation

Here we follow a quadratic transfer function matching approach. Let $G(\mathrm{e}^{\mathrm{j}\omega}; \theta)$ denote the transfer function of (1) with the system parameters θ as defined in (3). We choose θ to minimize the cost (w_l as in (38))

$$J_{2T}(\theta) = \sum_{l=0}^{(L_T/2)-1} \| \hat{\mathbf{g}}(\mathrm{e}^{\mathrm{j}\omega_l}) - \mathbf{g}(\mathrm{e}^{\mathrm{j}\omega_l}; \theta) \|^2_{\mathbf{W}(\omega_l)} \tag{52}$$

where $\hat{\mathbf{g}}(\mathrm{e}^{\mathrm{j}\omega_l}) = \mathrm{vec}\{\hat{G}(\mathrm{e}^{\mathrm{j}\omega_l})\}$, $\mathbf{g}(\mathrm{e}^{\mathrm{j}\omega_l}; \theta) = \mathrm{vec}\{G(\mathrm{e}^{\mathrm{j}\omega_l}; \theta)\}$, $\|\mathbf{e}\|^2_{\mathbf{W}(\omega_l)} = \mathbf{e}^{\mathscr{H}}\mathbf{W}(\omega_l)\mathbf{e}$, and $\mathbf{W}(\omega_l)$ is a positive-definite, Hermitian weighting matrix. This is a nonlinear iterative optimization problem which is initialized by the closed-form estimator (51).

A similar approach has been proposed in [14]; however, [14] has no counterpart to our Theorems 1 and 3 (see Sec. IV).

In light of the results of Sec. III.A, let us pick

$$\mathbf{W}(\omega_l) = \Sigma = \mathrm{cov}\{\hat{\mathbf{g}}(\mathrm{e}^{\mathrm{j}\omega_l}), \hat{\mathbf{g}}(\mathrm{e}^{\mathrm{j}\omega_l})\} \tag{53}$$

Then (52) leads to a (pseudo-) maximum likelihood (ML) parameter estimator (asymptotically). Since, in practice, Σ is unknown, we replace it with its consistent estimator $\hat{\Sigma}_T$ which is obtained from (42) by replacing all the quantities therein with their consistent estimators such as (30), (31), and (39).

IV. IDENTIFIABILITY AND CONSISTENCY

Here we will summarize some important results from [17]. In the following θ_0, n_{a0}, and n_{b0} denote the true values of θ, n_a, and n_b, respectively.

A. Identifiability

Let $G(e^{j\omega}; \theta_0)$ denote the transfer function of the true model (1) with $n_a = n_{a0}$, $n_b = n_{b0}$, and $\theta = \theta_0$. Consider the system of equations (45) rewritten as follows after replacing $G(e^{j\omega})$ with $G(e^{j\omega}; \theta_0)$:

$$-\sum_{i=1}^{n_a} G^T(e^{j\omega_k}; \theta_0) A_i^T e^{-j\omega_k i} + \sum_{i=1}^{n_b} B_i^T e^{-j\omega_k i} = G^T(e^{j\omega_k}; \theta_0) \tag{54}$$

Note that in (54) ω_k represents an arbitrary frequency, not necessarily equal to $2\pi k/L_T$.

The following result is proved in [17].

Theorem 1. *Given the transfer function* $G(e^{j\omega}; \theta_0)$ *of model (1) at* L *frequencies* $0 < \omega_1 < \omega_2 < \cdots < \omega_L < \pi$ *such that* $(2\pi - 1)n_a + n_b \leq 2L$ *and* $\min(n_a - n_{a0}, n_b - n_{b0}) \geq 0$, *then the set of solutions to (54) is such that* $G(e^{j\omega}; \theta) \equiv G(e^{j\omega}; \theta_0)$.

In [15], Theorem 3.1, a result similar to but more restrictive than our Theorem 1 has been proved; [15] requires measurements of the transfer function to be available on certain frequencies that are equidistant on $[0, \pi]$; in our formulation frequencies can be located arbitrarily on $(0, \pi)$.

1. Persistence of Excitation

Theorem 1 combined with the results of Theorems 2 and 3 (discussed later) motivate the following definition of *persistence of excitation* of inputs to linear systems for the purpose of consistent parameter estimation using integrated polyspectra. This is completely analogous to the concept of persistent excitation when using second-order statistics; see [3], Section 5.4. Similar to the second-order case (and also to the univariate-SISO case using polyspectra [9]), if the input is persistently exciting as defined below, the matrix $\mathbf{F}_\infty$ defined in the proof of Theorem 2 will have full rank, leading to parameter identifiability and consistant parameter estimation.

Definition. A stationary input $\{u(t)\}$ is *persistently exciting of order* L *w.r.t. the integrated polyspectral statistics* if its integrated polyspectrum matrix $S_{r_{lu}u}(\omega)$ is such that its inverse exists at L distinct frequencies in the interval $(0, \pi)$.

B. Consistency

Theorem 2. *Suppose that assumptions (AS1)–(AS4) hold true for the true model generating the data. Then the linear parameter estimator (51) is (strongly) consistent if (51) is solved at L distinct frequencies* $0 < \omega_1 < \omega_2 < \cdots < \omega_L < \pi$ *such that* $(2p-1)n_a + n_b \leq 2L$, $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, *and* $S_{r_l u u}(\omega_i)$ *exists for* $1 \leq i \leq L$.

Proof. See [17]. □

We now turn our attention to the fitting-error estimator. Define

$$\hat{\theta}_T := \arg\left\{ \inf_{\theta \in \Theta_C} J_{2T}(\theta) \right\} \tag{55}$$

where $J_{2T}(\theta)$ is given by (52) and $\Theta_C \subset \Theta$ is a compact set.

Theorem 3. *Suppose that assumptions (AS1)–(AS4) hold true for the true model generating the data. Assume also that* $J_{2T}(\theta)$ *utilizes at least L distinct frequencies* ω_l *in the interval* $(0, \pi)$ *such that* $(2p-1)n_a + n_b \leq 2L$, *and* $S^{-1}_{r_l u u}(\omega_l)$ *exists for* $1 \leq l \leq L$. *If* $\theta_0 \in \Theta_C$, *then* $\hat{\theta}_N$ *defined in (55) converges* w.p.1 *to a set* D_0 *as* $N \to \infty$ *where*

$$D_0 = \{\theta | A^{-1}(z;\theta) B(z;\theta) = A^{-1}(z;\theta_0) B(z;\theta_0)\}$$

Proof. See [17]. □

The following corollary follows from Theorem 3 and some standard results (see, e.g., [3], Section 6.3). The additional condition imposed in Corollary 1 ensures that the set D_0 consists of a single element θ_0.

Corollary 1. *Under the hypotheses of Theorem 3, if* $\min(n_a - n_{a0}, n_b - n_{b0}) = 0$, *then* $\hat{\theta}_N$ *converges* w.p.1 *to* θ_0 *as* $N \to \infty$.

V. PERFORMANCE ANALYSIS

Now we carry out asymptotic performance analysis of the parameter estimators presented in Section III under the assumption that the model orders n_a and n_b are known.

A. Equation Error Formulation

We first analyze the performance of (51). It will be convenient to reexpress it as

$$\hat{\vartheta}_T = (\mathcal{F}_T^{\mathcal{H}} \mathcal{F}_T)^{-1} \mathcal{F}_T^{\mathcal{H}} \mathbf{f}_{cT} \tag{56}$$

$$\mathbf{f}_{cT} = [\hat{G}(e^{j\omega_1}) \vdots \hat{G}(e^{j\omega_2}) \vdots \cdots \vdots \hat{G}(e^{j\omega_L})]^T \tag{57}$$

and $\mathcal{F}_T$ is a $(Lm) \times (n_a p + n_b m)$ matrix composed of appropriate elements from the left side of (45) with $G(\mathrm{e}^{\mathrm{j}\omega_k})$ replaced by $\hat{G}(\mathrm{e}^{\mathrm{j}\omega_k})$. Let ϑ_0 denote the true value of (49). Then we have

$$\bar{\vartheta}_T = \hat{\vartheta}_T - \vartheta_0 = (\mathcal{F}_T^{\mathcal{H}}\mathcal{F}_T)^{-1}\mathcal{F}_T^{\mathcal{H}}[\mathbf{f}_{cT} - \mathcal{F}_T\vartheta_0] = \bar{\mathcal{F}}_T[\mathbf{f}_{cT} - \mathcal{F}_T\vartheta_0] \tag{58}$$

where

$$\bar{\mathcal{F}}_T = (\mathcal{F}_T^{\mathcal{H}}\mathcal{F}_T)^{-1}\mathcal{F}_T^{\mathcal{H}} \tag{59}$$

Using (45) and (57) we have

$$[\mathbf{f}_{cT} - \mathcal{F}_T\vartheta_0]_{i\text{th block}} = \hat{G}^T(\mathrm{e}^{\mathrm{j}\omega_i})\mathcal{A}_0^T(\mathrm{e}^{\mathrm{j}\omega_i}) - \mathcal{B}_0^T(\mathrm{e}^{\mathrm{j}\omega_i}) \tag{60}$$

where

$$\mathcal{A}_0^T(\mathrm{e}^{\mathrm{j}\omega_i}) = I + \sum_{l=1}^{n_a} A_l^T(\vartheta_0)\,\mathrm{e}^{-\mathrm{j}\omega_i l} \tag{61}$$

$$\mathcal{B}_0^T(\mathrm{e}^{\mathrm{j}\omega_i}) = \sum_{l=1}^{n_b} B_l^T(\vartheta_0)\,\mathrm{e}^{-\mathrm{j}\omega_i l} \tag{62}$$

It follows then that asymptotically $\mathrm{vec}(\mathbf{f}_{cT} - \mathcal{F}_T\vartheta_0)$ is a zero-mean, (jointly) Gaussian random vector (cf. the distribution of $\hat{G}(\mathrm{e}^{\mathrm{j}\omega})$ discussed in Section III.A). It also follows from (40) that

$$\lim\nolimits_{T\to\infty}\bar{\mathcal{F}}_T \overset{\text{w.p.1}}{=} \bar{\mathcal{F}}_\infty := (\mathcal{F}_\infty^{\mathcal{H}}\mathcal{F}_\infty)^{-1}\mathcal{F}_\infty^{\mathcal{H}} \tag{63}$$

$$\lim\nolimits_{T\to\infty}\bar{\mathbf{f}}_{cT} \overset{w.p.1}{=} \bar{\mathbf{f}}_{c\infty} \tag{64}$$

where $\mathcal{F}_\infty$ and $\mathbf{f}_{c\infty}$ denote $\mathcal{F}_T$ and $\mathbf{f}_{cT}$, respectively, when the transfer function matrix $\hat{G}(\mathrm{e}^{\mathrm{j}\omega})$ is replaced with its true value $G(\mathrm{e}^{\mathrm{j}\omega};\vartheta_0)$ at various frequencies.

Using a perturbation expansion such as (41) for (58) and noting that $\mathbf{f}_{c\infty} - \mathcal{F}_\infty\vartheta_0 = 0$, it easily follows that

$$\bar{\vartheta}_T = E\{\bar{\mathcal{F}}_T\}[\mathbf{f}_{cT} - \mathcal{F}_T\vartheta_0] + o_p(\Delta_T^{-0.5}) = E\{\bar{\mathcal{F}}_T\}\,\mathbf{d}_T + o_p(\Delta_T^{-0.5}) \tag{65}$$

Therefore, it follows that

$$\begin{aligned}\mathrm{cov}(\mathrm{vec}(\bar{\vartheta}_T), \mathrm{vec}(\bar{\vartheta}_T)) &= \mathrm{cov}(\mathrm{vec}(\hat{\vartheta}_T), \mathrm{vec}(\hat{\vartheta}_T)) \\ &= (I \otimes \bar{\mathcal{F}}_\infty)\,\mathrm{cov}(\mathrm{vec}(\mathbf{d}_T), \mathrm{vec}(\mathbf{d}_T))(I \otimes \bar{\mathcal{F}}_\infty)^{\mathcal{H}}\end{aligned} \tag{66}$$

where we have used the fact that $\lim_{T\to\infty}E\{\bar{\mathcal{F}}_T\} = \bar{\mathcal{F}}_\infty$. Using (60) and (65) we have

$$\begin{aligned}\mathbf{d}_T &= \mathbf{f}_{cT} - \mathcal{F}_T\vartheta_0 \\ &= [\mathcal{A}_0(\mathrm{e}^{\mathrm{j}\omega_1})\,\hat{G}(\mathrm{e}^{\mathrm{j}\omega_1}) - \mathcal{B}_0(\mathrm{e}^{\mathrm{j}\omega_1}) \vdots \cdots \vdots \\ &\qquad \mathcal{A}_0(\mathrm{e}^{\mathrm{j}\omega_L})\,\hat{\mathrm{G}}(\mathrm{e}^{\mathrm{j}\omega_L}) - \mathcal{B}_0(\mathrm{e}^{\mathrm{j}\omega_L})]^T\end{aligned} \tag{67}$$

By [18] there exists a permutation matrix $\mathcal{U}$ (see Eqn. (4) and Table III in [18]) such that

$$\mathrm{vec}(\mathbf{d}_T) = \mathcal{U}\,\mathrm{vec}(\mathbf{d}_T^T) \tag{68}$$

Moreover, by (67) we have

$$\mathbf{d}_T^T = \mathcal{J}_A \hat{\mathcal{L}}_G - (\text{terms involving } \mathcal{B}_0) \tag{69}$$

where

$$\hat{\mathcal{L}}_G := \text{block diagonal}\{\hat{G}(e^{j\omega_1}), \hat{G}(e^{j\omega_2}), \ldots, \hat{G}(e^{j\omega_1})\} \tag{70}$$

$$\mathcal{J}_A := [\mathcal{A}_0(e^{j\omega_1}) \vdots \cdots \vdots \mathcal{A}_0(e^{j\omega_L})] \tag{71}$$

Finally using (42) and (65)–(70), we have

$$\begin{aligned}&\mathrm{cov}(\mathrm{vec}(\tilde{\vartheta}_T), \mathrm{vec}(\tilde{\vartheta}_T))\\ &= (I \otimes \bar{\mathcal{F}}_\infty)\mathcal{U}(I \otimes \mathcal{J}_A)\,\mathrm{cov}(\mathrm{vec}(\hat{\mathcal{L}}_G), \mathrm{vec}(\hat{\mathcal{L}}_G))(I \otimes \mathcal{J}_A)^{\mathcal{H}}\mathcal{U}(I \otimes \bar{\mathcal{F}}_\infty)^{\mathcal{H}}\end{aligned}$$

Covariance involving $\mathrm{vec}(\hat{\mathcal{L}}_G)$ follows from (42).

B. Fitting Error Formulation: Maximum Likelihood Estimator

Here we analyze the asymptotic performance of the estimator that minimizes (52) with weights satisfying (53). It is well known that in the case of maximum likelihood (ML) estimators the covariance of the ML parameter estimator is given by the inverse of the Fisher information matrix, provided that certain regularity conditions hold true.

Let ∇_θ denote the gradient operator w.r.t. vector θ. Define

$$\nabla_\theta J_{2T}(\theta_1) := \nabla_\theta J_{2T}(\theta)|_{\theta=\theta_1} \tag{72}$$

Using a Taylor series expansion around θ_0 (the true value of θ), it follows that

$$\begin{aligned}0 = \nabla_\theta J_{2T}(\hat{\theta}_T) &= \nabla_\theta J_{2T}(\theta_0) + \nabla^2_{\theta\theta} J_{2T}(\theta_0)(\hat{\theta}_T - \theta_0)\\ &\quad + o_p(\|(\hat{\theta}_T - \theta_0)\|)\end{aligned} \tag{73}$$

Therefore, it follows that

$$(\hat{\theta}_T - \theta_0) = -[\nabla^2_{\theta\theta} J_{2T}(\theta_0)]^{-1}\nabla_\theta J_{2T}(\theta_0) + o_p(\|(\hat{\theta}_T - \theta_0)\|) \tag{74}$$

It is easy to establish that

$$\nabla^2_{\theta\theta} J_{2T}(\theta_0) \xrightarrow{T\to\infty} E\{\nabla_\theta J_{2T}(\theta_0)(\nabla_\theta J_{2T}(\theta_0))^T\} \quad \text{w.p.1} \tag{75}$$

Using [3], Lemma B.4 it then follows that asymptotically

$$\mathrm{cov}(\hat{\theta}_T - \theta_0, \hat{\theta}_T - \theta_0) = [E\{\nabla_\theta J_{2T}(\theta_0)(\nabla_\theta J_{2T}(\theta_0))^T\}]^{-1} \tag{76}$$

Let θ_m denote the mth component of θ. Using (52) and (53) with Σ replaced by $\hat{\Sigma}_T$, we have

$$\frac{\partial J_{2T}}{\partial \theta_m} = \sum_{l=0}^{(L_T/2)-1} \Bigg\{ [\hat{\mathbf{g}}(e^{j\omega_l}) - \mathbf{g}(e^{j\omega_l};\theta)]^{\mathscr{H}} \hat{\Sigma}_T^{-1} \left[-\frac{\partial \mathbf{g}(e^{j\omega_l};\theta)}{\partial \theta_m} \right] + \left[\frac{\partial \mathbf{g}(e^{j\omega_l};\theta)}{\partial \theta_m} \right]^{\mathscr{H}} \hat{\Sigma}_T^{-1} [\hat{\mathbf{g}}(e^{j\omega_l}) - \mathbf{g}(e^{j\omega_l};\theta)] \Bigg\} \tag{77}$$

Using the results (42)–(44) and [3], Lemma B.4, it follows that asymptotically

$$E\left\{ \frac{\partial J_{2T}}{\partial \theta_m} \frac{\partial J_{2T}}{\varphi \theta_k} \right\} = \Delta_T \sum_{l=0}^{(L_T/2)-1} \Bigg\{ \left[\frac{\partial \mathbf{g}(e^{j\omega_l};\theta)}{\partial \theta_m} \right]^{\mathscr{H}} \bar{\Sigma}_\infty^{-1} \left[\frac{\partial \mathbf{g}(e^{j\omega_l};\theta)}{\varphi \theta_k} \right] + \left[\frac{\partial \mathbf{g}(e^{j\omega_l};\theta)}{\varphi \theta_k} \right]^{\mathscr{H}} \bar{\Sigma}_\infty^{-1} \left[\frac{\partial \mathbf{g}(e^{j\omega_l};\theta)}{\partial \theta_m} \right] \Bigg\} =: \Delta_T[\mathscr{J}_T]_{mk}, \tag{78}$$

where $[\mathscr{J}_T]_{mk}$ denotes the mkth element of the matrix $\mathscr{J}_T$ and where we have used the fact that $\lim_{T\to\infty}\Delta_T\Sigma_T = \bar{\Sigma}_\infty$ (= right side of (42)) w.p.1. Finally, using (75) and (78), it follows that

$$\operatorname{cov}(\hat{\theta}_T - \theta_0, \hat{\theta}_T - \theta_0) = \Delta_T^{-1}\mathscr{J}_T^{-1} = O(T^{-1}) \to 0 \quad \text{as } T\to\infty \tag{79}$$

VI. CONCLUSIONS

The parametric frequency domain approaches presented in [9] for estimation of the parameters of single-input single-output, linear errors-in-variables models have been extended to multivariable models. Two new classes of approaches were proposed and analyzed. An integrated polyspectrum-based persistence of excitation condition on system input was defined and related to parameter indentifiability of the multivariable system. Both classes of the parameter estimators were shown to be consistent in any measurement noise sequences with vanishing bispectra when integrated bispectrum-based approaches are used. The proposed parameter estimators were shown to be consistent in Gaussian measurement noise when integrated trispectrum-based approaches are used. The input to the system need not be a linear process but must have nonvanishing bispectrum or trispectrum. Asymptotic performance of the proposed estimators was analyzed and expressions for the variance/covariance matrix of the parameter estimators were derived.

For numerical examples (simulation results) involving SISO systems we refer the reader to [9]; MIMO systems are yet to be simulated.

ACKNOWLEDGMENTS

This work was supported, in part, by the National Science Foundation under Grant ECS-9504878.

REFERENCES

1. M Deistler. In H Nijmeijer, JM Schumacher, eds, Three Decades of Mathematical System Theory. Lecture Notes in Control & Information Sciences, Springer, 1989.
2. RE Kalman. In PR Krishnaiah, ed, Developments in Statstics, IV. Academic Press, 1983.
3. T Söderström, P Stoica. System Identification. London: Prentice Hall, 1989.
4. P Stoica, A Nehorai. Automatica 23:541–543, 1987.
5. JK Tugnait IEEE Trans Autom Contr AC-37:476–485, April, 1992.
6. Y Inouye, H Tsuchiya. Intern J Contr 53:1431–1448, 1991.
7. A Delopoulos, GB Giannakis. Automatica 30:1271–1294, 1994.
8. Y Inouye, Y Suga. Intern J Contr 59:1231–1253, May 1994.
9. JK Tugnait, Y Ye. IEEE Trans Autom Contr AC-40:670–683, 1995.
10. DR Brillinger. Time Series Data Analysis and Theory. New York: Holt, Rinehart and Winston, 1975.
11. BLS Prakasa Rao. Asymptotic Theory of Statistical Inference. New York: Wiley, 1987.
12. PL Lin, YC Wu. Trans ASME, J Dynamic Systems, Measurements, and Control 104:58–64, 1982.
13. CW Chen, JN Juang, G Lee. In Proc American Control Conf, San Francisco, CA, June 1993, pp 3057–3061.
14. DS Bayard. Automatica 30:1439–1444, 1994.
15. T McKelvey, H Akcay. In Proc 33rd IEEE Conf Decision & Control, Lake Buena Vista, FL, Dec 1994, pp 3348–3353.
16. T McKelvey, H Akcay. In Proc 33rd IEEE Conf Decision & Control, Lake Buena Vista, FL, Dec 1994, pp 3359–3364.
17. JK Tugnait. In Proc 34th IEEE Conf on Decision & Control, New Orleans, LA, Dec 13–15, 1995, pp 1246–1251.
18. JW Brewer. IEEE Trans Circuits & Systems CAS-25:772–781, 1978.
19. T Söderström. Automatica 17:713–725, 1981.
20. BDO Anderson, M Deistler. J Time Series Analysis 5:1–13, 1984.
21. BDO Anderson. Automatica 21:709–716, 1985.
22. M Deistler. In S Bittanti, ed, Time Series and Linear Systems. Lecture Notes in Control and Information Sciences, Vol. 86, 1986, pp 37–86.
23. M Deistler, BDO Anderson. J Econometrics 41:39–63, 1989.
24. M Green, BDO Anderson. IEEE Trans Automatic Control AC-31:467–471, 1986.
25. M Deistler. In J Gani, MB Priestley, Essays in Time Series and Allied Processes. J App Prob, Special volume 23A:23–40, 1986.

26. V Solo. In J Gani, MB Priestley, Essays in Time Series and Allied Processes. J App Prob, Special volume 23A:63–74, 1986.
27. JA Cadzow, OM Solomon, Jr. IEEE Trans Acoustics Speech Signal Processing ASSP-34:462–469, 1986.
28. H Akaike. Ann Inst Stat Math 18:269–276, 1966.
29. JMM Anderson, GB Giannakis. In Proc 25th Asilomar Conf Signals Systems Computers, Pacific Grove, CA, 1991, pp 608–612.
30. M Rosenblatt. Stationary Sequences and Random Fields. Boston: Birkhäuser, 1985.
31. AM Tekalp, AT Erdem. IEEE Trans Acoustics Speech Signal Processing ASSP-37:1537–1549, 1989.

14

Blind Deconvolution of Multichannel Linear Time-Invariant Systems of Nonminimum Phase

Yujiro Inouye

Osaka University, Osaka, Japan

I. INTRODUCTION

Blind deconvolution and equalization of linear time-invariant systems have widely received attention in various fields such as data communication, image processing, and geophysical data processing [1–5]. Recently, Shalvi and Weinstein proposed certain new criteria for blind deconvolution of nonminimum phase linear time-invariant systems [3]. Their work is, however, restricted to the single-channel (or scalar) case. Multichannel blind deconvolution is closely related to multichannel blind signal separation [5–9]. Multichannel blind deconvolution usually assumes that the components of input signals are white in the second- or higher-order sense [4], but multichannel blind separation lacks such an assumption as whiteness, and assumes that the components of input signals are mutually independent [5–9]. Most of the approaches proposed for blind signal separation are developed under the assumption that the linear systems involved are unknown constant gains [5–8]. Yellin and Weinstein proposed new criteria for multichannel blind signal separation [9]. However, their approach is confined to a restrictive case in which all the diagonal elements of the transfer function matrices involved are constant and equal to unity.

This chapter extends the Shalvi–Weinstein approach in [3] to the multichannel case [10]. We propose a multistage maximization criterion and a single-stage maximization criterion for attaining multichannel blind deconvolution. Then it is shown that the two maximization criteria are equivalent under the normalized condition of the input vector process.

Based on this fact, a necessary and sufficient condition is derived for multichannel blind deconvolution to be solvable. A simulation example is presented to demonstrate the effectiveness of the mutistage maximization criterion.

This chapter is organized as follows. In Section II, a precise formulation of the multichannel blind deconvolution problem under consideration is provided together with a fundamental formula for second- and fourth-order cumulants of interest in this chapter. A multistage maximization criterion and a single-stage maximization criterion are proposed in Section III for solving the multichannel blind deconvolution problem. Then it is shown that each of the two maximization criteria yields a solution to the multichannel blind deconvolution problem under the normalized condition of the input vector process. A necessary and sufficient condition is derived for multichannel blind deconvolution to be solvable. A simulation example is presented in Section IV for demonstrating the effectiveness of the multistage maximization criterion.

We use the following notation in this chapter. The ith element of a vector $\boldsymbol{a}$ is denoted by a_i, and the (i,j)th element of a matrix A is denoted by a_{ij}. The superscript * denotes the complex conjugate of a number, or the conjugate transpose of a vector or a matrix. Let $\|\cdot\|$ denote the Euclidean vector or matrix norm. Let Z denote the set of all integers. Let $E\{x\}$ denote the expectation of a random variable x, and $\text{cum}\{x_1, \ldots, x_n\}$ denote the nth-order (joint) cumulant of random variables $x_1, \ldots, x_n$ [11]. Let $c_{x_1,\ldots,x_n}(\tau_1, \tau_2, \ldots, \tau_{n-1})$ be the nth-order cumulant (function of $\tau_1, \ldots, \tau_{n-1}$) of jointly ($n$th-order) stationary random processes $\{x_1(t)\}, \ldots, \{x_n(t)\}$ defined by $c_{x_1,\ldots,x_n}(\tau_1, \tau_2, \ldots, \tau_{n-1}) = \text{cum}\{x_1(t+\tau_1), \ldots, x_{n-1}(t+\tau_{n-1}), x_n(t)\}$. Since we are only interested in second- and fourth-order cumulants in this chapter, we use the following simple notation for the variance σ_x^2 of $\{x(t)\}$, the correlation function $r_{x_1,x_2^*}(\tau)$ of $\{x_1(t)\}$ with $\{x_2^*(t)\}$, and the fourth-order cumulant $\kappa_{4,x}$ of $\{x(t)\}$ as $\sigma_x^2 := c_{2,x,x^*}(0)$, $r_{x_1,x_2^*}(\tau) := c_{2,x_1,x_2^*}(\tau)$, and $\kappa_{4,x} := c_{4,x,x^*,x,x^*}(0,0,0)$, respectively.

II. PROBLEM FORMULATION

We consider a multichannel, linear time-invariant (LTI), and generally noncausal system described by

$$y(t) = \sum_{k=-\infty}^{\infty} H(k)u(t-k) \tag{1}$$

where $y(t)$ is an n-column output vector, $\boldsymbol{u}(t)$ is an n-column input vector,

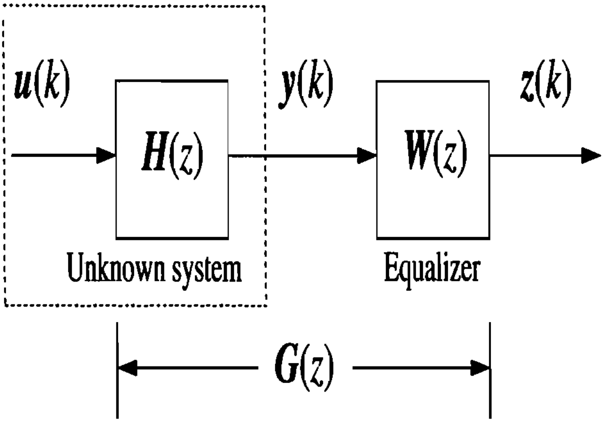


Figure 1 Unknown system and equalizer.

and $\{H(k)\}$ is an $n \times n$ matrix sequence called the *impulse response*. To retrieve the input signal, we precede the system with another multichannel LTI system called the *equalizer* given by

$$z(t) = \sum_{k=-\infty}^{\infty} W(k)y(t-k) \tag{2}$$

where $z(t)$ is an n-column vector called the *equalizer output*, and $\{W(k)\}$ is an $n \times n$ matrix sequence. The cascade connection of the two systems is illustrated in the schematic diagram in Fig. 1. We assume all the components of the input signals or the impulse responses are real or complex. This assumption is required for such an application as data communication using quadrature amplitude modulation (QAM) signals. Moreover, we make the following assumptions on the systems and signals involved.

A1. The system in (1), denoted by H, is unknown. It is stable, that is, the impulse response satisfies the absolute summability condition

$$\sum_{k=-\infty}^{\infty} \|H(k)\| < \infty \tag{3}$$

A2. The transfer function defined by

$$H(z) := \sum_{k=-\infty}^{\infty} H(k)z^k \tag{4}$$

is of full rank on the unit circle $|z| = 1$ (this implies that it has no zero on the unit circle).

A3. The input sequence $\{\boldsymbol{u}(t)\}$ is a zero-mean, non-Gaussian random vector process, whose component processes $\{u_i(t)\}$, $i = 1, \ldots, n$, are mutually independent. Moreover, each component process $\{u_i(t)\}$ is an independently and identically distributed (i.i.d.) process with nonzero variance $\sigma^2_{u_i} \neq 0$ and fourth-order cumulant $\kappa_{4,u_i} \neq 0$.

A4. The equalizer in (2), denoted by $\boldsymbol{W}$, is also stable.

We require the following two notions for stationary random vector processes throughout the chapter. It is said that a stationary random vector process $\{\boldsymbol{u}(t)\}$ satisfies the *normalized condition* if the variance of each component of the vector process $\{\boldsymbol{u}(t)\}$ is equal to unity. It is said to satisfy the *normalized whitening condition* if all the component processes $\{u_i(t)\}$, $i = 1, \ldots, n$ of the vector process $\{\boldsymbol{u}(t)\}$ are white random scalar processes with unit variance and if they are mutually uncorrelated. When the random process is zero-mean, the second condition is equivalent to $E\{\boldsymbol{u}(t+k)\boldsymbol{u}^*(t)\} = \boldsymbol{I}\delta(k)$, where $\boldsymbol{I}$ denotes the identity matrix and $\delta(k)$ denotes the Kronecker delta. Moreover, when the component processes $\{u_i(t)\}$, $i = 1, \ldots, n$, are mutually independent and they are all independently and identically distributed (i.i.d.) processes, the first condition is equivalent to the second one.

For the blind deconvolution and equalization of the unknown system $\boldsymbol{H}$, we cannot observe the inputs, but can observe only the outputs. This implies there are inherent ambiguities in the solution to the multichannel blind deconvolution problem as follows. In general, we cannot identify the order of the arrangement of the components $u_1(t), \ldots, u_n(t)$ of input vector $\boldsymbol{u}(t)$, the time origin of each component $u_i(t)$, and the magnitude (or the absolute value) of each component $u_i(t)$.

Taking these ambiguities into account, the multichannel blind deconvolution problem is formulated as follows. The problem is to find an equalizer $\boldsymbol{W}$ so that the transfer function $\boldsymbol{G}(z)$ of the combined system takes the form of

$$\boldsymbol{G}(z) = \boldsymbol{\Lambda}(z)\boldsymbol{D}\boldsymbol{P}, \tag{5}$$

where $\boldsymbol{\Lambda}(z)$, called a *shift matrix*, is a diagonal matrix with diagonal entries $\lambda_{ii}(z) = z^{l_i}$, $i = 1, \ldots, n$ (where l_i is an integer), $\boldsymbol{D}$ is a constant diagonal matrix, and $\boldsymbol{P}$ is a permutation matrix. We should note that $\boldsymbol{P}$, $\boldsymbol{\Lambda}(z)$, and $\boldsymbol{D}$ respectively correspond to the three types of ambiguities mentioned above, that is, the order ambiguity, the time-origin ambiguity, and the magnitude ambiguity of the components of the input vector $\boldsymbol{u}(t)$. In order to eliminate the magnitude ambiguity, we may assume at the outset that the input vector $\boldsymbol{u}(t)$ satisfies the normalized condition by dividing each component $u_i(t)$ by the square root of the variance $\sigma^2_{u_i}$. Under this normalized condition, the multichannel blind deconvolution problem is to find an equalizer $\boldsymbol{W}$ so that the transfer function $\boldsymbol{G}(z)$ of the combined system takes the form in (5) with the diagonal elements of the diagonal matrix $\boldsymbol{D}$ all being of unit magnitude, that is, $|d_{ii}| = 1$ for $i = 1, \ldots, n$.

By the multilinearity property of cumulants [11], we can derive the following formulas for the components of the equalizer output vector $z(t)$ from (1) and (2) with (A1)–(A4). Let $\{\boldsymbol{G}(t)\}$ be the impulse response of the cascade system in Fig. 1 for any $i_1, i_2 \in \{1, 2, \ldots, n\}$,

$$r_{z_{i_1}, z^*_{i_2}}(\tau_1) = \sum_{j=1}^{n} \sum_{r=-\infty}^{\infty} g_{i_1 j}(\tau + \tau_1) g^*_{i_2 j}(\tau) \sigma^2_{u_j} \tag{6}$$

For any $i_1, i_2, i_3, i_4 \in \{1, 2, \ldots, n\}$, we have

$$c_{4, z_{i_1}, z^*_{i_2}, z_{i_3}, z^*_{i_4}}(\tau_1, \tau_2, \tau_3) = \sum_{j=1}^{n} \sum_{\tau=-\infty}^{\infty} g_{i_1 j}(\tau + \tau_1) g^*_{i_2 j}(\tau + \tau_2) \times g_{i_3 j}(\tau + \tau_3) g^*_{i_4 j}(\tau) \kappa_{4, u_j} \tag{7}$$

III. BLIND DECONVOLUTION

To begin with, let us assume that the input process $\{\boldsymbol{u}(t)\}$ satisfies the normalized condition by dividing each component $\{u_i(t)\}$ by the square root of the variance $\sigma^2_{u_i}$ to eliminate the magnitude ambiguity.

First we consider the following multistage maximization criterion (A):

(Stage 1) Maximize $|\kappa_{4, z_1}|$ subject to $\sigma^2_{z_1} = 1$.
(Stage k) Maximize $|\kappa_{4, z_k}|$ subject to $\sigma^2_{z_k} = 1$ and $r_{z_k, z_i^*}(\tau) = 0$ for all $\tau \in Z$, and all $i = 1, 2, \ldots, k-1$. Here k moves successively from 2 to n.

Remark 1. The above successive procedures can be intepreted as follows. The first stage is just the same as the maximization procedure proposed by Shalvi and Weinstein [3]. In the context of real-valued signals, it corresponds to the minimum entropy deconvolution technique introduced by Wiggins [1] and analyzed rigorously by Donoho [2]. At the kth stage, the Shalvi–Weinstein approach is applied to the kth output $z_k(t)$ under the condition that the kth output $z_k(t)$ is uncorrelated in the second-order sense to the preceding outputs $z_i(t)$, $i = 1, \ldots, k-1$.

Then we obtain the following theorem.

Theorem 1. *Under the normalized condition of the input process, the multistage maximization criterion* (*A*) *gives a solution to the multichannel blind deconvolution problem.*

The proof of Theorem 1 is based on the followimg lemma.

Lemma 1. *Let $\{\lambda_i\}_{i=-\infty}^{\infty}$ be a bounded real-valued sequence with $\lambda := \max_{-\infty<i<\infty}|\lambda_i| > 0$, and $\{g_i\}_{i=-\infty}^{\infty}$ be a complex-valued sequence such that*

$$\sum_{i=-\infty}^{\infty} |g_i|^2 < \infty$$

Let I_0 be a set of integers defined by

$$I_0 = \{i \in Z \mid |\lambda_i| = \lambda\}$$

Then

$$\left|\sum_{i=-\infty}^{\infty} |g_i|^4 \lambda_i\right| \leq \left(\sum_{i=-\infty}^{\infty} |g_i|^2\right)^2 \lambda \tag{8}$$

where equality holds if and only if $\{g_i\}$ has at most one nonzero element g_{i_0} for some $i_0 \in I_0$. Thus, if $\Sigma_{i=-\infty}^{\infty}|g_i|^2 = 1$, then the following two claims hold true.

(a) $\left|\sum_{i=-\infty}^{\infty} |g_i|^4 \lambda_i\right| \leq \lambda$

(b) $\left|\sum_{i=-\infty}^{\infty} |g_i|^4 \lambda_i\right| = \lambda$ *if and only if $\{g_i\}$ has only a nonzero element g_{i_0} of magnitude 1 for some $i_0 \in I_0$.*

Proof. Let I_0^+ and I_0^- be two sets of integers defined respectively by

$$I_0^+ = \{i \in Z \mid \lambda_i = \lambda\}$$
$$I_0^- = \{i \in Z \mid \lambda_i = -\lambda\}$$

Then $I_0^+ \cup I_0^- = I_0$ and $I_0^+ \cap I_0^- = \emptyset$, where $\emptyset$ denotes the null set. First we find

$$-\sum_{i=-\infty}^{\infty} |g_i|^4 \lambda \leq \sum_{i=-\infty}^{\infty} |g_i|^4 \lambda_i \leq \sum_{i=-\infty}^{\infty} |g_i|^4 \lambda \tag{9}$$

where the first inequality holds with equality if and only if $g_i = 0$ for any $i \notin I_0^-$, and the second inequality holds with equality if and only if $g_i = 0$ for any $i \notin I_0^+$. Thus

$$\left|\sum_{i=-\infty}^{\infty} |g_i|^4 \lambda_i\right| \leq \sum_{i=-\infty}^{\infty} |g_i|^4 \lambda \tag{10}$$

where equality holds if and only if $g_i = 0$ for any $i \notin I_0^+$ or for any $i \notin I_0^-$. On the other hand, by a direct calculation we have

$$\begin{aligned}\left(\sum_{i=-\infty}^{\infty} |g_i|^2\right)^2 &= \left(\sum_{i=-\infty}^{\infty} |g_i|^2\right)\left(\sum_{k=-\infty}^{\infty} |g_k|^2\right)\\ &= \sum_{i=-\infty}^{\infty} |g_i|^4 + \sum_{i\neq k} |g_i|^2|g_k|^2\\ &\geq \sum_{i=-\infty}^{\infty} |g_i|^4\end{aligned}$$

with equality if and only if

$$\sum_{i\neq k} |g_i|^2|g_k|^2 = 0$$

which is true if and only if $\{g_i\}$ has at most one nonzero element. Thus, (8) holds true, where equality holds if and only if $\{g_i\}$ has at most one nonzero element g_{i_0} for some $i_0 \in I_0$. Moreover, we can easily obtain the two claims (a) and (b). □

We note that, in claim (b), $\{g_i\}$ can be represented as

$$g_i = c\delta(i - i_0)$$

for some integer $i_0 \in I_0$ and some constant c with $|c| = 1$.

Proof of Theorem I. Without loss of generality, we can assume that the elements $u_1(t), \ldots, u_n(t)$ of $\boldsymbol{u}(t)$ are arranged so as to satisfy the following decreasing sequence condition

$$|\kappa_{4,u_1}| \geq |\kappa_{4,u_2}| \geq \cdots \geq |\kappa_{4,u_n}| \tag{11}$$

When this is not the case and another arrangement $u_{k_1}(t), \ldots, u_{k_n}(t)$ satisfies the condition

$$|\kappa_{4,u_{k_1}}| \geq |\kappa_{4,u_{k_2}}| \geq \cdots \geq |\kappa_{4,u_{k_n}}|$$

let $\boldsymbol{P}$ be the permutation matrix corresponding to changing the integers 1 through n into the arrangement $k_1, \ldots, k_n$, that is, $\boldsymbol{P} = [p_{ij}]$ where $p_{ij} = 1$ if $j = k_i$, and $p_{ij} = 0$ if $j \neq k_i$. Put

$$\bar{\boldsymbol{u}}(t) = \boldsymbol{P}\boldsymbol{u}(t) \tag{12}$$

Then the elements $\bar{u}_1(t), \ldots, \bar{u}_n(t)$ of $\bar{\boldsymbol{u}}(t)$ satisfy the decreasing sequence condition.

Now we shall consider Stage 1 in the criterion (A). Let $G(z)$ be the transfer function of the cascade system shown in Fig. 1. Denote

$$K_1 := |\kappa_{4,z_1}| = |c_{4,z_1^* z_1^* z_1 z_1}(0,0,0)| \tag{13}$$

Then, using (7), we obtain

$$K_1 = \left| \sum_{j=1}^{n} \sum_{\tau=-\infty}^{\infty} |g_{1j}(\tau)|^4 \gamma_j \right| \tag{14}$$

where

$$\gamma_j := \kappa_{4,u_j} \neq 0 \tag{15}$$

Note here that the condition (11) means

$$|\gamma_1| \geq |\gamma_2| \geq \cdots \geq |\gamma_n| > 0 \tag{16}$$

Using (6) and $\sigma_{u_j}^2 = 1$ for $j = 1, 2, \ldots, n$, we have

$$\sigma_{z_1}^2 = \sum_{j=1}^{n} \sum_{\tau=-\infty}^{\infty} |g_{1j}(\tau)|^2 = 1 \tag{17}$$

where the second equality in (17) comes from the constraint $\sigma_{z_1}^2 = 1$ at Stage 1. Let J_0 be a set of integers defined by

$$J_0 = \{j \in Z \,|\, |\gamma_j| = |\gamma_1|\} \tag{18}$$

Applying Lemma 1 to maximizing K_1 subject to constraint (17), we obtain

$$K_1 \leq |\gamma_1| \tag{19}$$

where equality holds if and only if

$$g_{1j}(t) = \begin{cases} c_1 \delta(t - t_1), & \text{for some } j_0 \in J_0 \\ 0, & \text{for } j \neq j_0 \end{cases} \tag{20}$$

where $|c_1| = 1$ and t_1 is an integer. Here with no loss of generality, we can choose $j_0 = 1$ in (20) be rearranging the order in (16).

Next we shall consider Stage 2 in the criterion (A). We begin with considering the constraint $r_{z_1, z_2^*}(\tau) = 0$ for $\tau \in Z$. Using (6) with $\sigma_{u_j}^2 = 1$ for $j = 1, \ldots, n$, and taking account of (20) with $j_0 = 1$, we see that this constraint becomes

$$\begin{aligned} r_{z_1, z_2^*}(\tau) &= \sum_{j=1}^{n} \sum_{t=-\infty}^{\infty} g_{1j}(t+\tau) g_{2j}^*(t) \\ &= c_1 g_{21}^*(t_1 - \tau) \\ &= 0 \quad \text{for any } \tau \in Z \end{aligned}$$

which gives

$$g_{21}(t) = 0 \quad \text{for } t \in Z \tag{21}$$

Denote

$$K_2 := |\kappa_{4,z_2}| = |c_{4,z_2^* z_2^* z_2 z_2}(0,0,0)| \tag{22}$$

Then, using (6) and (7) together with (21), we obtain

$$K_2 = \left| \sum_{j=2}^{n} \sum_{\tau=-\infty}^{\infty} |g_{2j}(\tau)|^4 \gamma_j \right| \tag{23}$$

and

$$\sigma_{z_2}^2 = \sum_{j=2}^{n} \sum_{\tau=-\infty}^{\infty} |g_{2j}(\tau)|^2 = 1 \tag{24}$$

where we used $\sigma_{u_j}^2 = 1$ for $j = 2, \ldots, n$ and the constraint $\sigma_{z_2}^2 = 1$ in the first equality and the last equality, respectively. In the same way as before, applying Lemma 1 to maximizing K_2 subject to constraint (24), we obtain

$$K_2 \leq |\gamma_2| \tag{25}$$

where equality holds if and only if

$$g_{2j}(t) = \begin{cases} c_2 \delta(t - t_2) & \text{for } j = 2 \\ 0 & \text{for } j \neq 2 \end{cases} \tag{26}$$

where $|c_2| = 1$ and t_2 is an integer. Here we have chosen $j = 2$ in (26) without loss of generality by rearranging the order in (16) if necessary.

Continuing the same argument as above for the subsequent stages from $k = 3$ to $k = n$, we obtain at Stage i

$$K_i \leq |\gamma_i| \tag{27}$$

where equality holds if and only if

$$g_{ij}(t) = \begin{cases} c_i \delta(t - t_i) & \text{for } j = i \\ 0 & \text{for } j \neq i \end{cases} \tag{28}$$

where $|c_i| = 1$ and t_i is an integer. Thus the multistage maximization criterion (A) yields a solution given by

$$g_{ij}(t) = \begin{cases} c_i \delta(t - t_i) & \text{for } j = i \\ 0 & \text{for } j \neq i \end{cases} \tag{29}$$

which is equivalent to $\boldsymbol{G}(z) = \boldsymbol{\Lambda}(z)\boldsymbol{D}$ in the z-domain, where $\boldsymbol{\Lambda}(z) = \text{diag}[z^{-t_1}, \ldots, z^{-t_n}]$, and $\boldsymbol{D} = \text{diag}[c_1, \ldots, c_n]$ with $|c_i| = 1$ for

$i = 1, \ldots, n$. Taking account of the arrangement of the input components given in (12) when the input components do not necessarily satisfy the decreasing condition (11) or we need rearrangements of the order in (16), the maximization criterion (A) yields a solution given by

$$G(z) = \Lambda(z)DP \tag{30}$$

in the z-domain, where $\Lambda(z)$ is a shift matrix, D is a diagonal matrix with diagonal entries of magnitude 1, and P is a permutation matrix. □

Next we consider the following single-stage maximization criterion (B):

> Maximize $\Sigma_{i=1}^{n} |\kappa_{4,z_i}|^2$ subject to the constraints $r_{z_i,z_j}(\tau) = \delta(i-j)\delta(\tau)$ for all $\tau \in Z$ and all $i, j = 1, \ldots, n$.

We note that the constraints of criterion (B) require that the equalizer output process $\{z(t)\}$ satisfies the normalized whitening condition.

Remark 2. It can be seen below from the proof of Theorem 2 that we can replace the criterion function $\Sigma_{i=1}^{n} |\kappa_{4,z_i}|^2$ by the criterion function $\Sigma_{i=1}^{n} |\kappa_{4,z_i}|$.

Then we have also the following theorem.

Theorem 2. *Under the normalization condition of the input process, the single-stage maximization criterion (B) also yields a solution to the multichannel blind equalization problem.*

Before proceeding to prove Theorem 2, we provide the following two lemmas.

Lemma 2. *Under the normalization condition of the input process, i.e., $\sigma_{u_i}^2 = 1$ for $i = 1, \ldots, n$, the frequency response $\mathbf{G}(w)$ of any cascade system G satisfying the constraints of the criterion (B) is a unitary matrix for any frequency w, where $\mathbf{G}(w) := \mathbf{G}(e^{jw})$.*

Proof. Let $\{G(t)\}$ be the impulse response of the cascade system G. We can see from (6) together with the constraints of criterion (B) that

$$\sum_{\tau=-\infty}^{\infty} G(\tau+t)G^*(\tau) = \begin{bmatrix} \delta(t) & & 0 \\ & \ddots & \\ 0 & & \delta(t) \end{bmatrix}, \quad t \in Z \tag{31}$$

Since the Fourier transform of the Kronecker delta $\delta(t)$ is equal to one, (31) is equivalent to

$$G(w)G^*(w) = I \tag{32}$$

in the frequency domain. This means that $G(w)$ is a unitary matrix for any frequency w. □

We quote the next lemma from [8]. Here we introduce the following notation. Let $\bar{\boldsymbol{\gamma}}$ denote the vector obtained by replacing all elements of γ by their moduli (absolute values). Let $\bar{\boldsymbol{G}}$ denote the matrix obtained by replacing all elements of $\boldsymbol{G}$ by their moduli, and $\boldsymbol{G}^{(2)}$ denote the matrix obtained by replacing all elements of $\boldsymbol{G}$ by their powers of exponent 2. Therefore, if $\boldsymbol{G} = [g_{ij}]$, then $\bar{\boldsymbol{G}}^{(2)} = [|g_{ij}|^2]$.

Lemma 3. [8] *If* $\boldsymbol{G}$ *is an* $n \times n$ *unitary matrix, then*

$$\|\bar{\boldsymbol{G}}^{(2)}\boldsymbol{\gamma}\| \leq \|\boldsymbol{\gamma}\| \quad \textit{for any } \boldsymbol{\gamma} \in R^n \tag{33}$$

where R^n *denotes the real n-dimensional Euclidean space.*

Proof. Since $\boldsymbol{G}$ is unitary, $\bar{\boldsymbol{G}}^{(2)}$ becomes a doubly stochastic matrix (i.e., the sum of the elements in any row or column is equal to one). The Birkhoff theorem [12] states that the set of doubly stochastic matrices is a convex polyhedron whose vertices are permutation matrices. Thus, $\bar{\boldsymbol{G}}^{(2)}$ can be decomposed as

$$\bar{\boldsymbol{G}}^{(2)} = \sum_{i=1}^{N} \alpha_i \boldsymbol{P}_i \tag{34}$$

where N is a positive integer smaller than or equal to $n!$, each $\boldsymbol{P}_i$ is a permutation matrix, and $\alpha_i \geq 0$ such that

$$\sum_{i=1}^{N} \alpha_i = 1 \tag{35}$$

Then applying the triangular inequality, we have for any $\boldsymbol{\gamma} \in R^n$

$$\|\bar{\boldsymbol{G}}^{(2)}\boldsymbol{\gamma}\| \leq \|\bar{\boldsymbol{G}}^{(2)}\bar{\boldsymbol{\gamma}}\| \leq \sum_{i=1}^{N} \alpha_i \|\boldsymbol{P}_i \bar{\boldsymbol{\gamma}}\| = \|\bar{\boldsymbol{\gamma}}\| = \|\boldsymbol{\gamma}\| \qquad \square$$

Proof of Theorem 2. Since

$$\sigma_{z_i}^2 = \sum_{j=1}^{n} \sum_{\tau=-\infty}^{\infty} |g_{ij}(\tau)|^2 = 1, \quad i = 1, \ldots, n \tag{36}$$

we have $|g_{ij}(\tau)| \leq 1$ for any $\tau \in Z$ and any $i, j = 1, \ldots, n$. This implies

$$|g_{ij}(\tau)|^4 \leq |g_{ij}(\tau)|^2, \quad i, j = 1, \ldots, n \tag{37}$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{n}|K_i|^2 &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^4\gamma_j\right)^2\\ &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^4|\gamma_j|\right)^2\\ &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^2|\gamma_j|\right)^2 \end{aligned}\tag{38}$$

On the other hand, we shall show below that for any $\gamma \in R^n$

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^2|\gamma_j|\right)^2 \leq \sum_{j=1}^{n}|\gamma_j|^2 \tag{39}$$

Combining (38) with (39) gives

$$\begin{aligned}\sum_{i=1}^{n}|K_i|^2 &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^4|\gamma_j|\right)^2\\ &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^2|\gamma_j|\right)^2\\ &\leq \|\bar{\gamma}\|^2 = \|\gamma\|^2 \end{aligned}\tag{40}$$

Now assume equality in (40). Then, in particular, we have

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^4|\gamma_j|\right)^2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^2|\gamma_j|\right)^2 \tag{41}$$

Since $|g_{ij}(\tau)| \leq 1$ for any τ, i and j, (41) implies

$$|g_{ij}(\tau)|^4|\gamma_j| = |g_{ij}(\tau)|^2|\gamma_j| \tag{42}$$

for all values of τ and i, $j = 1, \ldots, n$. Since $|\gamma_j| \neq 0$ for all j, (42) gives

$$|g_{ij}(\tau)|^2(1 - |g_{ij}(\tau)|^2) = 0 \tag{43}$$

for all values of τ and $i, j = 1, \ldots, n$. Therefore, (43) is equivalent to the statement that $|g_{ij}(\tau)|$ is necessarily either zero or one for all values of τ and all $i, j = 1, \ldots, n$. Because

$$\sum_{\tau=-\infty}^{\infty} G(\tau)G^*(\tau) = I \tag{44}$$

(43) means that $G(t)$ can be represented as

$$G(t) = \mathrm{diag}[c_1\delta(t-t_1), \ldots, c_n\delta(t-t_n)]\boldsymbol{P} \tag{45}$$

with $|c_i| = 1$, $i = 1, \ldots, n$, where each t_k is an integer and $\boldsymbol{P}$ is a permutation matrix. In the z-domain, (45) is equivalent to

$$G(z) = \boldsymbol{C}\boldsymbol{\Lambda}(z)\boldsymbol{P} \tag{46}$$

where

$$\boldsymbol{C} = \mathrm{diag}[c_1, \ldots, c_n]$$

$$\boldsymbol{\Lambda}(z) = \mathrm{diag}[z^{t_1}, \ldots, z^{t_n}]$$

Here $\mathrm{diag}[c_1, \ldots, c_n]$ denotes an $n \times n$ diagonal matrix with diagonal elements $c_1, \ldots, c_n$. Thus, the cascade system $\boldsymbol{G}$, which is a solution of the single-stage maximization criterion (B), attains multichannel blind deconvolution. □

Proof of (39). Let $g_{ij}(w)$ be the Fourier transform of the sequence $\{g_{ij}(t)\}$. Then we have the Parseval identity

$$\sum_{\tau=-\infty}^{\infty} |g_{ij}(\tau)|^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |g_{ij}(w)|^2 \mathrm{d}w. \tag{47}$$

Therefore, for any $i = 1, \ldots, n$,

$$\begin{aligned}\left\{\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty} |g_{ij}(\tau)|^2|\gamma_j|\right\}^2 &= \left\{\sum_{j=1}^{n}\frac{1}{2\pi}\int_{-\pi}^{\pi} |g_{ij}(w)|^2|\gamma_j|\,\mathrm{d}w\right\}^2 \\ &= \left\{\frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{j=1}^{n} |g_{ij}(w)|^2|\gamma_j|\,\mathrm{d}w\right\}^2 \\ &\le \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sum_{j=1}^{n} |g_{ij}(w)|^2|\gamma_j|\right)^2 \mathrm{d}w \end{aligned} \tag{48}$$

where the last inequality comes from Schwarz's inequality. Therefore

$$\begin{aligned}\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty} |g_{ij}(\tau)|^2|\gamma_j|\right\}^2 &\le \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} |g_{ij}(w)|^2|\gamma_j|\right)^2 \mathrm{d}w \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \|\bar{G}^{(2)}(w)\bar{\gamma}\|^2 \mathrm{d}w \end{aligned} \tag{49}$$

Since $G(w)$ is unitary, applying Lemma 3 to the integrand in the right-hand side of (49) we have

$$\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^2|\gamma_j|\right\}^2 \leq \frac{1}{2\pi}\int_{-\pi}^{\pi}\|\bar{\gamma}\|^2\,\mathrm{d}w \qquad ,$$

$$= \|\bar{\gamma}\|^2 = \|\gamma\|^2 \tag{50}$$

which gives (39).

Remark 3. We can see below that we can use the criterion function $\Sigma_{i=1}^{n}|\kappa_{4,z_i}|$ instead of the criterion function $\Sigma_{i=1}^{n}|\kappa_{4,z_i}|^2$. Corresponding to (40), we have

$$\begin{aligned}\sum_{i=1}^{n}|K_i| &= \sum_{i=1}^{n}\left|\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^4\gamma_j\right| \\ &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^4|\gamma_j|\right) \\ &\leq \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\tau=-\infty}^{\infty}|g_{ij}(\tau)|^2\sqrt{|\gamma_j|}\right)^2 \\ &\leq \sum_{i=1}^{n}|\gamma_j|\end{aligned}$$

Using an argument similar to that in the derivation from (41) to (46), we can see that $G(z)$ takes the form of (46).

Based on Theorem 2, we can present a necessary and sufficient condition for the multichannel blind deconvolution of an LTI system as follows.

Theorem 3. *A necessary and sufficient condition for the multichannel blind deconvolution of the unknown system is that there exist n positive numbers* $\alpha_1, \ldots, \alpha_n$ *and a permutation* $(k_1, \ldots, k_n)$ *of the* $(1, \ldots, n)$ *such that the following* (a), (b), (c), *and* (d) *hold true.*

(a) $\sigma_{z_i}^2 = \alpha_i\sigma_{u_{k_i}}^2$ *for* $i = 1, \ldots, n$;
(b) $r_{z_i,z_i^*}(\tau) = 0$ *for all nonzero* $\tau \in Z$ *and all* $i = 1, \ldots, n$;
(c) $r_{z_i,z_j^*}(\tau) = 0$ *for all* $\tau \in Z$ *and all distinct* $i, j = 1, \ldots, n$; *and*
(d) $\kappa_{4,z_i} = \alpha_i^2\kappa_{4,u_{k_i}}$ *for* $i = 1, \ldots, n$.

Remark 4. Under the normalized condition of both the inputs and the equalizer outputs, i.e., $\sigma_{u_i}^2 = \sigma_{z_i}^2 = 1$ for $i = 1, \ldots, n$, we can choose all α_i to be equal to one in the above theorem. Thus, under the normalized condition of the inputs and the equalizer outputs, a necessary and sufficient condition for the multichannel blind deconvolution (with the gain factors d_i of unit magnitude) of the unknown system is that there exist a permuation $(k_1, \ldots, k_n)$ of $(1, \ldots, n)$ such that the following (a), (b) and (c) hold true:

(a) $r_{z_i,z_i^*}(\tau) = 0$ for all nonzero $\tau \in Z$ and all $i = 1, \ldots, n$;
(b) $r_{z_i,z_j^*}(\tau) = 0$ for all $\tau \in Z$ and all distinct $i, j = 1, \ldots, n$; and
(c) $\kappa_{4,z_i} = \kappa_{4,u_{k_i}}$ $i = 1, \ldots, n$.

Proof. For the *only if* part, let us suppose that the impulse response $\{G(t)\}$ of the cascade system takes the form of

$$G(t) = \mathrm{diag}[d_1 \delta(t - t_1), \ldots, d_n \delta(t - t_n)]\boldsymbol{P} \tag{51}$$

where $\boldsymbol{P}$ is a permutation matrix. Let $\boldsymbol{P}$ represent a permutation $(k_1, \ldots, k_n)$ of $(1, 2, \ldots, n)$. Then

$$\begin{bmatrix} u_{k_1} \\ \vdots \\ u_{k_n} \end{bmatrix} = \boldsymbol{P} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \tag{52}$$

Using the formula (5) and (6), we can easily obtain

$$\begin{aligned} \sigma_{z_i}^2 &= |d_i|^2 \sigma_{u_{k_i}}^2 && i = 1, \ldots, n; \\ r_{z_i,z_i^*}(\tau) &= 0 && \text{for all nonzero } \tau \in Z \text{ and all } i = 1, \ldots, n; \\ r_{z_i,z_j^*}(\tau) &= 0 && \text{for all } \tau \in Z \text{ and all distinct } i, j = 1, \ldots, n; \\ & && \text{and} \\ \kappa_{4,z_i} &= |d_i|^4 \kappa_{4,u_{k_i}} && i = 1, \ldots, n. \end{aligned}$$

Thus, by putting $\alpha_i = |d_i|^2 > 0$, they correspond to the relations (a), (b), (c), and (d) in Theorem 3.

For the *if* part, presume that there exist n positive numbers $\alpha_1, \ldots, \alpha_n$ and a permutation $(k_1, \ldots, k_n)$ of $(1, 2, \ldots, n)$ such that the relations (a), (b), (c), and (d) hold true. Let $\bar{u}_i$ and $\bar{z}_i$ be the normalized variables of u_i and z_i, respectively, that is,

$$\bar{u}_i = \frac{1}{\sigma_{u_i}} u_i \tag{53}$$

$$\bar{z}_i = \frac{1}{\sigma_{z_i}} z_i \tag{54}$$

Then we have the following relations for their second- and fourth-order cumulants:

$$\sigma_{\bar{u}_i}^2 = 1, \tag{55}$$

$$\sigma^2_{\tilde{z}_i} = 1, \tag{56}$$

$$\kappa_{4,\tilde{u}_i} = \frac{\kappa_{4,u_i}}{\sigma^4_{u_i}} \tag{57}$$

$$\kappa_{4,\tilde{z}_i} = \frac{\kappa_{4,z_i}}{\sigma^4_{z_i}} \tag{58}$$

Let

$$\tilde{\boldsymbol{u}} = (\tilde{u}_1, \ldots, \tilde{u}_n)^T$$
$$\tilde{\boldsymbol{z}} = (\tilde{z}_1, \ldots, \tilde{z}_n)^T$$

be the normalized vectors of $\boldsymbol{u}$ and $\boldsymbol{z}$, respectively. Then from (53) and (54)

$$\boldsymbol{u} = \mathrm{diag}[\sigma_{u_1}, \ldots, \sigma_{u_n}]\tilde{\boldsymbol{u}} \tag{59}$$

$$\tilde{\boldsymbol{z}} = \mathrm{diag}\left[\frac{1}{\sigma_{z_1}}, \ldots, \frac{1}{\sigma_{z_n}}\right]\boldsymbol{z} \tag{60}$$

We put $\boldsymbol{D}_1 = \mathrm{diag}[\sigma_{u_1}, \ldots, \sigma_{u_n}]$ and $\boldsymbol{D}_2 = \mathrm{diag}[1/\sigma_{z_1}, \ldots, 1/\sigma_{z_n}]$. Let us consider the cascade system with the normalized input and output shown in Fig. 2.

Let $\boldsymbol{G}'(z)$ be the transfer function of the normalized cascade system shown in Fig. 2. Then

$$\begin{aligned}\boldsymbol{G}'(z) &= \boldsymbol{D}_2\boldsymbol{G}(z)\boldsymbol{D}_1 \\ &= \mathrm{diag}\left[\frac{1}{\sigma_{z_1}}, \ldots, \frac{1}{\sigma_{z_n}}\right]\boldsymbol{G}(z)\mathrm{diag}[\sigma_{u_1}, \ldots, \sigma_{u_n}]\end{aligned} \tag{61}$$

From the relations (b) and (c) in Theorem 3 and Eq. (54) and (56), we have

$$r_{\tilde{z}_i,\tilde{z}_j^*}(\tau) = \delta(i-j)\delta(\tau) \quad \text{for all } \tau \in Z \text{ and all } i, j = 1, \ldots, n \tag{62}$$

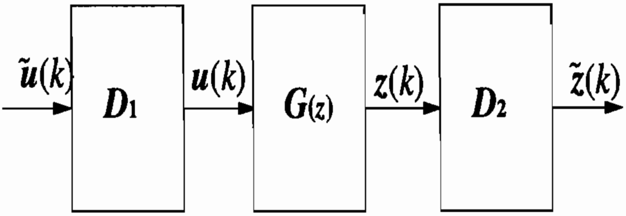


Figure 2 Normalized cascade system.

Therefore, under this normalized whitening condition of the vector process $\{z(t)\}$, from (38) and (40) in the proof of Theorem 2, we obtain the inequality

$$\sum_{i=1}^{n} |\kappa_{4,\tilde{z}_i}|^2 \leq \sum_{i=1}^{n} |\kappa_{4,\tilde{u}_i}|^2 \tag{63}$$

where equality holds if and only if

$$\boldsymbol{G}'(z) = \boldsymbol{D}\boldsymbol{\Lambda}(z)\boldsymbol{P} \tag{64}$$

where $\boldsymbol{D}$ is a diagonal matrix, $\boldsymbol{\Lambda}(z)$ is a shift matrix, and $\boldsymbol{P}$ is a permutation matrix. However, from (57), (58), and the relations (a) and (d) in Theorem 3, we obtain

$$\begin{aligned}\sum_{i} |\kappa_{4,\tilde{z}_i}|^2 &= \sum_{i} \left|\frac{\kappa_{4,z_i}}{\sigma_{z_i}^4}\right|^2 = \sum_{i} \left|\frac{\alpha_i^2 \kappa_{4,u_{k_i}}}{\alpha_i^2 \sigma_{u_{k_i}}^4}\right|^2 \\ &= \sum_{i} \left|\frac{\kappa_{4,u_{k_i}}}{\sigma_{u_{k_i}}^4}\right|^2 = \sum_{i} \left|\frac{\kappa_{4,u_i}}{\sigma_{u_i}^4}\right|^2 \\ &= \sum_{i} |\kappa_{4,\tilde{u}_i}|^2 \end{aligned} \tag{65}$$

Therefore, equality in (63) holds and thus (64) holds true. Combining (61) with (64), we have

$$\boldsymbol{D}_2\boldsymbol{G}(z)\boldsymbol{D}_1 = \boldsymbol{D}\boldsymbol{\Lambda}(z)\boldsymbol{P}$$

which gives

$$\boldsymbol{G}(z) = \boldsymbol{D}_2^{-1}\boldsymbol{D}\boldsymbol{\Lambda}(z)\boldsymbol{P}\boldsymbol{D}_1^{-1} = \boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{\Lambda}(z)(\boldsymbol{P}\boldsymbol{D}_1^{-1}\boldsymbol{P}^T)\boldsymbol{P} \tag{66}$$

Since $\boldsymbol{P}\boldsymbol{D}_1^{-1}\boldsymbol{P}^T$ is a diagonal matrix, the above equation becomes

$$\boldsymbol{G}(z) = \boldsymbol{D}'\boldsymbol{\Lambda}(z)\boldsymbol{P} \tag{67}$$

where

$$\boldsymbol{D}' = \boldsymbol{D}_2^{-1}\boldsymbol{D}(\boldsymbol{P}\boldsymbol{D}_1^{-1}\boldsymbol{P}^T) \tag{68}$$

Since $\boldsymbol{D}'$ is also a diagonal matrix, (67) means $\boldsymbol{G}(z)$ attains the multichannel blind deconvolution of the unknown system. □

IV. SIMULATION RESULTS

In order to see the effectiveness of the proposed criteria, we developed only a stochastic gradient algorithm for solving the multistage blind deconvolution problem by using the multistage maximization criterion (A). The algorithm requires PSK 4-(multichannel) spectral prewhitening of the output process of the unknown system as in [3]. We used a finite impulse response (FIR) system to approximate an equalizer.

We took the following system, which is a 2-input and 2-output all-pass system described by

$$\boldsymbol{H}(z) = \begin{pmatrix} \dfrac{0.5 + z^{-1}}{1 + 0.5z^{-1}} & 0 \\ 0 & \dfrac{0.2 + z^{-1}}{1 + 0.2z^{-1}} \end{pmatrix} \begin{pmatrix} \dfrac{1}{2} & -\dfrac{\sqrt{3}}{2} \\ \dfrac{\sqrt{3}}{2} & \dfrac{1}{2} \end{pmatrix} \tag{69}$$

We note that $\boldsymbol{H}(z)$ satisfies the all-pass condition $\boldsymbol{H}(\mathrm{e}^{\mathrm{j}\omega})\boldsymbol{H}^{*}(\mathrm{e}^{\mathrm{j}\omega}) = \boldsymbol{I}$. Hence we need not perform prewhitening in this case. The first input signals $u_1(t)$ were independent realizations for a 16-QAM source with unit variance, and the second input signals $u_2(t)$ were independent realizations for a PSK (phase-shift keying) source with unit variance. These signal constellations are shown in Figs (a) and (b) in Fig. 3.

In applying the stochastic gradient algorithm, we used an equalizer $\boldsymbol{W}$ of length 24 which was initialized to $\boldsymbol{W}^{(0)}$. The element values of the impulse response $\{\boldsymbol{W}^{(0)}(t);\ t = 1, \ldots, 24\}$ were set to be zero, except for $w_{11}^{(0)}(12) = w_{21}^{(0)}(12) = 1$, $w_{12}^{(0)}(12) = \frac{1}{2}$, and $w_{22}^{(0)}(12) = \frac{1}{6}$. The magnitudes of the elements of the impulse response of the initial cascade system $\boldsymbol{G}_0(z) = \boldsymbol{W}_0(z)\boldsymbol{H}(z)$ were shown in Fig. 4. It can be seen from Fig. 4 that the highest magnitude in responses $\{g_{11}^{(0)}(t)\}$ and $\{g_{12}^{(0)}(t)\}$ is $g_{12}^{(0)}(13)$ and the highest magnitude in responses $\{g_{21}^{(0)}(t)\}$ and $\{g_{22}^{(0)}(t)\}$ is $g_{21}^{(0)}(13)$. The algorithm contains stochastic expectation, and thus we used 50 data samples to calculate approximately stochastic expectation. As a measure of performance we used the *multichannel intersymbol interference* denoted by $\mathrm{M}_{\mathrm{ISI}}$, defined by

$$\mathrm{M}_{\mathrm{ISI}} := \sum_{i=1}^{n} \frac{\left|\sum_{j=1}^{n} \sum_{t=-\infty}^{\infty} |g_{ij}(t)|^2 - |g_{i\cdot}|_{\max}^2\right|}{|g_{i\cdot}(\cdot)|_{\max}^2} + \sum_{j=1}^{n} \frac{\left|\sum_{i=1}^{n} \sum_{t=1\infty}^{\infty} |g_{ij}(t)|^2 - |g_{\cdot j}|_{\max}^2\right|}{|g_{\cdot j}(\cdot)|_{\max}^2}$$

where $|g_{i\cdot}(\cdot)|_{\max}^2$ and $|g_{\cdot j}(\cdot)|_{\max}^2$ are respectively defined by

$$|g_{i\cdot}(\cdot)|_{\max}^2 := \mathrm{Max}_{j=1,\ldots,n} \mathrm{Max}_{-\infty<t<\infty} |g_{ij}(t)|^2$$

$$|g_{\cdot i}(\cdot)|_{\max}^2 := \mathrm{Max}_{i=1,\ldots,n} \mathrm{Max}_{-\infty<t<\infty} |g_{ij}(t)|^2$$

It can be seen easily that $\mathrm{M}_{\mathrm{ISI}} = 0$ if and only if $\boldsymbol{G}(z)$ is of the form (5).

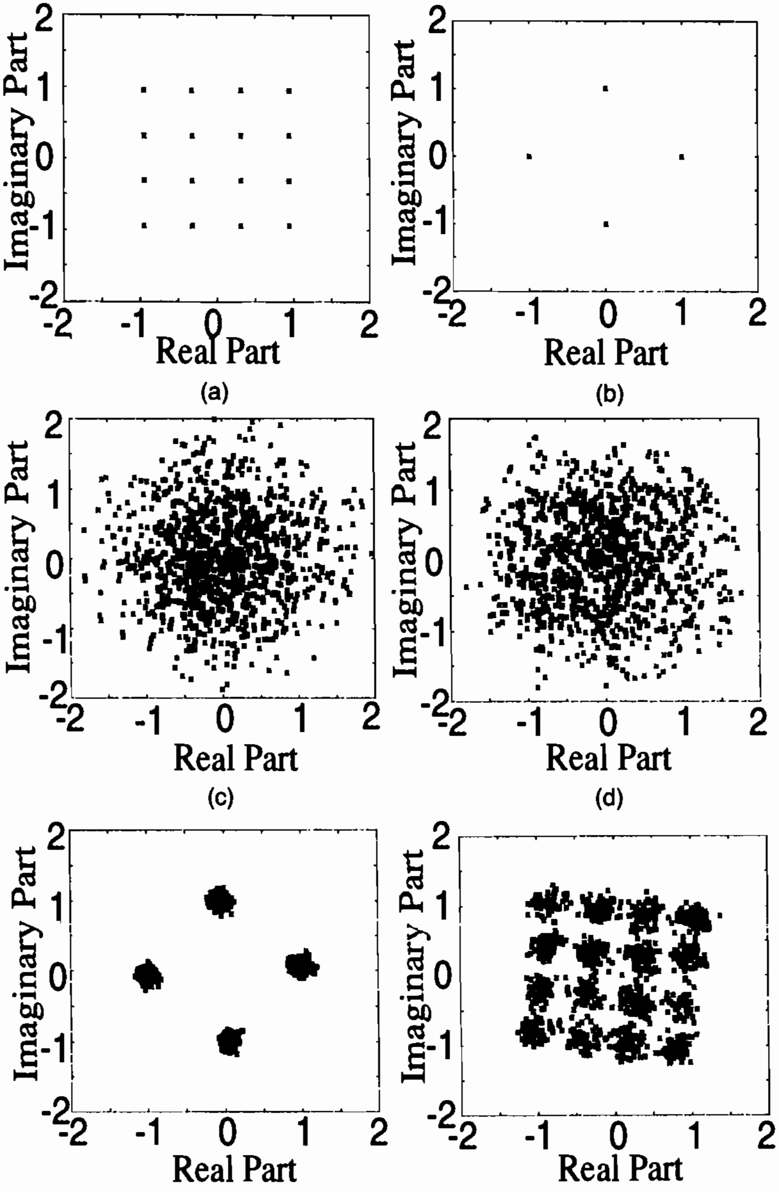
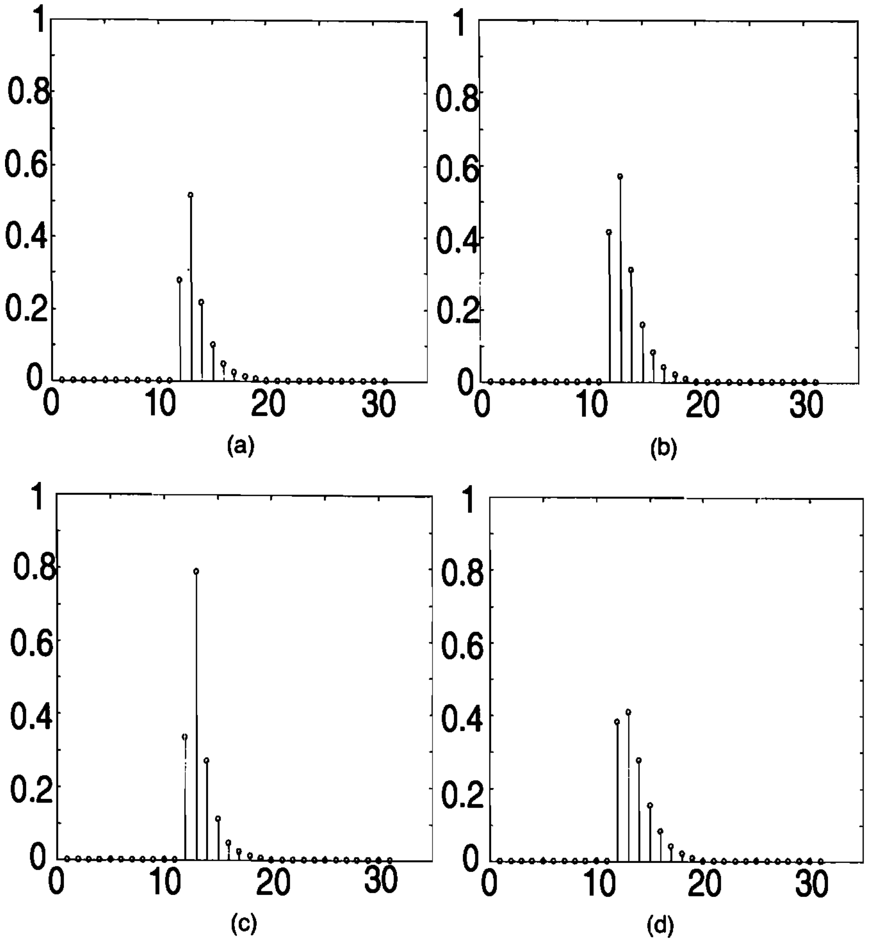


Figure 3 Signal constellations of channel inputs, channel outputs, and equalizer outputs: (a) inputs of channel 1; (b) inputs of channel 2; (c) unequalized outputs of channel 1; (d) unequalized outputs of channel 2; (e) equalized outputs of channel 1; (f) equalized outputs of channel 2.



(a)

(b)

(c)

(d)

Figure 4 Magnitudes of the impulse response $\{G_0(t)\}$ before equalization: (a) magnitude of $\{g_{11}^{(0)}(t)\}$; (b) magnitude of $\{g_{12}^{(0)}(t)\}$; (c) magnitude of $\{g_{21}^{(0)}(t)$; (d) magnitude of $\{g_{22}^{(0)}(t)\}$.

The initial M_{ISI} in the logarithmic (dB) scale was 8.0411 dB. The algorithm was tested in 10 Monte Carlo runs using 20,000 data samples at each of the two outputs. Figs (c), (d) and Figs (e), (f) in Fig. 3 show the channel output signal and the equalizer output signal constellations, respectively.

Since the absolute value of fourth-order cumulant of the 4-PSK source is greater than that of the 16-QAM source, the 4-PSK source was recovered

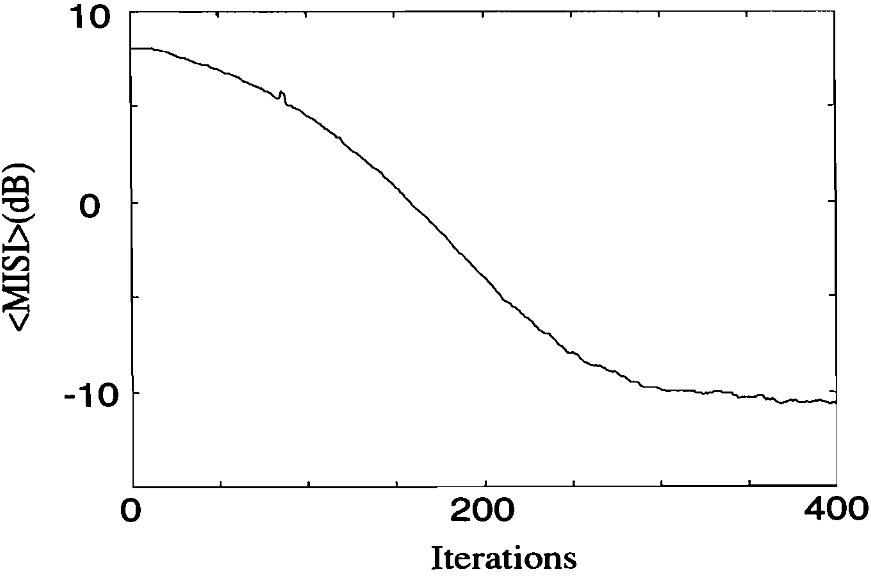


Figure 5 Performance of the algorithm for the 16-QAM source and the 4-PSK source.

as the output $z_1(t)$ at Stage 1 and the 16-QAM source as the output $z_2(t)$ at Stage 2. In Fig. 5, we plot the averaged M_{ISI}, denoted by $\langle M_{ISI} \rangle$, over 10 Monte Carlo runs. We see that the averaged M_{ISI} crosses the -10 dB level after 300 iterations. Figure 6 shows the magnitudes of the elements of the impulse response $\{G(t)\}$ after equalization.

V. CONCLUSIONS

We have proposed a multistage maximization criterion and a single-stage criterion for attaining multichannel blind deconvolution. It has been shown that they are equivalent under the normalization condition of the input process. A simulation example has been shown to illustrate the equalization results under the multistage maximization criterion (A). Stochastic gradient algorithms corresponding to the single-stage maximization criterion (B) are yet to be developed.

ACKNOWLEDGMENT

The author would like to thank Takehito Sato for his cooperation in the computer simulations in this chapter.

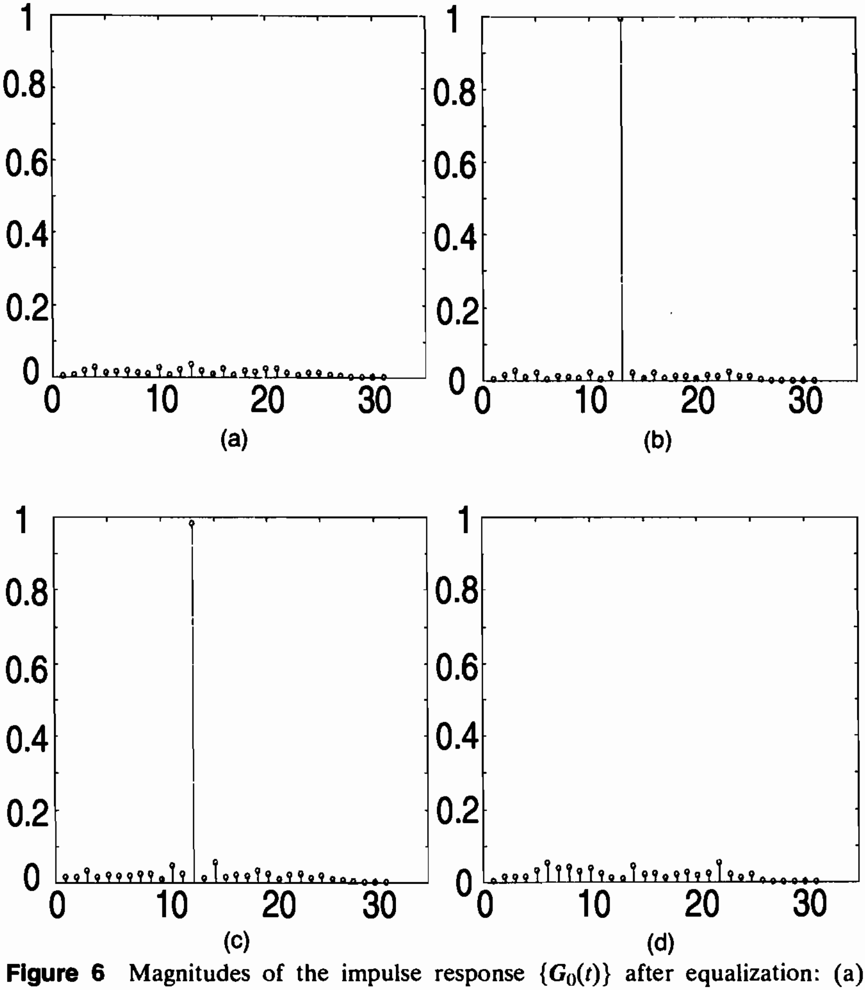


Figure 6 Magnitudes of the impulse response $\{G_0(t)\}$ after equalization: (a) magnitude of $\{g_{11}(t)\}$; (b) magnitude of $\{g_{12}(t)\}$; (c) magnitude of $\{g_{21}(t)\}$; (d) magnitude of $\{g_{22}(t)\}$.

REFERENCES

1. RA Wiggins. Geoexploration 16:21–35, 1978.
2. DL Donoho. In DF Findley, ed, Applied Times Series Analysis, II, New York: Academic Press, 1981, pp 565–608.
3. O Shalvi, E Weinstein. IEEE Trans Inform Theory 36:312–321, 1990.
4. H Oda, Y Sato. Proc IEEE Int Symp Inform Theory, 1993, pp 327.

5. JF Cardoso. Proc ICASSP 4:2109–2112, 1989.
6. P Comon, C Jutten, J Heraut. Signal Processing 24:1–10, 1991.
7. L Tong, Y Inouye, R Liu. IEEE Trans Signal Processing 41:2461–2470, 1993.
8. P Comon. Signal Processing 36:287–314, 1994.
9. D Yellin, E Weinstein. IEEE Trans Signal Processing 42:2158–2168, 1994.
10. Y Inouye, T Habe. Proc IEEE Signal Processing Workshop on Higher-Order Statistics, 1995, pp 96–100.
11. DR Brillinger. Time Series, Data Analysis and Theory. New York: McGraw Hill, 1981.
12. RA Horm, CR Johnson. Matrix Analysis. New York: Cambridge University Press, 1985.

This Page Intentionally Left Blank

15

Bayesian Approaches for Robust Array Signal Processing

A. Lee Swindlehurst

Brigham Young University, Provo, Utah

Mats Viberg

Chalmers University of Technology, Göteborg, Sweden

I. INTRODUCTION

A. Historical Background

The general problem addressed in this work is that of locating the sources of radiated energy received by an array of spatially distributed collectors, or sensors. It was, in fact, the desire to locate and track enemy aircraft using radar that gave birth to statistical signal processing as a subdiscipline of electrical engineering in the 1940s. In this early application, the parameters of interest were the direction, range, and radial velocity (doppler) of the target, all of which were to be estimated from the outputs of a single radar antenna. The direction to the target was the most important parameter to be estimated, since it had to be known before range and velocity could be determined. Target direction was usually determined by placing a collimating dish behind the antenna, sweeping the entire device through all possible target directions, and then finding the directions from which maximum energy was received. This simple system has been modified and refined over the years, and sophisticated algorithms have been developed to enhance radar direction-finding performance.

The problems of direction-finding and emitter localization have become important in many other areas besides radar. For example, acoustic signals received by hydrophones are used in underwater applications to detect and locate submarines and surface vessels. In oil exploration, explosive charges

This work was supported by the National Science Foundation under grant MIP-9408154.

are detonated below the earth's surface in order to create acoustic reflections off the various layers of the earth's crust. Measurement of the directions-of-arrival (DOAs) of these reflections helps determine the relative position and thickness of these layers. A common problem in both ground- and satellite-based communications is the cancelation of unwanted interference signals from one direction in favor of a desired signal arriving from another. Effective cancelation of the interference is highly dependent on knowledge of its DOA.

The applicability of source localization to a wide variety of problems has led to a correspondingly large amount of research and considerable refinement of the techniques used to obtain the location estimates. One of the very first refinements to the simple source location system described above was the use of an array of multiple collectors to increase sensitivity and spatial resolution. The first techniques that exploited this antenna array structure were referred to as *delay-and-sum* (DS) *beamformers*, since the individual antenna outputs were delayed, weighted, and summed together to form a single array output. By varying the delays and weights, the directional response of the array (sometimes called an antenna *beam*, hence the term *beamforming*) can be steered to various directions of interest, and the amount of energy impinging upon the array from these directions can be measured. Thus, not only are speed and reliability gained since the antenna is electronically rather than physically steered, but directionality is improved since the incoming energy is sampled over a much wider physical aperture.

Beamforming techniques have been widely used up to the present, primarily because they are simple and computationally efficient. Improvements have been made over the years to the original DS approach (see especially [1,2]), and adaptive implementations have been extensively studied (e.g., [3–5]). To shape the response of the array, some approaches have made use of the fact that, for a linear array, the DS and weighting operations are equivalent to a windowed discrete Fourier transform (DFT) in the spatial dimension. Although providing a link with the well-developed theory of time domain filter design, this connection points out that the resolution and convergence problems of the DFT are inherited by beamforming techniques. As an example, even with an infinite amount of noise-free data, DS beamforming will not generally give correct estimates of the location parameters of multiple emitters. This undesirable property is referred to as estimator *inconsistency*, and is the major drawback to use of beamforming algorithms in practice.

A number of methods borrowed from time series estimation have also been applied to the DOA problem. These techniques are typically based on parametric linear prediction or autoregressive (AR) models for the data

[6–9], and they tend to provide much higher resolution than conventional DS beamforming. Because these methods are derived for uniformly sampled time series, for the DOA estimation problem they are limited in application to uniform linear arrays. Their principal drawback stems from the fact that AR data models are not well matched to the signal-in-additive-noise model most often assumed for the array processing problem. The result is that even at moderate SNR values, the parameter estimates may, as in the case of beamforming, be inconsistent and subject to large biases.

B. Modern Techniques

As the field of source location research has matured, many have hearkened back to the statistical foundations of this parameter estimation problem and found maximum likelihood (ML) solutions [10–13]. The motivation for studying such methods is the desire to obtain unbiased, consistent location estimates of minimum variance. Since, in general, ML techniques require multidimensional nonlinear optimization procedures, most of the emphasis in recent years has been the development and analysis of suboptimal, but computationally simpler methods that achieve near-ML performance. Perhaps the biggest breakthrough came with the development of the so-called MUSIC algorithm (for MUltiple SIgnal Classification) by Schmidt in 1979 [14,15]. A similar approach was independently proposed by Bienvenu and Kopp that same year [16]. The MUSIC approach provided an elegant geometrical solution to the source location problem using the powerful mathematical tools of vector spaces. Although earlier work had hinted at this type of solution [17–19], the MUSIC algorithm was the first to fully exploit the underlying signal model for the antenna array problem.

Research in high-resolution DOA estimation has been dominated throughout the past ten years by analysis and extensions of the MUSIC algorithm. Its popularity is primarily due to the fact that it may be applied to antenna arrays of arbitrary geometry to estimate multiple parameters per source. Unlike beamforming and AR modeling techniques, MUSIC is a consistent, high-resolution estimator for situations involving multiple emitters and additive noise. MUSIC does, however, share some of the computational drawbacks of these algorithms. For example, the algorithm requires that the response of the array for all possible source parametrizations be known *a priori*. In addition, determining the MUSIC parameter estimates involves a potentially burdensome search procedure that is not guaranteed to have adequate resolution. For single-parameter, narrowband direction-finding applications, the recently introduced ESPRIT algorithm (for *Estimation of Signal Parameters via Rotational Invariance Techniques*) alleviates these difficulties by constraining the antenna array to be composed of two identical, but otherwise arbitrary, subarrays [20,21]. This special array geometry results in an elegant and computationally efficient solution that does not require prior knowledge of the subarray response.

Recent investigations of a special formulation of ESPRIT have demonstrated that beamforming, MUSIC, maximum likelihood, ESPRIT, and various other algorithms are all special cases of a general *subspace fitting* minimization procedure [22–24]. Unifying these techniques under a common framework sheds light on their geometrical relationships and facilitates comparative analyses. For example, the subspace fitting paradigm has been used to evaluate the *asymptotic* (in the data) performance of the above algorithms [25–29]. A byproduct of this analysis was the development of an optimal *weighted* subspace fitting algorithm that achieves parameter estimates of minimum variance [29–31].

C. Robustness Issues

All of the methods for direction-finding (DF) listed above rely on the availability of information about the array response, and assume that the signal wavefronts impinging on the array have perfect spatial coherence (e.g., perfect plane waves). The array response may be determined by either empirical measurements (a process referred to as array *calibration*), or by making certain assumptions about the sensors in the array and their geometry (e.g., identical sensors in known locations). Unfortunately, an array cannot be perfectly calibrated, and idealized assumptions made about the array geometry and wave propagation are never satisfied in practice. Due to changes in antenna location, temperature, and the surrounding environment, the response of the array may be significantly different than when it was last calibrated. Furthermore, the calibration measurements themselves are subject to gain and phase errors, and they can only be obtained for discrete DOA values (thus necessitating interpolation techniques for uncalibrated directions).

For the case of analytically calibrated arrays of nominally identical, identically oriented elements, errors result since the elements are not really identical and their locations are not precisely known. Furthermore, even if the calibration data were precisely known, inhomogeneous propagation effects cause the actual array response to a given signal to be different from the response at the time of calibration. Depending on the degree to which the actual antenna response differs from its nominal value, DF and beamformer performance may be significantly degraded.

To account for the effects described above, a slightly generalized model for the array response will be considered in this work. The response will be parametrized not only by the directions of arrival (DOAs) of the signals, but also by a vector of perturbation or "nuisance" parameters that describe deviations of the response from its nominal value. These parameters can include, for example, displacements of the antenna elements from their nominal positions, uncalibrated receiver gain and phase offsets, etc. With such a model, a natural approach is to attempt to estimate the unknown nuisance parameters simultaneously with the signal parameters. Such methods are referred to as *auto-calibration* techniques, and have been proposed by a number of authors, including [32–38] among many others.

When auto-calibration techniques are employed, it is critical to determine whether both the signal and nuisance parameters are identifiable. In certain cases they are not; for example, one cannot uniquely estimate both DOAs and sensor phase characteristics unless of course additional information is available, such as sources in known locations [39–41], cyclostationary signals with two or more known cycle frequencies [42], or partial information about the phase response of the array [43]. The identifiability problem can be alleviated if the perturbation parameters are assumed to be drawn from some known *a priori* distribution. While this itself represents a form of additional information, it has the advantage of allowing an optimal maximum *a posteriori* (MAP) solution to the problem to be formulated [35,37]. In [37] it is shown that, by using an asymptotically equivalent approximation to the resulting MAP criterion, the estimation of the signal and nuisance parameters can be decoupled, leading to a significant simplification of the problem. This result will be reviewed in the present work.

Presumably, any of the above auto-calibration methods would provide not only improved DOA estimates, but also calibration information that would be useful in beamformer implementation. Another goal of this paper is to investigate beamformer performance for the case where the optimal MAP perturbation parameter estimates of [37] are used to update the array calibration. Simulations demonstrate that such an approach can result in a significant performance improvement, measured using either interference rejection capability or mean-squared error. In addition, for simple additive unstructured calibration errors, the MAP approach is shown in certain cases to yield a beamformer similar to the *subspace corrected* algorithms described in [44–49].

Before moving on to a description of the MAP techniques described above, some additional background information is given in the next two sections. Section II outlines the mathematical model that will be assumed throughout this work, and introduces some important notation. A description of some simple array perturbation models will also be given in this section. Section III then focuses on several DOA estimation and beamforming algorithms that are relevant to subsequent discussions, and that will be used for comparison purposes in simulations later in the paper. The exact MAP estimator and an asymptotically equivalent approximation based on subspace fitting are described next in Section IV. The application of the MAP approach to minimum mean-square error beamforming is also presented, and some special cases are examined in detail. The paper concludes with a series of representative simulation examples in Section V.

II. MATHEMATICAL MODEL

For the most part, the standard narrowband modeling assumptions that have been used since the original work of Schmidt [14] and Bienvenu and Kopp [16] will be adopted. The term "narrowband" here refers to the assumption that the envelopes of the (bandpass) signals received by the array do not change appreciably as the signal wavefronts propagate across the array aperture. Furthermore, the signals are assumed to emanate from point source emitters, the array elements and receivers are linear devices (superposition applies), and the noise is additive with known second-order spatial statistics.

Consider an array of m sensors, having arbitrary positions and characteristics, that receives the waveforms of d narrowband point sources, where $d < m$. At time t, the output of each sensor is collected in the vector $\mathbf{x}(t)$, which is modeled as a superposition of the array's response to all d signals:

$$\mathbf{x}(t) = [a(\theta_1, \boldsymbol{\rho}) | \cdots | a(\theta_d, \boldsymbol{\rho})] \begin{bmatrix} s_1(t) \\ \vdots \\ s_d(t) \end{bmatrix} + \mathbf{n}(t) \tag{1}$$

$$= \mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho})\mathbf{s}(t) + \mathbf{n}(t) \tag{2}$$

The columns of the $m \times d$ matrix $\mathbf{A}$ are the so-called array *steering* vectors, denoted $a(\theta_i, \boldsymbol{\rho})$, $i = 1, \ldots, d$. These vectors describe the array response to a unit amplitude waveform with signal parameter(s) θ_i. The parameters in θ_i are typically associated with the location of the source, and can include azimuth angle of arrival, elevation angle, range, as well as other quantities related to polarization and doppler. Though not necessary, it is assumed for simplicity in our discussion that θ_i is a real-valued quantity equal to the DOA of signal i. The components of the d-vector $\boldsymbol{\theta}$ are the DOAs of the model, whereas the vector $\boldsymbol{\theta}_0$ represents their true values.

The above model also allows for a set of unknown perturbation* or nuisance parameters, collected in the real n-vector $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_n]^T$. As will be seen later, the elements of $\boldsymbol{\rho}$ can include both unstructured or structured parameters such as sensor gain, phase, position, mutual coupling, etc. Unlike $\boldsymbol{\theta}$, which is assumed to be deterministic, $\boldsymbol{\rho}$ will be modeled as a random vector drawn from a known Gaussian prior distribution. The mean of the distribution, denoted by $\boldsymbol{\rho}_0$, corresponds to the nominal array response with no perturbation. The covariance is defined by

$$\mathbf{E}[(\boldsymbol{\rho} - \boldsymbol{\rho}_0)(\boldsymbol{\rho} - \boldsymbol{\rho}_0)^T] = \boldsymbol{\Omega} \tag{3}$$

and is also assumed to be known and of full rank. The matrix $\boldsymbol{\Omega}$ could be determined, for example, using sample statistics from a number of independent, identical calibration experiments, or using tolerance data specified by the manufacturer of the sensors. Some examples of typical $\boldsymbol{\rho}$-parametrizations are given in Section II.A. The nominal array response is assumed to be unambiguous;* i.e., the matrix $[a(\theta_1, \boldsymbol{\rho}_0), \ldots, a(\theta_m, \boldsymbol{\rho}_0)]$ has full rank for any collection of distinct parameters $\theta_1, \ldots, \theta_m$.

The complex d-vector $\mathrm{s}(t)$ is composed of the emitter waveforms received at time t, and the m-vector $\mathbf{n}(t)$ accounts for additive measurement noise. Although the signal waveforms will be assumed to be white Gaussian random processes when deriving the exact MAP estimator, the properties of the proposed method have been analyzed under a less restrictive assumption in [37]. The covariance of the signals is assumed to satisfy

$$\mathbf{E}[\mathrm{s}(t)\mathrm{s}^*(s)] = \mathbf{P}\delta_{t,s} \tag{4}$$

$$\mathbf{E}[\mathrm{s}(t)\mathrm{s}^T(s)] = 0 \tag{5}$$

where $\delta_{t,s}$ is the Kronecker delta. The covariance matrix $\mathbf{P}$ is further assumed in this work to be positive definite (no perfectly coherent signals are present). Note that $\{\cdot\}^*$ denotes the complex conjugate transpose. The noise term, $\mathbf{n}(t)$, is modeled as a stationary, complex Gaussian random process that is uncorrelated with the signals. The noise is assumed to be

*More precisely, the array is assumed to be unambiguous for all perturbation parameters in an open neighborhood of ρ_0.

zero mean, circularly symmetric, as well as spatially and temporally white:

$$\mathbf{E}[\mathbf{n}(t)\mathbf{n}^*(s)] = \sigma^2 \mathbf{I}\delta_{t,s} \tag{6}$$

$$\mathbf{E}[\mathbf{n}(t)\mathbf{n}^T(s)] = 0 \tag{7}$$

Assuming N samples are taken from the array, the following data matrix may be defined:

$$\mathbf{X} = [\mathbf{x}(t_1) \cdots \mathbf{x}(t_N)] = \mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho})\mathbf{S} + \mathbf{N}, \tag{8}$$

where $\mathbf{S}$ and $\mathbf{N}$ are defined similarly to $\mathbf{X}$. Based on the measurements in $\mathbf{X}$, the problem of interest is to estimate the DOAs and individual waveforms of the sources. The number of signals, d, is assumed to be known.

A. Array Error Parametrizations

Since we are interested in studying the combined effects of finite sample errors and modeling errors, the size of the perturbations relative to the number of available data samples plays a crucial role. The variances of the estimated DOAs are known to be proportional to $1/N$ in the finite-sample-only case [22], whereas they are proportional to $\boldsymbol{\Omega}$ in the model-error-only case [27,28]. In the approximation introduced in Section IV.B, the relative contributions of the two error sources will be assumed to be of comparable magnitude, and the covariance matrix of the perturbation parameters will be expressed as

$$\boldsymbol{\Omega} = \bar{\boldsymbol{\Omega}}/N \tag{9}$$

where $\bar{\boldsymbol{\Omega}}$ is independent of N. An argument for the somewhat artificial assumption (9) is that if $\boldsymbol{\Omega} = o(1/N)$, then the effect of the modeling errors can be neglected and the methods designed for finite sample errors only are optimal. On the other hand, if $\boldsymbol{\Omega}^{-1} = o(1/N)$, the effect of the modeling errors dominates, thus rendering the methods designed solely for such errors optimal. Since the MAP approach presented herein is inherently more complicated than either of the methods that take only finite samples or modeling errors into account, the former should be avoided when one type of error dominates the other. A further assumption that will be made in Section IV when deriving the approximate MAP estimator is that the deviation $\boldsymbol{\rho} - \boldsymbol{\rho}_0$ is relatively small (i.e., second-order effects of the array perturbation will be ignored).

The reason for a *random* perturbation model as opposed to a *deterministic* one lies in the consideration of how one chooses to quantify the effects of the perturbation. In a given fixed scenario, of course, the presence of array errors will introduce a bias in the DF and signal estimates. Presumably, if one wanted to measure the magnitude of this bias, it would simply be a matter of directly computing the limiting ($N \rightarrow \infty$) estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{S}}$, and then subtracting $\boldsymbol{\theta}_0$ and $\mathbf{S}$. This procedure would obviously have to be repeated for every perturbation scenario considered, since the bias would be different in each case. The advantage of using a random model is that one can obtain a measure of the *average* effect of the array errors on estimation performance, measured now in terms of *variance* rather than bias, without being forced to adopt a particular perturbation scenario (which may be no more representative than any other similar perturbation).

It is useful to examine the form of $\boldsymbol{\rho}$ and its distribution for several commonly encountered perturbation models. Three such models are briefly described below.

1. Gain and Phase Errors

For arrays composed of nominally identical elements, a common approach used to describe deviations in the array response attempts to model the nonuniform gain and phase effects of the receiver electronics behind each antenna element. In this model, the nominal response is perturbed by an unknown complex diagonal matrix:

$$\mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho}) = \mathbf{G}\mathbf{A}(\boldsymbol{\theta}) \tag{10}$$

and

$$\boldsymbol{\rho} = \begin{bmatrix} \mathrm{Re}\{\mathbf{g}\} \\ \mathrm{Im}\{\mathbf{g}\} \end{bmatrix} \tag{11}$$

where $\mathbf{g} = \mathrm{diag}\{\mathbf{G}\}$. The mean of the distribution for $\boldsymbol{\rho}$ in this case is given by $\boldsymbol{\rho}_0 = [\mathbf{e}^T \, 0]^T$, where $\mathbf{e}$ is an $m \times 1$ vector of ones. For simplicity, it is often assumed that the covariance of $\boldsymbol{\rho}$ is given by $\boldsymbol{\Omega} = (\sigma_a^2/2)\mathbf{I}$, which implies that the individual gain and phase errors are all mutually independent and identically distributed. The effects of uncalibrated mutual coupling can be incorporated into the above model by assuming the matrix $\mathbf{G}$ has non-zero off-diagonal elements.

2. Sensor Position Errors

For this case, assume an arbitrary array of identical unit-gain omnidirectional sensors in the x-y plane with randomly perturbed sensor locations. The perturbed array response can be written as

$$a(\theta,\boldsymbol{\rho}) = \begin{bmatrix} \exp\left(j\frac{2\pi}{\lambda}[(x_1+\tilde{x}_1)\sin\theta+(y_1+\tilde{y}_1)\cos\theta]\right) \\ \vdots \\ \exp\left(j\frac{2\pi}{\lambda}[(x_m+\tilde{x}_m)\sin\theta+(y_m+\tilde{y}_m)\cos\theta]\right) \end{bmatrix}$$

where (x_i, y_i) are the nominal coordinates of the ith sensor and $(\tilde{x}_i, \tilde{y}_i)$ are the corresponding position errors. Defining

$$\boldsymbol{\rho} = \begin{bmatrix} \tilde{\mathbf{x}} \\ \tilde{\mathbf{y}} \end{bmatrix} \tag{12}$$

it is clear that $\boldsymbol{\rho}_0 = 0$ and that the structure of $\boldsymbol{\Omega}$ can be used to describe situations where the sensor location errors are related to one another. Such a situation for a towed-array application was presented in [37].

3. Unstructured Errors

In the models presented above, the array perturbation is based on physical insight into the structure of the array, assuming that one error source is dominant. In many applications, however, both of the effects described above (along with others, such as quantization effects, interpolation errors, etc.) are present simultaneously. In such cases, a model based on physical insight may be impractical and cumbersome. A pragmatic remedy to this situation is to assume a simple unstructured model for the perturbed array response, such as

$$\mathbf{A}(\boldsymbol{\theta},\boldsymbol{\rho}) = \mathbf{A}(\boldsymbol{\theta}) + \tilde{\mathbf{A}} \tag{13}$$

where

$$\boldsymbol{\rho} = \begin{bmatrix} \mathrm{Re}\{\mathrm{vec}(\tilde{\mathbf{A}})\} \\ \mathrm{Im}\{\mathrm{vec}(\tilde{\mathbf{A}})\} \end{bmatrix} \tag{14}$$

and, again, $\boldsymbol{\rho}_0 = 0$. Let the columns of $\tilde{\mathbf{A}}$, denoted $\tilde{a}_i$, be modeled as zero-mean Gaussian random vectors with moments

$$\mathbf{E}[\tilde{a}_i\tilde{a}_k^*] = \nu_{ik}\mathbf{I} \quad i,\, k = 1,\, \ldots,\, d \tag{15}$$

$$\mathbf{E}[\tilde{a}_i\tilde{a}_k^T] = 0 \qquad i,\, k = 1,\, \ldots,\, d \tag{16}$$

The above model corresponds to an additive, circularly symmetric, complex array perturbation that is uncorrelated from sensor to sensor, but possibly θ-dependent. It is easy to verify that under these assumptions, the covariance of $\boldsymbol{\rho}$ is given by

$$\boldsymbol{\Omega} = \frac{1}{2}\begin{bmatrix} \mathrm{Re}\{\mathbf{Y}\}\otimes\mathbf{I} & -\mathrm{Im}\{\mathbf{Y})\otimes\mathbf{I} \\ \mathrm{Im}\{\mathbf{Y}\}\otimes\mathbf{I} & \mathrm{Re}\{\mathbf{Y}\}\otimes\mathbf{I} \end{bmatrix} \tag{17}$$

where the *ik*th element of the matrix $\mathbf{Y}$ is ν_{ik}. Perturbation models similar to (13)–(16) have been used by a number of authors, primarily in the analysis of adaptive beamforming algorithms [50–52].

Although not realistic for all types of array perturbations, the above model is useful for purposes of algorithm analysis and comparison, and is reasonable for situations involving experimentally calibrated arrays, where the sources of error are often due to quantization effects in collecting the calibration data, interpolation errors in using a calibration grid, etc. The model of (13)–(16) is also related to the more realistic case considered in (10)–(11) involving gain and phase perturbations. To see this, suppose $\mathbf{Y} = \sigma_a^2\mathbf{I}$, the nominal response of the *k*th sensor is $a_k(\theta) = e^{j\phi_k}$, and let $\tilde{g}_k$ and $\tilde{\phi}_k$ represent the corresponding gain and phase perturbations. The *k*th diagonal element of $\mathbf{G}$ in (10) is thus $\mathbf{G}_k = (1+\tilde{g}_k)\,e^{j\tilde{\phi}_k}$, and

$$\hat{a}_k(\theta) = (1+\tilde{g}_k)\,e^{j(\phi_k+\tilde{\phi}_k)} \simeq a_k(\theta) + \tilde{a}_k(\theta)$$

when $|\tilde{a}_k(\theta)| \ll |a_k(\theta)|$, where

$$\tilde{a}_k(\theta) = (\tilde{g}_k + j\tilde{\phi}_k)\,e^{j\phi_k}$$

If we let σ_g^2 and σ_ϕ^2 represent the variances of $\tilde{g}_k$ and $\tilde{\phi}_k$, respectively, and if we assume that the gain and phase perturbations are zero-mean and independent, then

$$E[\tilde{a}_k\tilde{a}_k^*] = \sigma_g^2 + \sigma_\phi^2$$
$$E[\tilde{a}_k^2] = (\sigma_g^2 - \sigma_\phi^2)\,e^{j2\phi}$$

The model of (13)–(16) results if the gain and phase errors are roughly "of the same order" (σ_g^2 and σ_ϕ^2 are equal), with $\sigma_a^2 = \sigma_g^2 + \sigma_\phi^2$. Thus, while the model of (10)–(11) assumes angle independent gain and phase errors, the unstructured model in (13)–(16) can be thought of as representing gain and phase errors that are uncorrelated but angle dependent.

B. Subspace Notation

In Section IV, the exact MAP estimator is presented along with a less computationally demanding approximation. The proposed approximate technique is a *subspace-based* method, in that it relies heavily on the properties of the eigendecomposition of the array covariance. Under the above assumptions, the covariance matrix of the array output takes the form

$$\mathbf{R} = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} E[\mathbf{x}(t)\mathbf{x}^*(t)] = \mathbf{A}(\boldsymbol{\theta}_0, \boldsymbol{\rho}_0)\mathbf{P}\mathbf{A}^*(\boldsymbol{\theta}_0, \boldsymbol{\rho}_0) + \sigma^2\mathbf{I} \tag{18}$$

Since the matrix $\mathbf{APA}^*$ has rank d by assumption (the arguments of $\mathbf{A}$ will frequently be suppressed for notational convenience), σ^2 is an eigenvalue of $\mathbf{R}$ with multiplicity $m-d$, and the corresponding eigenvectors are all orthogonal to $\mathbf{A}$. The eigendecomposition of $\mathbf{R}$ thus takes the form

$$\mathbf{R} = \sum_{i=1}^{m} \lambda_i \mathbf{e}_i \mathbf{e}_i^* = \mathbf{E}_s \boldsymbol{\Lambda}_s \mathbf{E}_s^* + \sigma^2 \mathbf{E}_n \mathbf{E}_n^* \tag{19}$$

where $\boldsymbol{\Lambda}_s$ is a diagonal matrix containing the d largest eigenvalues, and the columns of the $m \times d$ matrix $\mathbf{E}_s$ are the corresponding unit-norm eigenvectors. Similarly, the columns of $\mathbf{E}_n$ are the $m-d$ eigenvectors corresponding to σ^2. Since $\mathbf{E}_n$ is orthogonal to $\mathbf{A}$, it follows that the range space of $\mathbf{E}_s$ coincides with that of $\mathbf{A}$. This observation forms the basis for all subspace-based estimation techniques, starting with the development of the popular MUSIC algorithm [14,16].

Assuming orthonormal eigenvectors, the orthogonal projection onto the range space of $\mathbf{A}$ is denoted

$$\boldsymbol{\Pi} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{A}\mathbf{A}^\dagger = \mathbf{E}_s\mathbf{E}_s^* \tag{20}$$

and its orthogonal complement is

$$\boldsymbol{\Pi}^\perp = \mathbf{I} - \mathbf{A}\mathbf{A}^\dagger = \mathbf{E}_n\mathbf{E}_n^* \tag{21}$$

Here, the Moore–Penrose pseudo-inverse of $\mathbf{A}$ is denoted

$$\mathbf{A}^\dagger = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* \tag{22}$$

Under the stated assumptions, the eigendecomposition of $\mathbf{R}$ and the above projection matrices can be consistently estimated by performing an eigendecomposition of the sample covariance matrix

$$\hat{\mathbf{R}} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{x}(t)\mathbf{x}^*(t) = \hat{\mathbf{E}}_s \hat{\boldsymbol{\Lambda}}_s \hat{\mathbf{E}}_s^* + \hat{\mathbf{E}}_n \hat{\boldsymbol{\Lambda}}_n \hat{\mathbf{E}}_n^* \tag{23}$$

where the partitioning of the eigenelements is similar to (19).

III. RELEVANT ALGORITHMS

In this section, a description is given of several algorithms for DOA estimation and beamforming that are referred to later. Although the approaches described here have not been specifically developed to handle errors in the array response, an indication will be given of how they may be made more robust to such errors. Consequently, for these algorithms the array response will be denoted using $\mathbf{A}(\boldsymbol{\theta})$ rather than the more general expression $\mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho})$. In all cases, $\mathbf{A}(\boldsymbol{\theta})$ is to be taken to mean the nominal response $\mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho}_0)$.

A. DOA Estimation

As mentioned earlier, most well-known DOA estimation algorithms have been shown to be special cases of the so-called (signal) subspace fitting (SSF) framework first presented in [22]. All algorithms in the SSF class can be shown to be equivalent (either algebraically or asymptotically in N) to the following minimization problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} V_{SSF}(\boldsymbol{\theta}) = \arg\min_{\theta} \operatorname{Tr}(\boldsymbol{\Pi}^{\perp}(\boldsymbol{\theta})\hat{\mathbf{E}}_s \mathbf{W} \hat{\mathbf{E}}_s^*) \tag{24}$$

where the choice of the weighting matrix $\mathbf{W}$ determines the particular algorithm in question. The rationale behind the SSF formulation is that the range space of the ideal signal subspace matrix $\mathbf{E}_s$ is contained in that of the true $\mathbf{A}$. Making the projection of $\hat{\mathbf{E}}_s$ onto the orthogonal complement of span $\mathbf{A}(\theta)$ small (in an appropriate metric) should therefore lead to a good estimate of $\boldsymbol{\theta}$. An equivalent *noise* subspace fitting (NSF) formulation can also be made [53], in which algorithms are shown to be special cases of

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} V_{NSF}(\boldsymbol{\theta}) = \arg\min_{\theta} \operatorname{Tr}(\mathbf{A}^*(\boldsymbol{\theta})\hat{\mathbf{E}}_n \hat{\mathbf{E}}_n^* \mathbf{A}(\boldsymbol{\theta})\mathbf{U}) \tag{25}$$

for certain choices of the weighting $\mathbf{U}$. The argument behind this formulation is that the true $\mathbf{A}$ is orthogonal to $\mathbf{E}_n$. The most well-known subspace-based algorithm is obtained by choosing $\mathbf{U}$ to be diagonal. The multidimensional search involved in (25) then decouples into d one-dimensional minimizations of the form

$$\hat{\theta}_k = \arg\min_{\theta} \mathbf{a}^*(\theta)\hat{\mathbf{E}}_n \hat{\mathbf{E}}_n^* \mathbf{a}(\theta) \tag{26}$$

The locations of the d deepest (distinct) minima of the criterion function in (26) define the MUSIC estimates [14].

One of the key results of [22] and [53,54] was the derivation of optimal values for both $\mathbf{W}$ and $\mathbf{U}$. The term optimal as used here means that, for a certain choice of $\mathbf{W}$ and $\mathbf{U}$, the SSF and NSF methods are large sample realizations of the maximum likelihood approach, and hence will yield

DOA estimates of minimum variance. In particular, the optimal weightings were shown to be

$$\mathbf{W}_{\mathrm{OPT}} = (\hat{\mathbf{\Lambda}}_s - \hat{\sigma}^2\mathbf{I})^2\hat{\mathbf{\Lambda}}_s^{-1} \tag{27}$$

$$\mathbf{U}_{\mathrm{OPT}} = \mathbf{A}^{\dagger}(\hat{\boldsymbol{\theta}}_0)\hat{\mathbf{E}}_s\mathbf{W}_{\mathrm{OPT}}\hat{\mathbf{E}}_s^*\mathbf{A}^{\dagger *}(\hat{\boldsymbol{\theta}}_0), \tag{28}$$

where $\hat{\sigma}^2$ and $\hat{\boldsymbol{\theta}}_0$ are consistent estimates of the noise power and DOAs, respectively. An excellent side-by-side derivation of the optimality of both the SSF and NSF methods can be found in [31].

The optimal weight matrices given in (27)–(28) were derived assuming that only the finite sample effects of noise and not those due to array model errors are present (i.e., $\boldsymbol{\rho} = \boldsymbol{\rho}_0$). A different set of weights results if model errors are accounted for and finite sample effects ignored [28]. More recently, it has been shown that for the more realistic case where both sources of error are present, the optimal weighting becomes a combination of the two weights obtained for the limiting cases [29]. It should be noted that the optimality of the approaches in [28,29] has only been established for the simple error model of (13)–(16).

B. Signal Estimation

As used here, *signal estimation* refers to the problem of estimating the samples of the signal waveforms **S** using the received data **X** from the array. This is most often accomplished by forming a linear combination of the array outputs (a spatial filter), and is often referred to as beamforming (although technically this is somewhat of a misnomer). The coefficients used in forming the linear combination are typically called the beamformer *weights*. The weights that minimize the mean squared error (MSE) of the signal estimate are simply the coefficients of a spatial Wiener filter:

$$\mathbf{W}_{\mathrm{MSE}} = \arg\min_{W} \|\mathbf{W}^*\mathbf{X} - \mathbf{S}\|_F^2 \tag{29}$$

$$= \mathbf{R}^{-1}\mathbf{R}_{xs} = \mathbf{R}^{-1}\mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho})\mathbf{P}, \tag{30}$$

where

$$\mathbf{R}_{xs} = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\mathbf{x}(t)\mathbf{s}^*(t) = \mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho})\mathbf{P}$$

When the desired signal is uncorrelated with the interference, **P** is diagonal and the minimum MSE solution is just a scaled version of the so-called minimum variance distortionless response (MVDR) beamformer [55]:

$$\mathbf{W} = \frac{\mathbf{R}^{-1}a(\hat{\theta})}{a^*(\hat{\theta})\mathbf{R}^{-1}a(\hat{\theta})} \tag{31}$$

In the general case where the signal and interference are correlated, the optimal MSE weights depend on the signals themselves through $\mathbf{R}_{xs}$ or $\mathbf{P}$, and thus they cannot be used directly (i.e., without a training sequence, for example). As a result, many techniques have been devised to approximate the Weiner solution. In the approach of [56], it is assumed that the DOAs of the signals have been estimated, and the quantities $\mathbf{P}$ and $\mathbf{R}$ in (30) are replaced by their *structured* maximum likelihood (ML) estimates:

$$\hat{\mathbf{P}}_s = \mathbf{A}^{\dagger}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0)(\hat{\mathbf{R}} - \hat{\sigma}^2\mathbf{I})\mathbf{A}^{\dagger *}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0)$$
$$\hat{\mathbf{R}}_s = \mathbf{A}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0)\hat{\mathbf{P}}_s\mathbf{A}^*(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0) + \hat{\sigma}^2\mathbf{I}$$

where $\hat{\mathbf{R}}$ is a sample estimate of $\mathbf{R}$, and $\hat{\boldsymbol{\theta}}$ is an ML estimate of the DOAs. Since calibration errors were not addressed in [56], the nominal model $\mathbf{A}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0)$ was used in (30) to calculate the beamformer weights. Nevertheless, the method performs well when calibration errors are present, as recently demonstrated in [57].

On the other hand, the MVDR approach is well known to be hyper-sensitive to array perturbations, especially at high SNR. While *ad hoc* methods employing artificial noise injection have been used to combat this problem, other techniques based on *subspace corrected* (SC) weights have found success in experimental systems [45,47]. In these approaches, the $\mathbf{R}^{-1}$ term in (31) is replaced by $\mathbf{E}_s\boldsymbol{\Lambda}_s^{-1}\mathbf{E}_s^*$. This is equivalent to projecting $\mathbf{a}(\hat{\theta})$ onto the signal subspace prior to forming the MVDR weights.

In [58], the improvement that results from using the method of [56] with $\mathbf{A}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}})$ rather than $\mathbf{A}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0)$ was investigated, where $\hat{\boldsymbol{\rho}}$ is obtained from the MAP estimator described in the next section. The results of this analysis will be briefly described below, and the resulting algorithm will be referred to as the MAP beamformer. An interesting connection can be made between the MAP beamformer and the SC-MVDR method. In particular, it will be shown below that for uncorrelated signals and simple unstructured array errors of the form (13)–(16), the SC-MVDR and MAP weights have a very similar form.

IV. MAP PARAMETER ESTIMATION

In this section, the exact MAP formulation of the problem is presented, along with a simplified, but asymptotically equivalent, approximation. Unlike the exact MAP approach, which in general requires nonlinear optimization over both the DOAs $\boldsymbol{\theta}$ and the nuisance parameters $\boldsymbol{\rho}$, the approximate MAP method is separable in $\boldsymbol{\rho}$. The resulting criterion requires a search only over the d DOA parameters in $\boldsymbol{\theta}$, which is a considerable simplification since the number of elements in $\boldsymbol{\rho}$ can be quite large. The MAP estimate of $\boldsymbol{\rho}$ is calculated directly given $\hat{\boldsymbol{\theta}}$, and can be used to perform an on-line calibration of the array.

A. Exact MAP Estimation

As mentioned earlier, the derivation of the MAP estimator given below assumes that the signals, noise, and array perturbation parameters are all Gaussian. The signals and noise are zero-mean with covariances $\mathbf{P}$ and $\sigma^2\mathbf{I}$, respectively, where $\mathbf{P}$ and σ^2 are unknown deterministic parameters (i.e., parameters with a noninformative prior distribution) that must be estimated. On the other hand, the mean $\boldsymbol{\rho}_0$ and covariance $\boldsymbol{\Omega}$ of $\boldsymbol{\rho}$ are assumed to be known. Besides $\mathbf{P}$ and σ^2, the other unknown parameters to be estimated are of course $\boldsymbol{\theta}$ (deterministic) and $\boldsymbol{\rho}$ (random).

For the moment, let $\boldsymbol{\eta} = \{\boldsymbol{\theta}, \mathbf{P}, \sigma^2\}$ represent the deterministic parameters to be estimated. As its name implies, the MAP estimator maximizes the probability density of the desired parameters given the received data:

$$\{\boldsymbol{\eta}, \boldsymbol{\rho}\}_{\text{MAP}} = \arg\max_{\eta,\rho}\, p_{\eta,\rho}(\boldsymbol{\eta}, \boldsymbol{\rho} | \mathbf{X}) \tag{32}$$

Using Bayes' rule with $\boldsymbol{\rho}$ and eliminating the prior on $\mathbf{X}$ yields the equivalent formulation

$$\{\boldsymbol{\eta}, \boldsymbol{\rho}\}_{\text{MAP}} = \arg\max_{\eta,\rho}\, p_{\eta,\rho}(\mathbf{X} | \boldsymbol{\eta}, \boldsymbol{\rho}) p_\rho(\boldsymbol{\rho}) \tag{33}$$

Since the densities in (33) are both Gaussian, it is simpler to work with their negative logarithms, in which case the MAP estimates are obtained by the following minimization problem:

$$\{\boldsymbol{\eta}, \boldsymbol{\rho}\}_{\text{MAP}} = \arg\min_{\eta,\rho}\, \{-\ln p_{\eta,\rho}(\mathbf{X} | \boldsymbol{\eta}, \boldsymbol{\rho}) - \ln p_\rho(\boldsymbol{\rho})\} \tag{34}$$

The $-\ln p_{\eta,\rho}(\mathbf{X} | \boldsymbol{\eta}, \boldsymbol{\rho})$ term in (34) is the standard negative log-likelihood function $V_{\text{ML}}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2)$, which is given by [59]

$$\begin{aligned} V_{\text{ML}}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2) = N(\log|\mathbf{R}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2)| \\ + \operatorname{Tr}\{\mathbf{R}^{-1}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2)\hat{\mathbf{R}}\}) + \text{const.} \end{aligned} \tag{35}$$

where $|\cdot|$ denotes the determinant. The other term involving the prior of $\boldsymbol{\rho}$ is easily shown to be

$$-\ln p_\rho(\boldsymbol{\rho}) = \tfrac{1}{2}(\boldsymbol{\rho} - \boldsymbol{\rho}_0)^T \boldsymbol{\Omega}^{-1}(\boldsymbol{\rho} - \boldsymbol{\rho}_0) \tag{36}$$

Consequently, the joint MAP estimate of $\boldsymbol{\theta}, \mathbf{P}, \sigma^2$ and $\boldsymbol{\rho}$ may be expressed as

$$\{\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}}, \hat{\mathbf{P}}, \hat{\sigma}^2\}_{\mathrm{MAP}} = \arg \min_{\theta,\rho,P,\sigma^2} V_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2) \tag{37}$$

$$V_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2) = V_{\mathrm{ML}}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2) + \tfrac{1}{2}(\boldsymbol{\rho} - \boldsymbol{\rho}_0)^T \boldsymbol{\Omega}^{-1}(\boldsymbol{\rho} - \boldsymbol{\rho}_0) \tag{38}$$

The ML criterion function is known to be separable in $\mathbf{P}$ and σ^2. For fixed $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho})$, the minimizing signal covariance matrix and noise power are [59]

$$\hat{\mathbf{P}} = \mathbf{A}^{\dagger}(\hat{\mathbf{R}} - \hat{\sigma}^2 \mathbf{I})\mathbf{A}^{\dagger *} \tag{39}$$

$$\hat{\sigma}^2 = \frac{1}{m-d} \operatorname{Tr}\{\boldsymbol{\Pi}^{\perp} \hat{\mathbf{R}}\} \tag{40}$$

Substituting (39)–(40) into (35) leads to [31]

$$V_{\mathrm{ML}}(\boldsymbol{\theta}, \boldsymbol{\rho}) = N \log |\mathbf{A}\hat{\mathbf{P}}\mathbf{A}^* + \hat{\sigma}^2 \mathbf{I}| + \text{const.} \tag{41}$$

Clearly, $V_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\rho}, \mathbf{P}, \sigma^2)$ is also separable in $\mathbf{P}$ and σ^2, and ignoring constant terms the concentrated MAP criterion function is

$$V_{\mathrm{MAP}}(\boldsymbol{\theta}, \boldsymbol{\rho}) = V_{\mathrm{ML}}(\boldsymbol{\theta}, \boldsymbol{\rho}) + \tfrac{1}{2}(\boldsymbol{\rho} - \boldsymbol{\rho}_0)^T \boldsymbol{\Omega}^{-1}(\boldsymbol{\rho} - \boldsymbol{\rho}_0) \tag{42}$$

This can be interpreted as a *regularized* ML criterion. That is, the effect of the prior distribution is to force $\hat{\boldsymbol{\rho}}_{\mathrm{MAP}}$ to be close to the nominal value, $\boldsymbol{\rho}_0$. If the perturbation parameters are identifiable, this effect is diminished as the number of snapshots, N, increases. Thus, the MAP estimate has the same asymptotic properties as the ML estimate (i.e., the pure auto-calibration technique). However, in many applications of interest, $\boldsymbol{\rho}$ cannot be consistently estimated along with the signal parameters. In such cases the prior distribution has a crucial influence on the asymptotic properties of the estimates of both $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$.

Even in the concentrated form of (42), the MAP criterion is still quite unwieldy due to the complicated dependence of V_{ML} on $\boldsymbol{\rho}$, and the fact that $\boldsymbol{\rho}$ can have a large number of elements ($2md$ for the model of (13)–(16)). In the next section, we show how $V_{\mathrm{ML}}(\boldsymbol{\theta}, \boldsymbol{\rho})$ can be replaced with an asymptotic approximation that is quadratic in $\boldsymbol{\rho}$. The resulting approximate MAP criterion will then also be quadratic (and hence separable) in $\boldsymbol{\rho}$, and can be further concentrated to depend only on the d elements of $\boldsymbol{\theta}$.

B. The MAP-NSF Method

As pointed out earlier, when the signal covariance has full rank and no array errors are present, the noise subspace fitting (NSF) cost function

$$V_{\text{NSF}} = N\,\text{Tr}\{\mathbf{A}^*\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*\mathbf{A}\mathbf{U}_{\text{OPT}}\} \tag{43}$$

is known to be asymptotically equivalent to the ML criterion when $\mathbf{U}_{\text{OPT}}$ is chosen as in (28). In the present case, the term "asymptotic equivalence" is to be interpreted as

$$\left.\frac{\partial V_{\text{MAP}}}{\partial \alpha}\right|_{\theta_0,\rho_0} = \left.\frac{\partial V_{\text{NSF}}}{\partial \alpha}\right|_{\theta_0,\rho_0} + o_p(1/\sqrt{N}) \tag{44}$$

where α refers to any component of $\boldsymbol{\theta}$ or $\boldsymbol{\rho}$, and the symbol $o_p(1/\sqrt{N})$ represents a term that tends to zero faster than $1/\sqrt{N}$ in probability. The extension of this result to the case where model errors are also present is immediate since the proof only depends on the fact that $\hat{\mathbf{R}} = \mathbf{R} + O_p(1/\sqrt{N})$. By standard first-order arguments, this implies that the MAP estimate is asymptotically (for large N) equivalent to the minimizing arguments of the following criterion function

$$V_{\text{NSF}}(\boldsymbol{\theta},\boldsymbol{\rho}) + \tfrac{1}{2}(\boldsymbol{\rho}-\boldsymbol{\rho}_0)^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\rho}-\boldsymbol{\rho}_0) \tag{45}$$

This criterion depends on its parameters in a simpler way than the exact MAP criterion (42). However, it still requires a nonlinear minimization over both $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$.

A further simplification of the criterion is possible that enables separation with respect to $\boldsymbol{\rho}$. Recall the following formulas for the vec$(\cdot)$ operator (vectorization of a matrix by stacking its columns) and the Kronecker product $\otimes$ (see [60,61]):

$$\text{Tr}\{\mathbf{ABCD}\} = \text{vec}(\mathbf{D}^T)^T(\mathbf{C}^T\otimes\mathbf{A})\,\text{vec}(\mathbf{B}) \tag{46}$$

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}^T\otimes\mathbf{A})\,\text{vec}(\mathbf{B}) \tag{47}$$

$$(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = ((\mathbf{AC})\otimes(\mathbf{BD})) \tag{48}$$

$$(\mathbf{A}\otimes\mathbf{B})^T = \mathbf{A}^T\otimes\mathbf{B}^T \tag{49}$$

Using (46), the NSF criterion can be rewritten as

$$N\,\text{Tr}\{\mathbf{A}^*\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*\mathbf{A}\mathbf{U}_{\text{OPT}}\} = N\,\mathbf{a}^*\hat{\mathbf{M}}\mathbf{a} \tag{50}$$

where

$$\mathbf{a} = \text{vec}(\mathbf{A}) \tag{51}$$

$$\hat{\mathbf{M}} = \mathbf{U}_{\text{OPT}}^T\otimes(\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*) \tag{52}$$

If the perturbation to the array is small enough so that second-order effects may be ignored, the vectorized steering matrix may be approximated locally around $\boldsymbol{\rho}_0$ as

$$\mathbf{a} = \mathbf{a}(\boldsymbol{\theta}, \boldsymbol{\rho}) \approx \mathbf{a}_0 + \mathbf{D}_{\rho}\bar{\boldsymbol{\rho}} \tag{53}$$

where

$$\mathbf{a}_0 = \mathbf{a}(\boldsymbol{\theta}, \boldsymbol{\rho}_0) \tag{54}$$

$$\mathbf{D}_{\rho} = \left[\frac{\partial \mathbf{a}(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \rho_1}, \ldots, \frac{\partial \mathbf{a}(\boldsymbol{\theta}, \boldsymbol{\rho})}{\partial \rho_n}\right]\Bigg|_{\boldsymbol{\theta}, \boldsymbol{\rho}_0} \tag{55}$$

$$\bar{\boldsymbol{\rho}} = \boldsymbol{\rho} - \boldsymbol{\rho}_0 \tag{56}$$

Note that, when evaluated at $\boldsymbol{\rho}_0$, the derivative of $\mathbf{a}$ with respect to $\boldsymbol{\theta}$ or $\boldsymbol{\rho}$ is identical to that of $\mathbf{a}_0 + \mathbf{D}_{\rho}\bar{\boldsymbol{\rho}}$. It follows that the minimizing arguments of (50) are asymptotically identical to the estimates obtained by minimizing the following approximate MAP-NSF criterion with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\rho}$:

$$(\mathbf{a}_0 + \mathbf{D}_{\rho}\bar{\boldsymbol{\rho}})^* \hat{\mathbf{M}}(\mathbf{a}_0 + \mathbf{D}_{\rho}\bar{\boldsymbol{\rho}}) + \tfrac{1}{2}\bar{\boldsymbol{\rho}}^T \bar{\boldsymbol{\Omega}}^{-1}\bar{\boldsymbol{\rho}} \tag{57}$$

where we have normalized by N and used (9).

Since the criterion function in (57) is quadratic in $\bar{\boldsymbol{\rho}}$, we easily obtain the minimum with respect to $\bar{\boldsymbol{\rho}}$ (for fixed $\boldsymbol{\theta}$) as

$$\hat{\boldsymbol{\rho}}_{\text{MAP-NSF}} = \boldsymbol{\rho}_0 - \boldsymbol{\Gamma}^{-1}\mathbf{f} \tag{58}$$

where

$$\boldsymbol{\Gamma} = \operatorname{Re}\{\mathbf{D}_{\rho}^* \hat{\mathbf{M}} \mathbf{D}_{\rho} + \tfrac{1}{2}\bar{\boldsymbol{\Omega}}^{-1}\} \tag{59}$$

$$\mathbf{f} = \operatorname{Re}\{\mathbf{D}_{\rho}^* \hat{\mathbf{M}} \mathbf{a}_0\} \tag{60}$$

Substituting (58) into (57) leads to the following separated criterion function

$$\mathbf{a}_0^* \hat{\mathbf{M}} \mathbf{a}_0 - \mathbf{f}^T \boldsymbol{\Gamma}^{-1}\mathbf{f} \tag{61}$$

Note that $\boldsymbol{\Gamma}$ and $\mathbf{f}$ depend on $\boldsymbol{\theta}$ through $\mathbf{D}_{\rho}$, and in principle $\mathbf{M}$ also depends on $\boldsymbol{\theta}$ through $\mathbf{U}_{\text{OPT}}$. However, it will be assumed that a consistent estimate of $\boldsymbol{\theta}$ is available to form the estimates $\hat{\mathbf{M}}$, $\hat{\mathbf{f}}$, and $\hat{\boldsymbol{\Gamma}}$. Under the stated assumptions, such an estimate can be obtained, for instance, by letting $\boldsymbol{\rho} = \boldsymbol{\rho}_0$ and $\mathbf{U}$ be diagonal in (25), which leads to the MUSIC algorithm (26). As shown in [37], the approximations made in forming $\hat{\boldsymbol{\Gamma}}$, $\hat{\mathbf{f}}$, and $\hat{\mathbf{M}}$ do not change the asymptotic properties of the final estimate.

The definitions of the quantities in the MAP-NSF cost function are repeated below for easy reference, followed by a summary of the proposed algorithm:

$$\mathbf{a}_0 = \operatorname{vec}(\mathbf{A}(\boldsymbol{\theta}, \boldsymbol{\rho}_0)) \tag{62}$$

$$\hat{\mathbf{M}} = \hat{\sigma}^{-2}(\hat{\mathbf{A}}^{\dagger}\hat{\mathbf{E}}_s(\hat{\mathbf{\Lambda}}_s - \hat{\sigma}^2\mathbf{I})^2\hat{\mathbf{\Lambda}}_s^{-1}\hat{\mathbf{E}}_s^*\hat{\mathbf{A}}^{\dagger *})^T \otimes (\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*) \tag{63}$$

$$\hat{\sigma}^2 = \frac{1}{m-d}\mathrm{Tr}\{(\mathbf{I} - \hat{\mathbf{A}}\hat{\mathbf{A}}^{\dagger})\hat{\mathbf{R}}\} \tag{64}$$

$$\hat{\mathbf{\Gamma}} = \mathrm{Re}\{\hat{\mathbf{D}}_{\boldsymbol{\rho}}^*\hat{\mathbf{M}}\hat{\mathbf{D}}_{\boldsymbol{\rho}} + \tfrac{1}{2}\bar{\mathbf{\Omega}}^{-1}\} \tag{65}$$

$$\hat{\mathbf{f}} = \mathrm{Re}\{\hat{\mathbf{D}}_{\boldsymbol{\rho}}^*\hat{\mathbf{M}}\mathbf{a}_0\} \tag{66}$$

The MAP-NSF Algorithm. Given the sample covariance $\hat{\mathbf{R}}$ and an initial estimate $\hat{\boldsymbol{\theta}}$ of the DOAs:

1. Compute the eigendecompostion $\hat{\mathbf{R}} = \hat{\mathbf{E}}_s\hat{\mathbf{\Lambda}}_s\hat{\mathbf{E}}_s^* + \hat{\mathbf{E}}_n\hat{\mathbf{\Lambda}}_n\hat{\mathbf{E}}_n^*$.
2. Compute the quantities (63)–(65).
3. Using $\hat{\boldsymbol{\theta}}$ as an initial guess, use a numerical method to solve the following optimization problem

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP-NSF}} = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) \tag{67}$$

$$V(\boldsymbol{\theta}) = \mathbf{a}_0^*\hat{\mathbf{M}}\mathbf{a}_0 - \hat{\mathbf{f}}^T\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{f}} \tag{68}$$

Let us finally point out that the MAP-NSF algorithm yields statistically efficient estimates of the signal and perturbation parameters, provided that no perfectly coherent signals are present. In [37] it was shown that the estimated parameters asymptotically achieve the Cramér–Rao lower bound for mixed deterministic and stochastic parameters. More details along with an expression for the asymptotic bound are provided in this reference. Although the MAP-NSF approach loses claim to asymptotic optimality only when the signals are 100% correlated, a performance degradation is evident at lower correlation levels, as will be demonstrated by several simulations in Section V. In such cases, relatively large values of N are required for MAP-NSF to achieve the CRB, and implementing the algorithm with smaller N can lead to a loss of signal resolution. However, as will also be demonstrated by the simulations, there is a wide variety of scenarios for which the MAP-NSF approach achieves the CRB for relatively small values of N.

C. Some Special Cases

The specific form taken on by the MAP-NSF criterion function depends of course on the parametrization chosen for $\boldsymbol{\rho}$. In the discussion that follows, the MAP-NSF criterion is examined for two of the three basic error models described in Section II.A. The unstructured model is studied first since the resulting equations are simplest, and there are some interesting connections to be made with some of the algorithms described in Section III. The case of independent gain and phase errors is also addressed.

1. Unstructured Errors

To begin with, note that for the model described by (13)–(16), $\boldsymbol{\rho}_0 = 0$ and $\hat{\mathbf{D}}_{\rho} = [\mathbf{I} \quad j\mathbf{I}]$, where $\mathbf{I}$ is $md \times md$. Thus, $\hat{\boldsymbol{\rho}} = -\boldsymbol{\Gamma}^{-1}\mathbf{f}$, and

$$\boldsymbol{\Gamma} = \begin{bmatrix} \mathrm{Re}(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}) & -\mathrm{Im}(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}) \\ \mathrm{Im}(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}) & \mathrm{Re}(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}) \end{bmatrix}$$

$$\mathbf{f} = \begin{bmatrix} \mathrm{Re}\{\hat{\mathbf{M}}\mathbf{a}_0\} \\ \mathrm{Im}\{\hat{\mathbf{M}}\mathbf{a}_0\} \end{bmatrix}$$

Using the fact that, for any invertible matrix $\mathbf{Z}$,

$$\begin{bmatrix} \mathrm{Re}\{\mathbf{Z}\} & -\mathrm{Im}\{\mathbf{Z}\} \\ \mathrm{Im}\{\mathbf{Z}\} & \mathrm{Re}\{\mathbf{Z}\} \end{bmatrix}^{-1} = \begin{bmatrix} \mathrm{Re}\{\mathbf{Z}^{-1}\} & -\mathrm{Im}\{\mathbf{Z}^{-1}\} \\ \mathrm{Im}\{\mathbf{Z}^{-1}\} & \mathrm{Re}\{\mathbf{Z}^{-1}\} \end{bmatrix} \tag{69}$$

it is easy to show that

$$\hat{\boldsymbol{\rho}} = -\begin{bmatrix} \mathrm{Re}\left\{\left(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}\right)^{-1}\hat{\mathbf{M}}\mathbf{a}_0\right\} \\ \mathrm{Im}\left\{\left(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}\right)^{-1}\hat{\mathbf{M}}\mathbf{a}_0\right\} \end{bmatrix} \tag{70}$$

A further simplification of (70) is possible that is quite revealing. Using the definition of $\hat{\mathbf{M}}$ in (63), note that

$$\left(\hat{\mathbf{M}} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes\mathbf{I}\right)^{-1} = \left((\mathbf{U}_{\mathrm{OPT}}^{T} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1})\otimes(\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*) + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\otimes(\hat{\mathbf{E}}_s\hat{\mathbf{E}}_s^*)\right)^{-1} \tag{71}$$

$$= \left(\mathbf{U}_{\mathrm{OPT}}^{T} + \frac{1}{N}\boldsymbol{\Upsilon}^{-1}\right)^{-1}\otimes(\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*) + N\boldsymbol{\Upsilon}\otimes(\hat{\mathbf{E}}_s\hat{\mathbf{E}}_s^*) \tag{72}$$

Multiplying the last equation above on the right by $\hat{\mathbf{M}}\mathbf{a}_0$ and simplifying then yields

$$\hat{\boldsymbol{\rho}} = -\begin{bmatrix} \mathrm{Re}\left\{\left[\left(\mathbf{I} + \frac{1}{N}(\boldsymbol{\Upsilon}\mathbf{U}_{\mathrm{OPT}}^{T})^{-1}\right)^{-1}\otimes(\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*)\right]\mathbf{a}_0\right\} \\ \mathrm{Im}\left\{\left[\left(\mathbf{I} + \frac{1}{N}(\boldsymbol{\Upsilon}\mathbf{U}_{\mathrm{OPT}}^{T})^{-1}\right)^{-1}\otimes(\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*)\right]\mathbf{a}_0\right\} \end{bmatrix} \tag{73}$$

Finally, using (13)–(14) and the properties of the Kronecker product, the MAP estimate of the array response becomes

$$\mathbf{A}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}}) = \mathbf{A}(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*\mathbf{A}(\hat{\boldsymbol{\theta}})\left(\mathbf{I} + \frac{1}{N}(\mathbf{\Upsilon}\mathbf{U}_{\mathrm{OPT}}^T)^{-1}\right)^{-1} \qquad (74)$$

The key point of interest is that, if $\mathbf{\Upsilon}^{-1}/N \rightarrow 0$, then the MAP estimate of the array response converges to a subspace corrected version of the nominal response:

$$\lim_{\mathbf{\Upsilon}^{-1}/N\rightarrow 0} \mathbf{A}(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\rho}}) = \hat{\mathbf{E}}_s\hat{\mathbf{E}}_s^*\mathbf{A}(\hat{\boldsymbol{\theta}})$$

Furthermore, if the estimated MAP array response is used in (31), the MVDR beamformer (31) will converge to the SC-MVDR approach. The condition $\mathbf{\Upsilon}^{-1}/N \rightarrow 0$ occurs either with a large data sample, or when the array perturbation is large. In either case, the information provided by the prior distribution of $\boldsymbol{\rho}$ is of little value, and is essentially ignored by the MAP criterion. This observation provides some theoretical justification for the SC-MVDR technique, which previously had been derived using *ad hoc* (but well-motivated) reasoning. However, in cases where the prior cannot be neglected, using SC response vectors for beamforming will not be optimal and significant degradation can result. This is seen in the simulation examples described in the next section.

In [37], it was shown that, for the unstructured array error model of (13)–(14), the concentrated MAP-NSF criterion

$$V(\boldsymbol{\theta}) = \mathbf{a}_0^*\hat{\mathbf{M}}\mathbf{a}_0 - \hat{\mathbf{f}}^T\hat{\mathbf{\Gamma}}^{-1}\hat{\mathbf{f}}$$

is asymptotically equivalent to the SSF algorithm

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} V_{\mathrm{SSF}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \mathrm{Tr}(\mathbf{\Pi}^{\perp}(\boldsymbol{\theta})\hat{\mathbf{E}}_s\mathbf{W}\hat{\mathbf{E}}_s^*) \qquad (75)$$

for a particular choice of $\mathbf{W}$. The optimal weighting was shown to be

$$\mathbf{W}_{\mathrm{OSF}} = \left(\hat{\mathbf{E}}_s^*\mathbf{A}(\hat{\boldsymbol{\theta}})^{\dagger *}\mathbf{\Upsilon}^T\mathbf{A}(\hat{\boldsymbol{\theta}})^{\dagger}\hat{\mathbf{E}}_s + \frac{\hat{\sigma}^2}{N}(\hat{\mathbf{\Lambda}}_s - \hat{\sigma}^2\mathbf{I})^{-2}\hat{\mathbf{\Lambda}}_s\right)^{-1} \qquad (76)$$

which is precisely the SSF weighting derived in [29] for the combined case where both the unstructured array errors and the finite sample effects of the noise are taken into account. For this particular model, the SSF solution is preferred over the MAP-NSF criterion since SSF can handle coherent emitters (rank deficient $\mathbf{P}$) and has been observed to have better numerical properties.

2. Gain and Phase Errors

The derivation of the MAP estimate of $\boldsymbol{\rho}$ and hence $\mathbf{g}$ for the model in (10)–(11) is straightforward but somewhat cumbersome, and thus will not be presented here. However, assuming $\boldsymbol{\Omega} = (\sigma_a^2/2)\mathbf{I}$, the result is quite simple and is given by

$$\hat{\mathbf{g}} = (\mathbf{I} + \sigma_a^2 N \hat{\mathbf{Z}}(\hat{\boldsymbol{\theta}}))^{-1}\mathbf{e} \tag{77}$$

where

$$\hat{\mathbf{Z}}(\boldsymbol{\theta}) = \left[\sum_{i,k=1}^{d} u_{ki}\bar{a}(\theta_i)a^T(\theta_k)\right] \odot (\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*) \tag{78}$$

u_{ki} is the kith element of $\mathbf{U}_{\mathrm{OPT}}$, $\overline{(\cdot)}$ denotes conjugation, and $\odot$ an element-wise (Hadamard) product. Note that for very small gain/phase errors where $\sigma_a \to 0$, $\hat{\mathbf{g}} \to \mathbf{e}$, and hence $\hat{\mathbf{G}} \to \mathbf{I}$ as expected.

To find a simpler expression for the concentrated MAP-NSF cost function, note that the NSF criterion may be written as

$$\mathrm{Tr}\{\mathbf{A}^*(\boldsymbol{\theta})\hat{\mathbf{E}}_n\hat{\mathbf{E}}_n^*\mathbf{A}(\boldsymbol{\theta})\mathbf{U}_{\mathrm{OPT}}\} = \mathbf{e}^T\mathbf{Z}(\boldsymbol{\theta})\mathbf{e} \tag{79}$$

Some simple algebra then yields

$$\hat{\boldsymbol{\theta}}_{\text{MAP-NSF}} = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} \mathbf{e}^T(\mathbf{Z}^{-1}(\boldsymbol{\theta}) + \sigma_a^2 N\mathbf{I})^{-1}\mathbf{e} \tag{80}$$

As $\sigma_a \to 0$, the standard NSF criterion of (25) is clearly retrieved. Although the large N behavior of (80) seems suspect, recall that σ_a^2 is assumed to be $O(1/N)$, and hence $\sigma_a^2 N = O(1)$. Again, as mentioned earlier, if finite sample effects are negligible compared to the array errors (as they would be for fixed σ_a^2 if $N \to \infty$), a different estimator should be used.

V. SIMULATION RESULTS

In this section, the performance of the MAP-NSF approach is studied by means of a number of simulation examples. Additional simulation results regarding parameter estimation accuracy can be found in [29,37]. The examples are divided into two parts, one involving DOA estimation performance (Section V.A), and the other focusing on beamforming results (Section V.B). All of the examples in this section assume a nominally unit-gain uniform linear array with half-wavelength interelement spacing.

A. DOA Estimation Results

The advantages provided by the robust MAP-based techniques developed in this work are most evident in difficult situations involving weak signals in the presence of strong interferers. In the three examples below, the output of a 10 element array was simulated using both a 0 dB SNR signal arriving from broadside and a 20 dB interferer. For the first two cases, the DOA of the interferer was 7°, and the array response was perturbed by an unstructured calibration error in the form of equation (13)–(17) with $\Upsilon = \sigma_a^2 \mathbf{I}$ and $\sigma_a = 0.05$. Figure 1 shows the RMS error of several DOA estimation algorithms based on 1000 trials for various sample sizes N, assuming the two signals have a correlation coefficient of -0.9. Figure 2 shows the performance of the same algorithms, this time with N fixed at 100 and the correlation varying from 0 to near 100% (the correlation phase was taken to be 180° relative to the first sensor in all of these simulations).

Since unstructured array errors were assumed in this example, the MAP-NSF algorithm is equivalent to the SSF approach of (75)–(76), and it was the SSF implementation that was used to obtain the results labeled MAP-SSF in the two figures. The WSF label refers to the SSF criterion implemented with the weighting matrix of (27) which was derived by only

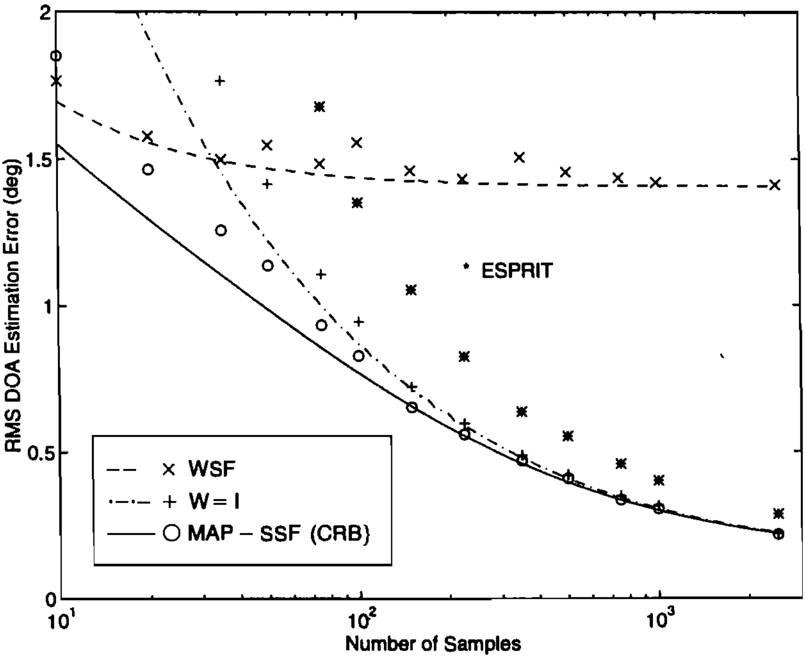


Figure 1 A comparison of DOA estimation performance, unstructured calibration errors, various sample sizes.

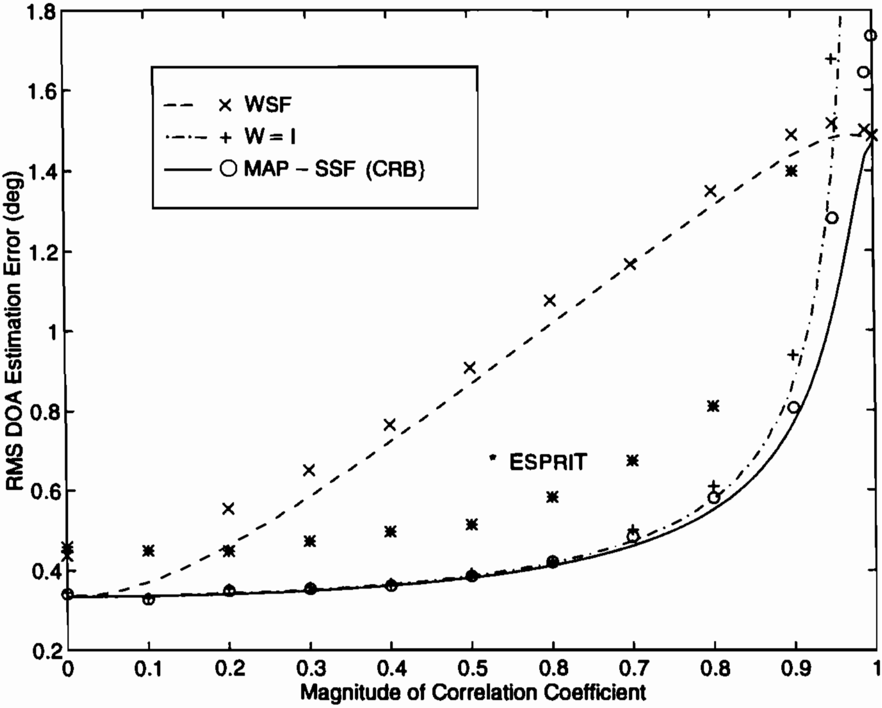


Figure 2 A comparison of DOA estimation performance, unstructured calibration errors, variable signal correlation.

taking the finite sample effects of noise into account. The results for no subspace weighting ($\mathbf{W} = \mathbf{I}$) and for the ESPRIT algorithm [21] are shown for comparison purposes. The continuous lines shown in the figures correspond to the (asymptotic) theoretical predictions of algorithm performance derived in [28,29,37], while the symbols ($\times, \circ, +, *$) represent the simulation results. Note that the predicted performance of the MAP-SSF approach corresponds with the CRB, shown as a solid line in the figures. The value of N required for the MAP-SSF algorithm to achieve the CRB depends on the difficulty of the scenario; for the case where the correlation is 90%, it appears that $N \geq 100$ is required. Larger levels of correlation would in turn require larger values of N. Except for the few cases where MAP-SSF has not yet reached its asymptotic performance limit, the algorithm significantly outperforms the other methods studied.

The results of a situation similar to that of Fig. 1 are plotted in Fig.

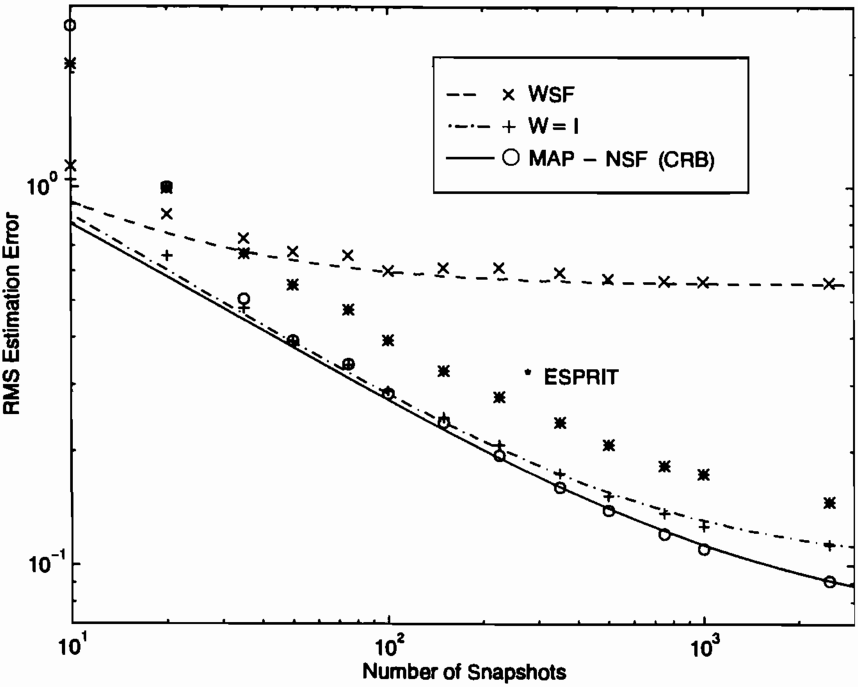


Figure 3 A comparison of DOA estimation performance, gain/phase calibration errors, various sample sizes.

3, the main difference being that the gain/phase perturbation model of (10)–(11) was used instead, also with $\sigma_a = 0.05$. In addition, the DOA of the interferer was chosen to be 10°, and the correlation coefficient was reduced to -0.6. A slightly easier situation was considered in this case to ensure that the MAP-NSF method would be able to resolve the two sources. As mentioned earlier, one of the disadvantages of NSF-based approaches is their loss of consistency for highly correlated signals. When the signals are closely spaced in angle, this performance degradation manifests itself in a loss of resolution ability. In fact, the bulk of the RMS error for MAP-NSF shown in Fig. 3 for $N = 10$ and $N = 20$ is due to the source at broadside not being resolved (i.e., both estimates occasionally converged to values near 10°). The SSF implementation of the MAP algorithm does not suffer from this deficiency, but has only been derived for the very simple error model of (13)–(17). The development of SSF-based MAP implementations for other error models is the subject of on-going research. In any case, when the MAP-NSF approach is able to resolve the two sources, its performance essentially achieves the CRB.

B. Beamforming Results

The first example involves a nominally unit-gain uniform linear array perturbed by an unstructured calibration error in the form of equations (13)–(17) with $\Upsilon = \sigma_a^2 \mathbf{I}$ and $\sigma_a = 0.2$. The array receives 100 samples of two 20 dB SNR uncorrelated Gaussian signals with arrival angles of 5° and 15°. Using DOA estimates from the optimal MAP estimator, the relative interference rejection capability of the MVDR, SC-MVDR, and MAP beamformers was calculated for various array sizes. The results are plotted in Fig. 4 based on 500 independent trials. The plot shows the gain of the beamformer weights for the 5° source in the direction of the 15° interferer (normalized for a unit gain response at 5°). The subspace correction

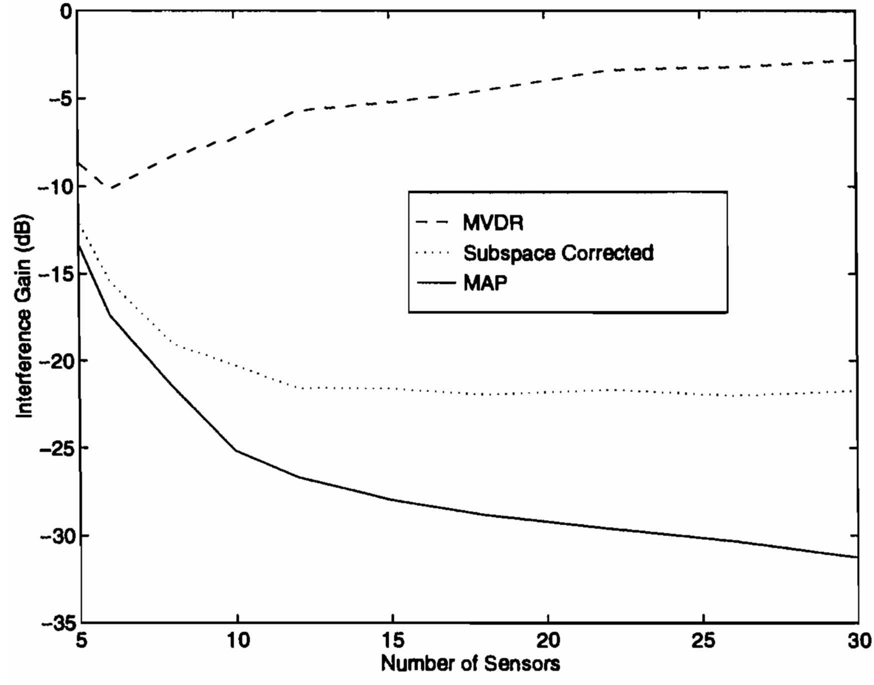


Figure 4 A comparison of beamformer performance, unstructured calibration errors.

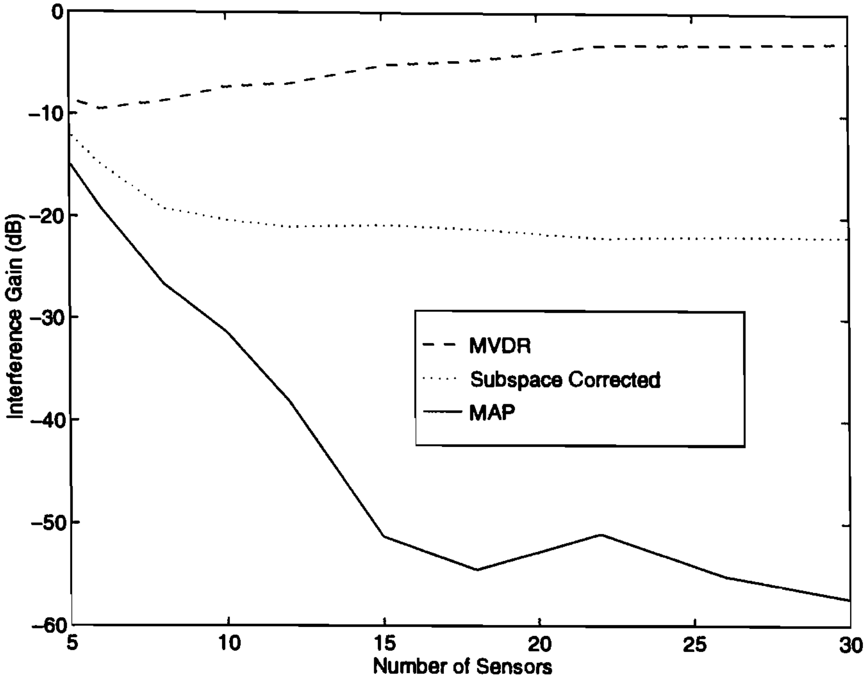


Figure 5 A comparison of beamformer performance, gain-phase calibration errors.

eliminates the signal cancelation effect of the MVDR approach, but the MAP beamformer provides a significant advantage, especially for larger arrays. The above simulation was repeated assuming receiver gain/phase errors described by (10)–(11), also with $\sigma_a = 0.2$, and the results are plotted in Fig. 5. The improvement for this type of calibration error is even more dramatic.

When the signals arriving at the array are highly correlated, interference rejection is no longer an appropriate performance criterion. In such cases, an optimal beamformer will attempt to combine correlated arrivals with the desired signal to improve the quality of the resulting estimate, as measured using (for example) mean-squared error. To examine beamformer performance for the case of correlated signals, a two-ray multipath channel was simulated for various relative delays. A mis-calibrated 5-element linear array was assumed to receive a random QPSK signal from −6°, as well as a slightly delayed copy of the signal from 6°. Both arrivals had an SNR of 0 dB, and the array was again perturbed according to

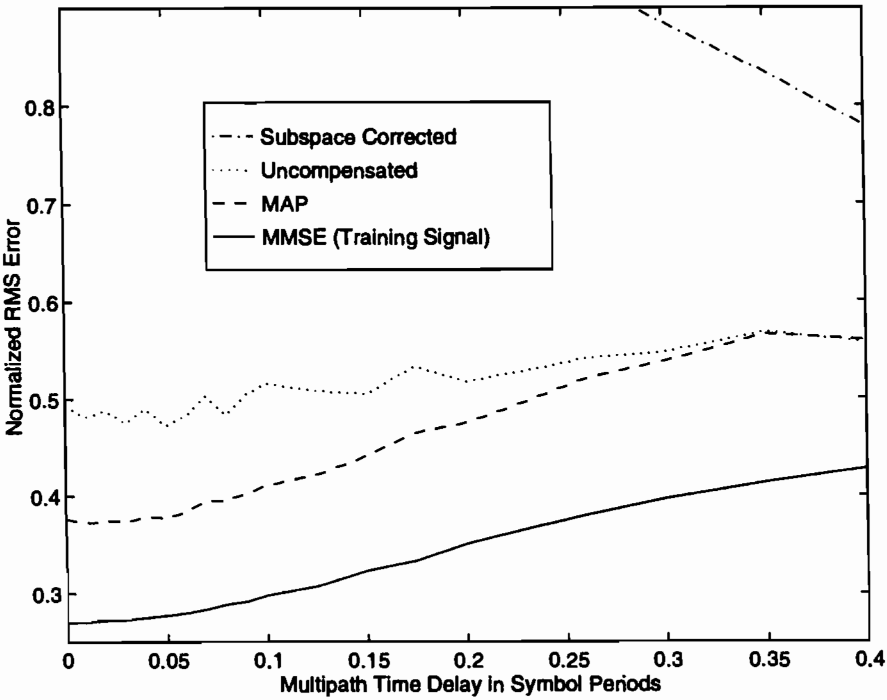


Figure 6 Root MSE performance of various beamformers for a multipath channel.

(13)–(17) with $\Upsilon = \sigma_a^2 \mathbf{I}$ and $\sigma_a = 0.15$. For each trial, MAP DOA estimates were obtained based on 75 samples from the array, and normalized RMS signal errors were computed. The results are plotted in Fig. 6 for various relative delays between the two arrivals. The "uncompensated" approach corresponds to the method of [56] implemented with $\mathbf{A}(\hat{\boldsymbol{\theta}}, \boldsymbol{\rho}_0)$ rather than $\mathbf{A}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\rho}})$ as in the MAP beamformer. The minimum MSE curve was obtained using a known 75-sample training sequence to compute the optimal weights, and was included to give an idea of the "best possible" performance.

While the SC-MVDR approach can to some degree compensate for array perturbations, it cannot eliminate signal cancelation due to the presence of a correlated arrival, and its performance in this case is quite poor. For small delays, correcting for calibration errors yields a 25–30% improvement in RMS error, which translates into a reduction in symbol error rate of approximately a factor of 6 (from .041 to .007) for this example.

VI. CONCLUSIONS

In this chapter, a Bayesian approach is presented for alleviating the sensitivity of parametric DOA estimation methods to imperfections in the data collection system. The exact MAP estimator is derived and found to be impractically complicated. Based on recent results on subspace-based ML approximations, a simpler method referred to as the MAP-NSF technique is derived. The method takes *a priori* information about the structure of the errors into account in an optimal way, and provides asymptotically minimum variance DOA estimates. Numerical examples and simulation studies have demonstrated a significant improvement as compared to traditional techniques. In particular, the examples of Section V demonstrated a substantial improvement in signal waveform estimation, resulting from appropriately utilizing the array perturbation model. This result is of great importance in, for example, communications applications. The price to be paid is an increased computational complexity. Despite the simplifications, the proposed method is notably more complex than DOA estimation methods that do not take the modeling errors into account. An exception is unstructured array errors, for which the MAP-NSF estimator reduces to the known signal subspace fitting (SSF) technique, but with a different choice of subspace weighting. An interesting alternative is therefore to approximate the array perturbations with an unstructured error model whenever possible. The resulting method is then no longer optimal, but can still yield substantially improved estimates compared to not taking the modeling errors into account.

REFERENCES

1. J Capon. Proc IEEE 57:1408–1418, 1969.
2. S Haykin. Array Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, 1984.
3. B Widrow, PE Mantey, LJ Griffiths, BB Goode. Proc IEEE 55:2143–2159, 1967.
4. RA Monzingo, TW Miller. Introduction to Adaptive Arrays. New York: Wiley Interscience, 1980.
5. RT Compton. Adaptive Antennas – Concepts and Performance. Prentice Hall, Englewood Cliffs, NJ: 1988.
6. JP Burg. Maximum Entropy Spectral Analysis. In Proc 37th Ann Int SEG Meeting, Oklahoma City, OK, 1967.
7. JP Burg. Maximum Entropy Spectral Analysis. PhD thesis, Stanford University, Stanford, CA, 1975.
8. E Parzen. Statistical Spectral Analysis (single Channel Case) in 1968. In Proc NATO ASI on Signal Processing, Enschede, Netherlands, 1968.

9. D Tufts, R Kumaresan. Proc IEEE 70:975–990, 1982.
10. WJ Bangs. Array Processing with Generalized Beamformers. PhD thesis, Yale University, New Haven, CT, 1971.
11. M Wax. Detection and Estimation of Superimposed Signals. PhD thesis, Stanford University, Stanford, CA, 1985.
12. JF Böhme. Estimation of Source Parameters by Maximum Likelihood and Nonlinear Regression. In Proc ICASSP '84, 1984, pp 7.3.1–7.3.4.
13. Y Bresler, A Macovski. IEEE Trans ASSP 34:1081–1089, 1986.
14. RO Schmidt. Multiple Emitter Location and Signal Parameter Estimation. In Proc RADC Spectrum Estimation Workshop, Griffiss AFB, NY, 1979, pp 243–258.
15. RO Schmidt. A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation. PhD thesis, Stanford University, Stanford, CA, 1981.
16. G Bienvenu, L Kopp. Principle de la Goniometrie Passive Adaptive. In Proc 7'eme Colloque GRESIT, Nice, France, 1979, pp 106/1–106/10.
17. DR Brillinger. A Frequency Approach to the Techniques of Principle Components, Factor Analysis and Canonical Variates in the Case of Stationary Series. In Roy Stat Soc Conf, Cardiff, Wales, 1964.
18. VF Pisarenko. Geophys J Roy Astron Soc 33:347–366, 1973.
19. AJ Berni. IEEE Trans AES 11:147–154, 1975.
20. RH Roy, ESPRIT – Estimation of Signal Parameters via Rotational Invariance Techniques. PhD thesis, Stanford University, Stanford, CA, 1987.
21. R Roy, T Kailath. IEEE Trans ASSP 37:984–995, 1989.
22. M Viberg, B Ottersten. IEEE Trans Sig Proc 39:1110–1121, 1991.
23. M Viberg. Subspace Fitting Concepts in Sensor Array Processing. PhD thesis, Linköping University, Linköping, Sweden, 1989.
24. BE Ottersten. Parametric Subspace Fitting Methods for Array Signal Processing. PhD thèsis, Stanford University, Stanford, CA, 1989.
25. M Viberg, B Ottersten, T Kailath. IEEE Trans Sig Proc 39:2436–2449, 1991.
26. B Ottersten, M Viberg, T Kailath. IEEE Trans Sig Proc 39:1122–1135, 1991.
27. A Swindlehurst, T Kailath. IEEE Trans Sig Proc 40:1758–1774, 1992.
28. A Swindlehurst, T Kailath. IEEE Trans Sig Proc 41:2882–2890, 1993.
29. M Viberg, A Swindlehurst. IEEE Trans Sig Proc 42:3073–3083, 1994.
30. B Ottersten, M Viberg, T Kailath. IEEE Trans Sig Proc 40:590–600, 1992.
31. B Ottersten, M Viberg, P Stoica, A Nehorai. In Haykin, Litva, and Shepherd, eds, Radar Array Processing, Berlin: Springer-Verlag, 1993, pp 99–151.
32. A Paulraj, T Kailath. In Proc IEEE ICASSP, Tampa, Fla, March 1985, pp 17.7.1–17.7.4.
33. Y Rockah, PM Schultheiss. IEEE Trans ASSP, 35:286–299, 1987.
34. AJ Weiss, B Friedlander. IEEE Trans ASSP 37:1958–1966, 1989.
35. B Wahlberg, B Ottersten, M Viberg. In Proc IEEE ICASSP, Toronto, Canada, 1991, pp 3277–3280.

36. M Wylie, S Roy, H Messer. IEEE Trans Sig Proc 42:3449–3459, 1994.
37. M Viberg, A Swindlehurst. IEEE Trans Sig Proc SP-42:3495–3507, 1994.
38. A Flieller, A Ferreol, P Larzabal, H Clergeot. In Proc ICASSP, Detroit, MI, 1995, pp 1884–1887.
39. J Lo, S Marple. In Proc ICASSP, Dallas, TX, 1987, pp 2260–2263.
40. B Ng, A Nehorai. In Proc ICASSP, Minneapolis, Minn, 1993, pp IV21–IV24.
41. M Koerber, D Fuhrmann. In Proc ICASSP, Minneapolis, Minn, 1993, pp IV340–IV343.
42. P Yip, Y Zhou. In Proc ICASSP, Detroit, MI, 1995, pp 1892–1895.
43. D McArthur, J Reilly. In Proc ICASSP, Adelaide, Australia, 1994, pp IV-201–IV-205.
44. D Tufts, I Kirteins, R Kumaresan. IEEE Trans Aero Elec Sys 19:313–316, 1983.
45. R Schmidt, R Franks. IEEE Trans Antenn Propag 34:281–290, 1986.
46. J Kim, C Un. Electronics Letters, 25:1076–1077, 1989.
47. B Wahlberg, I Mareels, I Webster. IEEE Trans Antenn Propag 39:21–28, 1991.
48. D Feldman, LJ Griffiths. In Proc IEEE ICASSP, Toronto, Canada, 1991, pp 1381–1384.
49. T Biedka. In Proc 27th Asilomar Conference on Signals, Systems, and Computers, Asilomar, CA, November 1993, pp 716–720.
50. RT Compton. IEEE Trans Aero Elec Sys AES-18:392–400, 1982.
51. LC Godara. IEEE Trans Aero Elec Sys AES-22:395–409, 1986.
52. NK Jablon. IEEE Trans Antenn Propag AP-34:996–1012, 1986.
53. P Stoica, K Sharman. IEEE Trans ASSP 38:1132–1143, 1990.
54. P Stoica, A Nehorai. IEEE Trans ASSP 38:1783–1795, 1990.
55. B Van Veen, K Buckley. IEEE ASSP Mag 5:4–24, 1988.
56. B Ottersten, R Roy, T Kailath. In Proc 23rd Asilomar Conference on Signals, Systems, and Computers, Asilomar, CA, November 1989, pp. 787–791.
57. J Yang, A Swindlehurst. IEEE Trans Sig Proc 43:2724–2732, 1995.
58. A Swindlehurst. A Maximum a Posteriori Approach to Beamforming in the Presence of Calibration Errors. In Proc 8th Workshop on Stat Sig Array Proc, Corfu, Greece, June 1996.
59. JF Böhme. Signal Proc 10:329–337, 1986.
60. A Graham. Kronecker Products and Matrix Calculus with Applications. Chichester, England: Ellis Horwood, 1981.
61. JR Magnus, H Neudecker. Matrix Differential Calculus with Applications in Statistics and Econometrics. Chichester, West Sussex: Wiley, 1988.

16

Selected Stochastic Methods and Signal Processing Used in Radar Systems

T. Sen Lee

National Space Program Office, Hsin-Chu, Taiwan

I. INTRODUCTION

This chapter reports the use of stochastic methods and signal processing techniques in tracking, imaging, and Doppler weather radar systems. The emphasis is on the practical aspects of how stochastic methods and signal processing techniques are used to formulate and solve radar system problems. On many occasions, only the descriptive versions of the problems and solutions are written. Most of the missing details can be found in MIT/Lincoln Laboratory reports cited in the reference section.

This chapter is divided into three parts. The first part (Section II) begins with a brief review of radar functions and mathematics to be included. The review intends to give just enough background for the reader to appreciate the connection between mathematics and practice. The second part (Sec. III) is devoted to radar problem formulations and their solutions. The third part lists all pertinent references.

II. THREE RADAR SYSTEMS

Stochastic methods and signal processing are required to optimize system performance under the constraint of limited available resource. This section first introduces potential radar system problems and required stochastic and signal processing techniques for system analysis. The applications to three different radar systems are then briefly described.

A. System Analysis

System analysis is needed during the entire project lifecycle [1]. The lifecycle starts with the concept definition. The second stage is design, implementation, and system integration. Successful operation usually is the ultimate goal of a project. Optimization and trade-off are performed to solve system problems. Factors employed in the analysis are system requirements, performance, schedule, budget, and risk. The risk consideration sometimes dominates over other factors. For example, in a defense system design, the risk of letting the threats leak through a defense system is weighted heavily in the trade-off analysis. In this situation, the analysis result must be conveyed clearly to the decision makers who are usually too busy to get into all the mathematical details. Reference 2 which is briefly reviewed below gives an example of this sort.

In [2], the probability of killing a conical target not invoking nuclear devices is derived as a function of endgame parameters. These endgame parameters are classified according to whether or not they are controllable by the defense. For a given set of uncontrollable endgame parameters, the probability of kill is optimized with respect to controllable endgame parameters. A tool has been developed for kill probability optimization and trade-off analyses of endgame parameters. Major conclusions found in [2] are:

1. Three ways to kill a reentry vehicle (RV) are hit-to-kill (HTK) interceptors, isotropic warheads, and aimed warheads. HTK requires the smallest miss distance but requires no fuzing. An aimed warhead can operate with the largest miss distance but requires the most sophisticated fuzing.
2. An HTK interceptor should always be targeted for zero miss distance; an aimed warhead should always have a standoff distance.
3. The requirements for a minimum strike angle for pellet impact impose constraints on the standoff distance and fuzing algorithm, particularly for small crossing angles.
4. Changing the angle of attack of the interceptor can change the effective warhead firing angle, which can improve the pellet strike angle.
5. Changes in RV orientation due to procession or angle-of-attack maneuvers will change the effective strike angle. The defense can enhance its performance by planning its intercepts at appropriate times in the RV precession cycle.

In the above example, complicated stochastic optimization and trade-off are performed. However, a clear message must be conveyed to the decision makers without the help of mathematics.

B. Tracking Radar

This type of radar systems usually performs three different functions: search, acquisition, and tracking. Optimal search strategies depend on many factors such as intended search areas or volumes as well as hardware and software limitations. If a big area/volume is to be searched, a dedicated search radar system usually is the optimal solution. In this situation, the search result will be sent to a tracking radar system for target acquisition. Stochastic techniques are needed to develop an optimal acquisition strategy.

Important considerations for the tracking function are maintaining tracks and tracking accuracy. To maintain tracks, a modern tracking radar usually uses Kalman filters which are susceptible for sudden target dynamic changes such as breakup and/or maneuvering. In these situations, *a priori* information utilization and/or adaptation become necessary.

In defending a ballistic missile attack, the interceptor has to be guided to a certain range of the target by the ground-based radar before its on-board sensors can guide itself towards the target. The previous sentence illustrates one of many reasons why tracking accuracy is important. Radar parameters that influence accuracy include beamwidth, bandwidth, alignment, etc. For an existing radar system, tracking accuracy can be maintained through routine metric calibration which has two different aspects described in some detail in Section III.

C. Imaging Radar

Target identification and surveillance are two main applications of this type of radar system. Since optimizing signal resolution is required, the radar system usually is equipped with most sophisticated signal processing techniques (see Section III). As such, the radar system depends on hand-over information supplied from tracking radar systems in general. In these situations, *a priori* information utilization and/or adaptation become necessary as well. Routine signature calibration is also called for to maintain accurate image generation.

D. Doppler Weather Radar

In the United States, the National Weather Service (NWS) and the Federal Aviation Administration (FAA) have begun deploying two Doppler weather radar networks. The NWS network called NEXRAD consists of many S-band Doppler weather radar systems that will replace the aging noncoherent systems. The FAA network called Terminal Doppler Weather Radar (TDWR) contains many C-band radar systems to be built near major airports to monitor hazardous weather and better manage the runway usage. Hazardous weather such a microbursts and gust fronts can be detected by a Doppler weather radar which is equipped with high-speed signal processors and sophisticated signal processing techniques [3]. Since these two coherent weather radar networks spread out over the entire United States, which covers the subtropical and mid-latitude regions, the Atlantic and Pacific coastlines, the great plains and mountain areas, a systematic approach to handle site adaptation is very important. The foundation for systematic site adaptation is based on statistical modeling of Doppler weather signals, which is described in some detail in Section III.

III. STOCHASTIC METHODS AND SIGNAL PROCESSING

The three subsections included here give examples of radar system problem forumulations and describe the use of stochastic and signal processing techniques in solving these problems. The first subsection is devoted to estimation. The second subsection reports resolution enhancement. The third subsection describes the utilization of *a priori* information and adaptation.

A. Estimation

The importance of estimation relating to radar system analysis arises from the need for improving system knowledge from noisy measurements. Unknowns in a radar system may appear as parameters, state variables, or functions. Consequently, this subsection will be divided into three parts. The emphasis is placed on subjects which may not yet be available in the open literature.

1. Parameter Estimation

A model with unknown parameters can be derived in order to translate radar measurements and error corrections into physical quantities. For examples, a Doppler weather radar measures coherent pulse sequences reflected back from precipitation. The coherent measurements can be translated into weather information such as reflectivity, wind velocity, and turbulence [3–5]. The other example related to tracking accuracy is described as follows.

As mentioned in the previous section, radar tracking accuracy can be maintained by a routine calibration procedure. A tracking radar system can provide range, range rate, azimuth, and elevation measurements.

Accuracy of the two angle measurements can be improved by a bias model correction. The angle bias model in its general form is described below.

(a) Azimuth angle

$$\begin{aligned}\Delta A = {} & a_0 + a_1 \sec(E) + a_2 \tan(E) + a_3 \sin(A - a_4) \tan(E) \\ & + a_5 V^2 \sin(E + a_6) \sin(A - \Theta) \sec(E) + a_7 \operatorname{sign}(\dot{A}) \\ & + a_{10} \cos(A - a_{11})\end{aligned}$$

(b) Elevation angle

$$\begin{aligned}\Delta E = {} & b_0 + b_1 \cos(E - b_2) + b_3 \cos(A - b_4) \\ & + b_5 V^2 \sin(E + b_6) \cos(A - \Theta) + b_7 E + b_8 \cos(2A - b_9)\end{aligned}$$

where

E = elevation encoder value,
A = azimuth encoder value,
V = wind speed,
Θ = wind direction,
$\dot{A}$ = azimuth rate, and
sign($\cdot$) = algebraic sign.

The encoder zeros sets are represented by a_0 and b_0. In general, the aximuth and elevation axes are not perfectly aligned. The alignment errors can be expressed in terms of tilt vectors projecting onto the azimuth and elevation platforms. The first-order projection gives rise to tilt amplitude terms a_3 and b_3 as well as the tilt phase terms a_4 and b_4. The second-order projection of the tilt vector onto the elevation platform gives rise to b_8 and b_9.

Unknown parameters described above can be estimated from angle measurements of sphere satellites whose orbits are cataloged and updated regularly. The sphere satellite ephemeris transferred into radar azimuth and elevation angles can be considered as the truth. The difference between the radar measurements and ephemeris-derived angles are the basis for parameter estimation [6]. Unknown parameters can be obtained by well-known least squares methods. Depending upon the geometry between the satellite orbit and the radar location, some unknowns may be difficult to estimate with high confidence. These parameters can be isolated by the technique described in the next subsection.

2. State-Vector Estimation

State-vector estimation has two different radar system applications. One is implemented for real-time use such as tracking. Post-mission data analysis for test performance evaluation is the second application. Both applications are very well documented in [7–11]. In the following, the relationship between observability and unbiased state-vector estimation is explored, and the application of a fixed-lag smoother is described.

A software package has been developed for post-mission test performance evaluation [12]. Major ingredients of the software package include three different coordinate systems, Newton and Euler equations, simplified aerodynamics, a simple earth magnetic field model, and a very complicated earth and geo-potential model. The program can take multiple sensor data collected from radar systems, optical sensors, and on-board instrumentation such as accelerometers, rate gyros, and magnetometers. The extended Kalman smoother which has forward and backward extended Kalman filters [7] is used. Estimated state-vector components include target position and velocity vectors, Euler angles and their rates, as well as aerodynamic coefficients. From the estimated state variables, the target angle of attack and coning motion can be derived. The derived quantities are important for vehicle reentry performance evaluation. Depending upon different combinations of multiple sensors and their relative geometry with respect to the target, radar system biases and radar site coordinate errors may be estimated. Observability of vehicle dynamics and radar measurement performance can be evaluated by the technique described below.

Let $x(t)$ be the state-vector satisfying the following nonlinear differential equation. Let $y(t)$ be the measurement vector.

$$\dot{x}(t) = f(x(t), t)$$

$$x(T) = x_1$$

$$y(t_k) = h(x(t_k)) + n(t_k)$$

$$t_0 \le t_1 \le \cdots \le t_K = T$$

where

x = n-dimensional state-vector,

f = n-dimensional vector function differentiable with respect to x and t,

T = terminal time

y = m-dimensional measurement vector,

t_k = kth discrete measurement time,

h = m-dimensional differentiable vector measurement function, and

$n(t_k)$ = m-dimensional zero-mean white Gaussian process.

The $n \times n$ observability Gramnian matrix is defined below.

$$M(a,K) = \sum_{k=1}^{K} \phi^T(t_k,t_K;a)H^T(t_k)R^{-1}H(t_k)\phi(t_k,t_K;\ a)$$

where

a = terminal state-vector which belongs to a convex set denoted by S,
K = number of measurements,
ϕ = transition matrix associated with f,
H = Jacobian matrix associated with h, and
R = noise process covariance matrix.

The existence of an unbiased terminal-state estimator for each vector in S implies that the observability Gramnian is positive-definite [13,14]. Therefore, if the observability Gramnian is nearly singular, then the terminal state estimate is biased. The following technique may be employed to find the unobservable components in the state-vector. The state-vector estimation problem can be reformulated to reduce the number of state variables.

The observability Gramnian has the following two decompositions.

$$M(a,K) = A^T(a,K)A(a,K)$$

and

$$M(a,K) = U(a,K)D(a,K)U^T(a,K)$$

where

$U = (u_1, \ldots, u_r)$,
u_i = n-dimensional column vector of U, and
D = r-dimensional diagonal matrix.

The diagonal elements of D are the singular values of M given by

$$\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_r^2$$

If r is smaller than n, then the null space of A contains state-vectors which are not observable from available measurements. The $n-r$ independent column vectors which are orthonormal to the column vectors of U span the null space of A. By examining the inner products of a given state-vector with respect to the $n-r$ independent column vectors, unobservable components in the state-vector can be identified. In addition to the $n-r$ independent column vectors, the column vectors associated with certain singular values which are smaller than a chosen threshold should also be included to form the unobservable space.

If the terminal state-vector estimation problem is formulated within the framework of an extended Kalman filter or smoother, the observability Gramnian can be approximated by the inverse of the derived error covariance matrix. Therefore, the computed error covariance can be used to investigate the observability condition.

The analysis technique described above is not suitable for real-time applications. However, recent defense radar systems strive to recognize subtle features for target identification. For example, in a theater missile defense scenario, the radar system has to identify the warhead early to give the interceptor enough time to fly to the target before it is too late. Using high-altitude drag estimates as a feature can be very effective in discriminating different objects among the decoy, fragment, and reentry vehicle (RV) because the decoy is lighter than the RV and fragments are not shaped for aerodynamic performance. Using a real-time Kalman filter which includes drag as one of the state variables can effectively distinguish fragments and the RV. However, a better technique than the Kalman filter such as a fixed-lag smoother [7] is needed to discriminate the decoy from the RV. The application of a fixed-lag smoother is reported in [15] which is briefly described as follows.

The backbone of a theater missile defense tracker is a Kalman filter. Though the outputs of the Kalman filter have many applications, Reference 15 focuses on their applications to slowdown and lateral acceleration estimation.

It is well known that the Kalman filter is vulnerable to sudden changes in target dynamics, such as oscillation in the angle of attack or the onset of a maneuver. In the worst case, the sudden change in dynamics may cause filter divergence that usually results in loss of the target track; in less catastrophic cases, the ballistic coefficient and lift estimates will lag the truth. The filter lag may cause the discrimination classifier to declare an incorrect target identification or the trajectory predictor to introduce a large pointing error between the threatening target and the interceptor in flight. In [15], it shows that a fixed-lag smoother can solve these problems at the expense of more computer processing and a slight real-time lag. The lag is not critical in decision-making when using the slowdown as a discrimination feature; however, it may be more critical in terms of supporting the interceptor. The slowdown computation and the concept of the fixed-lag smoother are described in [15]. In particular, an explanation is provided as to why the fixed-lag smoother removes the lag due to the Kalman filter. The results based on a simulated spiral maneuvering trajectory and real mission data which include target breakup measurements are shown.

3. Function Estimation

Much operational decision making is based on the result of statistical inference. Practical statistical inference depends on accurate probability density functions of radar signal, which are often difficult to obtain. Estimation of probability density functions can become too mathematical to be handled by engineers in general. A better engineering approach is to derive the function model of the desired probability density functions from physics, mathematics, and empirical reasoning. The desired probability density function can then be estimated from field measurements. An example of this approach is given below for a Doppler weather radar system.

Figure 1 shows a simplified block diagram of a Terminal Doppler Weather Radar system (TDWR). As introduced in Section II, the TDWRs are deployed near major airports for hazardous weather detection during airplane landing and taking-off. A TDWR scans the lowest portion of the sky to take weather measurements. Statistical decision making is required to discriminate weather against ground clutter from radar measurements.

The statistical modeling begins at the output node of the video detector. The digitized in-phase denoted by I, and quadrature denoted by Q returns are considered as sample functions of a two-dimensional noisy autoregressive (AR) process that admits a probability K-distribution [16,17]. The

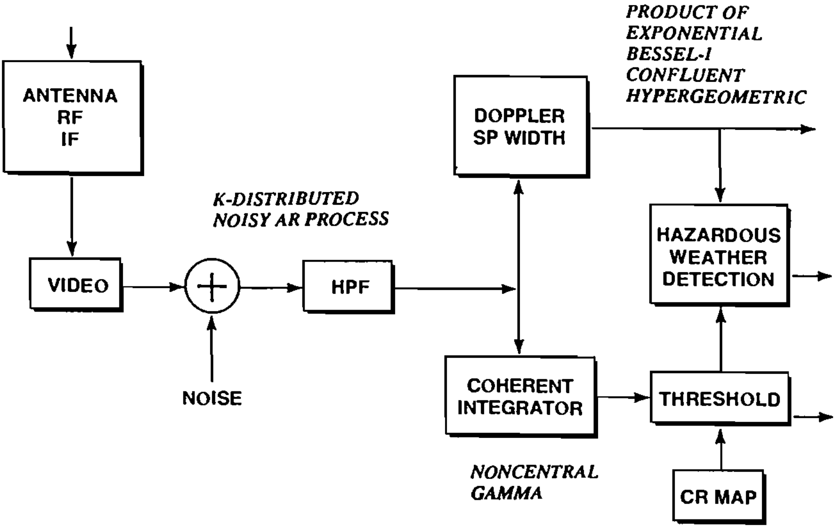


Figure 1 Doppler weather radar signal path.

K-distribution has the following form.

$$f_1(p) = \frac{2\alpha}{\Gamma(\alpha)\mu}\left(\frac{\alpha p}{\mu}\right)^{\alpha-1/2} K_{\alpha-1}\left(2\sqrt{\frac{\alpha p}{\mu}}\right)$$

where

p = signal power,
Γ = Gamma function,
μ = mean, and
K = modified Bessel function of the third kind.

The variance of the signal power p is controlled by α according to the following relationship.

$$\frac{\text{var}(p)}{\mu^2} = 1 + \frac{2}{\alpha}$$

As α approaches infinity, the K-distribution degenerates into an exponential function often used to model radar returns of Rayleigh scatterers. In practice, a normalized power with respect to the mean is used to simplify the model. The standard deviation of a normalized K-distribution is often referred to as the shape parameter.

High-pass filters [18,19] are used to suppress ground clutter. Clutter residue which is defined to be the output of the clutter filter preserves the K-distribution but with a different shape parameter. The power of the coherently integrated residue is shown to be Rician-distributed. The Rician density function is given below:

$$f_2(p) = \frac{1}{2\sigma^2} I_0\left(\sqrt{\frac{\lambda p}{\sigma^2}}\right)\exp\left(-0.5\left[\lambda + \frac{p}{\sigma^2}\right]\right)$$

where I_0 is the modified Bessel function of zeroth order, and σ is the associated second-order moment. The measured power can be used to infer radar reflectivity.

Most coherent weather radar systems employ correlated pulse pairs (PP) to compute Doppler [4]. The PP technique uses the first lag of the sample autocorrelation function of the measured complex-valued time series, $\{I(k) + jQ(k)\}$, to extract Doppler information. The joint probability density function of measured reflectivity and Doppler is a product of a Bessel function and a confluent hypergeometric function as shown below:

$$f_3(p, \theta) = f_2(p)\frac{1}{2\pi A\sqrt{|\Sigma|}}\exp(-0.5C)M\left(1, 0.5, \frac{B^2}{8A}\right)$$

where

θ = Doppler in units of radians which can be converted into meters per second by the following equation,

$$= \frac{4\pi\nu_r}{\lambda \mathrm{PRF}},$$

ν_r = Doppler in meters per second,

λ = wavelength,

PRF = pulse repetition frequency, and

Σ = covariance matrix of the first lag of the sample autocorrelation.

Other symbols in the above equation are described as follows. First, $|\cdot|$ denotes the determinant. The real and imaginary parts of the first lag of the sample autocorrelation, denoted by r_1 and r_2 respectively, are stacked as a two-dimensional vector. Therefore, Σ is a 2×2 covariance matrix whose components are denoted by Σ_{ij}. To explain the M-function, the following quantities are defined first:

$$a = \frac{\Sigma_{22}}{|\Sigma|}$$

$$b = -\frac{\Sigma_{12}}{|\Sigma|}$$

$$c = \frac{\Sigma_{11}}{|\Sigma|}$$

$\bar{r}_i$ = mean of r_i

$$A = a + b\sin(2\theta)$$

$$B = 2[a\bar{r}_1\cos(\theta) + b\bar{r}_1\sin(\theta) + b\bar{r}_2\cos(\theta) + c\bar{r}_2\sin(\theta)]$$

$$C = a\bar{r}_1^2 + c\bar{r}_2^2 + 2b\bar{r}_1\bar{r}_2$$

$M(\cdot,\cdot,t)$ is the confluent hypergeometric function which arises as the general solution of a time-varying differential equation of the second order such as Laplace's equation. The time-varying coefficients may be linear, quadratic, or simple functions of t. One of the simplest examples of a confluent hypergeometric function is $M(0.5, 0.5, t)$ which is e^t. Extensive tables are available in [20] for evaluating the confluent hypergeometric functions.

The probability density functions introduced above have been validated by field measurements taken by the MIT/Lincoln Laboratory portable Doppler weather radar system. The laboratory Doppler weather radar system is considered as the prototype unit of the TDWR built for the Federal Aviation Administration. It began as an S-band Doppler weather radar, then it was converted to C-band in the spring of 1990. The radar parameters can be found in [19]. Examples of field measurements can be found in [3] and [21].

B. Resolution Enhancement

The radar signal resolution limit is inversely proportional to the receiver bandwidth of the hardware. A number of modern estimation techniques have been studied over the past two decades with the goal of improving the range or Doppler resolution of a radar signal. These software techniques are generally called bandwidth expansion (BWE), because they can resolve scatterers that are closer than the hardware resolution limit. Six different BWE techniques are evaluated in [22], and they are briefly discussed as follows.

The target is assumed to consist of a sequence of point scatterers at range r_i, with radar cross-section (RCS) σ_i. The radar return is calculated from the monochromatic return at frequency f_k as

$$x_k = I(f_k) + jQ(f_k) = \sum_{i=0}^{n-1} \sqrt{\sigma_i} \exp\left(-jr_i \frac{4\pi f_k}{c}\right) + n_k \tag{1}$$

where

(I, Q) = in-phase and quadrature of frequency response,
n = number of scatters,
c = speed of light, and
n_k = zero-mean white Gaussian process.

The structure of the application of BWE techniques is depicted in Fig. 2. Each of the techniques operates on the measured frequency responses to extrapolate them beyond the measured values. The resulting wider frequency responses will result in finer resolution when transformed back into the range domain. The six techniques evaluated in [22] are listed below:

1. Fast Fourier Transform (FFT)
2. Maximum Entropy Method (MEM)
3. Forward-Backward Prediction Filter (FBPF) [23,24]
4. Tufts–Kumaresan Method (TK) [23–25]
5. Multiple Signal Classification (MUSIC) [26]
6. Least Squares Method (LS) [27,28]

Techniques 2 through 4 use autoregressive (AR) models to extrapolate the measured frequency responses. Both MUSIC and LS solve for scatterer RCSs and range separations directly.

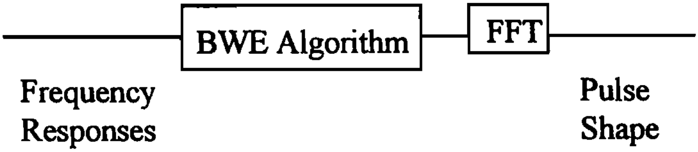


Figure 2 BWE data flow diagram.

1. AR-Model and Prediction Filter

The AR technique uses a finite impulse response (FIR) filter to characterize the measured frequency responses which are the inputs shown in Fig. 2. The FIR filter is specified by its transfer function given as follows.

$$H(\theta) = 1 + \sum_{k=1}^{L} a_k e^{-jk\theta}$$

where

L = filter order,
$a^T = (a_1, \ldots, a_L)$, filter coefficients, and
θ = discrete frequency, $2\pi \geq \theta \geq 0$.

The filter coefficients are sought to optimally annihilate the inputs. The corresponding FIR gives rise to the AR model used for extrapolating the measured frequency responses.

The optimization procedure is described as follows. Let $x_k, k = 0, \ldots, m-1$, be the measured frequency responses. The number, m, of measured frequency responses must be larger than the filter order L. The output vector obtained by feeding the measured frequency responses forward through the FIR filter is denoted by y and given by

$$y = A_1 a,$$

where A_1 is an $m \times L$ matrix given by

$$A_1 = \begin{vmatrix} x_L & \cdot\ \cdot & x_1 \\ x_{L+1} & \cdot\ \cdot & x_2 \\ \cdot & \cdot\ \cdot & \cdot \\ x_{m-1} & \cdot\ \cdot & x_{m-L} \end{vmatrix}$$

The backward output vector obtained by feeding the measured frequency responses in the reverse order through the FIR filter is denoted by z and is given as follows:

$$z = A_2 a$$

where A_2 is also an $m \times L$ matrix given by

$$A_2 = \begin{vmatrix} x_2^* & \cdot\ \cdot & x_{L+1}^* \\ x_3^* & \cdot\ \cdot & x_{L+2}^* \\ \cdot & \cdot\ \cdot & \cdot \\ x_{m-\mathrm{L}}^* & \cdot\ \cdot & x_m^* \end{vmatrix}$$

The objective is to seek the coefficient vector, a, that minimizes $y^* y + z^* z$. The superscript denotes conjugate transpose. The solution of the minimization problem is embedded in the following set of linear equations.

$$\begin{bmatrix} A_1 \\ A_2 \end{bmatrix} a = h$$

where h is given by

$$h^T = (x_{L+1}, x_{L+2}, \ldots, x_{L+m}, x_1^*, x_2^*, \ldots, x_{m-L}^*)$$

The MEM algorithm takes advantage of the well-known Levinson recursion to solve for the filter coefficients. The required number of numerical operations is proportional to m^2. Both FBPF and TK use singular value decomposition (SVD) to perform the minimization. The required number of operations is proportional to m^3. The FBPF retains all the singular values in solving for the filter coefficients, whereas the TK algorithm only uses singular values above noise power [23–25]. If the filter order, L, is large it will accurately represent the spectrum; however, using it to extrapolate may introduce noise sensitivity. If L is small, the filter may not be accurate. It has been suggested in [23] and [25] to limit L to within $m/2$ and $3m/4$.

2. MUSIC and LS

MUSIC and LS estimate parameters of the target scattered amplitude and location directly. The formulation of these two estimation algorithms is described as follows. Divide the m available frequency responses into p subsets. Each subset contains q pairs of in-phase and quadrature samples. Thus, the sample size m is assumed to be the product of p and q. The kth subset of frequency responses can be collected into a vector form denoted by u_k as

$$u_k^T = (x_{kq}, x_{kq+1}, \ldots, x_{(k+1)q-1})$$

Using the target signal model defined by Eq. (1), the above equation can be decomposed as

$$u_k = \Phi_k(r)s(\sigma) + w_k, \tag{2}$$

where Φ_k is a $q \times n$ matrix function of unknown range parameters given by

$$\Phi_k(r) = \begin{vmatrix} \exp\left(-jr_0 \frac{4\pi f_{kq}}{c}\right) & \cdot & \exp\left(-jr_{n-1} \frac{4\pi f_{kq}}{c}\right) \\ \cdot & \cdot & \cdot \\ \exp\left(-jr_0 \frac{4\pi f_{(k+1)q-1}}{c}\right) & \cdot & \exp\left(-jr_{n-1} \frac{4\pi f_{(k+1)q-1}}{c}\right) \end{vmatrix}$$

The n-vector s contains the unknowns RCS parameters given by

$$s^T(\sigma) = (\sqrt{\sigma_0}, \ldots, \sqrt{\sigma_{n-1}})$$

and the q-dimensional noise vector w_k contains q independent zero-mean white Gaussian processes defined in Eq. (1).

An LS algorithm iteratively searches for scatter RCSs and range separations that minimize error squares. The error is defined to be the difference of the frequency responses and the noise-free model described by Eq. (1). The LS algorithm developed based on [27] and [28] recognizes the particular structure of Eq. (2). First, the unknown parameters are separable. Secondly, the RCS parameters appear linearly in the equation. The rate of convergence is improved by taking advantage of the above two observations.

To describe MUSIC, a Fisher information matrix denoted by F is defined as follows.

$$F(r, \sigma) = \sum_{k=1}^{p} u_k u_k^* \tag{3}$$

The Fisher information matrix F is positive-definite with probability one because of the white Gaussian assumption. If w_k is zero, then the linear space generated by F has a special structure examined as follows. The rank of each outer product in Eq. 3 is one. If the number of frequency samples, m, is greater than n^2, then there exists n subsets of frequency responses such that the rank of F is n. Recall that n is the number of radar scatters. Thus, the linear space can be decomposed into signal and noise spaces. MUSIC searches for feasible range separations that define a signal space farthest away from the noise space. The detailed computational procedure can be found in [26].

Physically, certain discrete frequencies will excite a system of scatters differently. This discrete frequencies identify the best choice of subsets of frequency responses. The above observation implies that a coherent frequency-hop, nonchirping narrowband waveform is an economical way to improve range resolution. However, the sampling interval becomes nonuniform.

Both MUSIC and LS use SVD which imposes a heavy burden on the computing machine. MUSIC does not require an initial guess or iterations. Experience shows that MUSIC can be used to initiate the LS search.

The techniques which use SVD can enhance radar resolution more effectively in terms of noise rejection and bandwidth expansion. Computationally speaking, MEM is most efficient; however, the technique is susceptible to noise contamination. For example, the following conclusions are drawn from a simulation study based on a monostatic, plane wave, and point scatter model.

If the signal-to-noise ratio with respect to the lowest scatter (SNR) exceeds 20 dB, the LS techniques proposed in [27] and [28] can expand bandwidth more than six times. The technique proposed by Tufts and Kumaresan can still expand bandwidth three times even if the SNR is as low as 0 dB. The MEM can expand bandwidth three times if the SNR exceeds 30 dB, 2.5 times if the SNR exceeds 20 dB, and 1.5 times if the SNR exceeds 5 dB.

For the AR-based techniques, the selection of the model order has a tremendous effect on the outcome. The order is related to the number of radar scatters. The search-based techniques such as MUSIC and LS also require one to guess the number of scatters to enhance the efficiency of the search procedure. Practically, the number of scatters is part of the solution to be sought. Recent results reported in [29] may be useful for estimating the number of scatters from measurements.

C. Using *a priori* Information and Adaptation

The techniques introduced in the previous two sections are effective only to a certain degree. For example, an engineer may need to estimate parameters which may be unobservable from available measurements. Resolution enhancement techniques require knowing the number of radar scatters to be effective. In tracking problems, sudden target dynamic changes such as maneuvering and breakup often cause the tracking system to lose track. Two possible ways to overcome these situations are to incorporate *a priori* information in the techniques or set up a self-learning procedure in the techniques based on observed data. The following two subsections describe some specific problems and their solutions using these two different approaches. The intent is to show how mathematics is used to incorporate available information to solve problems.

1. Using *a priori* Information

In solving both estimation and statistical decision problems, one often seeks the *a posteriori* conditional probability density function of measurements since it contains all statistical information of interest. Statistically, the probability density function of the unknowns is the *a priori* information which can be incorporated into the *a posteriori* conditional probability density function according to Baysian theory.

In a practical setting, instead of knowing the *a priori* probability density function, the unknowns are known to within certain physical bounds which are deterministic in nature. The unknowns may be unobservable from available measurements. However, the engineer often needs to find a best guess of the unknowns in order to solve the problem at hand.

(a) Nonmodel-Based Estimation. An approach taken in [30] is to formulate the guessing by solving a linear or nonlinear inequality problem. A set of inequalities is given below:

$$A(x)x \leq b$$

where the matrix, A, defines the physical bounds of the unknown vector x, and b is the bounding vector. If A is independent of x, then the above equation defines a polyhedron. The geometric center of the polyhedron may be viewed as the best guess of the unknown vector. The center of the bounded domain defined by the inequality can be found efficiently by an algorithm invented by Khachian [30,31].

(b) Tracking Breakup Targets. A well-designed reentry vehicle will survive through the atmosphere. A poorly designed reentry vehicle will break up during reentry. The breakup event causes sudden target dynamic changes. Many experiments have been conducted in the late 1960s and recently to examine the breakup phenomenon for developing a ground-based defense system. Information about deliberately breaking up targets may only be obtained from intelligence gathering. Naturally breaking up events are well known from physics and past experiments. Gathered information can be incorporated into the tracker design and it is briefly described as follows.

If target breakup is expected, the tracker had better be formulated within the framework of a multitarget tracker [9,10] to reduce the probability of target leakage. Breakup detection should be incorporated within the tracker. Detection can be based on simple but effective features derived from the radar signal. Features such as monopulse voltage ratios, integrated radar cross-sections within the range window, or Doppler measurements have been proved to be effective for breakup detection. The idea is to measure these features and detect sudden changes of these features. The decision logic usually is simple but not straightforward in order to cover all possibilities derived from the *a priori* information. Once the breakup event is detected, the multitarget tracker has to be kicked in, in conjunction with target identification. The multitarget tracker should assign heavier weights on track files identified to belong to threatening targets. A narrow-band waveform is preferred once the breakup event is detected. The radar bandwidth may gradually increase to enhance the target identification capability.

2. Adaptation

Like the subject of state-vector estimation, practical adaptive algorithms have two different versions depending on whether they are for real-time use or not. An example of a real-time adaptive algorithm is maneuvering target tracking [8]. Target maneuvering can be detected by observing the significance of estimated target lift. If the list estimate is significant, the target is most likely maneuvering. The significant test of the lift estimate has to be very simple because the target velocity usually exceeds a few kilometers per second depending upon the trajectory. A real-time adaptive algorithm for maneuvering target tracking may be implemented with the following principles.

1. Include the lift component in the tracking filter.
2. Shut-off lift estimation in the exo-atmospheric region by setting the associated initial condition and process noise covariance to zero.
3. In the endo-atmospheric region, set the associated process noise covariance proportional to the significance of the lift estimate.

Another example of a real-time adaptive algorithm implemented for a real-time imaging radar is an adaptive fading memory Kalman filter [32]. In [32], the age-weighting factor used in the fading memory Kalman filter [7] is tuned in real-time to adjust the filter bandwidth. If the filter bandwidth is tuned narrowly enough, the fading memory Kalman filter can filter out targer range variations due to its rotational motion. The range outputs can be converted into the scattered phases of the target center of gravity, called range-derived phases. The target image is constructed based on phase variations induced by target rotational motion, which are obtained by subtracting out the range-derived phases from the phase measurements. However, if the bandwidth is tuned too narrowly, the target will be lost. The above two observations illustrate the importance of real-time adaptation which guarantees filter stability.

Stability analysis and sample properties of an adaptive fading memory Kalman filter are also reported in [32].

The other aspect of adaptation is learning from accumulating past experience in order to improve future performance. An example of this sort is the making of clutter residue maps which are implemented in Terminal Doppler Weather Radar (TDWR) systems [21]. As introduced in Section II, TDWR systems are deployed near major airports for hazardous weather detection during airplane landing and taking-off. A TDWR scans the lowest portion of the sky to take weather measurements. The measurements are compared with the clutter residue maps. If the measurements exceed the map values, then the precipitation is significant. The making of a clutter residue map is described in [3], [19], and [21] which include discussions on how to select site-specific parameters such as the number of maps needed, map resolution, and threshold values. These site-specific parameters are selected based on past performance of the clutter residue maps in terms of false hazardous weather alarms and leakage statistics.

IV. SUMMARY

The stochastic methods and signal processing techniques described in this chapter include parameter estimation, real-time tracking, post-mission trajectory estimation, statistical modeling of radar signal, resolution enhancement, using *a priori* information, and adaptation. These techniques have applications in tracking, imaging, and Doppler weather radar systems for defending ballistic missile attacks and hazardous weather detection and prediction.

The emphasis of this chapter is on the practical aspects of how stochastic methods and signal processing techniques are used to formulate and solve radar system problems. Only descriptive versions of the problem and solutions are given. Most of the missing details can be found in MIT/Lincoln Laboratory reports cited in the reference section.

REFERENCES

1. NASA System Engineering Handbook, 1984.
2. TS Lee, JF Baldassini. Endgame Analysis of NNK Engagement. MIT/Lincoln Laboratory Project Report DA-34, April 1993.
3. RJ Doviak, DS Zrnic. Doppler Radar and Weather Observations. New York: Academic Press, 1993.
4. DS Zrnic. IEEE Trans Aerosp Electron Syst AES-13: 1977.

5. M Labitt. Coordinated Radar and Aircraft Observations of Turbulence. MIT/Lincoln Laboratory Project Report ATC 108, 1981.
6. EM Gaposchkin. Calibration of Low Altitude Kwajalein Radars. MIT/Lincoln Laboratory Technical Report 748, August 1986.
7. A Gelb, ed. Applied Optimal Estimation. Cambridge, MA: The MIT Press, 1974.
8. CB Chang, JA Tabaczynski. IEEE, AC-29, 1984.
9. SS Blackman. Multiple-Target Tracking with Radar Applications. Dedham, MA: Artech House, Inc., 1986.
10. Y Bar-Shalom, ed. Multitarget-Multisensor Tracking: Advanced Applications. Norwood, MA: Artech House, Inc., 1990.
11. G Biernson. Optimal Radar Tracking Systems. New York: Wiley, 1990.
12. TS Lee. Lincoln Orbit Determination Program: General Description. MIT/Lincoln Laboratory Project Report SDP-265, August 1984.
13. TS Lee, KP Dunn, CB Chang. On Observability and Unbiased Estimation of Nonlinear Systems. Proc 10 IFIP Conference on System Modeling and Optimization. New York: Springer-Verlag, 1982.
14. TS Lee, KP Dunn. On Mean Kullback–Leibler's Information, Maximum Likelihood Principle and Uncertain Nonlinear Systems. MIT/Lincoln Laboratory Technical Report 674, November 1983.
15. TS Lee, LC Youens. TMD Application of a Fixed-lag Smoother. MIT/Lincoln Laboratory Project Report-384, February 1994.
16. E Jakeman, PN Pusey. IEEE Trans Antenna Propagat AP-24, 1976.
17. JK Jao. IEEE Trans Antennas Propagat AP-32, 1984.
18. T Parks, J McClellan. IEEE Trans Circuit Th CT-19: 1972.
19. JE Evan, WH Drury, DP Hynek, TS Lee, BH Stevens. Doppler Weather Radar Engineering for Airport Wind Shear Detection. IEEE International Conference, Arlington, VA, 1990.
20. M Abramowitz, IA Stegun. Handbook of Mathematical Functions. New York: Dover, 1970.
21. TS Lee. Statistical Model of Doppler Weather Radar Signal. Proc 26 ISCIE International Symposium on Stochastic Systems Theory and Its Applications. Osaka, Japan, October 1994.
22. TS Lee. Comparison of Six Bandwidth Extrapolation Techniques. Project Report SDP-378, MIT/Lincoln Laboratory, June 1994.
23. K Kuomo. Bandwidth Extrapolation Technique for Improved Range Resolution of Coherent Radar Data. Project Report CJP-60, MIT/Lincoln Laboratory, January 1991.
24. SL Marple. Digital Spectral Analysis with Applications. New York: Prentice Hall, 1987.
25. R Kumaresan. Estimating the Parameters of Exponentially Damped or Undamped Sinusoidal Signal in Noise. PhD Dissertation, Electrical Engineering Department, University of Rhode Island, RI, 1982.
26. RO Schmidt. Multiple Emitter Location and Signal Parameter Estimation. Proc RADC Spectrum Estimation Workshop, October 1979.

27. GH Golub, V Pereyra. SIAM Num Anal 10: 1973.
28. A Ruhe, PA Wedin. SIAM Rev 22: 1980.
29. M Viberg, B Ottersten, T Kailath. IEEE Trans Signal Processing 39:2436–2449, 1991.
30. TS Lee. Non-Model Based Estimation with Applications to Sensor Calibration Using Multiple Sensors. MIT/Lincoln Laboratory Report, TR-714, April 1985.
31. LG Khachian. Sov Math Dukl 20:191, 1979.
32. TS Lee. IEEE Trans Circuits Systems 35: 1988.

This Page Intentionally Left Blank

17

Statistical Methods for Robust Change Detection in Dynamical Systems with Model Uncertainty

Kousuke Kumamaru, Jinglu Hu, and Katsuhiro Inoue

Kyushu Institute of Technology, Iizuka, Japan

Torsten Söderström

Uppsala University, Uppsala, Sweden

I. INTRODUCTION

In recent years, model-based fault detection and isolation (FDI) have received considerable attention, and many FDI techniques have been proposed, see, e.g., [1–4]. There is a crucial problem in practice that the model-based FDI methods may be very sensitive to modeling errors. Therefore it will be important to improve their robustness to unmodeled dynamics.

A typical scheme for robust FDI is proposed by Frank and Wunnenberg (1989) [5], and Patton and Kangethe (1989) [6] based on observer design approaches, in which the robustness is achieved by appropriately designing the observers. In their approaches, however, the distribution matrices of fault and disturbance should be known in the state-space representation.

Most failures in dynamical systems can be represented as unexpected variations in system operation modes which are caused by the changes in system configuration parameters. Based on this assumption, Kumamaru and Söderström (1986) [7] have proposed a change detection system (CDS), in which change detection is carried out by detecting model distortion using Kullback discrimination information (KDI). In this paper, design of the CDS will be extended to cases of unmodeled dynamics to achieve robust change detection. For this purpose, the following three aspects will be considered:

1. *Robust identification.* In the robust identification literature, unmodeled dynamics has been described as a random quantity with *soft bound* or a deterministic quantity with *hard bound* [8]. Though the hard bound description has been widely used in robust control theory, it seems that the soft bound description is more suitable for designing robust CDS based on the KDI, since the stochastic property of the unmodeled dynamics can easily be incorporated into the KDI. It is also found that to describe unmodeled dynamics as a bound does not seem to be essential, because the size of unmodeled dynamics plays a more important role than the bounds in the design of robust CDS. From these viewpoints, two approaches to robust identification will be considered. One is a soft bound approach where a soft bound of the unmodeled dynamics is introduced based on Goodwin's stochastic embedding approach [9]. The other is a nonbound approach where the unmodeled dynamics is described by a difference between a complex model and a simple model.
2. *The KDI analysis.* To evaluate the KDI in a feasible way for finite but fairly large data sets, an algorithm is derived, incorporating the descriptions of unmodeled dynamics explicitly. Furthermore, indexes for unmodeled dynamics are also introduced.
3. *Change detection scheme.* To achieve robustness of the CDS, decision-making schemes will be developed so that the change detection can be carried out on the basis of the KDI, the indexes for unmodeled dynamics, and other prior information about the system.

Figure 1 shows the basic structure of the robust CDS.

This paper is organized as follows: in Section II, we summarize the KDI-based change detection method in the ideal case where there is no unmodeled dynamics. The method is extended to cases where unmodeled dynamics exists in Section III, and robust change detection schemes are developed based on soft bound and nonbound approaches. In Section IV, two numerical simulations are carried out to demonstrate the effectiveness of the proposed methods. Section V is finally devoted to discussions and conclusions.

II. CHANGE DETECTION VIA THE KDI

In this section we summarize change detection methods using the KDI; for more details see [10,11,7].

Consider a discrete-time linear SISO system described by

$$\mathcal{S}\colon\quad y(t) = G_0(q^{-1})u(t) + H_0(q^{-1})e_0(t)$$
$$e_0(t) \in N(0, \sigma_0^2) \tag{1}$$

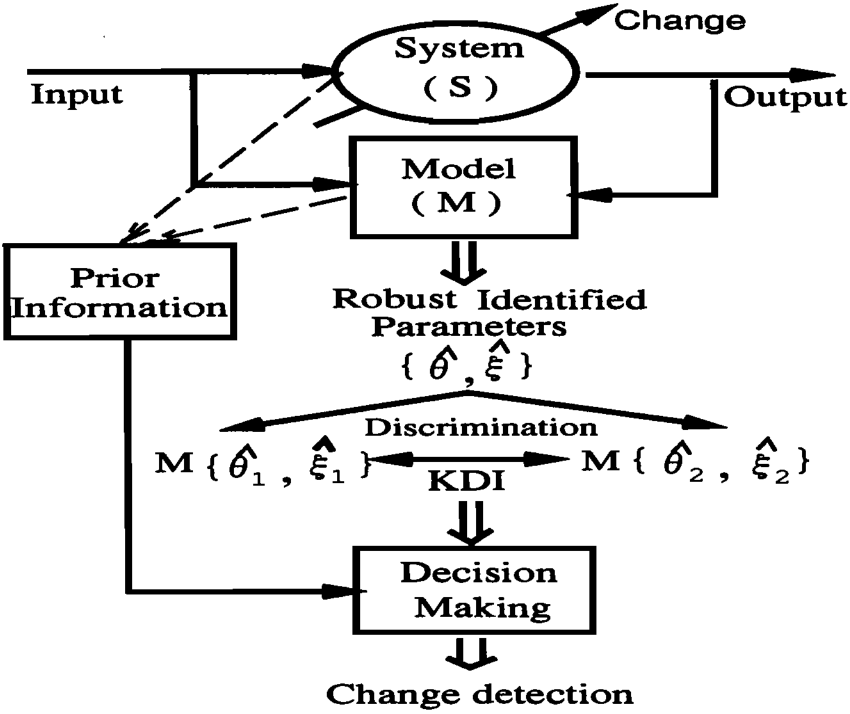


Figure 1 Structure of the change detection system (CDS).

where $y(t)$ is the output at time t $(t = 1, 2, \ldots)$, $u(t)$ the input, and $e_0(t)$ the white Gaussian noise. $G_0(q^{-1})$, $H_0(q^{-1})$ are scalar rational functions in the backward shift operator q^{-1}. Without loss of generality, we will assume that $G_0(q^{-1})$ and $H_0(q^{-1})$ meet the following conditions:

- $G_0(q^{-1})$, $H_0(q^{-1})$ are causal
- $G_0(0) = 0$, $H_0(0) = 1$
- $G_0(q^{-1})$, $H_0(q^{-1})$ are asymptotically stable.

For identification, we use the parametric models of the form

$$\begin{aligned} \mathcal{M}(\Theta):\quad & y(t) = G(q^{-1}, \Theta)u(t) + H(q^{-1}, \Theta)e(t) \\ & e(t) \in N(0, \sigma^2) \end{aligned} \tag{2}$$

where Θ and σ^2 denote unknown parameters.

Assume then that data from the system are available from two disjoint time intervals I_1 and I_2, see Fig. 2, and denote the estimates of the unknown parameter vector Θ using the data obtained from these intervals

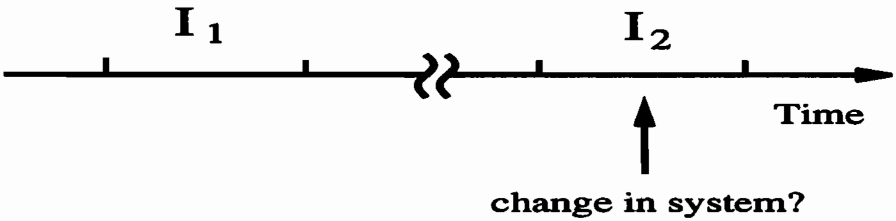


Figure 2 Data from two disjoint time intervals.

by $\hat{\Theta}_1$ and $\hat{\Theta}_2$, respectively. The discrimination of identified models $\mathcal{M}\{\hat{\Theta}_1\}$ and $\mathcal{M}\{\hat{\Theta}_2\}$ can be executed via the KDI.

Applying the KDI to the likelihood functions of the identified models, $p(Y_{t+1}|\hat{\Theta}_i, U_t)$, $(i = 1, 2)$, we get

$$I_{t+1}[1,2] = \int p(Y_{t+1}|\hat{\Theta}_1, U_t) \log \frac{p(Y_{t+1}|\hat{\Theta}_1, U_t)}{p(Y_{t+1}|\hat{\Theta}_2, U_t)} \mathrm{d}Y_{t+1} \tag{3}$$

where $U_t = [u(1)\, u(2) \cdots u(t)]^T$ and $Y_{t+1} = [y(1)\, y(2) \cdots y(t+1)]^T$ are the input–output data sets from the interval I_1. The index in (3) hence indicates how well the model using $\hat{\Theta}_2$ describes the data in the interval I_1. In other words the criterion $I_N[1,2]$ is an index for discriminating the models $\mathcal{M}\{\hat{\Theta}_1\}$ and $\mathcal{M}\{\hat{\Theta}_2\}$ via the difference in the corresponding likelihood functions.

Now applying Bayesian analysis of the likelihood function to (3) based on the Gaussian assumption of $e(t)$ in (2), we can derive the following expression of the KDI, $I_N[1,2]$ (see [7,10])

$$I_N[1,2] = \sum_{i=0}^{N-1} I_{i+1}[1,2] = I_N^{(1)}[1,2] + I_N^{(2)}[1,2] + I_N^{(3)}[1,2] \tag{4}$$

where

$$I_N^{(1)}[1,2] = \frac{N}{2}[\hat{\sigma}_1^2/\hat{\sigma}_2^2 - 1) - \log(\hat{\sigma}_1^2/\hat{\sigma}_2^2)] \tag{5}$$

$$I_N^{(2)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1} \|H_2^{-1}(G_1 - G_2)u(k+1)\|^2_{\hat{\sigma}_2^{-2}} \tag{6}$$

$$I_N^{(3)}[1,2] = \frac{N}{2}\left\{\frac{\hat{\sigma}_2^{-2}}{2\pi i} \oint (H_2^{-1}(z)H_1(z) - 1)\hat{\sigma}_1^2(H_1(z^{-1})\right.$$

$$\left. H_2^{-1}(z^{-1}) - 1)\frac{\mathrm{d}z}{z}\right\} \tag{7}$$

$$G_i = G(q^{-1}, \hat{\Theta}_i), \qquad H_i(z) = H(z, \hat{\Theta}_i), \qquad i = 1, 2 \tag{8}$$

where N is the length of data sets in the interval I_1. The obtained KDI (4) consists of three terms, all of which are nonnegative. The first term (5) expresses the deviation of $\hat{\sigma}_2^2$ from $\hat{\sigma}_1^2$; the second term (6) expresses the difference between G_2 and G_1; the third term (7) describes the difference between H_2 and H_1. Thus the difference between the identified models can be evaluated by the KDI $I_N[1,2]$ in a feasible way for finite but fairly large data sets N.

We consider the case where $\hat{\Theta}$ represents the true system parameter $\mathcal{P}$, namely we assume that $\hat{\Theta}$ does converge towards the true parameter $\mathcal{P}$ via some identification algorithm. In such an ideal case where there is no unmodeled dynamics, the distortion of $\mathcal{M}\{\hat{\Theta}_1\}$ and $\mathcal{M}\{\hat{\Theta}_2\}$ implies a change of system parameters which may be caused by a system failure. This change can thus be detected by applying the thresholding approach to the KDI

$$I_N[1,2] \gtrless \eta \Rightarrow \begin{cases} \text{change} \\ \text{no change} \end{cases} \tag{9}$$

The threshold η is usually determined according to the false-alarm rate, in which the statistical properties of $I_N[1,2]$ provide useful information. The interested reader is referred to [10,7] for the details of the statistical properties.

III. CHANGE DETECTION WITH MODEL UNCERTAINTY

The method described in the previous section is based on one important assumption: there is no model uncertainty in the identification. However, in practical applications, model uncertainties are inevitable for the following reasons: (1) the system is usually unknown; (2) a low-order model is often used for convenience. Therefore it is crucial and highly motivated to extend the method to the case where there exists model uncertainty.

To deal with model uncertainty, we assume that the plant is described by

$$G_0(q^{-1}) = G(q^{-1},\Theta) + G_\Delta(q^{-1}) \tag{10}$$

where $G_\Delta(q^{-1})$ denotes unmodeled dynamics. Here we do not introduce the unmodeled dynamics into the noise model, so that $H_0(q^{-1}) = H(q^{-1},\Theta)$. The filters $G(q^{-1},\Theta)$ and $H(q^{-1},\Theta)$ are scalar rational functions in the backward shift operator q^{-1} with appropriate orders. Then the model (2) considering model uncertainty has the form

$$\begin{aligned} \mathcal{M}\colon\quad & y(t) = G(q^{-1},\Theta)u(t) + G_\Delta(q^{-1})u(t) + H(q^{-1},\Theta)e(t) \\ & e(t) \in N(0,\sigma^2) \end{aligned} \tag{11}$$

The identification of model (11) depends on the form of $G_\Delta(q^{-1})$. In the robust identification literature, two types of bounds are considered for $G_\Delta(q^{-1})$; soft bound and hard bound. As mentioned in Section I, we will consider two extensions in this section. One is based on a soft bound approach, while the other is based on a nonbound approach. For clarity in discussion, our extensions will be constrained to specified models. The principles may, however, be applied to other models.

A. Extension Based on a *Soft Bound* Approach

We consider a discrete-time linear SISO ARMAX system described by

$$\mathcal{S}:\quad A_0(q^{-1})y(t) = B_0(q^{-1})u(t) + C_0(q^{-1})e_0(t)$$
$$e_0(t) \in N(0, \sigma_0^2) \tag{12}$$

where $A_0(q^{-1})$, $B_0(q^{-1})$, and $C_0(q^{-1})$ are scalar polynomials in the backward shift operator q^{-1} with appropriate orders. We assume that A_0, B_0, and C_0 are described by

$$\begin{aligned} A_0(q^{-1}) &= A(q^{-1},\theta) + A_\Delta(q^{-1}) \\ B_0(q^{-1}) &= B(q^{-1},\theta) + B_\Delta(q^{-1}) \\ C_0(q^{-1}) &= C(q^{-1},\xi) \end{aligned} \tag{13}$$

where $A_\Delta(q^{-1})$ and $B_\Delta(q^{-1})$ denote unmodeled dynamics. The $A(q^{-1},\theta)$, $B(q^{-1},\theta)$, and $C(q^{-1},\xi)$ are scalar polynomials in the backward shift operator q^{-1} with orders n, m, and l, respectively

$$\begin{aligned} A(q^{-1},\theta) &= 1 + a_1q^{-1} + \cdots + a_nq^{-n} \\ B(q^{-1},\theta) &= b_1q^{-1} + \cdots + b_mq^{-m} \\ C(q^{-1},\xi) &= 1 + c_1q^{-1} + \cdots + c_lq^{-1} \end{aligned} \tag{14}$$

Now according to the basic idea of Goodwin's stochastic embedding approach [9], a stochastic prior model may be introduced in order to define the distribution of unmodeled dynamics

$$A_\Delta(q^{-1}) = \sum_{i=1}^{L} \eta_{ai}q^{-i} \qquad B_\Delta(q^{-1}) = \sum_{i=1}^{M} \eta_{bi}q^{-i} \tag{15}$$

$$\eta = [\eta_{a1} \cdots \eta_{aL}\, \eta_{b1} \cdots \eta_{bM}]^T \in N(0, C_\eta) \tag{16}$$

where η_{ai} $(i = 1, \ldots, L)$ and η_{bi} $(i = 1, \ldots, M)$ are assumed to be mutually independent.

It therefore follows that the parametric model is given by

$$\mathcal{M}(\theta,\xi):\quad A(q^{-1},\theta)y(t) = B(q^{-1},\theta)u(t) + \omega(t) \tag{17}$$

$$\omega(t) = \psi(t)\eta + C(q^{-1},\xi)e(t) \tag{18}$$

where

$$\begin{aligned} \theta &= [a_1 \cdots a_n \, b_1 \cdots b_m]^T \\ \xi &= [\mathrm{diag}\{C_\eta\} \, \sigma^2 \, c_1 \cdots c_l]^T \\ \psi(t) &= [-y(t-1) \cdots -y(t-L) \, u(t-1) \cdots u(t-M)] \end{aligned} \tag{19}$$

Comparing the model (17)–(18) with the general model (11), we see that this accounts for the case of

$$G(q^{-1}, \Theta) = B(q^{-1}, \theta)/A(q^{-1}, \theta) \tag{20}$$

$$G_\Delta(q^{-1}) u(t) = \psi(t) \eta / A(q^{-1}, \theta) \tag{21}$$

$$H(q^{-1}, \Theta) = C(q^{-1}, \xi)/A(q^{-1}, \theta) \tag{22}$$

1. Identification of the Model

Since the term $\omega(t)$ in (17) is a Gaussian correlated noise, it may be treated as an MA noise process. In this way, the vector θ can be estimated by using the prediction error method (PEM), while the vector ξ can be estimated by applying the maximum likelihood (ML) method in (18) with the data $\hat{\omega}(t)$ calculated from $\hat{\theta}$.

In order to identify C_η, let us introduce the following parametrization similar to [9]:

$$E\{\eta_{ai}^2\} = \alpha_a(\lambda_a)^i \qquad E\{\eta_{bi}^2\} = \alpha_b(\lambda_b)^i \tag{23}$$

The parameter ξ thus becomes

$$\xi = [\alpha_a, \lambda_a, \alpha_b, \lambda_b, \sigma^2, c_1, \ldots, c_l]^T \tag{24}$$

Now introduce a vector defined by

$$W = [\hat{\omega}(1) \, \hat{\omega}(2) \cdots \hat{\omega}(N)]^T \tag{25}$$

Then the corresponding likelihood function $p(W|\xi)$ is subject to $N(0, \Sigma)$ because of the assumptions of (12) and (16). The estimation of ξ can be obtained by maximizing the log-likelihood function $l(W|\xi)$

$$\hat{\xi} = \arg \mathrm{Max}\{l(W|\xi)\} \tag{26}$$

where

$$l(W|\xi) = -\frac{1}{2} \log \det \Sigma - \frac{1}{2} W^T \Sigma^{-1} W + \mathrm{const} \tag{27}$$

with

$$\begin{aligned}\Sigma &= \Psi C_\eta \Psi^T + E\{VV^T\}\\ V &= [\nu_1, \nu_2 \cdots \nu_N]^T\\ \Psi &= [\psi(1)^T \psi(2)^T \cdots \psi(N)^T]^T\end{aligned} \tag{28}$$

and $\nu_t = C(q^{-1}, \xi)e(t)$.

When optimization-based methods such as PEM and ML are used, the parameter estimates are determined as the global optimal point of the criterion function. However, there is a potential risk that an optimization-based method is stuck at a local optimum when the criterion function is not unimodal [12]. Moreover, the risk has been found to increase in the presence of unmodeled dynamics [13]. This problem can be solved by using our proposed robust system identification method obtained by combining a genetic algorithm (GA) with an optimization-based method. The robust identification is executed by the optimization-based method using a good initial value searched by a nonstandard GA (NSGA). The NSGA evolves not only in a probabilistic manner (e.g., crossover and mutation) but also in a nonprobabilistic manner (e.g., development). Figure 3 shows the basic idea, see [14] for more details.

2. Calculation of the KDI

To calculate the KDI, we follow the procedure of the previous section. Here the unknown parameter vector is $\{\theta, \xi\}$ whose estimates from the two disjoint intervals I_1 and I_2 will be denoted as $\{\hat{\theta}_1\, \hat{\xi}_1\}$ and $\{\hat{\theta}_2\, \hat{\xi}_2\}$, respectively. The KDI is then used to detect the distortion of identified models $M\{\hat{\theta}_1\, \hat{\xi}_1\}$ and $M\{\hat{\theta}_2\, \hat{\xi}_2\}$.

Applying the KDI to the likelihood functions of the identified models, $p(Y_{t+1} | \theta_i, \xi_t, U_t)$ $(i = 1, 2)$, we get

$$I_{t+1}[1,2] = \int p(Y_{t+1} | \hat{\theta}_1, \hat{\xi}_1, U_t) \log \frac{p(Y_{t+1} | \hat{\theta}_1, \hat{\xi}_1, U_t)}{p(Y_{t+1} | \hat{\theta}_2, \hat{\xi}_2, U_t)} \mathrm{d}Y_{t+1} \tag{29}$$

where $U_t = [u(1)\, u(2) \cdots u(t)]^T$ and $Y_{t+1} = [y(1)\, y(2) \cdots y(t+1)]^T$ are the input–output data sets from interval I_1. Because of the Gaussian assumptions of (12) and (16), the likelihood functions $p(Y_{t+1} | \hat{\theta}_i, \hat{\xi}_t, U_t)$ are Gaussian distributed. Then based on Bayesian analysis, the KDI (29) can again be developed into an explicit form, see [15,13,16] for details,

$$I_N[1,2] = I_N^{(1)}[1,2] + I_N^{(2)}[1,2] + I_N^{(3)}[1,2] + I_N^{(4)}[1,2] \tag{30}$$

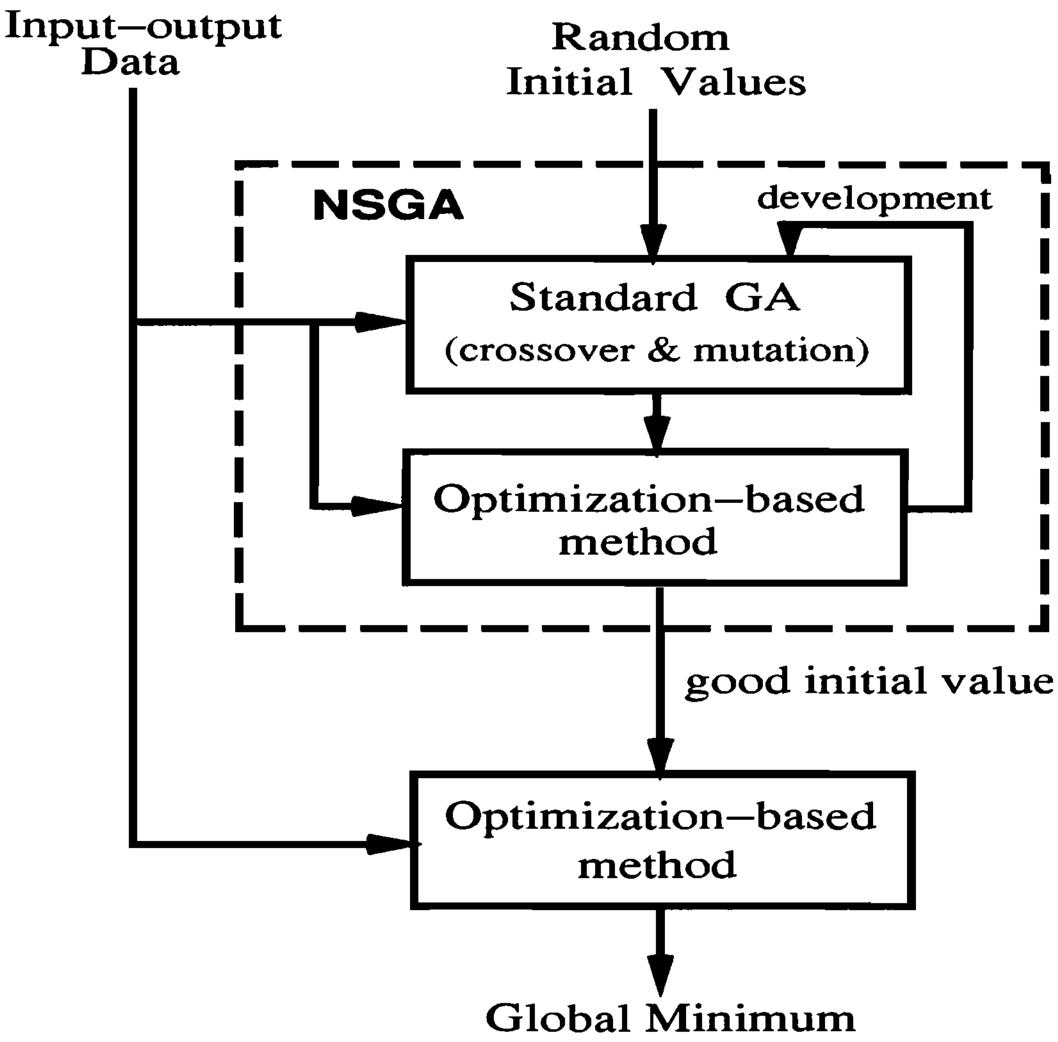


Figure 3 Basic idea of the robust identification method based on an effective combination of the GA and an optimization-based method.

where

$$I_N^{(1)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1}(\hat{\Sigma}_2(k)^{-1}\hat{\Sigma}_1(k) - 1) - \frac{1}{2}\sum_{k=0}^{N-1}\log(\hat{\Sigma}_1(k)/\hat{\Sigma}_2(k)) \tag{31}$$

$$I_N^{(2)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1}\|H_2^{-1}(G_1 - G_2)u(k+1)\|^2_{\hat{\Sigma}_2(k)^{-1}} \tag{32}$$

$$\begin{aligned} I_N^{(3)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1}\Bigg\{\hat{\Sigma}_2(k)^{-1}\frac{1}{2\pi \mathrm{i}}\oint (H_2^{-1}(z)H_1(z) - 1) \\ \times \hat{\sigma}_1^2(H_1(z^{-1})H_2^{-1}(z^{-1}) - 1)\frac{\mathrm{d}z}{z}\Bigg\} \end{aligned} \tag{33}$$

$$I_N^{(4)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1}\left\{\hat{\Sigma}_2(k)^{-1}\frac{1}{2\pi\mathrm{i}}\oint (H_2^{-1}(z) - H_1^{-1}(z)) \right.$$
$$\left. \times Q_{k+1,1}\hat{C}_{\eta 1}Q_{k+1,1}^T(H_2^{-1}(z^{-1}) - H_1^{-1}(z^{-1}))\frac{\mathrm{d}z}{z}\right\} \tag{34}$$

and where

$$\hat{\Sigma}_\iota(k) = H_\iota^{-1}Q_{k+1,i}\hat{C}_{\eta i}Q_{k+1,i}^T H_i^{-1} + \hat{\sigma}_i^2 \tag{35}$$

$$G_i = \frac{B(q^{-1}, \hat{\theta}_\iota)}{A(q^{-1}, \hat{\theta}_i)} \qquad H_\iota = \frac{C(q^{-1}, \hat{\xi}_i)}{A(q^{-1}, \hat{\theta}_i)} \qquad Q_{k,i} = \frac{\psi(k)}{A(q^{-1}, \hat{\theta}_\iota)} \tag{36}$$

Note that an extra term (34) is induced from unmodeled dynamics and is related mainly to the difference between H_2 and H_1. Moreover, in this case unmodeled dynamics affects all the terms in (31)–(34).

As an index for evaluating model uncertainty, we introduce

$$I_{C_\eta}[j] = \frac{1}{N}\sum_{k=1}^{N}\Delta\hat{\Sigma}_j(k) = \frac{1}{N}\sum_{k=1}^{N}(H_j^{-1}Q_{k,j}\hat{C}_{\eta j}Q_{k,j}^T H_j^{-1}) \tag{37}$$

From the expressions of H_j and $Q_{k,j}$, we know that this index depends on the unmodeled dynamics only and is not related to θ.

B. Extension Based on a *Nonbound* Approach

In the previous subsection, we borrowed the idea of a stochastic description of unmodeled dynamics and incorporated it into the KDI. However, in the case of black-box modeling, a comparatively simple model is occasionally used for convenience. The unmodeled dynamics resulting from such a simple model will contain some characteristics, which can no longer be described by the stationary stochastic process model. This fact motivates us to develop an alternative approach for describing the unmodeled dynamics.

Note that if we constrain our discussion to the cases where data records are long enough for the estimation, the modeling can be achieved with desired accuracy by merely increasing the model complexity. Moreover, if a complex model is chosen so that it includes the simple model as its submodel, the unmodeled dynamics may be described by using the *difference* between the simple model and the complex model. Because this error description does not belong to a bound approach, we here call it nonbound approach.

Figure 4 shows the basic idea of the nonbound approach, in which the system is represented by a *complex model* $\mathcal{M}\{\theta, \xi\}$ with desired accuracy and it includes a *simple model* $\mathcal{M}_s\{\theta\}$ as its submodel. Obviously, the

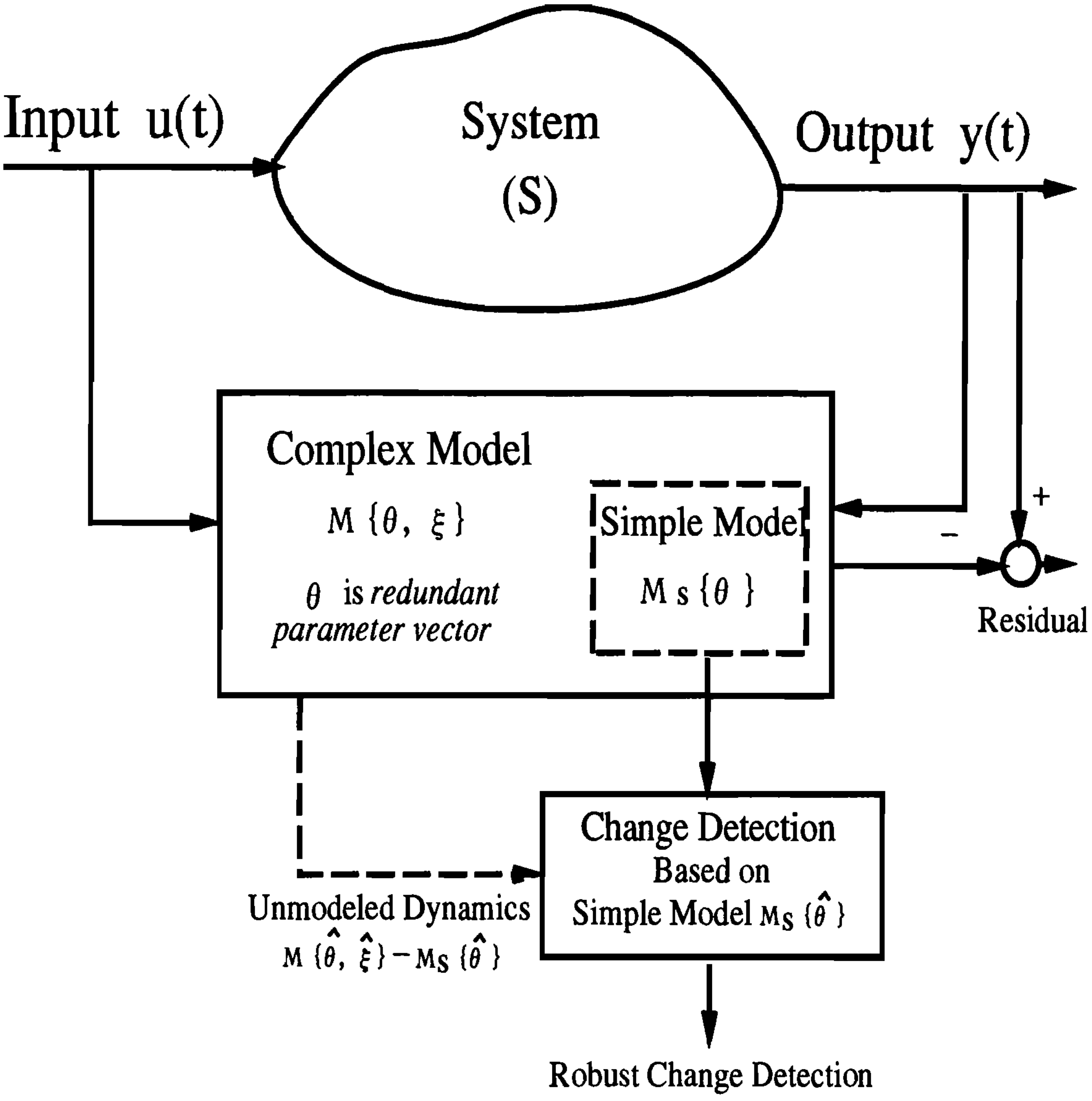


Figure 4 A nonbound approach based on hierarchical modeling.

residual generated from it can be used for change detection of the system. Many existing change detection methods are based on the residual, see, e.g., [1]. On the other hand, the simple model $\mathcal{M}_s\{\theta\}$ embedded in $\mathcal{M}\{\theta, \xi\}$ will be identified such that its parameter vector θ describes the main characteristics of the system. In this way, the KDI-based robust change detection can be applied to the simple model, in which the unmodeled dynamics is estimated by the nonbound approach.

In order to realize the above robust change detection, we should consider two key conditions. The first is that the common parameter vector θ of the simple model and the complex model should be chosen to be a *redundant parameter vector.** The other is that the identification of $\mathcal{M}\{\hat{\theta}, \hat{\xi}\}$ should be carried out so that the obtained results $\{\hat{\theta}, \hat{\xi}\}$ can give the identification of $\mathcal{M}_s\{\hat{\theta}\}$ as well as the estimates of the unmodeled dynamics. In the following subsections, we will propose a hierarchical modeling and identification algorithm for this purpose.

1. Linear Case: ARMAX Model

In order to make the nonbound approach easily understood, we first discuss the linear case where the complex model is simply a high-order linear ARMAX model.

Consider a system described by (12) where the degrees of $A_0(q^{-1})$, $B_0(q^{-1})$, and $C_0(q^{-1})$ are assumed to be n_0, m_0, and l_0, respectively. When the system is identified, the model used is assumed to have the form:

$$\mathcal{M}\text{:}\quad \bar{A}(q^{-1},\theta,\xi)y(t) = \bar{B}(q^{-1},\theta,\xi)u(t) + \bar{C}(q^{-1},\theta,\xi)e(t)$$
$$e(t) \in N(0,\sigma^2) \tag{38}$$

where $\bar{A}(q^{-1},\theta,\xi)$, $\bar{B}(q^{-1},\theta,\xi)$, $\bar{C}(q^{-1},\theta,\xi)$ are scalar polynomial functions in q^{-1} with orders n_d, m_d and l, respectively.

$$\begin{aligned}
\bar{A}(q^{-1},\theta,\xi) &= 1+\bar{a}_1 q^{-1} + \cdots + \bar{a}_{n_d} q^{-n_d} \\
\bar{B}(q^{-1},\theta,\xi) &= \bar{b}_1 \text{q}^{-1} + \cdots + \bar{b}_{m_d} q^{-m_d} \\
\bar{C}(q^{-1},\theta,\xi) &= 1+\bar{c}_1 q^{-1} + \cdots + \bar{c}_l q^{-l}
\end{aligned} \tag{39}$$

Here the order of the model (n_d, m_d, and l) will be chosen as close to that of the system (n_0, m_0, and l_0) as possible to achieve the expected modeling accuracy. In order to conduct a hierarchical modeling, we assume that $\bar{a}_i$, $\bar{b}_i$, and $\bar{c}_i$ in (39) consist of two parts

$$\bar{a}_i = \begin{cases} a_i + \delta_{a_i} & 1 \le i \le n \\ \delta_{a_i} & n < i \le n_d \end{cases} \tag{40}$$

$$\bar{b}_i = \begin{cases} b_i + \delta_{b_i} & 1 \le i \le m \\ \delta_{b_i} & m < i \le m_d \end{cases} \tag{41}$$

$$\bar{c}_i = c_i + \delta_{c_i} \tag{42}$$

and denote θ and ξ as

$$\begin{aligned}
\theta &= [a_1 \cdots a_n, b_1 \cdots b_m, c_1 \cdots c_l]^T \\
\xi &= [\delta_{a_1} \cdots \delta_{a_{n_d}}, \delta_{b_1} \cdots \delta_{b_{m_d}}, \delta_{c_1} \cdots \delta_{c_l}]^T
\end{aligned} \tag{43}$$

***Redundant parameter vector*: Let θ and ξ be two parameter vectors of the complex model $\mathcal{M}\{\theta,\xi\}$. The vector θ is said to be a redundant parameter vector if for any given θ, ξ, and θ_0, there exists a ξ' that $\mathcal{M}\{\theta,\xi\} \equiv \mathcal{M}\{\theta_0,\xi'\}$ holds.

Now consider θ to be the parameter vector of another ARMAX model $\mathcal{M}_s\{\theta\}$, that is, the *simple model*, whose order is described by n, m, and l. Here $n \le n_d$ and $m \le m_d$ will be chosen so that the simple model is of low order and the parameters of the simple model have useful physical interpretations. Moreover, it is clear that θ introduced here is a redundant parameter vector of model (38).

Introduce the notation:

$$G = B(q^{-1}, \theta)/A(q^{-1}, \theta) \qquad H = \bar{C}(q^{-1}, \theta, \xi)/A(q^{-1}, \theta) \tag{44}$$

$$\Delta G_t = \frac{-A_\Delta(q^{-1}, \xi)y(t) + B_\Delta(q^{-1}, \xi)u(t)}{A(q^{-1}, \theta)u(t)} \tag{45}$$

where $A(q^{-1}, \theta)$, $B(q^{-1}, \theta)$ have the forms described in (14) and

$$\begin{aligned} A_\Delta(q^{-1}, \xi) &= \delta_{a_1} q^{-1} + \cdots + \delta_{a_{nd}} q^{-n_d} \\ B_\Delta(q^{-1}, \xi) &= \delta_{b_1} q^{-1} + \cdots + \delta_{b_{md}} q^{-m_d} \end{aligned} \tag{46}$$

Then we may rewrite the model (38) as

$$y(t) = Gu(t) + \Delta G_t u(t) + He(t) \tag{47}$$

2. Nonlinear Case: Quasi-ARMAX Model

Consider SISO nonlinear discrete-time systems which can be described by the general Kolmogorov–Gabor [17] polynomial of appropriate order:

$$\begin{aligned} y(t) = &\sum_{i=1}^{\gamma} \alpha_i x_i(t) + \sum_{i=1}^{\gamma} \sum_{j=1}^{\gamma} \alpha_{ij} x_i(t) x_j(t) \\ &+ \sum_{i=1}^{\gamma} \sum_{j=1}^{\gamma} \sum_{l=1}^{\gamma} \alpha_{ijl} x_i(t) x_j(t) x_l(t) + \cdots \end{aligned} \tag{48}$$

in which the elements of $[x_i(t) \quad i = 1, \ldots, \gamma]$ are assumed to be past input–outputs of the system,

$$\left.\begin{aligned} &x_i(t) = y(t-i); \quad i = 1, \ldots, n \\ &x_{j+n}(t) = u(t-j); \quad j = 1, \ldots, m \\ &(\gamma = n + m) \end{aligned}\right\} \tag{49}$$

Then from (48) and (49), we can get

$$y(t) = \sum_{i=1}^{n} \bar{a}_{i,t} y(t-i) + \sum_{i=1}^{m} \bar{b}_{i,t} u(t-i) \tag{50}$$

where

$$\bar{a}_{i,t} = a_i + \Delta a_{i,t} \qquad \bar{b}_{i,t} = b_i + \Delta b_{i,t} \tag{51}$$

$$\Delta a_{i,t} = \sum_{j=1}^{n} \alpha_{ij} y(t-j) + \frac{1}{2} \sum_{j=n+1}^{\gamma} \alpha_{ij} u(t-j+n) + \cdots \tag{52}$$

$$\Delta b_{i,t} = \sum_{j=n+1}^{\gamma} \alpha_{ij} u(t-j+n) + \frac{1}{2} \sum_{j=1}^{n} \alpha_{ij} y(t-j) + \cdots \tag{53}$$

In (51), the α_i has been replaced by a_i and b_i for simplicity. When a moving average (MA) noise model is used, we have a form of quasi-ARMAX and name it a quasi-ARMAX model

$$y(t) = \sum_{i=1}^{n} \bar{a}_{i,t} y(t-i) + \sum_{i=1}^{m} \bar{b}_{i,t} u(t-i) + \sum_{i=1}^{l} c_i e(t-i) + e(t) \tag{54}$$

The quasi-ARMAX model (54) can be considered as a generalized ARMAX model, whose coefficients consist of two parts: (1) the constant parameters a_i and b_i describing the linear part, and (2) the complicated functions $\Delta a_{i,k}$ and $\Delta b_{i,k}$ denoting the nonlinear part.

(a) Hybrid Quasi-ARMAX Model. Nonlinear nonparametric models have attracted much interest due to their ability to encompass truly nonlinear behavior of dynamical systems. However, they do not have a "good" structure easily used for control design or system analysis. On the other hand, a linear ARMAX model has a "good" structure, but it does not have the ability to encompass nonlinear behavior. Due to these complementary properties of a linear ARMAX model and a nonlinear nonparametric model, a model consisting of an effective combination of a linear ARMAX model and nonlinear nonparametric models may have a "good" structure and flexibility as well. Based on this motivation, we will propose such a hybrid model by representing the nonlinear terms ($\Delta a_{i,t}$ and $\Delta b_{i,t}$) in the quasi-ARMAX model (54) using a set of nonlinear nonparametric models.

From (52), (53), we know that as the order of the system nonlinearity increases, the $\Delta a_{i,t}$ and $\Delta b_{i,t}$ may become very complicated. Hence it is difficult to represent them using explicit functional forms. Nonlinear nonparametric models will however be good candidates for representing $\Delta a_{i,t}$ and $\Delta b_{i,t}$. Let

$$\begin{aligned} \Delta a_{i,t} &= f_{i,t} \quad (i = 1, \ldots, n) \\ \Delta b_{j,t} &= f_{j+n,t} \quad (j = 1, \ldots, m) \end{aligned} \tag{55}$$

where $f_{i,t}$ ($i = 1, \ldots, n + m$) are realized using nonlinear nonparametric models described by

$$f_{i,t} = \sum_{j=1}^{M} \omega_{ij} \mathcal{N}_f(p_{ij}, X(t)) \tag{56}$$

where $\mathcal{N}_f(x)$ are the "basis functions," ω_{ij} and p_{ij} are the parameters of the nonparametric model, and $X(t) = [u(t-1) \cdots u(t-\tilde{m}) y(t-1) \cdots y(t-\tilde{n})]$ ($\tilde{n} \leq n, \tilde{m} \leq m$) is its input vector. Note here that the coordinate parameters ω_{ij} are to be estimated, while the scale and position parameter vectors p_{ij} are determined based on knowledge information.

The hybrid model described by (54)–(56) is formed by an effective combination of a linear ARMAX part and certain nonlinear nonparametric models and is named a *hybrid quasi-ARMAX model*. It has the following distinctive features [18,19].

1. *Structure* It has a "good" structure which is similar to a linear ARMAX model. This linear structure may be taken as an advantage for dealing with nonlinear systems under the framework of linear system theory.
2. *Flexibility* Since it incorporates a group of certain nonlinear nonparametric models into the coefficients, it has flexibility to describe general nonlinear systems.
3. *Simplicity* By using knowledge information efficiently for determining p_{ij}, the hybrid quasi-ARMAX model is built to be linear in the parameters to be estimated. It thus has simplicity for estimation.

In this section, we will apply the hybrid quasi-ARMAX model to robust change detection. For simplicity, we rewrite it as

$$\begin{aligned} \mathcal{M}\text{:} \quad \bar{A}(q^{-1}, t, \theta, \xi) y(t) &= \bar{B}(q^{-1}, t, \theta, \xi) u(t) + \bar{C}(q^{-1}, \theta, \xi) e(t) \\ e(t) &\in N(0, \sigma^2) \end{aligned} \tag{57}$$

where $\bar{A}(q^{-1}, t, \theta, \xi)$, $\bar{B}(q^{-1}, t, \theta, \xi)$, and $\bar{C}(q^{-1}, \theta, \xi)$ are scalar polynomial function in q^{-1} with orders n, m, and l, respectively

$$\begin{aligned} \bar{A}(q^{-1}, t, \theta, \xi) &= 1 + \bar{a}_{1,t}\mathrm{q}^{-1} + \cdots \bar{a}_{n,t}q^{-n} \\ \bar{B}(q^{-1}, t, \theta, \xi) &= \bar{b}_{1,t}q^{-1} + \cdots + \bar{b}_{m,t}q^{-m} \\ \bar{C}(q^{-1}, \theta, \xi) &= 1 + \bar{c}_1 q^{-1} + \cdots + \bar{c}_l q^{-l} \end{aligned} \tag{58}$$

The coefficients of the polynomials (58) contain two parts: constant parameters and nonlinear terms. (Note: we do not consider the nonlinearity in the noise model. However, $\overline{c}_i$ is assumed to have two parts so that the vector θ is a redundant parameter vector of model (57).)

$$\begin{aligned}\overline{a}_{i,t} &= a_i + f_{i,t}\\ \overline{b}_{i,t} &= b_i + f_{i+n,t}\\ \overline{c}_i &= c_i + \delta_{c_i}\end{aligned} \tag{59}$$

where a_i, b_i are constant parameters and $f_{i,t}$ $(i = 1, \ldots, n+m)$ are nonlinear functions which will be represented using a set of nonlinear nonparametric models described by (56). For example, if an adapative fuzzy system is used, we have

$$f_{i,t} = \frac{\sum_{j=1}^{M} \omega_{ij}(\wedge_{k=1}^{r} \mu_{A_k^j}(x_k(t)))}{\sum_{j=1}^{M} (\wedge_{k=1}^{r} \mu_{A_k^j}(x_k(t)))} \tag{60}$$

where $\wedge$ is the minimum operator, $\mu_{A_k^j}$ is the membership function of the fuzzy set A_k^j, M is the number of rules, and ω_{ij} are adjustable parameters. In this way, the parameters to be estimated for the hybrid quasi-ARMAX model are defined as follows:

$$\begin{aligned}\theta &= [a_1 \cdots a_n\, b_1 \cdots b_m\, c_1 \cdots c_l]^T\\ \xi &= [\omega_{11} \cdots \omega_{n1} \cdots \omega_{(n+m)1} \cdots \omega_{(n+m)M}\, \delta_{c_1} \cdots \delta_{c_l}]^T\end{aligned} \tag{61}$$

Note that it is easy to show that the parameter θ defined in (60) and (61) is a redundant parameter vector of $\mathcal{M}\{\theta, \xi\}$. Now if we introduce the following notation

$$G = B(q^{-1}, \theta)/A(q^{-1}, \theta) \qquad H = \bar{C}(q^{-1}, \theta, \xi)/A(q^{-1}, \theta) \tag{62}$$

$$\Delta G_t = \frac{-A_\Delta(q^{-1}, t, \xi)\, y(t) + B_\Delta(q^{-1}, t, \xi)\, u(t)}{A(q^{-1}, \theta)\, u(t)} \tag{63}$$

where

$$\begin{aligned}A_\Delta(q^{-1}, t, \xi) &= f_{1,t} q^{-1} + \cdots + f_{n,t} q^{-n}\\ B_\Delta(q^{-1}, t, \xi) &= f_{(1+n),t} q^{-1} + \cdots + f_{(m+n),t} q^{-m}\end{aligned} \tag{64}$$

then the hybrid quasi-ARMAX model can be rewritten in the same form as (47).

The model (47) is similar in form to the general linear model (11). Therefore, identification based on the hybrid quasi-ARMAX model may be explained as identifying the nonlinear system using a linear ARMAX model $\mathcal{M}\{\hat{\theta}\}$, while the unmodeled dynamics is described by a set of nonparametric models $\Delta G_t(\hat{\theta}, \hat{\xi})$.

3. Identification of the Model

The θ and ξ are two parameter vectors of the complex model $\mathcal{M}\{\theta, \xi\}$. As discussed above, the identification should be done such that the estimates $\hat{\theta}, \hat{\xi}$ have the interpretation that $\hat{\theta}$ describes the linear ARMAX model, while $\hat{\xi}$ describes the unmodeled dynamics.

Note that the fact that the vector θ is a redundant parameter vector of the complex model $\mathcal{M}\{\theta, \xi\}$, that is, for any given θ, ξ, and θ_0, there exists a ξ' such that $\mathcal{M}\{\theta, \xi\} \equiv \mathcal{M}\{\theta_0, \xi'\}$ holds. Therefore, the identification can be implemented in the following two steps.

Step 1. Estimation of θ

Let $\xi = 0$, then the models (38) and (57) become

$$A(q^{-1}, \theta)y(t) = B(q^{-1}, \theta)u(t) + C(q^{-1}, \theta)e(t) \tag{65}$$

The estimate $\hat{\theta}$ can thus be obtained by matching this linear ARMAX model to the input–output data of the system.

Step 2. Estimation of ξ

Set $\theta = \hat{\theta}$ as constant, and use $\xi = 0$ as initial value. The vector ξ is estimated by applying the prediction error method (PEM) to the model (38) or (57).

Although the models (38), (57), and (65) are linear in the parameters, the criteria in the identifications are not always unimodal because there exists an MA noise model $C(q^{-1}, \theta)e(t)$. To prevent the estimations being stuck in a local minimum, the identifications may be carried out using a guaranteed identification algorithm obtained by combining the genetic algorithm with an optimization-based method [14].

4. Calculation of the KDI

Corresponding to the model (38) and (57), the KDI (29) can be converted into an explicit form based on the assumption of Gaussian distribution and Bayesian analysis. The result is

$$I_N[1,2] = I_N^{(1)}[1,2] + I_N^{(2)}[1,2] + I_N^{(3)}[1,2] + I_N^{(4)}[1,2] \tag{66}$$

where

$$I_N^{(1)}[1,2] = \frac{N}{2}[(\hat{\sigma}_1^2/\hat{\sigma}_2^2 - 1) - \log(\hat{\sigma}_1^2/\hat{\sigma}_2^2)] \tag{67}$$

$$I_N^{(2)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1} \| H_2^{-1}(G_1 - G_2)u(k+1)\|^2_{\hat{\sigma}_2^{-2}} \tag{68}$$

$$\begin{aligned} I_N^{(3)}[1,2] = \frac{N}{2}\left\{\frac{\hat{\sigma}_2^{-2}}{2\pi \mathrm{i}} \oint (H_2^{-1}(z)H_1(z) - 1)\hat{\sigma}_1^2(H_1(z^{-1}) \cdot \right. \\ \left. \times H_2^{-1}(z^{-1}) - 1)\frac{\mathrm{d}z}{z}\right\} \end{aligned} \tag{69}$$

$$\begin{aligned} I_N^{(4)}[1,2] = \frac{1}{2}\sum_{k=0}^{N-1} H_2^{-1}(2G_1 - 2G_2 + \Delta G_{1,k+1} - \Delta G_{2,k+1})u(k+1) \\ \times H_2^{-1}(\Delta G_{1,k+1} - \Delta G_{2,k+1})u(k+1)/\sigma_2^2 \end{aligned} \tag{70}$$

$$\Delta G_{i,k+1} \triangleq \Delta G_{k+1}/\theta = \hat{\theta}_i, \quad \xi = \hat{\xi}_i \quad i = 1,2 \tag{71}$$

We notice that the first three terms in (66)–(70) are the same as (5)–(7) of the ideal case. Furthermore, the contribution of the unmodeled dynamics gives a new term (70), which does not affect the other terms. We may therefore use the first three terms as change detection indexes and use the fourth term to evaluate the contribution of unmodeled dynamics.

5. Index of Unmodeled Dynamics

In order to build a change detection system robust to the unmodeled dynamics, an index which is sensitive to the unmodeled dynamics is very important. We will introduce two such indexes.

The first one is based on the fourth term of the KDI. Since the unmodeled dynamics ΔG_t is rather sensitive to the noise realization, $I_N^{(4)}[1,2]$ is too sensitive to noise when it is directly used as an index of unmodeled dynamics. In order to derive a less sensitive index, we consider the case where change detection is implemented on-line. The KDIs (66)–(70) are calculated on-line for each estimated $\hat{\theta}_t$ and $\hat{\xi}_t$, namely we compute $I_N[1,t]$, $I_N^{(1)}[1,t]$, $I_N^{(2)}[1,t]$, $I_N^{(3)}[1,t]$, $I_N^{(4)}[1,t]$ at time t. Then the mean value of $I_N^{(4)}[1,t]$ in a moving window of length $\mathscr{W}$ can be used as an index of the unmodeled dynamics

$$I_{\Delta G}^{(1)}(t) = \frac{1}{\mathscr{W}}\sum_{j=t-\mathscr{W}}^{t} I_N^{(4)}[1,j] \tag{72}$$

Since in the nonbound approach, the unmodeled dynamics ΔG_t is identified directly, it is reasonable to use it as an index of unmodeled dynamics. Because the ΔG_t is a function of the past input–output of system, we introduce the second index of unmodeled dynamics using the mean value of $|\Delta G_k(\hat{\theta}_t, \hat{\xi}_t|, k = 0, \ldots, N-1$, calculated from the input–output data sets in internal I_1

$$I^{(2)}_{\Delta G}(t) = \frac{1}{N}\sum_{k=0}^{N-1} |\Delta G_k(\hat{\theta}_t, \hat{\xi}_t)| \tag{73}$$

C. Change Detection Scheme

In the ideal case where no unmodeled dynamics exists, the change detection can be executed by applying the thresholding approach to the KDI, see (9). In the presence of unmodeled dynamics, the identified model parameters may change depending on the noise realization, resulting in a fluctuation of the calculated KDI. When the unmodeled dynamics and the noise are large, the fluctuation might be so large that the method of applying the thresholding approach to the KDI becomes infeasible. Robust decision-making schemes should be developed for the change detection in such cases.

Scheme 1. It should be noticed that the identified model parameters have different sensitivity to the noise. For example, it is found that the fluctuations of $\hat{a}_i$ are typically not so large as those of $\hat{b}_i$. On the other hand, the KDI consists of four different terms, each of which has its particular relationship with the model parameters. Therefore, if a scheme can be developed to use the information of the four terms of the KDI and the index of unmodeled dynamics effectively, it is possible to achieve a robust change detection. As a choice for such a scheme, a neural network approach can be considered. However, only simple logic is used here as an example to show the possibility.

Considering that $I^{(1)}_N[1,2]$, which expresses the discrimination of the prediction errors, is not related directly to the system failure, we only use $I^{(i)}_N[1,2]$ $(i = 2,3,4)$ to develop the robust decision making. Introduce a binary transform function as follows

$$f_i = \mathcal{T}_i(I^{(i)}_N[1,2]) = \begin{cases} 1 & \text{when } I^{(i)}_N > h_i \\ 0 & \text{when } I^{(i)}_N < h_i \end{cases} \qquad i = 2,3,4 \tag{74}$$

where the h_is $(i = 2,3,4)$ are thresholds which may be determined by using the index of unmodeled dynamics I_{C_η} and other prior information (see

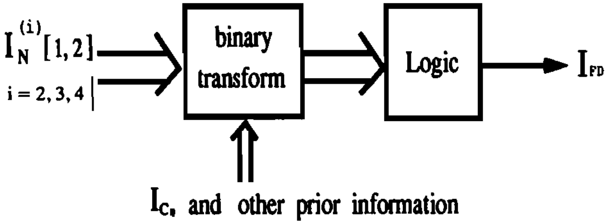


Figure 5 An example of a robust change detection scheme.

Fig. 5). Then the change detection index I_{FD} can be obtained by applying a logic to the f_is

$$I_{FD} = \mathcal{D}(f_i, i = 2,3,4) \tag{75}$$

where $\mathcal{D}$ may simply be a majority voting rule.

Scheme 2. Let us consider the case where the change detection is implemented on-line. The KDI $I_N[1,t]$ will be calculated for each identified $\hat{\theta}_t$ and $\hat{\xi}_t$. Then a robust decision-making strategy may be obtained based on a statistical test of $I_N[1,t]$. Because the fourth term of $I_N[1,t]$ is mainly contributed from the unmodeled dynamics which is rather sensitive to noise realization, we introduce such a robust scheme in the following way

$$\bar{I}_N[1,t] = \frac{1}{\mathcal{L}} \sum_{i=t-\mathcal{L}}^{t} (I_N^{(1)}[1,i] + I_N^{(2)}[1,i] + I_N^{(3)}[1,i]) \tag{76}$$

and execute change detection by

$$\bar{I}_N[1,t] \gtrless \eta \Rightarrow \begin{cases} \text{change} \\ \text{no change} \end{cases} \tag{77}$$

where $\mathcal{L}$ and η may be determined based on the indexes of unmodeled dynamics and other information about the system.

IV. NUMERICAL SIMULATIONS

In this section, we will carry out some numerical simulations to demonstrate the effectiveness of the methods discussed in the previous sections.

A. Implementation of Simulations

Figure 6 shows the implementation of simulations. The data are available from two disjoint intervals I_1 and I_2. It is assumed that the system is under normal mode in the interval I_1, and is to be monitored in the interval I_2. Now, from I_1 the unknown paramter vector $\{\theta, \xi\}$ is estimated as $\{\hat{\theta}_1, \hat{\xi}_1\}$, and from I_2 it is estimated on-line as $\{\hat{\theta}_t, \hat{\xi}_t\}$, in which the subscript t denotes the time. Note that when a soft bound approach is used, the identification can only be implemented in a limited on-line way, that is, for each estimated θ, the vector ξ is estimated using a batch identification algorithm, because only a batch identification algorithm is so far available for estimation of the soft bound of the unmodeled dynamics. If a change occurs in the interval I_2, the effect will be reflected as a difference between the identified models $\mathcal{M}\{\hat{\theta}_1, \hat{\xi}_1\}$ and $\mathcal{M}\{\hat{\theta}_t, \hat{\xi}_t\}$. The discrimination of $\mathcal{M}\{\hat{\theta}_1, \hat{\xi}_1\}$ and $\mathcal{M}\{\hat{\theta}_t, \hat{\xi}_t\}$ is performed via the KDI and the change detection will be realized using the change detection schemes.

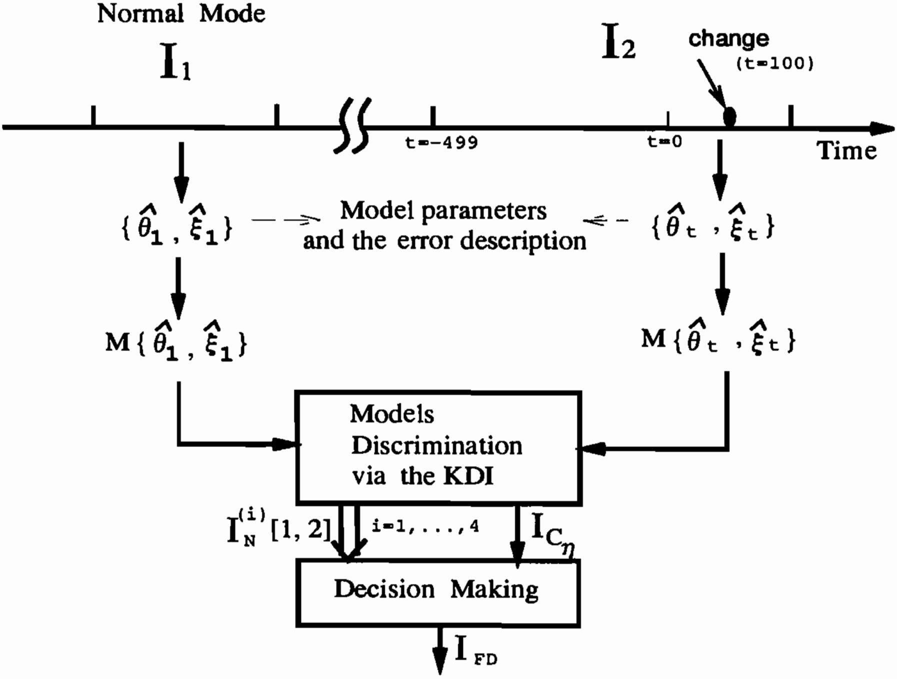

Normal Mode
I_1
I_2
change
(t=100)
t=-499
t=0
Time
$\{\hat{\theta}_1, \hat{\xi}_1\}$
Model parameters
and the error description
$\{\hat{\theta}_t, \hat{\xi}_t\}$
$M\{\hat{\theta}_1, \hat{\xi}_1\}$
$M\{\hat{\theta}_t, \hat{\xi}_t\}$
Models
Discrimination
via the KDI
$I_N^{(i)}[1,2]$
i=1, . . . ,4
I_{C_η}
Decision Making
I_{FD}

Figure 6 Implementation of the simulations.

B. Example 1: A Nonlinear System

Let us consider a nonlinear system described by (78) as the true system

$$\begin{cases} z(t) = -a_1 z(t-1) - a_2 z(t-2) + b_1 u(t-1) + b_2 u(t-2) + e_0(t) \\ y(t) = \tanh(\beta z(t)) \end{cases} \tag{78}$$

where $e_0(t) \in N\,(0, 0.01)$ is a white noise sequence and the input of system $u(t)$ is a PRBS (pseudo-random binary sequence). In the normal mode, the parameter vector is $\theta_p = [a_1\ a_2\ b_1\ b_2]^T = [-1.4\ 0.6\ 0.1\ 0.09]^T$ and it is assumed that the parameter vector is changed to $\theta_p = [-1.16\ 0.3\ 0.08\ 0.06]^T$ at $t = 100$. Furthermore, from (78) we know that as β increases, the nonlinearity of the system will increase. In the simulations, the cases where $\beta = 0.5$, 1.0, 2.5 are considered.

The hybrid quasi-ARMAX model described by (57)–(60) is used as a model to identify the system, in which $n = 2$, $m = 2$, $l = 0$, $M = 16$, and $X(t) = [y(t-1)\, y(t-2)]$.

Obviously, the residual between the model and the system can be used for the change detection, for which many existing approaches can be used, see, e.g., [1]. Here we will however execute the change detection using the nonbound approach described in the paper.

As shown in Fig. 6, the input–output data are available from two intervals I_1 and I_2. The interval I_1 contains 500 data sets, from which $\{\hat{\theta}_1, \hat{\xi}_1\}$ is estimated, while the interval I_2 contains 700 data sets, from which $\{\hat{\theta}_t, \hat{\xi}_t\}$ $(t = 0, \ldots, 200)$ is estimated on-line. The initial values $\{\hat{\theta}_0, \hat{\xi}_0\}$ for the on-line estimation are obtained from the first 500 data sets.

Three cases are considered, in which $\beta = 0.5, 1.0, 2.5$. When β is large, the unmodeled dynamics is expected to be large because the nonlinearity of the system increases. For each case, the output of the KDI $I_N[1, t]$ is calculated from (66)–(70). The output of the change detection index $\overline{I}_N[1, t]$ is calculated from (76), and the indexes of unmodeled dynamics $I^{(1)}_{\Delta G}(t)$ and $I^{(2)}_{\Delta G}(t)$ are calculated from (72) with $\mathscr{W} = 60$ and (73), respectively.

Table 1 shows the chosen values of $\mathscr{L}$ and η and the values of indexes of unmodeled dynamics. We can see that $I^{(1)}_{\Delta G}(t)$ and $I^{(2)}_{\Delta G}(t)$ increase as β increases, and when $I^{(1)}_{\Delta G}(t)$ and $I^{(2)}_{\Delta G}(t)$ increase, $\mathscr{L}$ should be chosen larger while η should be smaller.

Figure 7 shows the results for $\beta = 1.0$: (a) the sum of the first three terms of the KDI; (b) the output of the change detection index $\overline{I}_N[1, t]$; (c) the index of unmodeled dynamics $I^{(1)}_{\Delta G}(t)$; (d) the index of unmodeled dynamics $I^{(2)}_{\Delta G}(t)$. From Fig. 7b, we can see that the change is detected correctly.

Table 1 $\measuredangle$, η and the Indexes of Unmodeled Dynamics

β	Mean value of $I^{(1)}_{\Delta G}(t)$	Mean value of $I^{(2)}_{\Delta G}(t)$	Window width $\measuredangle$	Threshold η
0.5	0.4388	0.0207	2	8
1.0	2.7142	0.1312	5	4
2.5	4.3372	0.2316	15	3

C. Example 2: A Linear System with Time Delay

As the true system, consider the third-order SISO system

$$G(s) = \frac{ke^{-\tau s}}{s(T_0 s + 1)(T_1 s + 1) + k} \tag{79}$$

where $k = 2.5$, $T_0 = 0.1$, $T_1 = 0.5$, and $\tau = 0.25$. After being sampled with a period of 0.25, the system is corrupted by a white noise sequence $e_0(t) \in N(0, \sigma_0^2)$. The input to the system is a PRBS.

We will apply the soft bound approach to this example. Therefore, the model used to identify the system is described by (17) and (18) where the degrees of polynomials A, B, and C are chosen to be $n = 2$, $m = 2$, and $l = 0$, respectively.

The input–output data available from two intervals I_1 and I_2 are the same as those in Example 1. The interval I_1 contains 500 data sets, from which $\{\hat{\theta}_1, \hat{\xi}_1\}$ is estimated, while the interval I_2 contains 700 data sets, from which $\{\hat{\theta}_t, \hat{\xi}_t\}$ $(t = 1, \ldots, 200)$ is estimated. Since an on-line identification algorithm is not so far available for $\hat{\xi}_t$, the $\hat{\theta}_t$ is identified on-line while $\hat{\xi}_t$ is identified using a batch algorithm. The initial value $\hat{\theta}_0$ is obtained from the first 500 data sets.

The simulations are carried out for various values of the noise variance σ_0^2 ranging from 0.01 to 0.5. It is assumed that the physical parameters are changed from their normal values to $k = 0.5$, $T_0 = 0.05$, $T_1 = 0.2$ at $t = 100$. In each of the simulations, the $\{\hat{\theta}_t, \hat{\xi}_t\}$ are estimated using the method described by (23)–(28), and the terms of the KDIs, $I_N^{(i)}[1, t]$, $(i = 1, \ldots, 4; t = 1, \ldots, 200)$ and $I_N[1, t]$ are calculated by using (30) and (31)–(34). Finally, the change detection index I_{FD} is calculated by using (74) and (75). For simplicity, in Fig. 8 we only show I_{FD} and the second term of KDI which expresses the difference between the system models G_2 and G_1.

Figures 8a and b show the outputs of the second term of the KDI, $I_N^{(2)}[1, t]$, for $\sigma_0^2 = 0.01$ and $\sigma_0^2 = 0.05$, respectively. We can see that the

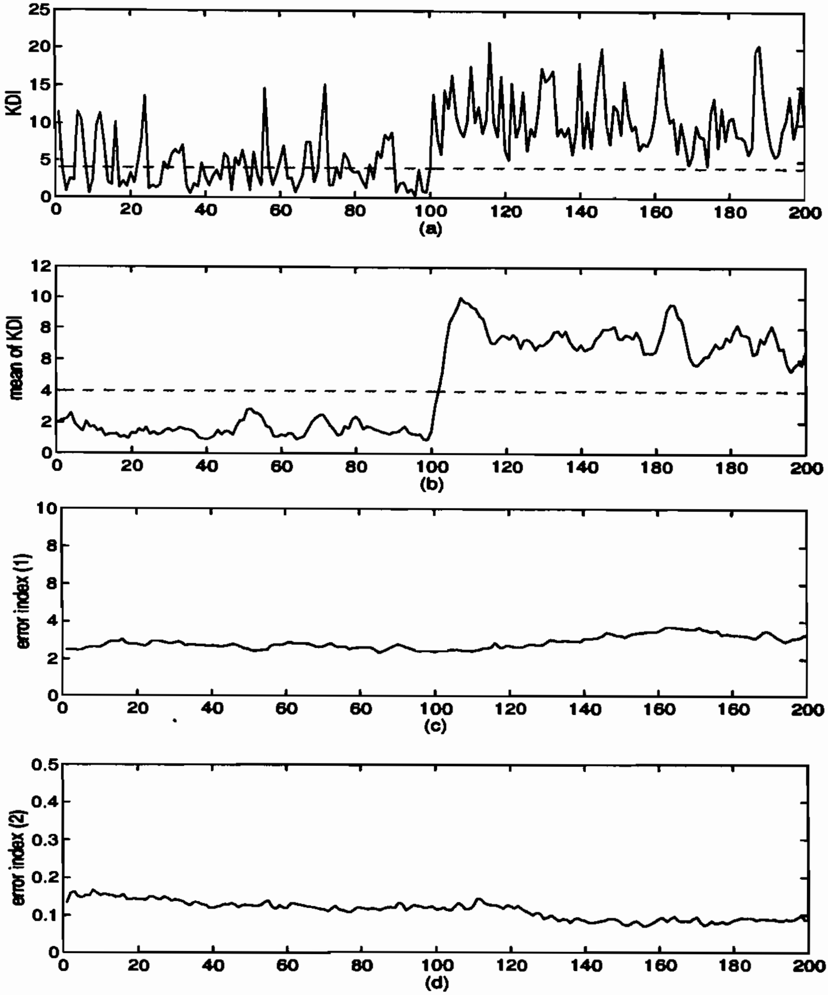


Figure 7 The results of Example 1: (a) the sum of the first three terms of the KDI; (b) the change detection index $I_N[1,t]$; (c) the index of unmodeled dynamics $I^{(1)}_{\Delta G}(t)$; (d) the index of unmodeled dynamics $I^{(2)}_{\Delta G}(t)$.

fluctuation of the KDI increases as the noise increases. The fluctuation of the KDI might become so large that the thresholding approach becomes invalid. Figure 8c shows the output of the robust decision-making I_{FD} calculated by using (74) and (75) where h_2, h_3, and h_4 were determined based on the index of unmodeled dynamics I_{C_η} and other prior information

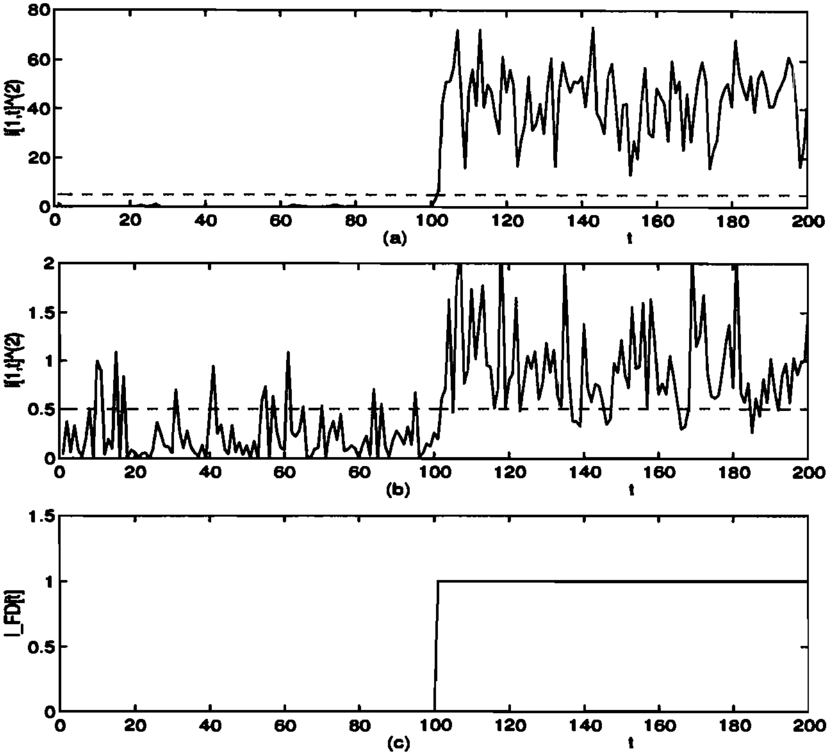


Figure 8 The results of Example 2: (a) the second term of the KDI for $\sigma_0^2 = 0.01$; (b) the second term of the KDI for $\sigma_0^2 = 0.5$; (c) the output of the robust decision-making scheme.

such as the variance of the noise. In this case, $h_2 = 0.5$, $h_3 = 10$, and $h_4 = 0.5$ were chosen for $\sigma_0^2 = 0.5$. We can see that the decision making responds correctly to the system change.

For $\sigma_0^2 = 0.01$, the estimated parameters $\hat{\theta}_1, \hat{\xi}_1$, and $I_{C_{\eta 1}}$ from the interval I_1 in this case were

$$\begin{aligned}
\hat{\theta}_1 &= [a_1, a_2, b_1, b_2] \\
&= [-1.5869, 0.8176, -0.0042, 0.0641] \\
\hat{\xi}_1 &= [\alpha_a, \lambda_a, \alpha_b, \lambda_b, \sigma^2] \\
&= [0.1366, 0.3553, 0.0075, 0.1430, 0.0122] \\
I_{C_{\eta 1}} &= 0.0322
\end{aligned}$$

V. DISCUSSIONS AND CONCLUSIONS

We have discussed some statistical methods for robust change detection in dynamical systems with model uncertainty. These methods have been developed based on two extensions of the CDS via the KDI to systems with partly unmodeled dynamics. In the first extension a soft bound has been introduced for describing unmodeled dynamics based on a stochastic embedding approach. The second extension is based on a nonbound approach. The effectiveness of the proposed approaches has been confirmed through numerical simulations.

In the soft bound approach, however, it should be noted that the soft bound based on the stochastic embedding approach is identified by maximizing the log-likelihood function $l(W|\xi)$. Therefore, the following open problems are to be solved: (1) an on-line algorithm is not so far available for the identification; (2) the existing batch algorithm is rather time-consuming. Thus further research is needed to develop an effective on-line identification algorithm.

On the other hand, the nonbound approach has the following advantages compared with the soft bound approach: (1) the existing on-line algorithms can be applied for the identification; (2) the KDI can be analyzed in the form in which the unmodeled dynamics is explicitly expressed in a separated term; (3) it has wide applications including general nonlinear systems; (4) the residual information obtained from the identified model may be used for change detection as well. Therefore from the viewpoint of practical applications, the nonbound approach seems to be rather suitable for change detection in dynamical systems.

REFERENCES

1. M Basseville. Automatica 24:309–326, 1988.
2. PM Frank. Proc IMACS/SICE Symp '92, Kobe, Japan, 1992, pp 777–784.
3. R Isermann. Automatica 29:815–835, 1993.
4. K Kumamaru, K Inoue, S Nonaka, H Ono, T Söderström. Proc 10th IFAC Symp Identification, Copenhagen, vol. 2, 1994, pp 303–308.
5. PM Frank, J Wunnenberg. In: RJ Patton, PM Frank, R Clark, eds, Fault Diagnosis in Dynamic Systems: Theory and Application. Prentice Hall, 1989, pp 47–98.
6. RJ Patton, SM Kangethe. In: RJ Patton, PM Frank, R Clark, eds, Fault Diagnosis in Dynamic Systems: Theory and Application. Prentice Hall, 1989, pp 99–154.
7. K Kumamaru, T Söderström. Trans Soc Instrum Contr Eng 22:1135–1140, 1986 (in Japanese).
8. B Ninness, G Goodwin. Automatica 31:1771–1797, 1995.

9. GC Goodwin, M Gevers, B Ninness. IEEE Trans Autom Contr 37:913–928, 1992.
10. T Söderström, K Kumamaru. On the use of Kullback discrimination index for model validation and fault detection, Tech. Rep. UPTEC85 20 R, Institute of Technology, Uppsala University, 1985.
11. K Kumamaru, S Sagara, T Söderström. In: RJ Patton, PM Frank, R Clark, eds, Fault Diagnosis in Dynamic Systems: Theory and Application. Prentice Hall, 1989, pp 439–475.
12. T Söderström, P Stoica. System Identification. Hemel Hempstead: Prentice Hall, 1989.
13. K Kumamaru, J Hu, K Inoue, H Ono. Proc 26th ISCIE International Symposium on Stochastic Systems Theory and Its Applications, Osaka, Japan, 1994, pp 173–178.
14. J Hu, K Kumamaru, K Inoue. Trans Soc Instrum Contr Eng 32:714–721, 1996.
15. K Kumamaru, S Sagara, A Nakai, T Söderström. Proc 9th IFAC/IFORS Symp Identification and System Parameter Estimation, 1991, pp 657–662.
16. K Kumamaru, J Hu, K Inoue, T Söderström. Robust fault detection using index of Kullback discrimination information. Proc 13th IFAC World Congress, San Francisco, 1996, pp 205–210.
17. P Eykoff. System Identification. New York: Wiley, 1974.
18. J Hu, K Kumamaru. Proc 34th SICE Annual Conference (international session), Sapporo, 1995, pp 1211–1216.
19. J Hu, K Kumamaru, K Inuoe. A hybrid quasi-ARMAX Modeling Scheme for identification and control of nonlinear systems. Proc 35th IEEE Conference on Control and Decision, Kobe, 1996, pp 1413–1418.

This Page Intentionally Left Blank

18

Detecting Changes in Acting Stochastic Models and Model Implementation via Stochastic Binary Neural Networks

Anthony Burrell and P. Papantoni-Kazakos

University of Alabama, Tuscaloosa, Alabama

Achilles G. Kogiantis

University of Southwestern Louisiana, Lafayette, Louisiana

I. INTRODUCTION AND BACKGROUND

Neural networks are distributed structures designed to perform specific tasks. Thus, their design should be task or objective oriented, where then concrete analytic relationships between various network weights and parameters may be developed, with the subsequent dramatic reduction in the free parameters to be "learned" during the network training process (see Pados *et al.* [27–31]). When training is attained via interaction with the environment it corresponds to *supervised learning*, where in objective-oriented network designs supervised learning algorithms should be dictated by the performance criterion pertinent to the objective (see Pados *et al.* [27–31]).

When neural networks operate in stochastically described environments, supervised learning becomes a statistical sequential estimation problem dealt with by stochastic approximation methods. There is a rich literature in such methods represented by the works of Abdelhamid [1], Beran [6], Blum [8], Fabian [13], Fisher [14], Gerencser [15], Kashyap *et al.* [17,18], Kiefer *et al.* [20], Kushner [22], Kushner *et al.* [23], Ljung [24], Ljung *et al.* [25], Robbins *et al.* [33], and Young *et al.* [39].

In the neural networks literature, supervised learning has been basically limited to techniques arising from the Robbins–Monro [33] method and its extensions, with performance criterion the least squares error. The representative works on the subject are those by Barron *et al.* [5], Elman *et al.* [12], Gorman *et al.* [16], Minsky *et al.* [26], Rosenblatt [34], Werbos [35], White [36], Widrow [37], and Widrow *et al.* [38]. Literature in the area, when the performance criterion is, instead, the Kullback–Leibler distance (see Blahut *et al.* [7] and Kazakos *et al.* [19]) and the techniques used do not necessarily arise from the Robbins/Monro method, is represented by the works of Ackley *et al.* [2], Amari *et al.* [3], Pados *et al.* [27,29,31], and Kogiantis *et al.* [21].

In this chapter, we consider the neural implementation of a sequential algorithm for the detection of change in the acting stochastic environmental model, when the model alternatives must be learned themselves in interaction with the environment. The original algorithm for parametrically designed stochastic models can be found in Bansal *et al.* [4] and its parametric extension can be found in Burrell [9] and Burrell *et al.* [10,11]. The applications of the algorithm are numerous, ranging from quality control, to the detection of edges in images, to the recognition of failures in network lines, to the dynamic capacity allocation in multimedia ATM networks. For the latter application, see Burrell [9] and Burrell *et al.* [10].

II. THE ALGORITHM AND DISCRETE ROBUST MAPPINGS

The extended parametric problem in Burrell [9] and Burrell *et al.* [10,11] is as follows. Let x_1^n denote the sequence $x_1, \ldots, x_n$ of n observations after time zero. Let the process which initially generates the data sequence be the process μ_0. Let it be possible that a shift to any one of $m-1$ independent processes μ_i, $i = 1, \ldots, m-1$, may occur at any point in time, where if a $\mu_0 \to \mu_i$ shift occurs, then the process μ_i remains active thereafter. The objective is to detect the occurrence of a $\mu_0 \to \mu_i$ shift as accurately and as timely as possible, including the detection of the process μ_i which μ_0 changed to. Let us denote by f_i, $i = 0, \ldots, m-1$, the density or probability functions induced by the processes μ_i, $i = 1, \ldots, m-1$, and let $f_i(x_n | x_1^{n-1})$ denote the density/probability function of x_n conditioned on the sequence x_1^{n-1}. Then, the *parametric* algorithm in [9] is as follows:

1. Select a threshold $\delta_0 > 0$.
2. Have $m-1$ parallel algorithms operating. The *i*th algorithm $i = 1, \ldots, m-1$ is monitoring a $\mu_0 \to \mu_i$ shift. $T_n^{0i}(x_1^n)$ denotes the operating

value of the ith algorithm at time n, given the observation sequence x_1^n. The operating value $T_n^{0i}(x_1^n)$ is updated as follows:

$$T_0^{0i} \equiv 0$$

$$T_n^{0i}(x_1^n) = \max\left(0, T_{n-1}^{0i}(x_1^{n-1}) + \log \frac{f_i(x_n|x_1^{n-1})}{f_0(x_n|x_1^{n-1})}\right) \tag{1}$$

3. The algorithmic system stops the first time n when either one of the $m-1$ parallel algorithms crosses the common threshold δ_0. If the ith algorithm is the one that first crosses the threshold, then it is declared that a $\mu_0 \rightarrow \mu_i$ shift has occurred.

Let N_{0i} denote the extended stopping variable induced by the ith algorithm in the system; that is, $N_{0i} \triangleq \inf\{n: T_n^{0i}(x_1^n) \geq \delta_0\}$. Let us define the following quantities.

$$L_n^{ij}(x_1^n) \triangleq n^{-1} \log \frac{f_j(x_1^n)}{f_i(x_1^n)} \quad i,j=0,1,\ldots,m-1$$

$$I_{ij} \triangleq \lim_{n\rightarrow\infty} L_n^{ij}(x_1^n) \quad i,j=0,1,\ldots,m-1$$

$$p_n^{ij}(\nu) \triangleq P(L_n^{ij}(x_1^n) < \nu | \mu_j) \quad i,j=0,1,\ldots,m-1 \tag{2}$$

Consider, then, the following conditions:

(A) $I_{ij}; i,j=0,1,\ldots,m-1$ exist and
$I_{ij} = E\{I_{ij}|\mu_j\}$, a.s. (P_{μ_j})

(B) For $i,j=0,1,\ldots,m-1$ and for $\nu \in (0, I_{ij})$.

$$\lim_{n\rightarrow\infty} np_n^{ij}(\nu) = 0 \text{ and } \sum_{n\geq 1} p_n^{ij}(\nu) < \infty$$

Conditions (A) state the existence of the generalized Kullback–Leibler information numbers $\{I_{ij}\}$ while conditions (B) ensure the convergence of the large-deviation probabilities $\{p_n^{ij}(\nu)\}$ fast enough. The following theorem and corollary can now be expressed, whose proof can be found in Burrell *et al.* [11].

Theorem 1. *Let the processes $\{\mu_j; j=0,1,\ldots,m-1\}$ be stationary, ergodic, mutually independent, and satisfying conditions (A) and (B), then,*

$$\textit{As} \quad \delta_0 \rightarrow \infty, \quad \underset{\substack{j=0,1,\ldots,m-1\\ i=1,\ldots,m-1}}{E\{N_{0i}|\mu_j\}}: \begin{cases} \sim [I_{0j} - I_{ij}]^{-1} \log \delta_0 & \text{if } I_{0j} \geq I_{ij} \\ \geq 2^{-1}\delta_0 & \text{if } I_{0j} < I_{ij} \end{cases} \tag{3}$$

Corollary. *Given that the process μ_j, $j=1, \ldots, m-1$, is acting throughout, the algorithmic system will asymptotically detect the $\mu_0 \rightarrow \mu_j$*

shift correctly, in the expected stopping time sense. That is:

$$E\{N_{0j}|\mu_j\} \sim I_{0j}^{-1}\log\delta_0\text{:}$$
$$\begin{cases} <E\{N_{0i}|\mu_j\} \sim [I_{0j}-I_{ij}]^{-1}\log\delta_0 & \forall i\neq j: I_{0j}>I_{ij} \\ <E\{N_{0i}|\mu_j\} \geq 2^{-1}\delta_0 & \forall i\neq j: I_{0j}<I_{ij} \end{cases} \tag{4}$$

Given that the process μ_j, $j=1, \ldots, m-1$, is acting throughout, the asymptotic expected stopping time of the algorithmic system is $E\{N_{0j}|\mu_j\} \sim I_{0j}^{-1}\log\delta_0$.

Remarks. (a) The algorithm is asymptotically optimal in the sense that, as $\delta_0 \rightarrow \infty$ and for any extended stopping variables $\{N'_{0j}; j=1, \ldots, m-1\}$ such that $E\{N'_{0j}|\mu_0\} \geq 2^{-1}\delta_0$, then $E\{N_{0j}|\mu_j\} \leq E\{N'_{0j}|\mu_j\}$.

(b) From Theorem 1 and its corollary, the importance of the generalized Kullback–Leibler information numbers $\{I_{ij}\}$ is clear. The larger these numbers are and the closer to each other, the better is the performance of the extended algorithm, in the sense that the smaller are then the ratios $E\{N_{0j}|\mu_j\}/E\{N_{0i}|\mu_j\}$; $i\neq j$; $i, j=1, \ldots, m-1$; as $\delta_0 \rightarrow \infty$.

(c) The asymptotic correct detections induced by the extended algorithm, as stated by the corollary, are due to the single common threshold δ_0 used. The latter is basically associated with the starting process μ_0.

The implicit assumption in the algorithm in (1) is that the conditional densities/probabilities $\{f_i(x_n|x_1^{n-1})\}$ are analytically known. In reality, these densities/probabilities may have to be estimated and/or may not possess analytic forms. In addition, it is generally desirable that the environment be mapped into a set of distinct representations for computability. We thus propose the deployment of m neural networks that are *pretrained* to reproduce discrete versions of the conditional probabilities/densities $\{f_i(x_n|x_1^{n-1})\}$. The latter networks will then provide the appropriate inputs to the algorithm in (1).

A. Discrete Finite Memory Robust Mappings

Let $x_1, \ldots, x_n$ denote a sequence of observations, representing the environment. Then, given the ith environmental model, given $x_1, \ldots, x_n$, the objective of the discrete mapping is to predict which one of M distinct regions the observation x_{n+1} is going to be in. Denoting these regions A_j, $j=1, \ldots, M$, a high-performance predictive encoding operation requires in fact the computation of the conditional probabilities $\{p_{ij}(x_1, \ldots, x_n) \triangleq P(x_{n+1}\in A_j/x_1, \ldots, x_n)\}_{1\leq j\leq M}$, which are used to map an observed sequence $\{x_1, \ldots, x_n\}$ onto each of the regions $\{A_j\}$, with corresponding probabilities $\{p_{ij}(x_1, \ldots, x_n)\}$. Two problems arise immediately:

1. Exploding computational load, due to the increasing memory represented by the sequences $(x_1, \ldots, x_n)$.
2. Statistical information on the sequences $(x_1, \ldots, x_n)$ needed for the computation of the probabilities $\{p_{ij}(x_1, \ldots, x_n)\}$.

The first problem is resolved if the increasing memory is approximated by finite, say size-l memory. That is, the increasing computational load is, instead, bounded if the process that generates the observations is approximated by an l-order Markov process. Then, the information loss is minimized when the process is Gaussian (see Blahut [7]).

Thus, to reduce the exploding computational load due to increasing data memory, we may *initially* model the process that generates the environmental data or observations by an l-order Gaussian Markov process, whose autocorrelation $m \times m$ matrix Q_i, has components identical to the corresponding components of the original process. We name this initial (Gaussian and Markov) process, the *nominal* process.

Starting with our nominal process, but incorporating then statistical uncertainties in our model, we are led to a powerful (qualitatively) robust formalization, which results in a stochastic mapping (see Papantoni-Kazakos *et al.* [32]), as follows.

Given observations $(x_1, \ldots, x_n)$, use the l most recent observations for the prediction of the next datum x_{n+1}, and defining $y_m^T = [x_{n-m+1}, \ldots, x_n]$, decide that $x_{n+1} \in A_j$ with probability $q_{ij}(y_m)$, defined as follows,

$$q_{ij}(y_m) = \frac{1}{M} r_i(y_m) + [1 - r_i(y_m)] p_{ij}(y_m) \tag{5}$$

where $p_{ij}(y_m)$ is the conditional probability of $x_{n+1} \in A_j$, given $y_m^T = [x_{n-m+1}, \ldots, x_n]$, under the ith model, as induced by the Gaussian and Markov nominal process, and where, for some positive finite constant λ_i,

$$r_i(y_m) \triangleq 1 - \min\left[1, \frac{\lambda_i}{\sqrt{y_m^T Q_i^{-1} y_m}}\right] \tag{6}$$

The value of the constant λ_i in (6) represents the level of confidence to the "purity" of the data vector y_m, in terms of it being generated by the nominal Gaussian process: the higher the value of λ_i, the higher the level of confidence, where as λ_i decreases, increased weight on purely random mappings (represented by the probability $1/M$ per region) is induced.

In addition, robust estimation of the autocorrelation matrix Q_i will be required. The components of the autocorrelation matrix Q_i should emerge from the statistics of the nominal Gaussian process. It is thus necessary to provide a scheme for the robust estimation of the matrix Q, in which observations generated from the outlier process are rejected in the estimation of the components of the matrix (see Kazakos *et al.* [19]). Special attention should be paid to allow for the existence of the inverse autocorrelation matrix Q_i^{-1} from the estimates of its components.

The robust prediction expression in (5) was based on a Gaussian assumption for the nominal process that generates the data in the environment, where the latter assumption was the result of an information-theoretical approach to the reduction of the computational load caused by increased past memeory. The important robust effects induced by the mapping in (5) remain unaltered, however, when instead the probability $p_{ij}(y_m)$ in (5) arises from an arbitrary non-Gaussian process, and when its conditioning on y_m is substituted by conditioning on quantized values of the scalar quantity $y_m^T Q_i^{-1} y_m$. When quantized values are involved, the implementation of the mapping in (5) requires the following stages.

1. *Preprocessing*. This stage corresponds to long-term memory and involves the robust preestimation (see Kazakos *et al.* [19]), and storage of the matrix Q_i^{-1}.
2. *Processing*. This stage corresponds to short-term memory. It uses the matrix Q_i^{-1} from the preprocessing step and the observation vector y_m to: (i) first compute the quadratic expression $y_m^T Q_i^{-1} y_m$, (ii) then represent $y_m^T Q_i^{-1} y_m$ in a quantized form comprising N distinct values, (iii) finally, use the quantized values in (ii) to compute the corresponding value of the function $r_i(y_m)$ in (6).
3. *Predictive mapping*. This stage involves the estimation of the probabilities $\{p_{ij}(y_m)\}$ and the computation of the probabilities $\{q_{ij}(y_m)\}$ in (5) using inputs from the processing stage, and the subsequent implementation of the prediction mappings.

The three different stages above are performed sequentially by separate but connected neural structures, named the *preprocessing layer*, *processing layer*, and *predictive mapping layer*, respectively. Our focus in this paper is on the latter layer: its structure and its operations. Towards that direction, we first note that, due to the quantization operations at the processing layer, the expression in (5) takes the following form:

$$q_{ijp} = \frac{1}{M} r_{ip} + [1 - r_{ip}] p_{ijp}; \qquad \text{for } y_m^T Q_i^{-1} y_m \rightarrow R_p; \quad p = 1, \ldots, N \qquad (7)$$

where q_{ijp}, p_{ijp}, and r_{ip} denote, respectively, the probabilities $q_{ij}(y_m)$ and $p_{ij}(y_m)$ and the number $r_i(y_m)$, when the quantized value of $y_m^T Q_i^{-1} y_m$ equals R_p.

III. THE PREDICTIVE MAPPING LAYER

Consider the integer M in (7), and let s be a unique positive integer, such that $2^{s-1} < M \leq 2^s$. Then, in modulo-2 arithmetic, each state j, $j = 1, \ldots, M$, can be represented by an s-length 0-1 binary sequence $x_1 \cdots x_s$. The state R_ρ is provided as an input to the prediction layer by the processing layer, and the former produces a binary sequence $x_1 \ldots x_s$ as a prediction mapping. Given the state R_ρ, the operations of the prediction layer must be such that a given prediction sequence $x_1 \cdots x_s$ is produced stochastically with probability

$$q_i(x_1 \cdots x_s/R_\rho) = \frac{1}{M} r_{i\rho} + [1 - r_{i\rho}] p_i(x_1 \cdots x_s/R_\rho) \tag{8}$$

where expression (8) is the same as expression (7) when the binary representation of the positive integer j in the latter is $x_1 \cdots x_s$, and where $p_i(x_1 \cdots x_s/R_\rho)$ is the prediction mapping generated by the nominal process that represents the *actual data environment*. Due to the stochastic nature of the rule in (8), such is also the nature of the predictive mapping layer, whose neural representation corresponds then to a stochastic neural network, first developed by Kogiantis *et al.* [21], when the response of each neuron is limited to binary. We proceed with the description of the latter representation.

Let us temporarily assume that the probabilities $p_i(x_1 \cdots x_s/R_\rho)$ have been "learned" and are known. Without lack in generality, let us also assume that $M = 2^s$. The original constraint of binary firing per neuron in the prediction layer leads us to the digital representation of the future states $\{x_1, \ldots, x_s\}$. The design can be accommodated easily in a binary tree structure. In detail, given the observed state R_ρ and the resulting $R_{i\rho}$ value, the mapping $x_1 \cdots x_s$ can be obtained via a stochastic binary tree search, on the 2^s-leave tree, as follows. (a) With probability $r_{i\rho}$ a fair tree-search is activated, where the tree-node $x_1, x_1 = 0,1$ is visited with probability 0.5, and each of the two tree-nodes branching off a visited tree-node $x_1 \cdots x_k$, $1 \leq k \leq s-1$ is also visited with probability 0.5. (b) With probability $1 - r_{i\rho}$ a generally biased tree-search is activated, where the tree-node x_1 is visited with probability $p_i(x_1/R_\rho)$, while from a visited tree-node $x_1 \cdots x_k$, $1 \leq k \leq s-1$ the tree-node $x_1 \cdots x_k x_{k+1}$ is visited with probability.

$$p_i(x_{k+1}/x_1 \cdots x_k, R_\rho) \triangleq p_i(x_1 \cdots x_k x_{k+1}/x_1 \cdots x_k, R_\rho)$$

where

$$p_i(x_1 \cdots x_s/R_\rho) = p_i(x_1/R_\rho) \cdots p_i(x_{k+1}/x_1 \cdots x_k, R_\rho) \cdots \\ p_i(x_s/x_1 \cdots x_{s-1}, R_\rho) \tag{9}$$

Thus, the predictive mapping layer may be viewed as being comprised of a fair-search binary tree and a number of biased-search binary trees, each of the latter corresponding to a specific observation state R_ρ. Given R_ρ, the common fair-search binary tree is activated with probability $r_{i\rho}$, while, with probability $1-r_{i\rho}$, the biased-search binary tree that corresponds to the state R_ρ is activated, instead; we name the latter tree the R_ρ-tree. The nodes of each of the above binary trees are neurons that "fire," if the corresponding tree-nodes are "visited." Given R_ρ, a specific mapping $x_1 \cdots x_s$ is generated either equiprobably from the fair-search binary tree with probability $r_{i\rho}$, or from the R_ρ-tree via the sequential stochastic representation in (9) with probability $1-r_{i\rho}$. It is thus in the R_ρ-tree that the probabilities which generate the data of the environment must be "learned" and then used to generate prediction mappings.

Given the observation state R_ρ, consider the R_ρ-tree in conjunction with the sequential stochastic representation in (9) of the corresponding prediction mappings, as generated by the process representing the actual data environment. Let $u_{x_1 \cdots x_k}, 1 \le k \le s$, represent the binary random output of the neuron that corresponds to the node $x_1 \cdots x_k$ of the R_ρ-tree. Then, $u_{x_1 \cdots x_k} = 1$ if and only if $u_{x_1 \cdots x_i} = 1, \forall i \le k$. Thus, the output $u_{x_1 \cdots x_k}$ may be viewed as generated by a product, $W_{x_1} W_{x_2/x_1} \cdots W_{x_k/x_1 \cdots x_{k-1}}$, of mutually independent binary random variables $\{W_{x_i/x_1 \cdots x_{i-1}}\}_{1 \le i \le k}$, whose distributions at the operational stage of the R_ρ-tree must be as follows (in view of (9)):

$$\begin{aligned} P(u_{x_1 \cdots x_k} = 1) &= P(W_{x_1} W_{x_2/x_1} \cdots W_{x_k/x_1 \cdots x_{k-1}} = 1) \\ &= p_i(x_1 \cdots x_k/R_\rho) = p_i(x_1/R_\rho) \cdots p_i(x_k/x_1 \cdots x_{k-1}, R_\rho) \\ &= P(W_{x_1} = 1) P(W_{x_2/x_1} = 1) \cdots P(W_{x_k/x_1 \cdots x_{k-1}} = 1); \\ &\quad 2 \le k \le s \\ P(u_{x_1} = 1) &= P(W_{x_1} = 1) = p_i(x_1/R_\rho), \end{aligned} \tag{10}$$

where

$$P(W_{x_k/x_1 \cdots x_{k-1}} = 1) = p_i(x_k/x_1 \cdots x_{k-1}, R_\rho); \qquad 2 \le k \le s \tag{11}$$

The above logical arguments and expressions lead to the following neural structure of the R_ρ-tree. (a) The neuron corresponding to the tree-node $x_1; x_1 = 0,1$ has a binary random variable W_{x_1} built in, where $W_0 = 1 - W_1$. At the operational stage, the neuron must be activated with probability $p_i(x_1/R_\rho)$; thus, $P(W_{x_1} = 1) = p_i(x_1/R_\rho)$ then, where $P(W_1 = 1) = 1 - P(W_0 = 1)$. (b) For $k \ge 2$, the neuron corresponding to the tree-node $x_1 \cdots x_k$ has a binary random variable $W_{x_k/x_1 \cdots x_{k-1}}$ built in and fires if and only if the latter variable takes the value 1 and simultaneously the neuron corresponding to the tree-node $x_1 \cdots x_{k-1}$ fires as well. Thus, the binary neural output $u_{x_1 \cdots x_k}$ is formed as a product $u_{x_1 \cdots x_{k-1}} W_{x_k/x_1 \cdots x_{k-1}}$, where

$$P(u_{x_1 \cdots x_k} = 1) = P(u_{x_1 \cdots x_{k-1}} W_{x_k/x_1 \cdots x_{k-1}} = 1) = P(u_{x_1 \cdots x_{k-1}} = 1) \times P(W_{x_k/x_1 \cdots x_{k-1}} = 1) \qquad (12)$$

and where, at the operational stage of the R_ρ-tree, the probability $P(W_{x_k/x_1 \cdots x_{k-1}} = 1)$ must be as in (11). We note that

$$W_{1/x_1 \cdots x_{k-1}} = 1 - W_{0/x_1 \cdots x_{k-1}}, \qquad \forall k \geq 2, \qquad \forall x_1 \cdots x_{k-1} \qquad (13)$$

and thus

$$P(W_{1/x_1 \cdots x_{k-1}} = 1) = 1 - P(W_{0/x_1 \cdots x_{k-1}} = 1), \qquad \forall k \geq 2, \quad \forall x_1 \cdots x_{k-1}$$

As is clear from the derivations and arguments in this section, the parameters of interest in the R_ρ-tree neural network consist of the independent binary random variables W_1 and $\{W_{1/x_1 \cdots x_{k-1}}; x_i = 0, 1; 1 \leq i \leq k-1\}_{2 \leq k \leq s}$, whose distributions $p_i(1/R_\rho)$ and $\{p_i(1/x_1 \cdots x_{k-1}, R_\rho); x_i = 0, 1; 1 \leq i \leq k-1\}_{2 \leq k \leq s}$, must be "learned" in advance, via interaction with the environment which represents the ith model.

A. Learning at the Predictive Mapping Layer

Given the R_ρ-tree, we observe that, due to (10), any adaptations of the probability $P(u_{x_1 \cdots x_s} = 1)$ *backpropagate* to adaptations of each of the probabilities $P(W_{x_1} = 1), \ldots, P(W_{x_s/x_1 \cdots x_{s-1}} = 1)$. It thus suffices to focus on the learning of the probabilities $\{P(u_{x_1 \cdots x_s} = 1)\}$ for the various binary sequences $x_1 \ldots x_s$, which correspond to the responses of the output or "visible" neurons in the R_ρ-tree network. For ease of presentation, let us now consider a fixed sequence $x_1 \cdots x_s$ (in conjunction with the fixed observed state R_ρ that represents the R_ρ-tree). Let then p denote the value of the probability $p(x_1 \cdots x_s/R_\rho)$, as induced by the environment, and let q denote the value of the probability $P(u_{x_1 \cdots x_s} = 1)$. Let the natural number n denote the discrete observation time from the beginning of the learning stage, and let $\hat{p}_n$ and $\hat{q}_n$ denote estimates at time n of the probability values p and q, respectively. Finally, let the random variable V_n be defined as equal to 1, if the environmental event $\{x_1 \cdots x_s/R_\rho\}$ occurs at time n, and as equal to 0, otherwise, and let

$$Z_\rho \triangleq V_\rho(1 - W_\rho), \qquad W_\rho = \begin{cases} 0 & \text{w.p.} 1 - r_{i\rho} \\ 1 & \text{w.p.} r_{i\rho} \end{cases}$$

In Kogiantis *et al.* [21], a Kullback–Leibler matching criterion between p and q was used in conjunction with Newton's iterative numerical method to develop the supervised learning algorithm stated below.

Algorithm. Initial values: Select an initial value $\hat{q}_1 > 0$, while $\hat{p}_1 = v_1$.

Computational steps:

(a) Given the computed value $\hat{p}_n$, and z_{n+1}, as in (17), compute $\hat{p}_{n+1}$ as follows:

$$\hat{p}_{n+1} = \hat{p}_n + \frac{z_{n+1}[1 - r_{i\rho}]^{-1} - \hat{p}_n}{n+1} \tag{14}$$

For some small positive value δ, the value $\hat{p}_{n+1}$ is corrected to δ if $\hat{p}_{n+1} < \delta$, and is corrected to $1 - \delta$ if $\hat{p}_{n+1} > 1 - \delta$.

(b) Given the computed values $(\hat{q}_n, \hat{p}_n)$, given z_{n+1}, compute $\hat{q}_{n+1}$ as follows:

$$\begin{aligned}\hat{q}_{n+1} = \hat{q}_n &- \frac{(\hat{q}_n - \hat{p}_n)\hat{q}_n(1 - \hat{q}_n)}{(\hat{q}_n - \hat{p}_n)^2 + \hat{p}_n(1 - \hat{p}_n)} \\ &+ \frac{z_{n+1}[1 - r_{i\rho}]^{-1} - \hat{p}_n}{n+1} \\ &\times \left[\frac{\hat{q}_n(1 - \hat{q}_n)}{(\hat{q}_n - \hat{p}_n)^2 + \hat{p}_n(1 - \hat{p}_n)}\right]^2\end{aligned} \tag{15}$$

where $\hat{p}_{n+1} - \hat{p}_n = (z_{n+1}[1 - r_{i\rho}]^{-1} - \hat{p}_n)/(n+1)$ from (14). For some small positive value ζ, the value $\hat{q}_{n+1}$ is corrected to ζ if $\hat{q}_{n+1} < \zeta$, and is corrected to $1 - \zeta$ if $\hat{q}_{n+1} > 1 - \zeta$.

In Kogiantis *et al.* [21], the following theorem was proved.

Theorem 2. *Let the process which generates the observed data in the environment be ergodic. Let then s denote the probability of the event* $\{x_1 \cdots x_s/R_\rho\}$, *as induced by the latter process. Then, the supervised learning algorithm converges to the probability s, almost surely, with rate inversely proportional to the sample/iteration size n.* □

We note that in Theorem 2, if the process that generates the observed data in the environment is ergodic, and if $\{s(x_1 \cdots x_s/R_\rho)\}$ denote the prediction mappings induced by the latter process, then, via the learning algorithm and with almost sure convergence, the prediction mappings produced by the predictive mapping layer are governed by the probabilities

$$\{q'(x_1 \cdots x_s/R_\rho) \triangleq (1 - r_{i\rho})s(x_1 \cdots x_s/R_\rho) + M^{-1} r_{i\rho}\}$$

In Kogiantis *et al.* [21], it was found that the learning algorithm converges rapidly to predictive probability mappings that are close to those induced by the environment, even under mismatch network conditions. Specifically, when past dependence decays fast with distance, then, even when the network order is less than the order of the Markovian environmental model, convergence to almost the true process is attained in less than fifty iterations in most cases.

IV. SOME NUMERICAL EVALUATIONS

The outputs of the pretrained neural networks provide the predictive mappings needed in the operation of the sequential algorithm in Section II. To study the nonasymptotic performance characteristics of the latter algorithm, we introduce the following performance metrics for given common threshold δ_0:

$\beta_i(n)$, $i = 1, \ldots, m-1$: The probability that a $\mu_0 \rightarrow \mu_i$ shift is decided before or at time n, given that the process μ_i is acting throughout.

$\alpha_i(n)$, $i = 1, \ldots, m-1$: The probability that before or at time n a $\mu_0 \rightarrow \mu_i$ false alarm occurs; that is, it is decided that μ_0 has shifted to process μ_i, while μ_0 has never changed.

Let us also define the following probabilities

$p_{i0j}(n)$, $i = 0, 1, \ldots, m-1$: The probability that the algorithm which
$j = 1, \ldots, m-1$ monitors a $\mu_0 \rightarrow \mu_j$ shift stops at n, given that the process μ_i acts throughout.

Then, we easily conclude that the following expressions hold.

$$\beta_i(n) = \sum_{l \leq n} p_{i0i}(l) \prod_{j \neq i} \left(1 - \sum_{k \leq l} p_{i0j}(k)\right), \quad i = 1, \ldots, m-1$$

$$\alpha_i(n) = \sum_{l \leq n} p_{00i}(l) \prod_{j \neq i} \left(1 - \sum_{k \leq l} p_{00j}(k)\right), \quad i = 1, \ldots, m-1 \tag{16}$$

The performance metrics $\{\beta_i(n)\}$ and $\{\alpha_i(n)\}$ can be used for the appropriate nonasymptotic selection of the common threshold δ_0. Qualitatively speaking, a threshold δ_0 is sought such that, for relatively small n values, the probabilities $\{\beta_i(n)\}$ are sufficiently large, while the probabilities $\{\alpha_i(n)\}$ may simultaneously be below specified desirable levels. Note that all probabilities $\{\beta_i(n)\}$ and $\{\alpha_i(n)\}$ are monotonically nondecreasing with increasing n. These probabilities converge to values that are generally less than one. As functions of n, the probabilities $\{\beta_i(n)\}$ and $\{\alpha_i(n)\}$ represent power and false alarm curves, respectively, which can be generally computed numerically.

For our numerical evaluations, we adopted the following stochastic models. We assume that the processes $\{\mu_i\}$ are homogeneous Poisson, with corresponding rates $\{r_i\}$. Then, directly from (1) in Section II, we have:

$$T_0^{0i} \equiv 0$$

$$T_n^{0i}(x_1^n) = \max\left(0, T_{n-1}^{0i}(x_1^{n-1}) + [r_0 - r_i] + x_n \log\frac{r_i}{r_0}\right) \tag{17}$$

where x_n denotes the number of arrivals at time n.

Let us define

$$\xi(r_0, r_i) \triangleq [r_0 - r_i]\left[\log\left(\frac{r_0}{r_i}\right)\right]^{-1} \tag{18}$$

where $\min(r_0, r_i) < \xi(r_0, r_i) < \max(r_0, r_i)$. We select rates $\{r_i\}$ such that $\xi(r_0, r_i)$ are rational numbers for all i, and then we define the integers t_{0i} and $s_{0i}; t_{0i} < s_{0i}$ as follows:

$$\xi(r_0, r_i) = \frac{t_{0i}}{s_{0i}} \tag{19}$$

The algorithmic thresholds can then be selected as positive integers, without lack in generality, and the algorithmic adaptation in (17) can be transformed as follows:

$$T_n^{0i}(x_1^n) = \max\{0, T_{n-1}^{0i}(x_1^{n-1}) + (-1)^{\text{ime}(0,i)}[s_{0i}x_n - t_{0i}]\} \tag{20}$$

where

$$\text{ime}(0, i) = \begin{cases} 0 & \text{if } r_i > r_0 \\ 1 & \text{if } r_i < r_0 \end{cases}$$

The pertinent Kullback–Leibler numbers $\{I_{ij}\}$ are as follows:

$$I_{ij} = r_i\left[1 - \frac{r_j}{r_i} + \frac{r_j}{r_i}\ln\left(\frac{r_j}{r_i}\right)\right] \tag{21}$$

For the computation of the power curves $\{\beta_i(n)\}$ and the false alarm curves of $\{\alpha_i(n)\}$ in (16), the following probabilities are used:

$P_{i0j}(n, y)$: The probability that at time n the algorithm that monitors a $\mu_0 \rightarrow \mu_j$ shift has not crossed its threshold and its operating value equals y while the process μ_i is generating the data throughout.

For the Poisson processes, let ν_0 denote the integer threshold coresponding to the threshold δ_0. Then, the following recursions are easily found.

$$p_{i0j}(n,y) = \sum_{m:\, 0\le y-(-1)^{\mathrm{imc}(0,j)}[s_{0j}m-t_{0j}]} \mathrm{e}^{(-r_i)}\ \frac{(r_i)^m}{m!}$$

$$p_{i0j}(n-1, y-(-1)^{(0,j)}[s_{0j}m-t_{0j}]) \quad 1 \le y \le \nu_0$$

$$p_{i0j}(n,0) = \sum_{0\le w\le \nu_0 - 1} p_{i0j}(n-1,w) \sum_{\substack{m:\, w+(-1)^{\mathrm{ime}(0,j)}[s_{0j}m-t_{0j}]\gtrless \\ \le 0}}$$

$$\mathrm{e}^{(-r_i)}\ \frac{(r_i)^m}{m!}, \quad n \ge 1 \tag{22}$$

with initial conditions $p_{i0j}(0,0) = 1$.

Let $p_{i0j}(n)$ denote the probability that the $r_0 \to r_j$ monitoring algorithm first crosses its threshold ν_0 in n steps after the beginning of its operation, given that the rate r_i is acting. Then,

$$p_{i0j}(n) = \sum_{0\le y\le \nu_0 - 1} p_{i0j}(n-1,y) \sum_{m:\, y+(-1)^{\mathrm{ime}(0,j)}[s_{0j}m-t_{0j}]\ge \nu_0}$$

$$\mathrm{e}^{(-r_i)}\ \frac{(r_i)^m}{m!}, \quad n \ge 1 \tag{23}$$

A. Numerical Results and Their Evaluations

We selected the Poisson rates $r_1 = 0.08$, $r_2 = 0.12$, $r_3 = 0.24$, and $r_4 = 0.32$, and evaluated the sequential algorithm for testing the shift r_2 to (r_1, r_3, r_4). Specifically, we computed then the power and false alarm curves $\{\beta_i(n)\}$ and $\{\alpha_i(n)\}$. Our results are exhibited in Fig. 1, where the converging $\{\beta_i(n)\}$ and $\{\alpha_i(n)\}$ values are plotted for different threshold values, and where the Kullback–Leibler numbers $\{I_{ij}\}$ are included as well. From the figure, we notice the changing performance behavior as the threshold values and Kullback–Leibler numbers change. As the Kullback–Leibler numbers increase, the performance generally improves. When the Kullback–Leibler numbers are very small, much higher thresholds are needed for distinguishing between the corresponding processes, while for smaller thresholds, the false alarm may even exceed the power then. We emphasize, however, that the role of the Kullback–Leibler numbers in the performance of the $\{\alpha_i(n)\}$ and $\{\beta_i(n)\}$ performance metrics is not as clear

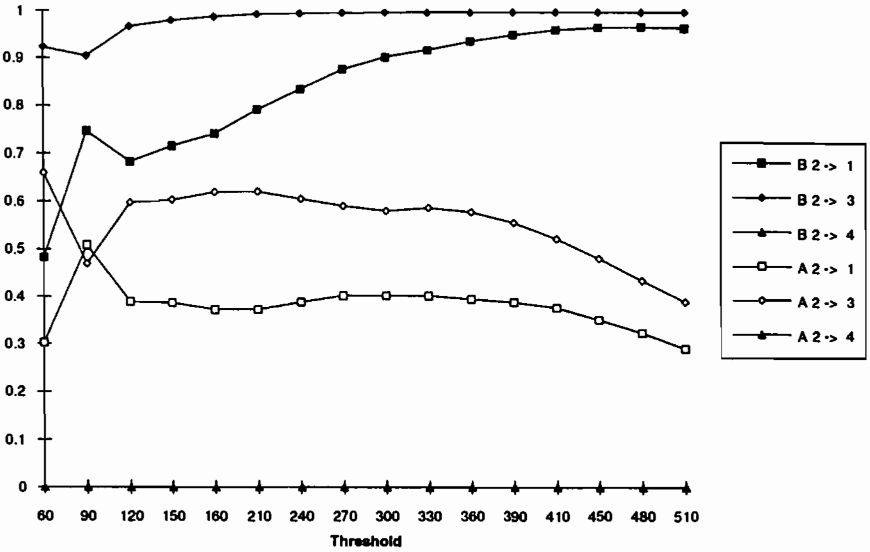


Figure 1 $\{\beta_i(n)\}$ and $\{\alpha_i(n)\}$ converging values against different thresholds for the extended algorithm: $B_j \rightarrow i$: converging $\beta_i(n)$ value, for starting rate r_j; $A_j \rightarrow i$: converging $\alpha_i(n)$ value, for starting rate r_j; B2 → 4 and A2 → 4 overlap; starting rate: r_2; $I_{21} = 0.0075627912$; $I_{23} = 0.0463553223$; $I_{24} = 0.1138653584$.

as it is when a single $r_i \rightarrow r_j$ algorithm is isolated. Indeed, in the $\{\alpha_i(n)\}$ and $\{\beta_i(n)\}$ metrics, the performance characteristics of all the parallel algorithms are involved, instead. That is, the $\{\alpha_i(n)\}$ and $\{\beta_i(n)\}$ curves represent performance characteristics of a whole interactive algorithmic system (the system of parallel algorithms).

To isolate some basic properties of the sequential algorithms, we tested it when only two Poisson rates are involved. For this case, we selected the system that consists of the rate pair (0.12, 0.24) since it represents the least favorable pair among the parallel algorithms in the r_2 to (r_1, r_3, r_4) system (in the Kullback–Leibler sense).

In Fig. 2, we plot the power and false alarm curves for the $r_1 \rightarrow r_2$ algorithm in the (0.12, 0.24) system. From Fig. 2, we observe that power almost equal to one and false alarm equal to about 0.01 is attained in about 100 data points.

We note that for the undertaken Poisson processes the pretraining of the neural networks that represent the predictive mappings of the various models require less than fifty iterations, where any network order is appropriate here.

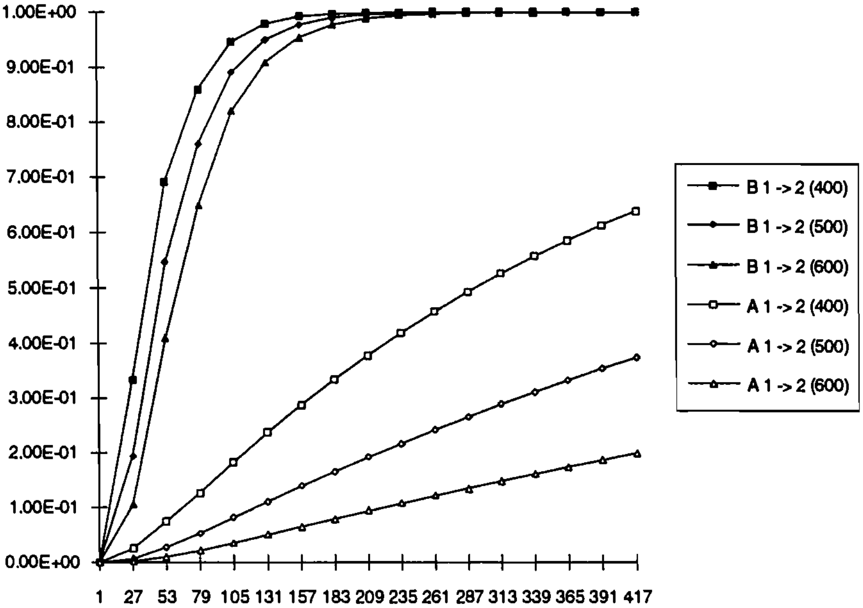


Figure 2 The $r_1 = 0.12$ to $r_2 = 0.24$ system: $\beta_2(n)$ and $\alpha_2(n)$ values against n for the extended algorithm and for different thresholds (xxx). B$j \rightarrow i$ (xxx): $\beta_i(n)$ values, for the $r_j \rightarrow r_i$ algorithm with threshold (xxx) A$j \rightarrow i$ (xxx): $\alpha_i(n)$ values, for the $r_j \rightarrow r_i$ algorithm with threshold (xxx) $I_{12} = 0.0463553223$.

V. CONCLUSIONS

We considered a sequential algorithm for the detection of change from a given environmental model to a number of alternatives, when the latter models need to be learned via supervised environmental observations. We assumed discretized observations and stochastic neural networks for the supervised learning of the models. We also adopted a backpropagating supervised learning algorithm for the pretraining of these networks that guarantees almost sure and rapid convergence under general model conditions. The overall system is efficient and accurate, as well as robust, and has numerous applications.

REFERENCES

1. F Abdelhamid. Ann Stat 1:1158–1174, 1973.
2. DH Ackley, GE Hinton, TJ Sejnowski. Cognitive Science 9:147–169, 1985.

3. S Amari, K Kurata, H Nagoaka. IEEE Trans Neural Networks 3:260–271, 1992.
4. RK Bansal, P Papantoni-Kazakos. IEEE Trans Inform Theory IT-32:227–235, 1986.
5. AR Barron, FW va Straten, RL Barron. Proc IEEE Int Conf Cybernetics and Society, 1977, pp 724–727.
6. R Beran. Annals Instit Statist Math 28:77–89, 1976.
7. RE Blahut. IEEE Trans Inform Theory IT-20:405–417, 1987.
8. JR Blum. Ann Math Stat 22:737–744, 1954.
9. AT Burrell. Traffic Management in Broadband Integrated Services Digital Networks. PhD Dissertation, University of Virginia, 1994.
10. AT Burrell, D Makrakis, P Papantoni-Kazakos. Traffic Monitoring for Capacity Allocation of Multimedia Traffic in Broadband Networks. In review.
11. AT Burrell, P Papantoni-Kazakos. Extended Sequential Algorithm for Detecting Changes in Acting Stochastic Processes. In review.
12. J Elman, D Zipser. J Acoust Soc Amer 83:1615–1626, 1988.
13. V Fabian. Ann Math Stat 39:1327–1332, 1968.
14. RA Fisher. J Roy Stat Soc 85:597i, 1922.
15. L Gerencser. In: CI Byrnes, A Lindquist, eds, Modeling, Identification and Robust Controls. Amsterdam: North-Holland, 1986, pp 581–594.
16. P Gorman, T Sejnowski. Neural Networks 1:75–90, 1988.
17. RL Kashyap, CC Blaydon. Proc IEEE 54:1127–1129, 1966.
18. RL Kashyap, C Blaydon, KS Fu. In: JM Mendel, KS Fu, eds, Adaptive Learning and Pattern Recognition Systems. New York: Academic Press, 1970, pp 329–355.
19. D Kazakos, P Papantoni-Kazakos. Detection and Estimation. Computer Science Press, 1989.
20. J Kiefer, J Wolfowitz. Ann Math Statist 23:462–466, 1952.
21. AG Kogiantis, P Papantoni-Kazakos. IEEE Trans Syst Man Cybern 27 (Part B): June, 1997, to appear.
22. H Kushner. SIAM J Appl Math 47:169–185, 1987.
23. H Kushner, D Clark. Stochastic Approximation Methods for Constrained and Unconstrained Systems. Berlin: Springer-Verlag, 1978.
24. L Ljung. IEEE Trans Automat Contr 22:551–575, 1977.
25. L Ljung, T Söderström. Theory and Practice of Recursive Identification. Cambridge, MA: MIT Press, 1983.
26. M Minsky, S Papert. Perceptrons. Cambridge, MA: MIT Press, 1969.
27. D Pados, P Papantoni-Kazakos. IEEE Trans Neural Networks 6:596–609, 1995.
28. D Pados, KW Halford, D Kazakos, P Papantoni-Kazakos. IEEE Trans Syst Man Cybern 25:21–42, 1995.
29. D Pados, P Papantoni-Kazakos, D Kazakos, A Kogiantis. IEEE Trans Syst Man Cybern 24:1519–1531, 1994.
30. D Pados, P Papantoni-Kazakos. A Note on the Estimation of the Generaliza-

tion Error and the Prevention of Overfitting. Invited paper in the IEEE International Conference on Neural Networks, Orlando, FL, June 1994.

31. D Pados, P Papantoni-Kazakos. A Class of Neyman Pearson and Bayes Learning Algorithms for Neural Classification, IEEE Int Symp Inform Theory, June 27–July 1, 1994, Trondheim, Norway.
32. P Papantoni-Kazakos, D Kazakos, K Birimwal. Inform. Comput 98: 56–98, 1992.
33. H Robbins, S Monro. Ann Math Stat 22:400–407, 1951.
34. F Rosenblatt. The Perceptron: A Perceiving and Recognizing Automaton. Report 85-60-1, Cornell Aeronautical Laboratory, Buffalo, New York, 1957.
35. P Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD Dissertation, Harvard University, Dept. of Applied Mathematics, 1974.
36. H White. Amer Stat Assoc 84:1003–1013, 1989.
37. B Widrow, ME Hoff. Adaptive Switching Circuits. 1960 IRE WESCON Convention Record, 96-104, 1960.
38. B Widrow. In: MD Yovits, GT Jacobi, GD Goldstein, eds, Self-Organizing Systems. Washington, DC: Spartan Books, 1962, pp 435–461.
39. TY Young, RA Westerberg. IEEE Trans Inform Theory IT-18:518–519, 1972.

This Page Intentionally Left Blank

19

Invariant Features Associated with a Conditional Distribution Induced by Self-Similar Patterns

Kohji Kamejima

Osaka Institute of Technology, Osaka, Japan

I. INTRODUCTORY REMARKS

Despite continuous deformation, visual information can be articulated into a set: an *alphabet of patterns*, perhaps consisting of simple figures [5] or fractal patterns [15]. The articulation is the first step in image analysis for interactive design [11,17], cooperative decision [9,18], and propositional scene analysis [10,16]. Imagery consisting of simple *artificial* patterns can be articulated into a system of *a priori* alphabets via statistical and/or syntactic analysis [3,4]. For instance, 3D contours are matched for locating objects in a noisy image [1,7], and 2D lines are extracted for discriminating objects in the *background* [8]. However, these *rigid* alphabets are too susceptive to visual deformation in describing complicated *natural* imagery.

Recently, fractal theory demonstrated that the complexity of various natural images arises from self-similarity [12,14]. For example, the texture on a roadway yields a distribution of similar patterns in an image (Fig. 1). Although the position of each diamond figure is dependent on the viewpoint, a random parameter, the proportion of pattern is well regulated through perspective projection. Figure 2 shows the relation between the size and the location of diamond patterns in photographs taken along the roadway. As indicated in this figure, the size of texture image linearly decreases in accordance with the increase of the height in the image plane. This proportionality is a statistical representation of the roadway scene in terms of visual complexity. Since this visual complexity is not explicitly dependent on a specific texture patterns, we can discriminate the roadway

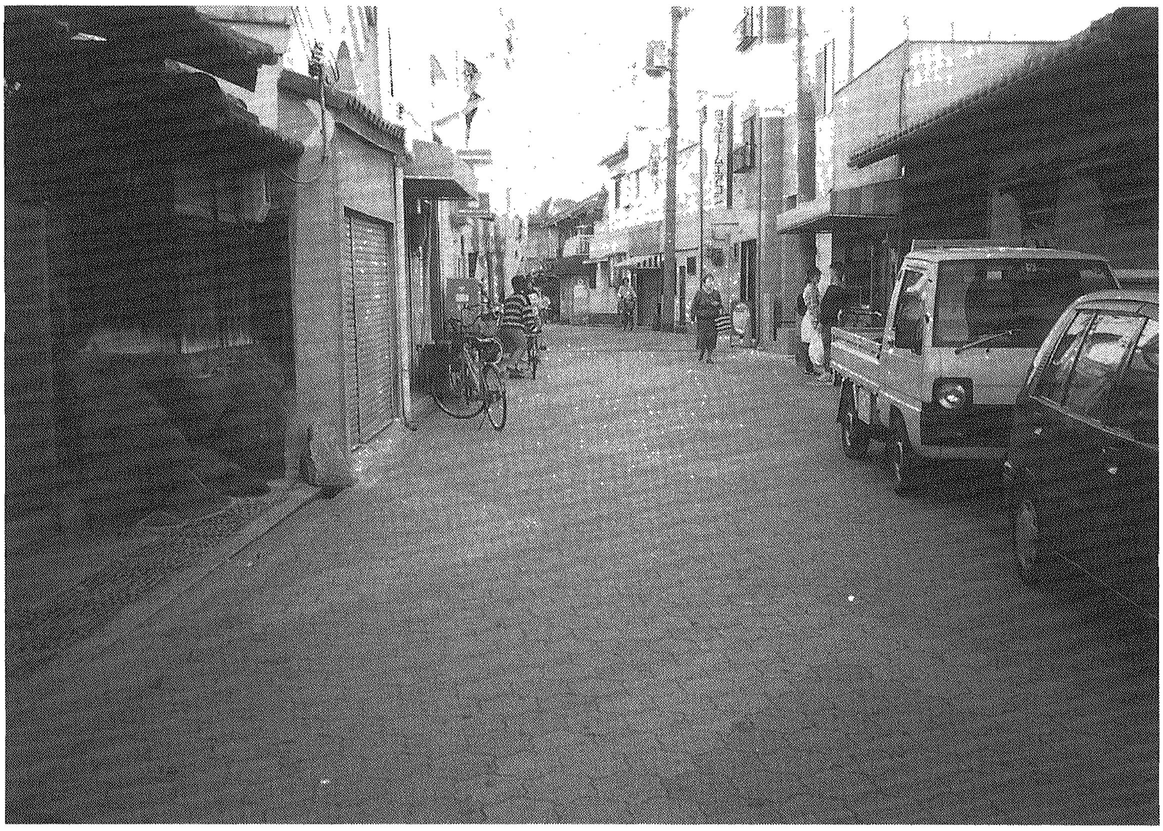


Figure 1 Shopping street.

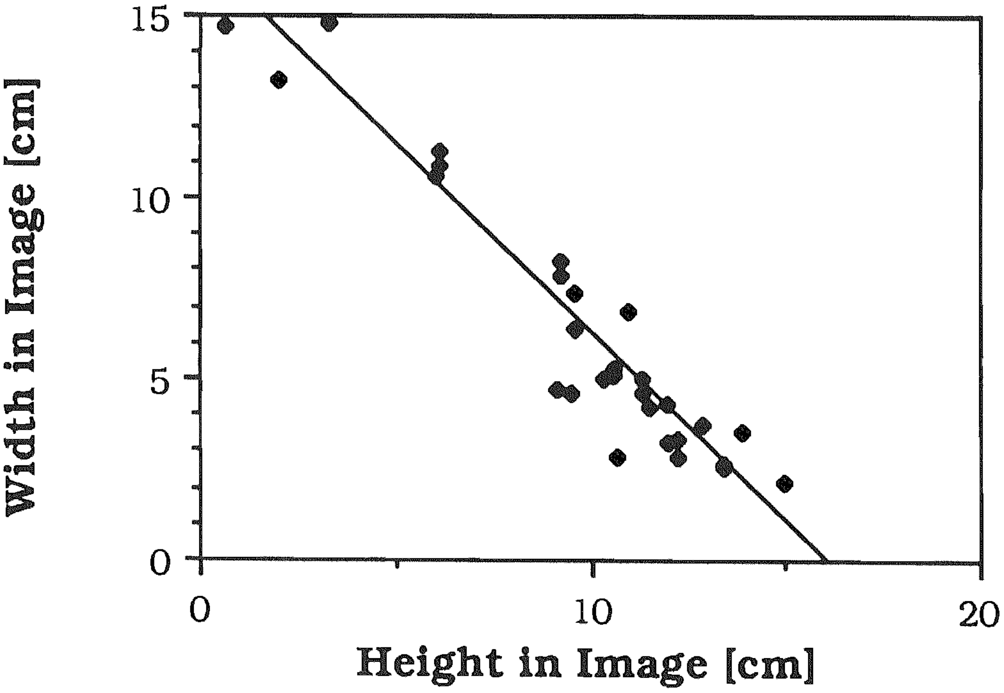


Figure 2 Simple underlying rule in scene image.

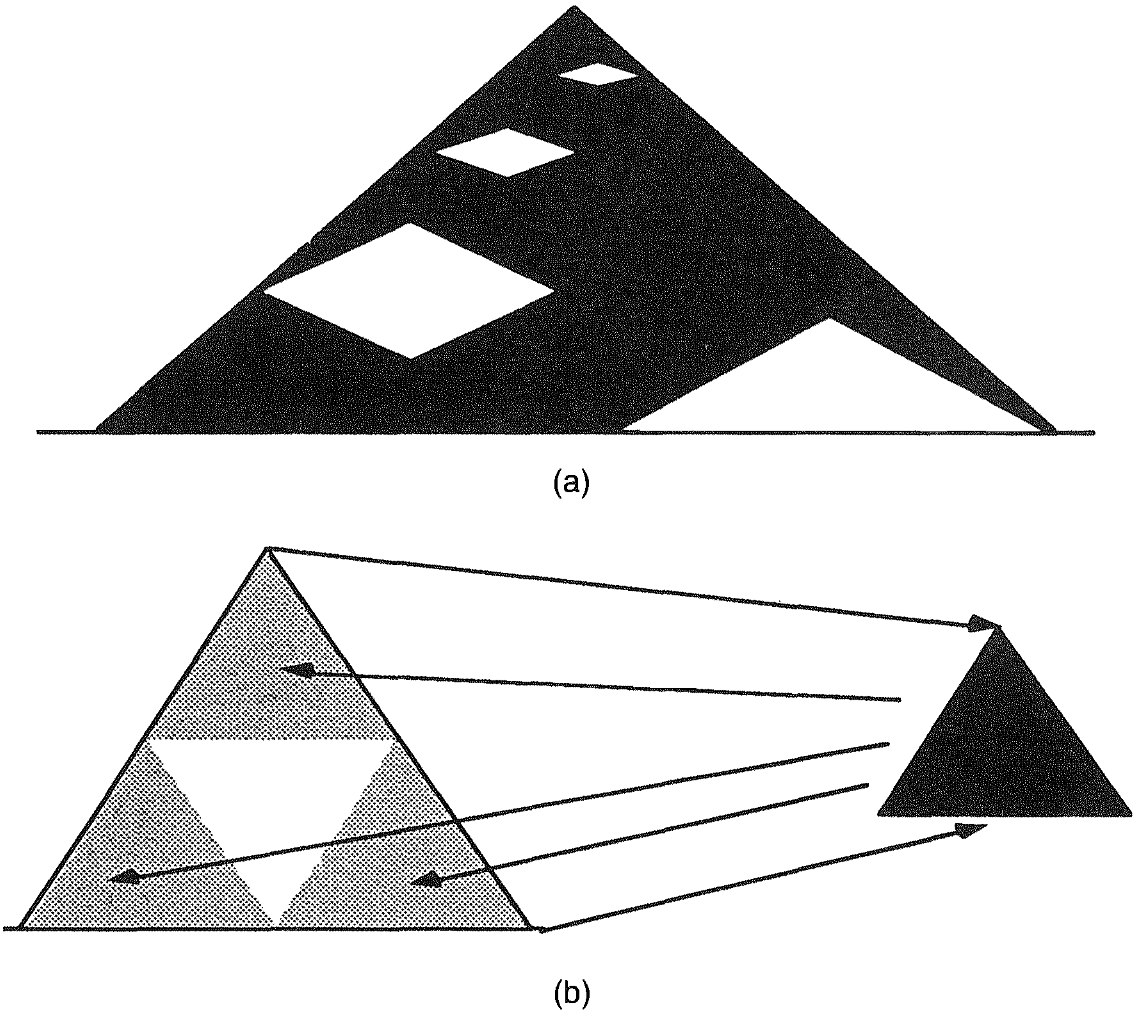


Figure 3 Imaging process: (a) image feature distribution; (b) self-similarity.

pattern from a not-yet-identified texture by testing the proportionality in the observed image. The underlying proportionality in scene imagery, as shown in Fig. 3, implies a dynamical system that generates a self-similar pattern called a "Sierpinski gasket." Thus, we have a set of self-similarity mappings as a generic alphabet for various roadway imagery.

Noticing that the self-similarity can be extended to various natural patterns [2], let the imaging processes be "programmed" in terms of the alphabet $\mu_i, i = 1, 2, 3, \ldots$, where μ_i is the mapping from a bounded image plane $\Omega \subset R^2$ into itself. This program specifies a sequence of patterns $\{\Xi_t\}, \Xi_t = \{\omega_\tau \in \Omega, \tau \leq t\}$, through the explorative expansion:

$$\omega_{t+1} \in \bigcup_{i=1}^{m} \mu_i(\Xi_t) \tag{1a}$$

$$\Xi_0 = \Theta \tag{1b}$$

The imaging process (1) well generates *ordered-by-results* patterns, if they exist, through an *avalanche of exploration* triggered by the preassigned *seeds* Θ. By restricting the *programming language* $\{\mu_t\}$ in the class M, the totality of contraction mappings from Ω into itself, the iteration process (1) has a limit point called an attractor [6]. Thus, it is not so restrictive to identify the object image Λ with the uniquely generated attractor Ξ of the mapping set $\nu = \{\mu_t\}$, i.e.,

$$\Lambda \approx \Xi = \lim_{t \to \infty} \Xi_t \tag{2}$$

The association (2) implies that a class of random patterns can be coded on the basis of contraction mappings. The recognition of such a pattern results in the identification of a mapping set $\nu = \{\mu_1, \mu_2, \mu_3, \ldots, \mu_m\}, \mu_i \in M$, that controls the imaging process (1). Due to the logical operations, however, the coordinate of each point ω_t is not additive with respect to the code $\nu = \{\mu_t\}$. This essential nonlinearity implies that the code $\nu = \{\mu_t\}$ should be determined based on the entire pattern to be generated. By this self-reference, the coding process potentially yields an explosion of decisions. Deterministic computation schemes enumerate this potential explosion through successive explicit search. Furthermore, the generation of the mapping image $\mu_i(\Xi_t)$ requires intolerable computation resources in processing practical imagery. Because of this twofold difficulty, it is not easy to design coding systems within the framework of a deterministic computation scheme.

Despite the nonlinearity of attractors, however, each transient set Ξ_{t+1} is logically additive with respect to μ_t for fixed Ξ_t. This implies that the nonlinear system (1) yields a measurable representation on image plane Ω. Let the class of "visibles" be restricted to $\mathcal{F} = \mathcal{B}[\Omega]$, the Borel field of subsets of Ω, and define $\mathcal{N} = \mathcal{B}[M]$ likewise. Then the generation rule (1) induces the following equivalence:

$$\Lambda \approx \Xi \in \mathcal{F} \Leftrightarrow \nu \in \mathcal{N} \tag{3}$$

This equivalence (3) suggests that the mapping μ_i can be analyzed as an $\mathcal{F}$-measurable random variable taking its values in M. For this purpose, in this chapter, a parallel distributed scheme is introduced to generate a conditional distribution on $(\Omega, \mathcal{F})$. This conditional distribution yields a version of discrete information called feature points. The feature points are shown to have invariant subsets with respect to self-similarity mappings. Thus, the identification of imaging process (1) results in the detection of invariant subsets of feature points.

II. STOCHASTICS IN MORPHOLOGICAL COMPUTATION PROCESS

Consider the similarity evaluation of visible patterns $\Lambda, \Xi \in \mathfrak{F}$. Since visible patterns consist of sample points in the image plane Ω, the disparity between a specific reference point $\lambda \in \Lambda$ and another point set Ξ should be indexed as a point–set function $\Omega \times \mathfrak{F} \rightarrow R^1$. Simultaneously, noticing that the visibles $\Lambda, \Xi \in \mathfrak{F}$ should be matched without explicit dependence on the coordinate system, the similarity of Λ and Ξ is required to be measured in terms of a set–set function $\mathfrak{F} \times \mathfrak{F} \rightarrow R^1$. To maintain the consistency of these indices, the set–set function must be represented as a functional of the point–set function distributed in Ω.

As a common basis of point–pattern similarity and pattern–pattern similarity, define

$$\mathrm{d}P(\omega) = \frac{\mathrm{d}\omega}{\displaystyle\int_{\Omega} \mathrm{d}\omega} \tag{4}$$

where $\mathrm{d}\omega = \mathrm{d}\omega_1 \mathrm{d}\omega_2$ is the volume element in a properly chosen rectangular coordinate system (ω_1, ω_2), and consider the similarity evaluation as a measurable function on the probability space $(\Omega, \mathfrak{F}, P)$. Noting that the contour pattern provides a cue for visual perception [13], let a similarity function φ of a point $\omega \in \Omega$ to a pattern $\Lambda \in \mathfrak{F}$ be represented in terms of the solution to the following boundary value problem:

$$\Delta\varphi(\omega) = \rho\varphi(\omega), \qquad \omega \in (\Lambda - \partial\Lambda) \tag{5a}$$

$$\varphi(\omega) = 0, \qquad \omega \in \Omega - \Lambda \tag{5b}$$

$$\varphi(\omega') = 1, \qquad \omega' \in \partial\Lambda \tag{5c}$$

where $\Delta = \partial^2/\partial\omega_1^2 + \partial^2/\partial\omega_2^2$ is the Laplace operator and ρ is a positive constant. In the distributed parameter system, the observed contour pattern $\partial\Lambda$ specifies the boundary condition (5c) and diffuses inwards in the domain through the local interaction (5a). Thus, the pattern Λ induces a "distribution conditioned by $\partial\Lambda$", $\varphi(\omega | \partial\Lambda)$. Since

$$\Delta\varphi(\omega | \partial\Lambda) \geq 0, \quad \text{a.e. in } \Omega \tag{6}$$

for $\rho > 0$, $\varphi(\omega | \partial\Lambda)$, called the *conditional distribution*, is subharmonic in Λ, i.e.,

$$0 \leq \varphi(\omega | \partial\Lambda) \leq \varphi(\omega' | \partial\Lambda) < \infty, \quad \text{a.e.} \tag{7}$$

for any $\omega \in \Lambda$ and $\omega' \in \partial\Lambda$. By virtue of the external constraint (5b), the maximal principle (7) is extended to Ω. This implies that, in the context of image analysis, the conditional distribution $\varphi(\omega|\partial\Lambda)$ provides an evaluation for the proximity of a point $\omega \in \Omega$ to the boundary $\partial\Lambda$ (see Fig. 4).

Adding to this pointwise proximity indexing, the solution to the system (5) yields a probability measure $P(\sigma|\Lambda)$ given by

$$P(\sigma|\Lambda) = \frac{\displaystyle\int_\sigma \varphi(\omega|\partial\Lambda)\,\mathrm{d}P(\omega)}{\displaystyle\int_\Lambda \varphi(\omega|\partial\Lambda)\,\mathrm{d}P(\omega)} \tag{8}$$

for fixed Λ. By definition, $P(\sigma|\Lambda)$ is a functional of the conditional distribution $\varphi(\omega|\partial\Lambda)$. Furthermore $P(\sigma|\Lambda)$ obeys the following conditions:

$$0 \le P(\sigma|\Lambda) \le P(\Lambda|\Lambda) = 1, \qquad \text{for } \sigma, \Lambda \in \mathcal{F} \tag{9a}$$

$$P(\cup\sigma_i|\Lambda) = \sum_i P(\sigma_i|\Lambda), \qquad \text{for } \sigma_i \in \mathcal{F}, \quad \sigma_i \cap \sigma_j = \varnothing, \tag{9b}$$

$$P(\sigma|\cup\Lambda_i)\,P(\cup\Lambda_i) = \sum_i P(\sigma|\Lambda_i)\,P(\Lambda_i), \qquad \text{for } \Lambda_i \in \mathcal{F},$$
$$\Lambda_i \cap \Lambda_j = \varnothing, \quad P(\Lambda_i) = P(\Lambda_j) \tag{9c}$$

Thus, we have a *conditional probability*, $P(\sigma|\Lambda)$, as an index of the coverage of a given domain Λ by a *model* σ (see Fig. 5). This implies that the distributed information $\varphi(\omega|\partial\Lambda)$ yields a cue for analyzing patterns from the viewpoints of contour similarity as well as domain consistency.

Consider the computable representation of the conditional distribution $\varphi(\omega|\partial\Lambda)$. For simplicity, first, suppose that there exists a contraction mapping $\mu \in M$ that generates Λ, i.e.,

$$\Lambda = \{\omega \in \Omega \,|\, \omega \in \mu(\Omega)\}, \quad \mu \in M \tag{10}$$

In this case, by the combination of the continuity of φ with the maximum principle (7), we have a nonempty set $\bar{\Theta}_\Lambda$ given by:

$$\bar{\Theta}_\Lambda = \{\theta \in \Lambda \,|\, \varphi(\theta|\partial\Lambda) \le \varphi(\omega|\partial\Lambda)\}, \quad \omega \in \Lambda \tag{11}$$

By definition, the set $\bar{\Theta}_\Lambda$ is a representation of the most distant point from the boundary. Since the distribution $\varphi(\omega|\partial\Lambda)$ is uniquely determined for fixed ρ, $\bar{\Theta}_\Lambda$ is a representation for the origin associated with a given pattern Λ. For the convenience of parallel computation, let the *origin* $\bar{\Theta}_\Lambda$ be

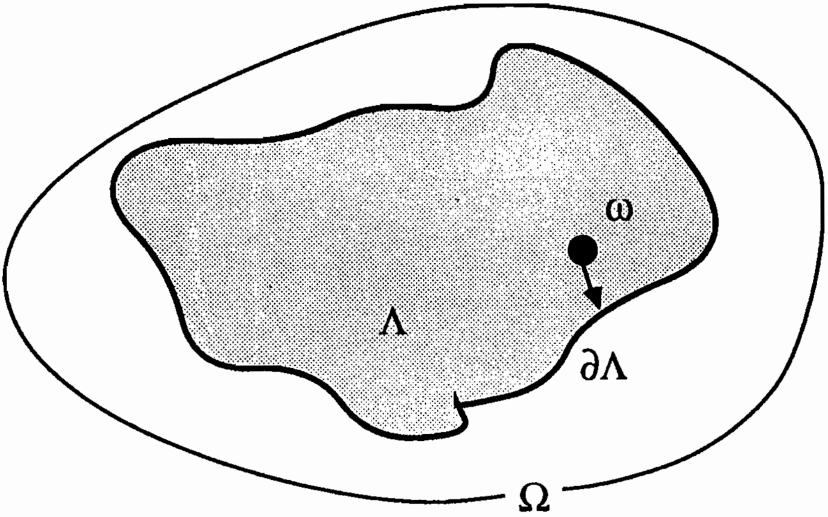


Figure 4 Boundary evaluation.

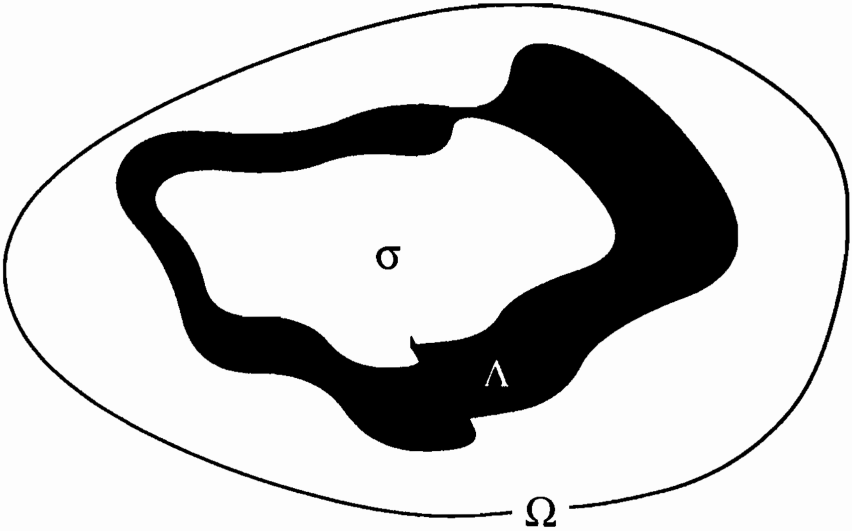


Figure 5 Domain consistency.

substituted by a *feature pattern*, Θ_Λ, given by

$$\Theta_\Lambda = \{\theta \in \Lambda \,|\, \nabla\varphi(\theta|\partial\Lambda) = 0, \varphi(\theta|\partial\Lambda) > 0\} \tag{12a}$$

where

$$\nabla\varphi(\theta|\partial\Lambda) = \begin{bmatrix} \dfrac{\partial\varphi(\theta|\partial\Lambda)}{\partial\omega_1} \\ \dfrac{\partial\varphi(\theta|\partial\Lambda)}{\partial\omega_2} \end{bmatrix} \tag{12b}$$

Noting that $\bar{\Theta}_\Lambda \subset \Theta_\Lambda$, the feature pattern Θ_Λ is nonempty. Furthermore, since the requirement $\varphi(\theta|\partial\Lambda) > 0$ implies $\Delta\varphi(\theta|\partial\Lambda) > 0$, by Eq. (5a), the set Θ_Λ involves a discrete subset consisting of the local minimum points of φ. Thus, the feature pattern Θ_Λ can be exploited as the observation of the origin associated with the image Λ.

The representation (12) can also be applied to a *static* pattern generation process. Let $\Lambda \in \mathcal{F}$ be a fixed observation pattern and suppose that $\nu^0 = \{\mu_i^0 \in M\}$ is a fixed set of contraction mappings satisfying the following condition:

$$\Lambda_i \cap \Lambda_j = \varnothing, \quad \text{for } i \neq j \tag{13}$$

where $\Lambda_i = \mu_i^0(\Lambda)$. Define a mapping $\tilde{\mu}_i$ by

$$\tilde{\mu}_i(\omega) = \omega, \quad \text{for } \omega \in \Lambda_i \tag{14a}$$

$$P\{\tilde{\mu}_i(\Omega - \Lambda_i)\} = P\{\tilde{\mu}_i(\omega) \,|\, \omega \in \Omega - \Lambda_i\} = 0 \tag{14b}$$

for each μ_i^0. Then we have a projector $\tilde{\nu}[\nu^0] = \{\tilde{\mu}_i\}$ associated with ν^0. Define $\Xi^0 = \cup_{\nu^0}\Lambda_i$. Since $\omega \in \cup_{\nu^0}\Lambda_i$ implies $\omega \in \cup_{\tilde{\nu}}\tilde{\mu}_i(\Xi^0)$, we have the recursive representation of the set Ξ^0:

$$\Xi^0 = \left\{\omega \in \bigcup_{\nu^0} \Lambda_i \,\middle|\, \omega \in \bigcup_{\tilde{\nu}} \tilde{\mu}_i(\Xi^0)\right\} \tag{15}$$

Thus, the projector $\tilde{\nu}$ can be identified with the trivial invariant set Ξ^0. In addition, we have a trivial invariant feature associated with a mapping class $\tilde{\nu}$:

$$\Theta^0 = \left\{\theta \in \bigcup_{\nu^0} \Theta^0_{\Lambda_i} \,\middle|\, \theta \in \bigcup_{\tilde{\nu}} \tilde{\mu}_i(\Theta^0)\right\} \tag{16a}$$

where

$$\Theta^0_{\Lambda_i} = \left\{\theta_i \in \bigcup_{\nu^0} \Lambda_i \,\middle|\, \nabla\varphi(\theta_i|\partial\Lambda_i) = 0, \varphi(\theta_i|\partial\Lambda_i) > 0\right\} \tag{16b}$$

In Eq. (16b), $\partial\Lambda_i$ denotes the boundary of the associated domain Λ_i.

III. EXTENSION TO SELF-SIMILAR PATTERNS

As shown in Eqs. (15) and (16), the projector $\bar{\nu}[\nu^0]$ generates the invariant feature Θ^0 associated with the trivial invariant set Ξ^0. This implies that we can verify the generator ν^0 through a computation process on the discrete information Θ^0. This computability can be extended to generalized invariant sets. Let the code $\nu = \{\mu_i \in M\}$ be a set of contraction mappings satisfying the open set condition [6]: i.e., there exists an open set $O \in \mathfrak{F}$ satisfying the following conditions

$$\bigcup_{\nu} \mu_i(O) \subset O \tag{17a}$$

with

$$\mu_i(O) \cap \mu_j(O) = \varnothing \quad \text{for any } \mu_i, \mu_j \in \nu \tag{17b}$$

Suppose that Ξ is an associated invariant set:

$$\Xi = \left\{ \omega \in \Omega \,\middle|\, \omega \in \bigcup_{\nu} \mu_i(\Xi) \right\} \tag{18}$$

By the open set condition (17), the generated attractor Ξ has fractal dimension. For such a singular pattern, define a smoothed region by

$$\Xi^{\varepsilon} = \left\{ \omega \in \Omega \,\middle|\, \int_{\Omega} m(\omega - \xi)\,\mathrm{d}\chi(\xi) > 0 \right\}, \tag{19a}$$

where $\chi[\cdot]$ is the invariant measure associated with the attractor $(\cdot)$ and $m(\omega)$ denotes a mollifier:

$$m(\omega) = \begin{cases} m_0 \cdot \exp\left(-\dfrac{1}{1 - |\omega|^2} \right) & \text{for } |\omega| < 1; \\ 0 & \text{for } |\omega| \geq 1 \end{cases} \tag{19b}$$

Since the support of the invariant measure $\chi[\Xi]$ is identical to the attractor Ξ [2], it follows that

$$\int_{\Omega} m(\omega - \xi)\,\mathrm{d}\chi(\xi) > 0 \quad \text{for } \min_{\xi \in \mathrm{X}} |\omega - \xi| < \varepsilon \tag{20}$$

Thus, the domain Ξ^{ε} is a slightly expanded smooth region around the attractor Ξ. By solving the boundary value problem (5) in the smoothed domain Ξ^{ε}, we have the following discrete feature pattern

$$\bar{\Theta} = \{ \bar{\theta} \in \Xi^{\varepsilon} \,|\, \nabla\varphi(\bar{\theta}\,|\,\partial\Xi^{\varepsilon}) = 0, \varphi(\bar{\theta}\,|\,\partial\Xi^{\varepsilon}) > 0 \} \tag{21}$$

On the basis of this observable $\bar{\Theta}$, we can specify the sufficient condition

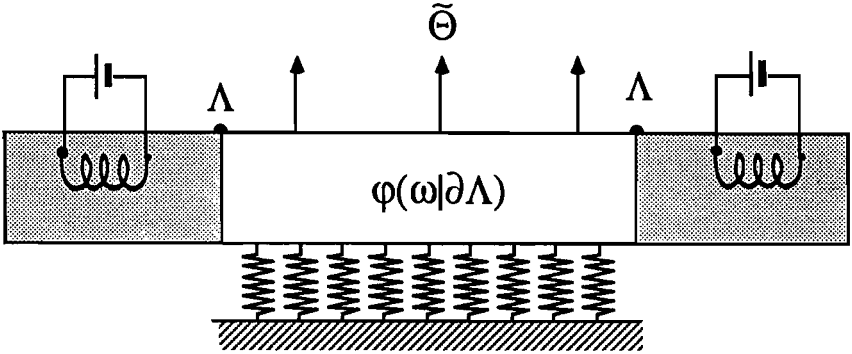


Figure 6 Feature pattern detector.

for the existence of an invariant feature associated with ν:

$$\Theta = \left\{ \{\theta \in \tilde{\Theta} \,\middle|\, \theta \in \bigcup_{\nu} \mu_i(\Theta)\} \right. \tag{22}$$

The existence of the invariant feature is presented in the next section. The definition of conditional distribution combined with the description of feature patterns implies the existence of a parallel distributed system as shown in Fig. 6. This system *filters* the random pattern Λ to yield the feature pattern $\tilde{\Theta}$. Although this filtering process is described without any explicit information on the imaging process to be identified, the resulting feature pattern provides a version of the observables of the constraint (22). Thus, the feature pattern detector can be exploited as the preprocessor of the computation process for the joint detection of the contraction code $\nu = \{\mu_i \in M\}$ and the invariant feature Θ.

IV. EXISTENCE OF AN INVARIANT FEATURE

By a system of coordinate transforms $\nu = \{\mu_i \in M\}$, the observable $\tilde{\Theta}$ is transferred in a neighborhood of $\tilde{\Theta}$ in Ω. This suggests that the self-similarity domain transformation $\cup_\nu \mu_i(\Xi) \to \Xi$ yields the range $\cup_\nu \mu_i(\tilde{\Theta})$ covering $\tilde{\Theta}$. However, the distribution $\varphi(\omega | \partial\mu_i(\Xi^\varepsilon))$ is not identical with the restriction of $\varphi(\omega | \partial\Xi^\varepsilon)$ into the domain $\mu_i(\Xi^\varepsilon)$ due to the disparity of the boundary condition (Fig. 7). In this section, the trajectory of the observables $\tilde{\Theta}$ is investigated under a *boundary preserving contraction process*.

First, we notice the continuity of the contraction mapping, i.e.,

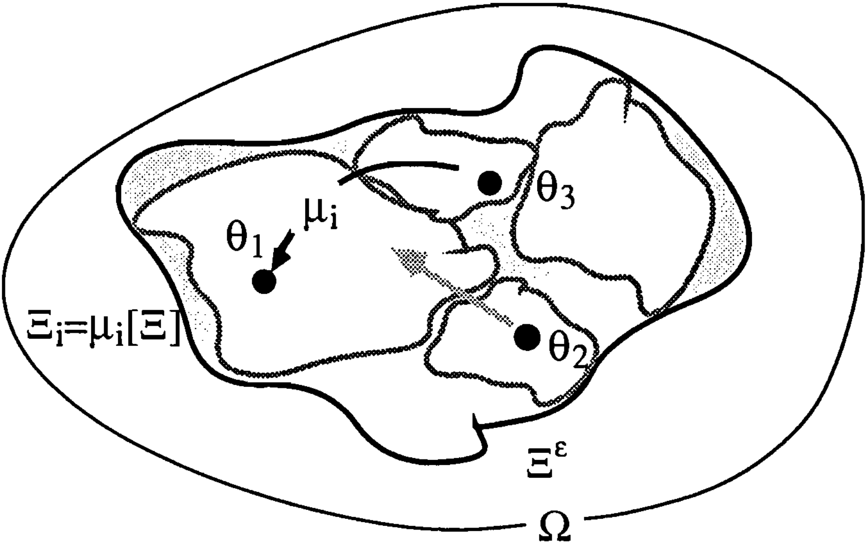


Figure 7 Existence of an invariant feature.

Proposition 1 (Fig. 8). *Suppose that μ is a contraction mapping of Ω into itself. Then, for any $\delta > 0$, we have*

$$\mu[B_\delta(\omega)] \subset B_\delta(\mu[\omega]) \tag{23}$$

where $B_\delta(\omega)$ is a closed ball around ω of radius δ.

For the smoothed region Ξ^ϵ with feature pattern $\bar{\Theta}$ given by (21), consider a simple contraction process. Noting that the continuity of the function φ is preserved by a continuous mapping, by Proposition 1, we have

Proposition 2 (Fig. 9). *Let μ be a contraction mapping of Ξ^ε into itself and suppose that $\bar{\Xi}_\mu$ are feature points associated with the reduced region $\Xi_\mu = \mu(\Xi^\varepsilon)$, i.e.,*

$$\bar{\Theta}_\mu = \{\bar{\theta} \in \Xi_\mu \mid \nabla\varphi_\mu = 0, \varphi_\mu > 0\} \tag{24}$$

where

$$\Delta\varphi_\mu(\omega) = \rho\varphi_\mu(\omega), \quad \omega \in \Xi_\mu \tag{25a}$$
$$\varphi_\mu(\omega) = 0, \quad \omega \in \Omega - [\Xi_\mu \cup \partial\Xi_\mu] \tag{25b}$$
$$\varphi_\mu(\omega') = 1, \quad \omega' \in \mu(\partial\Xi_\mu) \tag{25c}$$

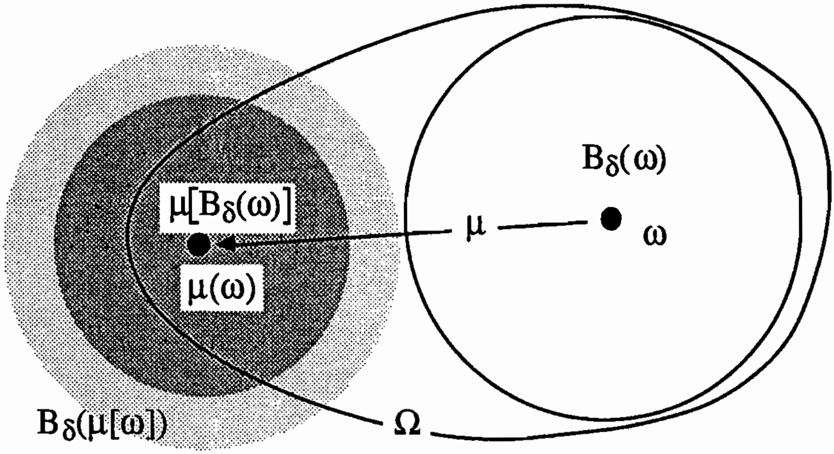


Figure 8 Domain image.

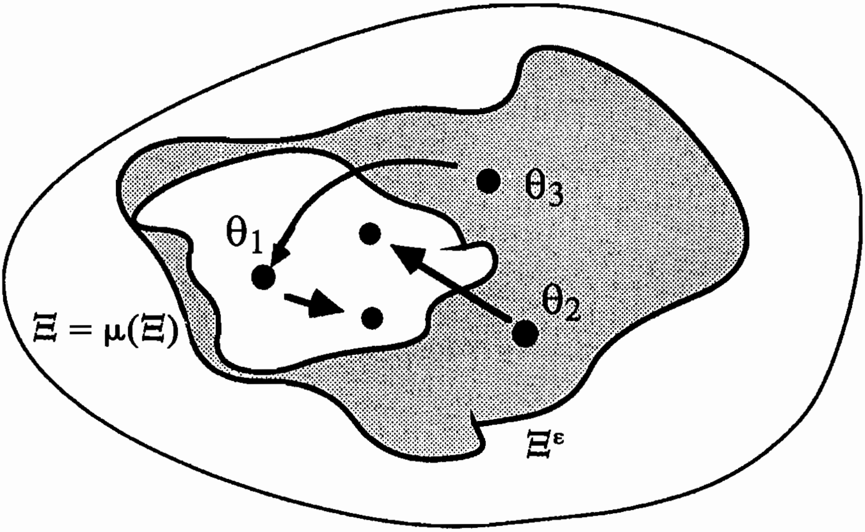


Figure 9 Feature regeneration.

Then

$$\nabla\varphi_\mu[\mu(\bar\theta)] = 0, \quad \textit{for any } \bar\theta \in \bar\Theta, \tag{26a}$$

or, equivalently,

$$\mu(\tilde\Theta) = \tilde\Theta_\mu \tag{26b}$$

For tracing the trajectory under boundary preserving contractions, define $\partial_i = \mu_i(\partial\Xi^\varepsilon)$ and consider the solution $\{\varphi_i, i = 1, 2, \ldots\}$ to the following system:

$$\Delta\varphi_i(\omega) = \rho\varphi_i(\omega), \quad \omega \in \mu_i(\Xi^\varepsilon) \tag{27a}$$

$$\varphi_i(\omega) = 0, \quad \omega \in \Omega - [\mu_i(\Xi^\varepsilon) \cup \partial_i] \tag{27b}$$

$$\varphi_i(\omega') = 1, \quad \omega' \in \partial_i \tag{27c}$$

For this system, let the feature distribution $\cup_\nu \mu_i(\tilde{\Theta})$ be transferred to the feature pattern associated with a *boundary preserving map* through the continuation process (Fig. 10). In this process, first, the domain is transferred with associated boundary, as in the system (25), then the mapped boundary vanishes smoothly. This continuation process is described by the following one-parameter group of distributed parameter system:

$$\Delta\varphi^\alpha(\omega) = \rho\varphi^\alpha(\omega), \quad \omega \in \Xi^\varepsilon - \bigcup_\nu \partial_i \tag{28a}$$

$$\varphi^\alpha(\omega^\circ) = (1-\alpha) + \alpha\varphi(\omega^\circ), \quad \omega^\circ \in \bigcup_\nu \partial_i \subset \Xi^\varepsilon \tag{28b}$$

$$\varphi^\alpha(\omega) = 0, \quad \omega \in \Omega - [\Xi^\varepsilon \cup \partial\Xi^\varepsilon] \tag{28c}$$

$$\varphi^\alpha(\omega') = 1, \quad \omega' \in \partial\Xi^\varepsilon \tag{28d}$$

where $\alpha \in [0,1]$. Obviously the solution φ^α is a homotopy, i.e.,

$$\varphi^\alpha(\omega) : \varphi_i(\omega) \rightarrow \varphi(\omega) \quad \text{for } \alpha : 0 \rightarrow 1 \text{ and } \omega \in \mu_i(\Xi^\varepsilon) \tag{29}$$

By the continuity of the solution φ^α with respect to the boundary value $\varphi^\alpha(\omega^\circ)$, the variation of the gradient field $\nabla\varphi^\alpha$ is evaluated as follows:

$$\frac{\mathrm{d}\nabla\varphi^\alpha}{\mathrm{d}\alpha} \approx \frac{\partial\nabla\varphi^\alpha}{\partial\alpha} + [\Delta\varphi^\alpha] \frac{\mathrm{d}\tilde{\theta}^\alpha}{\mathrm{d}\alpha} \tag{30}$$

where

$$\tilde{\theta}^\alpha : \nabla\varphi^\alpha(\bar{\theta}^\alpha) = 0 \quad \text{and} \quad \varphi^\alpha(\tilde{\theta}^\alpha) > 0 \tag{31}$$

By letting

$$\frac{\mathrm{d}\nabla\varphi^\alpha(\omega^\circ)}{\mathrm{d}\alpha} = 0, \quad \text{at } \omega^\circ \in \bigcup_\nu \partial_i \subset \Xi^\varepsilon \tag{32}$$

we have a formal description of a curve Q that governs the movement of

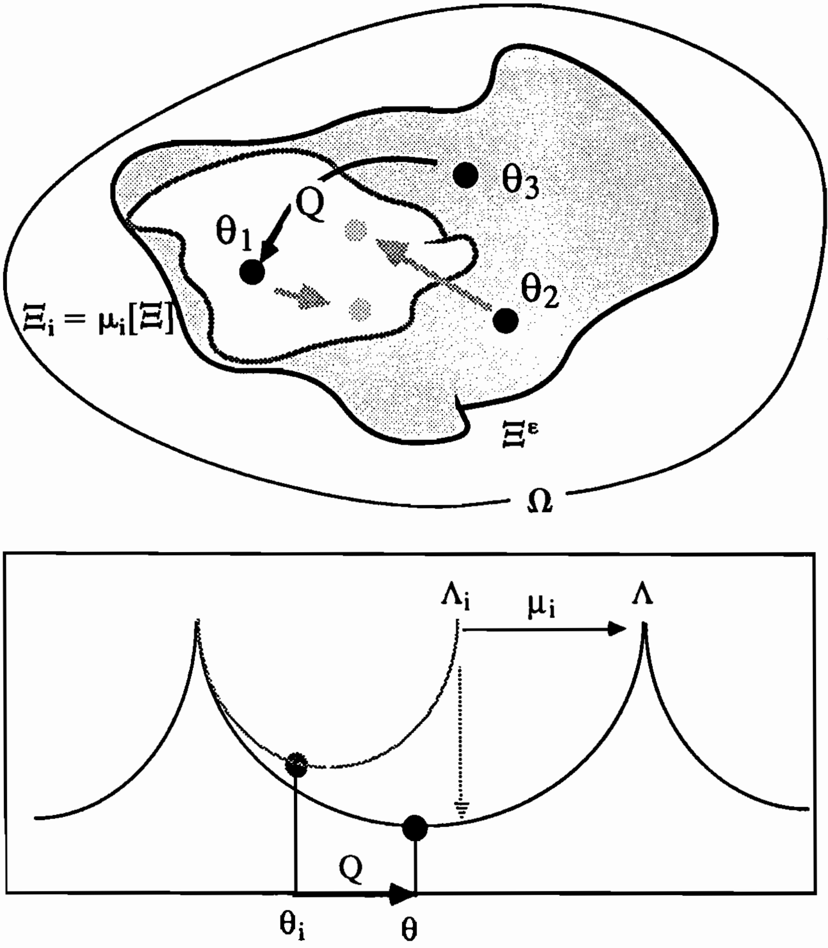


Figure 10 Homotopy associated with self-similar mappings.

the feature point $\bar{\theta}^{\alpha}$ as follows:

$$Q:\frac{\mathrm{d}\bar{\theta}^{\alpha}}{\mathrm{d}\alpha}=-\left(\frac{1}{\rho\varphi^{\alpha}}\right)\frac{\partial\nabla\varphi^{\alpha}}{\partial\alpha},\quad \text{for } 0\leq\alpha\leq 1 \tag{33}$$

Since $\varphi^{\alpha}(\bar{\theta}^{\alpha}|\partial_i)>0$ in each $\mu_i(\Xi^{\varepsilon})$, the ordinary differential equation (33) with initial condition $\nabla\varphi^{0}(\bar{\theta}^{0}|\partial_i)=0$ has a unique solution. By the existence of the uniquely determined curve Q, we have the following

Proposition 3 (Maximality of homotopy). *The feature point $\bar{\theta}^{\alpha}$, $0\leq\alpha\leq 1$, lies in a curve Q.*

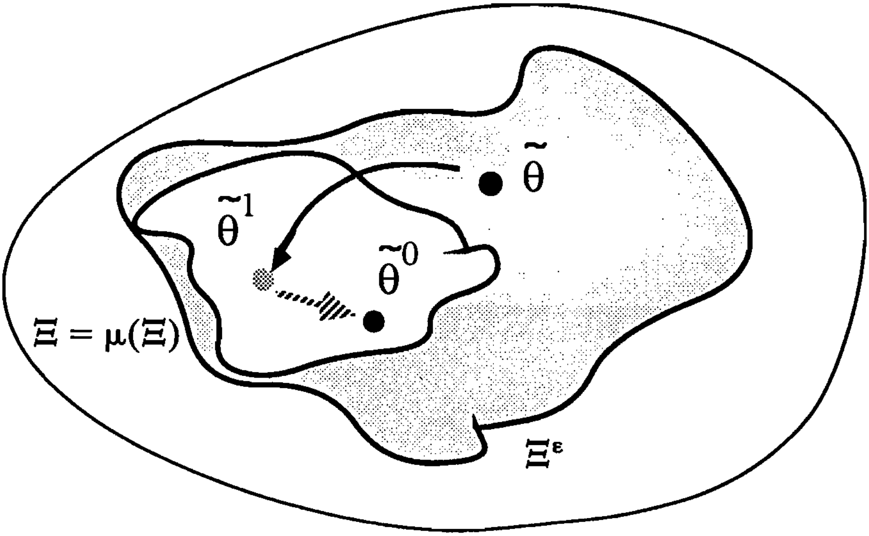


Figure 11 Variation of homotopic path.

The smoothness of the homotopic path is combined with the continuous dependence of the distribution φ on the boundary value to evaluate the deviation of the feature pattern as follows.

Proposition 4 (Fig. 11). *The variation of the homotopic path satisfies the following inequality*:

$$|\bar{\theta}^1 - \bar{\theta}^0| \le C\varepsilon \tag{34}$$

for some constant C.

Proof. From Eq. (33), it follows that

$$|\bar{\theta}^1 - \bar{\theta}^0| = \left| \int_0^1 \frac{\nabla\varphi_i(\bar{\theta}^\alpha) - \nabla\varphi(\bar{\theta}^\alpha)}{\rho\Theta^\alpha(\omega)} \mathrm{d}\alpha \right| \le \frac{\sup\limits_{\omega\in\mu_i(\mathrm{X}^\varepsilon)} |\nabla\varphi_i(\omega) - \nabla\varphi(\omega)|}{\rho \inf\limits_{\omega\in\mu_i(\mathrm{X}^\varepsilon)} \varphi^\alpha(\omega)} \tag{35}$$

along the path $\bar{\theta}^\alpha$, $0 \le \alpha \le 1$. Since $\varphi^\alpha(\omega)$ is positive definite, i.e.,

$$\inf_{\omega\in\mu_i(\mathrm{X}^\varepsilon)} \varphi^\alpha(\omega) \ge \frac{1}{C_0} \tag{36}$$

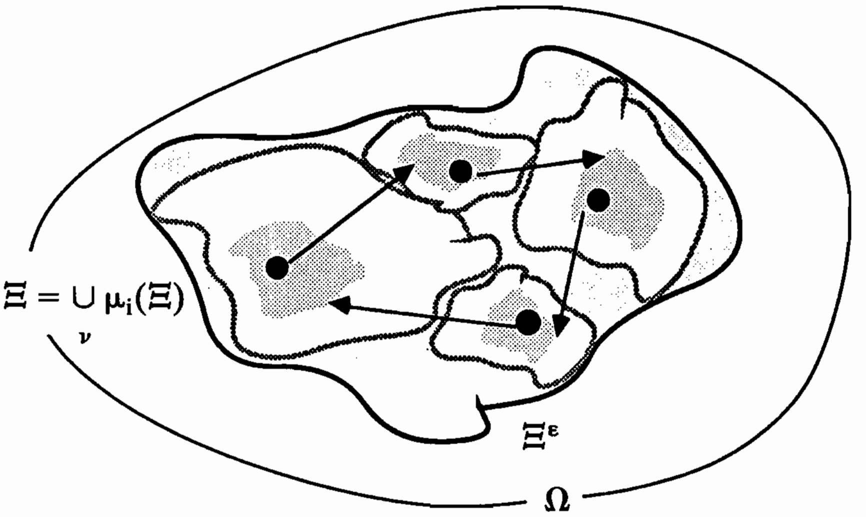


Figure 12 Approximately invariant feature.

for some constant C_0, it follows that

$$|\bar{\theta}^1 - \bar{\theta}^0| \le \frac{C_0}{\rho} \sup_{\omega^\circ \in \partial\mu_i(X^\varepsilon)} |\nabla\varphi_i(\omega^\circ) - \nabla\varphi(\omega^\circ)| \tag{37}$$

Noticing that the solution to the elliptic equation is infinitely differentiable and continuously dependent on the boundary value, we have

$$\sup_{\omega \in \mu_i(X^\varepsilon)} |\nabla\varphi_i(\omega) - \nabla\varphi(\omega)| \le C_1 \sup_{\omega^\circ \in \partial\mu_i(X^\varepsilon)} |1 - \varphi(\omega^\circ)| \tag{38}$$

$$|1 - \varphi(\omega^\circ)| \le C_2 \inf_{\substack{\omega' \in \partial X^\varepsilon \\ \omega^\circ \in \partial\mu_i(X^\varepsilon)}} |\omega' - \omega^\circ| \tag{39}$$

for constants C_1 and C_2. The combination of (38) and (39) yields the following evaluation

$$|\bar{\theta}^1 - \bar{\theta}^0| \le \frac{C_0}{\rho} C_1 C_2 \varepsilon \tag{40}$$

as was to be proved.

The existence and evaluation of the homotopic path are combined to prove the existence of an invariant feature. Assume that the expansion parameter

ε is so small that all feature points $\tilde{\theta} \in \tilde{\Theta}$ are included in the attractor Ξ. Then we have the following results.

Existence Theorem (Fig. 12). *Suppose that Ξ is the attractor of a set of contraction mappings $\nu \in \mathcal{N}$ and assume that the associated smoothed region Ξ^ε yields the feature pattern $\tilde{\Theta}$ satisfying the following condition*:

$$\tilde{\Theta} \subset (\Xi - \partial\Xi) \tag{41}$$

Then there exists a subset of $\tilde{\Theta}$, designated by the invariant feature, satisfying the following condition:

$$\Theta = \left\{ \theta \in \tilde{\Theta} \,\middle|\, \theta \in \bigcup_{\nu} \langle \mu_i(\Theta) \rangle_\delta \right\} \tag{42}$$

for some constant $\delta > 0$. In Eq. (42), $\langle \mu_i(\Theta) \rangle_\delta$ denotes the region in Ω given by

$$\langle \mu_i(\Theta) \rangle_\delta = \bigcup_{\theta \in 6} \langle \mu_i(\theta) \rangle_\delta \tag{43a}$$

where

$$\langle \theta \rangle_\delta = \{ \omega \in \Omega \,|\, \varphi(\theta) < \varphi(\omega) < \varphi(\theta) + \delta \} \tag{43b}$$

for $\theta \in \Theta$.

Proof. Consider the ν-Q transformation specified by the mappings ν and homotopic path Q. By assumption (41) combined with Proposition 2, we have

$$\bigcup_{\nu} \mu_i(\tilde{\Theta}) \subset \Xi \tag{44}$$

By invoking Proposition 4, the mapping image $\cup_\nu \mu_i(\tilde{\Theta})$ generates a set of feature points, if they exist, in an associated $C\varepsilon$-neighborhood given by

$$\bigcup_{\nu} B_{C\varepsilon}(\mu_i(\tilde{\Theta})) = \left\{ \omega \in \Xi^\varepsilon \,\middle|\, |\omega - \tilde{\theta}| < C\varepsilon, \tilde{\theta} \in \bigcup_{\nu} \mu_i(\tilde{\Theta}) \right\} \tag{45}$$

The existence of nonvanishing homotopic points is ensured by the self-similarity of the domain Ξ: under the consistency condition (41), the nonempty set $\Theta \subset \Xi$ must be regenerated via this ν-Q process. In other words, there exists a homotopic path Q with a nonvanishing end point of the homotopic path $Q, \tilde{\theta}^\circ \in \tilde{\Theta}$, such that

$$\tilde{\theta} \in \bigcup_{\nu} B_{C\varepsilon}(\mu_i(\tilde{\theta}^\circ)) \tag{46}$$

for each feature point $\tilde{\theta}\in\tilde{\Theta}\neq\varnothing$. The remainder of this proof is devoted to the φ-evaluation of the $C\varepsilon$-neighborhood. Since $\Delta\varphi(\omega)>0$ for arbitrary $\omega\in\Xi^{\varepsilon}$, the region $\langle\tilde{\theta}\rangle_{\delta}$ is nonempty for any $\delta>0$ and $\tilde{\theta}\in\tilde{\Theta}$. Select the smoothing parameter $\varepsilon>0$ so that the following condition is satisfied:

$$B_{C\varepsilon}(\tilde{\theta})\subset\langle\tilde{\theta}\rangle_{\delta} \tag{47}$$

for arbitrary $\tilde{\theta}\in\tilde{\Theta}$. For such a slightly smoothed pattern Ξ^{ε}, the associated feature points $\tilde{\Theta}$ are transferred into $\cup_{\nu}B_{C\varepsilon}(\mu_i(\tilde{\Theta}))\subset\cup_{\nu}\langle\mu_i(\tilde{\Theta})\rangle_{\delta}$ via the ν-Q process. Thus, the combination of the expansion estimation (47) with the regenerativity (46) yields that

$$\tilde{\Theta}\subset\bigcup_{\nu}\langle\mu_i(\tilde{\Theta})\rangle_{\delta} \tag{48}$$

as was to be proved.

Remark. Due to the disparity of the boundary location, the feature pattern is not invariant with respect to the contraction code ν, i.e., the mapping image of the feature points is not ensured to be on the feature points. This theorem claims the existence of an approximately invariant subset if the feature pattern satisfies the partially consistent condition (41). Furthermore, this existence theorem provides a practical method for estimating the order of approximation in terms of the field information φ.

V. EXPERIMENTS

The coding scheme was verified through simulation studies. The patterns to be observed are generated through a Monte Carlo simulation [2] and identified within a preassigned set of attractors, called the *dictionary*. An example of an observed pattern, called "ROAD", is shown in Fig. 13 and the dictionary is illustrated in Fig. 14. These fractal patterns are generated based on the set of contraction mappings; the alphabet is of the following form: ((REDUCTION (*rh rv*)) (ROTATION *rt*) (shift (*sh sv*))), where *rh*, *rv*, *rt*, *sh* and *sv* denote pattern manipulation for horizontal reduction, vertical reduction, counterclockwise rotation, horizontal shift, and vertical shift, respectively. For example, the "ROAD" pattern is generated via the following set of reduced-and-shift mappings:

((REDUCTION (50% 50%)) (ROTATION 0) (shift (0% 0%)))
((REDUCTION (50% 50%)) (ROTATION 0) (shift (−45% −25%)))
((REDUCTION (50% 50%)) (ROTATION 0) (shift (45% −25%)))
((REDUCTION (50% −50%)) (ROTATION 0) (shift (0% −50%)))

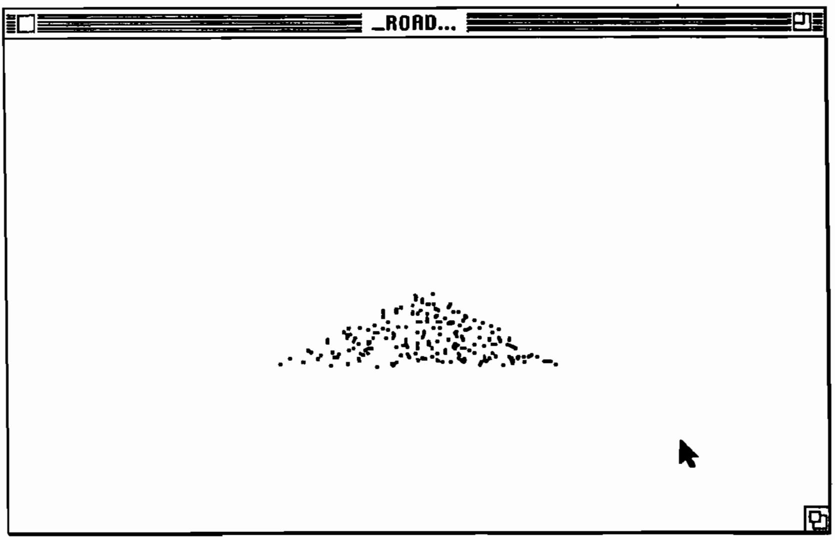
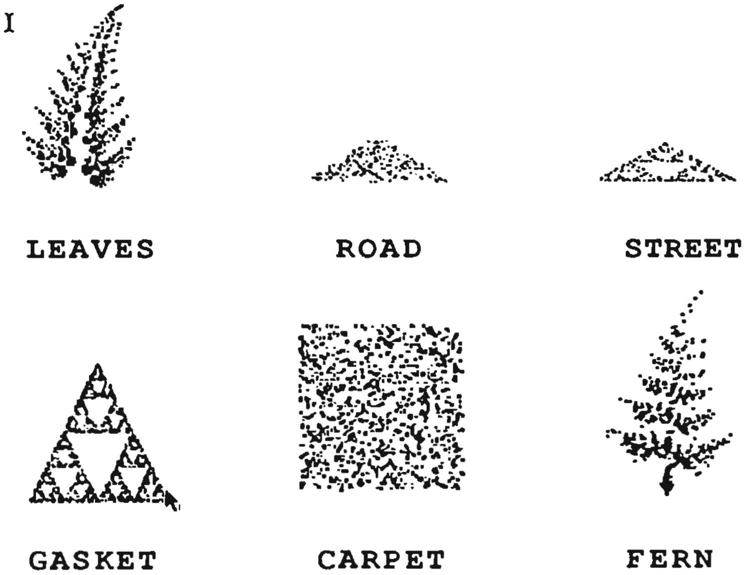


Figure 13 Pattern to be detected.

where a negative vertical reduction means the "FLIP-VERTICAL" operation. On the other hand, the "FERN" pattern in Fig. 14 is coced in terms of the following reduced-rotation-and-shift alphabets:

```
((REDUCTION (84% 84%)) (ROTATION −3deg) (shift (4% 17%)))
((REDUCTION (−33% 33%)) (ROTATION 65deg) (shift (30% 61%)))
((REDUCTION (33% 33%)) (ROTATION 65deg) (shift (−30% −52%)))
((REDUCTION (9% 18%)) (ROTATION 0) (shift (0% −80%)))
```

The detection scheme was implemented on a digitized image plane with 32×32 resolution. In what follows, each image point is identified with the coordinate (i, j), $0 \leq i, j \leq 31$ in the digital image plane. In detecting the invariant feature, a given pattern Λ is assumed to be well approximated by an attractor Ξ, i.e., in the coding processes, each pattern Λ is identified with a slightly smoothed expansion Ξ^{ϵ} as a digital image. Simulation results for the "ROAD" pattern are shown in Fig. 15 where the observed pattern $\Lambda = \Xi^{\epsilon}$ and the extracted feature patterns $\tilde{\Theta}$ are indicated in views entitled "Observables" and "Features", respectively. For extracting the feature pattern $\tilde{\Theta}$, a sequence of conditional distributions $\{\varphi_t(i, j \mid \partial\Lambda),$ $t = 1, 2, 3, \ldots\}$ was generated in the domain Λ through the following



LEAVES ROAD STREET

GASKET CARPET FERN

Figure 14 Dictionary of patterns.

successive approximation scheme:

$$\varphi_{t+1}(i,j\,|\,\partial\Lambda) = \begin{cases} \dfrac{1}{1+\rho/4}\,\overline{\varphi}_t(i,j\,|\,\partial\Lambda) & \text{for } (i,j)\in\Lambda \\ 1 & \text{for } (i,j)\in\Omega-\Lambda \end{cases} \tag{49a}$$

where $\overline{\varphi}_t(i,j\,|\,\partial\Lambda)$ denotes the local average defined by

$$\begin{aligned}\overline{\varphi}_t(i,j\,|\,\partial\Lambda) = \frac{1}{4}[&\varphi_t(i-1,j\,|\,\partial\Lambda) + \varphi_t(i,j-1\,|\,\partial\Lambda) + \varphi_t(i,j+1\,|\,\partial\Lambda) \\ &+ \varphi_t(i+1,j\,|\,\partial\Lambda)]\end{aligned} \tag{49b}$$

In this simulation, the parameter ρ was chosen so that the "viability" factor $1/(1+\rho/4)$ is adjusted to 1/2 and the initial condition was given by

$$\varphi_0(i,j\,|\,\partial\Lambda) = \begin{cases} 0 & \text{for } (i,j)\in\Lambda \\ 1 & \text{for } (i,j)\in\Omega-\Lambda \end{cases} \tag{49c}$$

The iteration process (49) was halted when the following condition was satisfied:

$$\max_{\substack{0\le i\le 31\\ 0\le j\le 31}} |\varphi_t(i,j\,|\,\partial\Lambda) - \varphi_{t-1}(i,j\,|\,\partial\Lambda)| < 0.01 \tag{50}$$

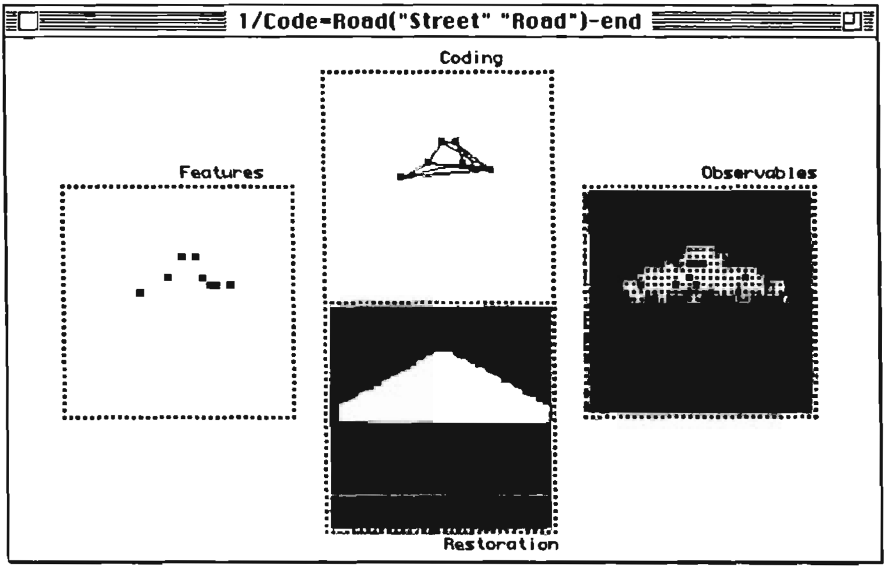


Figure 15 Coding results.

By considering $\varphi(i,j|\partial\Lambda) \approx \varphi_{t^*}(i,j|\partial\Lambda)$, where t^* is the halting time satisfying the inequality (50), the feature pattern was extracted via the following local analysis:

$$\begin{aligned}\tilde{\Theta} = \{(i,j), 0 \le i,j \le 31 \,|\, &[\varphi(i-1,j|\partial\Lambda) - \varphi(i,j|\partial\Lambda)] \\ &\times [\varphi(i+1,j|\partial\Lambda) - \varphi(i,j|\partial\Lambda)] > 0\, [\varphi(i,j+1\,|\,\partial\Lambda) - \varphi(i,j|\,\partial\Lambda] \\ &\times [\varphi(i,j-1\,|\,\partial\Lambda) - \varphi(i,j|\partial\Lambda)] > 0\} \qquad (51)\end{aligned}$$

The consistency of each dictionary pattern was tested through invariant feature detection in this feature pattern $\tilde{\Theta}$. The invariant feature Θ was computed through the following successive algorithm:

$$\Theta_{\tau+1} = \left\{ \theta \in \tilde{\Theta} \,\middle|\, \theta \in \bigcup_{\nu} \mu_i(\Theta_\tau) \right\} \qquad (52a)$$

with initial condition

$$\Theta_0 = \tilde{\Theta} \qquad (52b)$$

where $\nu = \{\mu_i\}$ is selected in the dictionary. Since the scheme reduces the finite set $\tilde{\Theta}$ monotonically for fixed ν, the sequence Θ_τ, $\tau = 1,2,\ldots$ converges to the invariant feature Θ within finite steps if $\tilde{\Theta} \cap (\Xi - \partial\Xi) \ne \varnothing$ and $\Theta \ne \varnothing$. The contraction code ν detected via the successive scheme

(52) and resulting attractor are shown in "Coding" and "Restoration" views, respectively. This "Coding" view exhibits the totality of connections that is generated by the detected contraction code ν in the feature pattern $\tilde{\Theta}$.

As shown in Fig. 15, the coding scheme successfully selects the generator of a given almost self-similar pattern Λ in the dictionary. Generally, the imaging process (1) is monotonic with respect to contraction codes in the following sense:

$$\text{If } \nu_1 \subset \nu_2 \text{ then } \Xi_1 \subset \Xi_2 \tag{53}$$

where Ξ_i, $i = 1, 2$, are attractors associated with the contraction codes ν_i, $i = 1, 2$, respectively. Let $\tilde{\Theta}_i$, $i = 1, 2$, be restrictions of the feature pattern $\tilde{\Theta}$ given by

$$\tilde{\Theta}_i = \tilde{\Theta} \cap \Xi_i, \quad i = 1, 2 \tag{54}$$

Obviously $\tilde{\Theta}_1 \subset \tilde{\Theta}_2$ for the attractors Ξ_1 and Ξ_2 satisfying $\Xi_1 \subset \Xi_2$. Since $\Theta_i \subset \Theta_i$, $i = 1, 2$, where Θ_i, $i = 1, 2$, are invariant features associated with the contraction codes ν_i, $i = 1, 2$, we can introduce a natural hierarchy for the invariant coding:

$$\begin{array}{l}\text{If } \nu_1 \subset \nu_2 \text{ then } \Theta_1 \text{ is more easily detected than } \Theta_2 \text{ in the}\\ \text{fixed observation } \Lambda\end{array} \tag{55}$$

In this simulation the contraction code "STREET", which is a subdescription of the generator "ROAD", is simultaneously detected. The comparison between the invariant features associated with these two codes is also displayed in Fig. 16 with the associated attractors.

The ambiguity of the computational coding can be removed by evaluating the conditional probability (8) on the digital image as follows:

$$P(\sigma | \Lambda) = \frac{\sum_{(i,j) \in \sigma} \varphi(i, j | \partial \Lambda)}{\sum_{(i,j) \in \Lambda} \varphi(i, j | \partial \Lambda)} \tag{56a}$$

where σ denotes the attractor associated with the detected code ν

$$\sigma = \bigcup_{\nu} \mu_i(\sigma) \tag{56b}$$

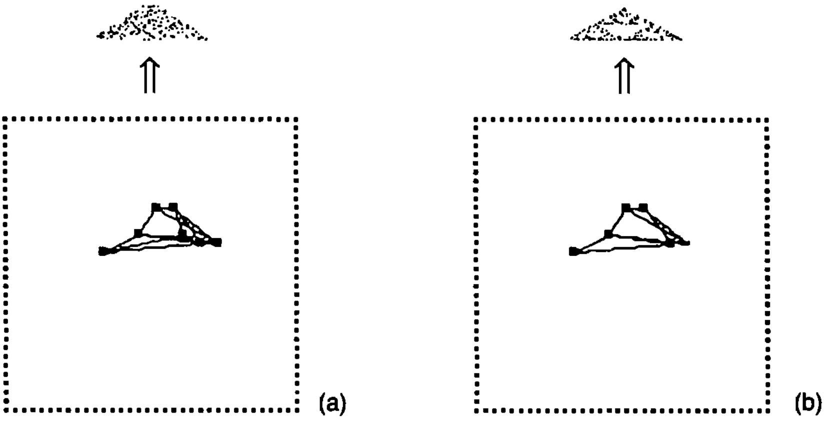


Figure 16 Alphabet of patterns: (a) road; (b) street.

Following this probabilistic evaluation, a contraction code "ROAD" is selected in the dictionary as shown in Fig. 15.

Another example of simulation results for a "FERN" pattern is illustrated in Figs. 17–19. In this simulation, the computation algorithm (52) was applied to the feature pattern detected on the fractal pattern shown in Fig. 17. The extracted invariant feature and associated attractor are shown in Fig. 18. In this case, a subdescription, called "LEAVES," was extracted with target pattern "FERN." The totality of connections generated by the contraction codes is compared in Fig. 19. In this simulation, the best-fit code "FERN" was selected based on the probabilistic evaluation (56) as indicated in Fig. 18.

VI. CONCLUDING REMARKS

A conditional distribution was introduced for evaluating the similarity between complicated patterns. This conditional distribution is represented on a parallel distributed system. As the basis of computation, discrete feature pattern can be extracted through local analysis of the conditional distribution. Combining the self-similarity of the imaging process with the smoothness of the conditional distribution, the feature pattern was proved to yield an invariant subset. This existence theorem claims that the generation process of self-similar patterns can be detected in a preassigned dictionary through finite computation steps. The detection scheme was verified through computer simulation.

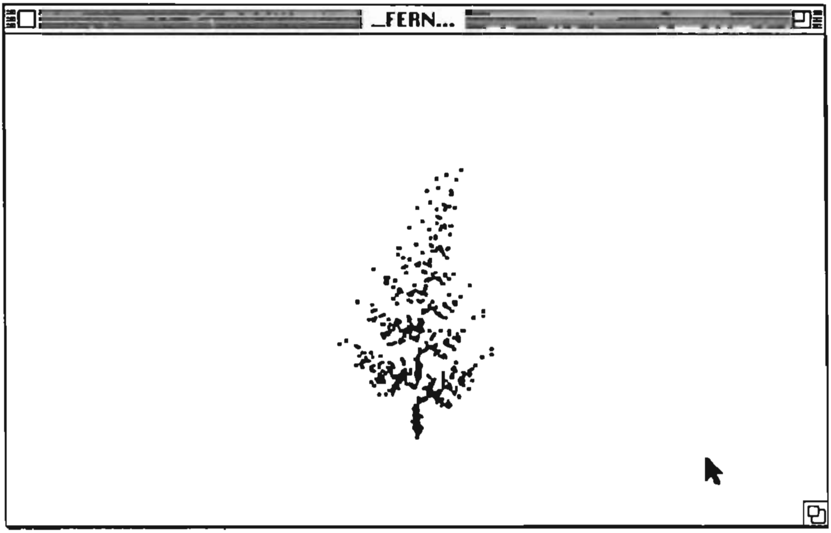


Figure 17 Pattern to be detected.

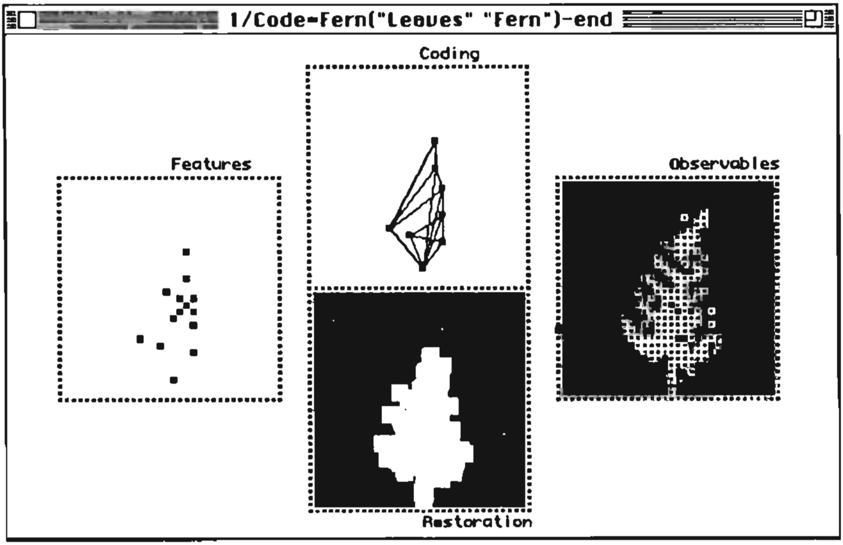


Figure 18 Coding results.

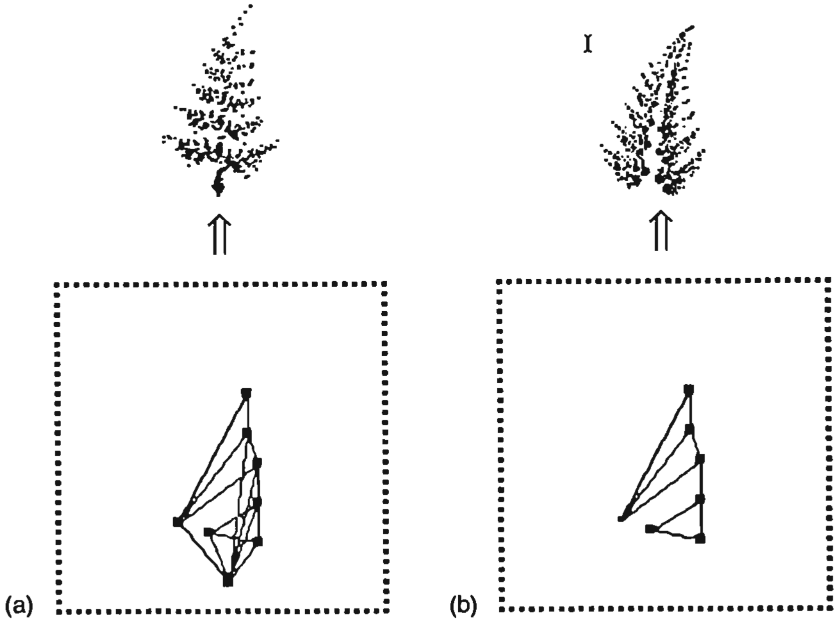


Figure 19 Alphabet of patterns: (a) fern; (b) leaves.

ACKNOWLEDGMENTS

I would like to thank Ms. Yatsuka Yoshida for her help in producing and analyzing the photographs.

REFERENCES

1. T Aoki, K Kamejima. Proc 25th ISCIE Symp on Stochastic Systems Theory and Its Applications, Osaka, Japan, November 1993, pp 139–144.
2. MF Barnsley. Fractals Everywhere. San Diego, California: Academic Press, 1988.
3. RO Duda, PE Hart, Pattern Classification and Scene Analysis. New York: John Wiley, 1979.
4. KS Fu. Syntactic Pattern Recognition and Applications. Englewood Cliffs, New Jersey: Prentice-Hall, 1982.
5. I Fujita, K Tanaka, M Ito, K Cheng. Nature 360:343–346, 1992.
6. JE Hutchinson. Indiana Univ Math J 30:713–747, 1981.
7. K Kamejima, YC Ogawa, Y Nakano. Proc IEEE-SICE IECON '84, Tokyo, Japan, October 1984, 52–57.

8. K Kamejima, YC Ogawa, Y Nakano. Proc IEEE IECON '86, Milwaukee, Wisconsin, September 1986, pp 713–718.
9. K Kamejima, I Takeuchi, T Hamada, YC Watanabe. Proc IEEE Int Workshop on Intelligent Robots and Systems. Tokyo, Japan, October 1988, pp 673–678.
10. K Kamejima, T Hamada, M Tsuchiya, YC Watanabe. In: I Masaki, ed, Vision-based Vehicle Guidance. New York: Springer Verlag, 1992, pp 173–203.
11. J Lansdown. In: AJ Crilly, RA Earnshaw, H Jones, eds, Fractals and Chaos. New York: Springer-Verlag, 1991, pp 211–224.
12. BB Mandelbrot. The Fractal Geometry of Nature. New York: Freeman, 1983.
13. D Marr, E Hildreth, Proc R Soc London B- 207:187–217, 1980.
14. M McGuire. An Eye for Fractals – a graphic and photographic essay, Redwood City, California: Addison Wesley, 1991.
15. Y Miyashita, HS Chang. Nature 331:568–570, 1988.
16. YC Ogawa, K Kamejima, Y Nakano. Proc IEEE Int Workshop on Industrial Applications of Machine Vision and Machine Intelligence, Tokyo, Japan, February 1987, pp 266–271.
17. HF Smith. In H-O Peitgen, JM Henriques, LF Penedo, eds, Fractals in the Fundamental and Applied Science. Amsterdam: North-Holland, 1991, pp 407–424.
18. SL Tanimoto. Proc 4th International Workshop on Parallel Image Analysis, Lyon, France, December, 1995, pp 159–168.

20
Gibbs Random Image Models and Sampling

Masaki Suwa* **and Sueo Sugimoto**

Ritsumeikan University, Shiga, Japan

I. INTRODUCTION

Markov-type two-dimensional (2-D) random processes (or random fields) have been widely applied to many aspects of practical image processing such as image modeling [1–6], image restoration [7–13], image segmentation [14–20], and texture analysis and synthesis [21–25]. An excellent review by Derin and Kelly [26] discusses many Markov-type 2-D random field models and their interrelations applied to image processing.

In this chapter, we consider Markov random image fields, especially Gibbs random fields and their sampling methods [4]. Gibbs random fields (GRF) (or Gibbs distributions) have been extensively studied in statistical mechanics since the 1920s [27], and recently in image and signal-processing applications. Sampling is a kind of generating process for the realization of the random field, for example, generating a texture image from a MRF model. To obtain such a realization, we have to calculate the Gibbs distribution which requires a large computational cost because of the existence of the normalizing constant, called the partition function, in the Gibbs distribution.

Instead of computing the normalizing constant, the relaxation-type algorithms such as the Metropolis algorithm [28,29] or Gibbs sampler [7,4] are used. But these sampling algorithms sometimes exhibit a phase transition phenomenon [4], namely, continuing a simulation leading to an image in which the synthesized image is going to be the same gray level and be dominated by one or two gray levels. Such a GRF model would certainly not be appropriate in texture modeling [4].

*Current affiliation: OMRON Corporation, Kyoto, Japan.

The exchange algorithm [22] which is a sort of Metropolis algorithm can avoid such a phenomenon, but this algorithm does not satisfy the so-called positivity condition of MRF.

Avoiding the phase transition, we discuss here a new relaxation-type sampling algorithm based on the Metropolis algorithm. Our proposed sampling algorithm contains the *a priori* information on the relative frequencies of gray levels in the original image, and its sampling process can be described by a finite homogeneous Markov chain.

In Section II, MRF models, specially GRF models as well as the Metropolis sampling algorithm, are presented. A Metropolis-like sampling algorithm of the GRF model is proposed for avoiding the phase transition phenomenon, and the stationary distribution of a finite homogeneous Markov chain associated with the proposed relaxation-type sampling is discussed in Section III. Preliminary results regarding Markov chains are presented in Section III.B. In Section III.C, we prove that the homogeneous Markov chain associated with a new sampling algorithm has a stationary distribution which is identical to the original Gibbs distribution. Furthermore, in Section IV, we perform computer experiments and show that the phase transition phenomenon can be avoided by applying our new sampling algorithm.

II. GIBBS RANDOM IMAGE MODELS

We describe the original discrete (finite lattice) image at time t as a random variable matrix $\mathbf{X}(t)$ of size $M \times N$ and denote its realization of the image as $\mathbf{x}$. Let S be a set of sites s, and let $\eta^{(m)} = \{\eta_s^{(m)}, s \in S\}$ be a mth-order neighborhood system for S [2,7]. We denote $\mathbf{X} = \{X_s, s \in S\}$ as any family of random variables indexed by s and assume a common state space; $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ called "gray levels." Let Ω be the set of all possible configurations of an image, namely, $\Omega = \{\mathbf{x}^{(i)} | i = 1, \ldots, K^{MN}\}$ and the number of elements of Ω is K^{MN} (denote $|\Omega| = K^{MN}$).

Here, we apply the equivalence between an MRF and the Gibbs distribution, from the so-called Hammersley and Clifford theorem [2,26] stated as "$\mathbf{X}$ is an MRF with respect to $\eta^{(m)}$ and $P(\mathbf{X} = \mathbf{x}) > 0$ for all $\mathbf{x}$, called the positivity condition, if and only if $\mathbf{X}$ is a Gibbs random field with respect to $\eta^{(m)}$ and the associated cliques."

Therefore the joint probability (mass function) of the MRF image $\mathbf{x}$ can be represented by the Gibbs distribution,

$$\mathrm{Prob}[\mathbf{X}(t) = \mathbf{x}] = p_t(x_{s_1}, x_{s_2}, \ldots, x_{s_{MN}}) = p_t(\mathbf{x}) = \frac{1}{Z}\exp\left\{-\frac{U(\mathbf{x}; \Theta)}{T(t)}\right\} \quad (1)$$

$T(t)$ is a function of t that means the physical temperature and Θ is a set of Gibbsian parameters. For the present purpose of sampling the GRF, we assume that the Gibbs distribution is not dependent on t, namely, $p_t(\mathbf{x}) \equiv p(\mathbf{x})$ by setting $T(t) = 1$; and $U(\mathbf{x}; \Theta)$, called the energy function, determines the energy of the GRF, and is of the form:

$$U(\mathbf{x}; \Theta) \equiv \sum_{c \in C} V_c(\mathbf{x}; \Theta) \tag{2}$$

where C denotes the set of cliques [2,7], and the family $\{V_c, c \in C\}$ is called a potential. The partition function (or normalizing constant), denoted as Z, is of the form:

$$Z = \sum_{\mathbf{x} \in \Omega} \exp\{-U((\mathbf{x}; \Theta))\} \tag{3}$$

For the purpose of generating a sample from the Gibbs random image, we have to calculate the Gibbs distribution (1). But it requires a huge computational cost to calculate the probabilities of the image $\{\mathbf{x}\}$ due to the partition function Z in (1) and (3); namely, computation of K^{MN} probabilities is required. To eliminate calculating Z, relaxation-type algorithms are often used. Here, we show the so-called Metropolis algorithm.

The Metropolis algorithm is realized by the following finite homogeneous Markov chain with transition probability $\bar{\mathbf{P}}$. Namely the (i, j) element of $\bar{\mathbf{P}}$ is given by

$$[\bar{\mathbf{P}}]_{ij} = \mathrm{Prob}[\mathbf{X}(t+1) = \mathbf{x}^{(j)} | \mathbf{X}(t) = \mathbf{x}^{(i)}] \equiv \begin{cases} [\bar{\mathbf{G}}]_{ij}[\bar{\mathbf{A}}]_{ij} & \mathbf{x}^{(j)} \in \Omega_i \\ 1 - \sum_{k \neq j} [\bar{\mathbf{P}}]_{ik} & \mathbf{x}^{(j)} = \mathbf{x}^{(i)} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\bar{\mathbf{G}}$ is called the "generation probability" matrix and $\bar{\mathbf{A}}$ is called the "acceptance probability" matrix. The (i, j) element of the matrix $\bar{\mathbf{G}}$ is given by the following definition,

$$[\tilde{\mathbf{G}}]_{ij} \equiv \begin{cases} \dfrac{1}{|\Omega_i|} & \mathbf{x}^{(j)} \in \Omega_i \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where Ω_i is the set $\{\mathbf{x}^{(j)} \in \Omega\}$ which differs from $\mathbf{x}^{(i)}$ in only one site, namely, for some site $\hat{s}$, $x_{\hat{s}}^{(j)} \neq x_{\hat{s}}^{(i)}$, $x_s^{(j)} = x_s^{(i)}$ $(s \neq \hat{s})$. Clearly, $\mathbf{x}^{(j)} \in \Omega_i \Leftrightarrow \mathbf{x}^{(i)} \in \Omega_j, \mathbf{x}^{(i)} \notin \Omega_i$. $|\Omega_i|$ is the number of all states $\mathbf{x}$ contained in Ω_i, i.e. $|\Omega_i| = KMN - 1$ for a randomly chosen site $\hat{s}$, or $|\Omega_i| = K - 1$ for choosing $\hat{s}$ by raster-scanning.

The (i, j) element of the acceptance probability matrix $\bar{\mathbf{A}}$ is defined as follows:

$$[\bar{\mathbf{A}}]_{ij} \equiv \begin{cases} \min\left[1, \dfrac{p(\mathbf{x}^{(j)})}{p(\mathbf{x}^{(i)})}\right] & \mathbf{x}^{(j)} \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

It is noted that the likelihood ratio of $p(\mathbf{x}^{(j)})/p(\mathbf{x}^{(i)})$ can be evaluated in practice because the ratio does not depend on the partition function Z in (3), and only those cliques involving site $\hat{s}$ where the gray level differs from $\mathbf{x}^{(i)}$. Therefore we can summarize the Metropolis sampling algorithm as follows [4,29]:

Step 1 Initialization (assign gray levels of all sites; $s = 1, \ldots, MN$, randomly).

Step 2 Choose an abitrary site $\hat{s}$ in $\mathbf{x}^{(i)}$; usually the site $\hat{s}$ is chosen by raster-scanning. Denote $x_{\hat{s}}^{(i)} = \lambda_k$.

Step 3 Choose a gray level $\lambda_m \neq \lambda_k$ with the probability $[\bar{\mathbf{G}}]_{ij}$.

Step 4 Replace $\mathbf{x}^{(i)}$ by $\mathbf{x}^{(j)}$ with probability $[\bar{\mathbf{A}}]_{ij}$, where $x_{\hat{s}}^{(j)} = \lambda_m$, $x_s^{(j)} = x_s^{(i)}$ $(s \neq \hat{s})$.

Step 5 Go to Step 2, until the number of iterations ($N \times M$(: one raster-scan) $\times N_L$), where N_L is a number of raster-scans in a large loop.

III. A METROPOLIS-LIKE SAMPLING ALGORITHM

Sampling algorithms such as the Metropolis algorithm as well as the Gibbs sampler sometimes causes a phase transition. To overcome such a phenomenon, we propose here a new Metropolis-like sampling algorithm with a new transition probability matrix $\mathbf{P}$ consisting of the following generation and acceptance probability matrices.

The (i, j) element of the generation probability $\mathbf{G}$ is defined by

$$[\mathbf{G}]_{ij} \equiv \begin{cases} \dfrac{1}{\alpha^{(j)}} f_g(\pi_o(x_{\hat{s}}^{(j)}) - \pi(x_{\hat{s}}^{(j)}; \mathbf{x}^{(j)})) & \mathbf{x}^{(j)} \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\hat{s}$ is chosen randomly from S or in regular order by raster-scanning, and $\pi(x_{\hat{s}}^{(j)}; \mathbf{x}^{(i)})$ is the relative frequency of the gray level $x_{\hat{s}}^{(j)}$ for image $\mathbf{x}^{(j)}$, which is computed as follows.

Let $n_k^{(i)}$ by the number of pixels taking the gray level λ_k in the image $\mathbf{x}^{(i)}$. Then the relative frequency of the gray level λ_k in the image $\mathbf{x}^{(i)}$ can be computed as $n_k^{(i)}/MN$. Let $x_{\hat{s}}^{(i)} = \lambda_{k_{\hat{s}}^{(i)}}$; then $\pi(x_{\hat{s}}^{(j)}; \mathbf{x}^{(j)}) = (n_k^{(i)} + 1)/MN$; if $x_{\hat{s}}^{(j)} \neq \lambda_{k_{\hat{s}}^{(i)}}$.

Also $\alpha^{(j)}$ means the normalizing constant, i.e.,

$$\alpha^{(j)} = \sum_{x_s^{(j)}=\lambda_1, x_s^{(j)} \neq \lambda_{k_s^{(i)}}}^{\lambda_K} f_g(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)})),$$

and $f_g(y)$ is a monotone increasing positive function for $-1 \leq y \leq 1$. $\pi_o(x_s^{(j)})$ is the relative frequency of the gray level $x_s^{(j)}$ for the original image, or an invariant probability of the gray levels λ for the GRF model in (1) is adopted if such a probability measure exists and can be evaluated from the GRF model. Here we explain what an invariant probability $\pi_o(\lambda)$ of the gray levels λ for the GRF model in (1) is, as follows.

Define the cumulative probability of x_{s_i} as

$$\pi_{o,s_i}(x_{s_i}) \equiv \sum_{x_{s1}=\lambda_1}^{\lambda_K} \cdots \sum_{\{\text{except}\, x_{s_i}\}} \cdots \sum_{x_{sMN}=\lambda_1}^{\lambda_K} p(x_{s_1}, x_{s_2}, \ldots, x_{s_{MN}})$$

Then, if a probability π_o exists such that, for any s_i, $\pi_o(\lambda) = \pi_{o,s_i}(\lambda)$, then we call $\pi_o(\lambda)$ the invariant probability of the gray level λ.

Due to the assumption of the function f_g, therefore, the generation probability of $x_s^{(j)} = \lambda_k$ is larger than the one of $x_s^{(j)} = \lambda_m$ if $\pi_o(\lambda_k) - \pi(\lambda_k; \mathbf{x}^{(j)}) > \pi_o(\lambda_m) - \pi(\lambda_m; \mathbf{x}^{(j)})$. This means that the gray level λ_k is more possibly chosen if the "desired" relative frequency $\pi_o(\lambda_k)$ is larger than the relative frequency $\pi(\lambda_k; \mathbf{x}^{(j)})$ in the generation probability step. Consequently, in the generation probability step, λ_k is more possibly generated such that $\pi(\lambda_k; \mathbf{x}^{(j)})$ may tend to the "desired" relative frequency $\pi_o(\lambda_k)$.

The (i, j) element of the acceptance probability matrix $\mathbf{A}$ is also defined by

$$[\mathbf{A}]_{ij} \equiv \begin{cases} \min\left[1, \dfrac{p(\mathbf{x}^{(j)}) f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))\, \alpha^{(j)}}{p(\mathbf{x}^{(i)}) f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))\, \alpha^{(i)}}\right] & \mathbf{x}^{(j)} \in \Omega_i \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $f_a(y)$ is defined as a monotone decreasing positive function for $-1 \leq y \leq 1$. We also assume that the following relation holds together with $f_g(y)$,

$$f_g(y) \cdot f_a(y) = 1 \tag{9}$$

In comparison with the acceptance probability (6) in the Metropolis algorithm, the ratio $[f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))]/[f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))]$ appears in the new acceptance probability (8). The existence of the term shows that if $\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}) \leq \pi_o(x_s^{(i)} - \pi(x_s^{(i)}; \mathbf{x}^{(i)})$, then $[f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))]/[f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))] \geq 1$, thus $x_s^{(j)}$ is more acceptable such that $\pi(x_s^{(j)};\mathbf{x}^{(j)})$ may tend to the "desired" relative frequency $\pi_o(x_s^{(j)})$.

Finally the transition probability **P** is defined similarly to (4), with (7) and (8):

$$[\mathbf{P}]_{ij} \equiv \mathrm{Prob}[\mathbf{X}(t+1) = \mathbf{x}^{(j)} | \mathbf{X}(t) = \mathbf{x}^{(i)}] \equiv \begin{cases} [\mathbf{G}]_{ij}[\mathbf{A}]_{ij} & \mathbf{x}^{(j)} \in \Omega_i \\ 1 - \sum\limits_{k \neq j} [\mathbf{P}]_{jk} & \mathbf{x}^{(i)} = \mathbf{x}^{(j)} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Comparing (7), (8) with (5), (6), respectively, the question will arise: what is the stationary distribution of the finite homogeneous Markov chain with the transition probability (10)? To reply to this question, we show the following theorem regarding the stationary distribution for the transition probability in (10).

Theorem 1. *The finite homogeneous Markov chain of the transition matrix (10) with (7) and (8) has a stationary distribution which is identical to the one given by the Gibbs distribution in (1).*

Before we show the proof of this theorem, we prepare preliminary results for finite Markov chains, according to [29,30].

A. Preliminary Results for Markov Chains

Definition 1. A Markov chain with transition matrix P is irreducible, if for any pair of states $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)} \in \Omega$, there is a positive probability of reaching $\mathbf{x}^{(j)}$ from $\mathbf{x}^{(i)}$ in a finite number of trials, i.e.

$$\forall \mathbf{x}^{(i)}, \mathbf{x}^{(j)},\ \exists n \geq 1:\ [P^n]_{ij} > 0$$

Definition 2. A Markov chain with transition matrix P is aperiodic, if for each state $\mathbf{x}^{(i)} \in \Omega$, the greatest common divisor $\gcd(D_i) = 1$, where the set of D_i consists of all integers $n > 1$, with $[P^n]_{ii} > 0$. The integer $\gcd(D_i)$ is called the period of state $\mathbf{x}^{(i)}$. Thus, aperiodicity requires all states to have period one.

Lemma 1. *An irreducible Markov chain with transition matrix P is aperiodic if*

$$\exists \mathbf{x}^{(j)} \in \Omega:\ [P]_{jj} > 0 \quad .$$

Lemma 2. *Suppose that a finite homogeneous Markov chain with transition matrix P be irreducible and aperiodic. Then there exists a* *stochastic vector* q whose elements $[q]_i$ are uniquely determined by the following equation:*

$$[q]_i = \sum_j [q]_j [P]_{ji} \tag{11}$$

Lemma 3. *Let a finite homogeneous Markov chain with transition matrix P be irreducible and aperiodic. If the elements of a stochastic vector q satisfy the following equation:*

$$[q]_i [P]_{ij} = [q]_j [P]_{ji} \tag{12}$$

then the stochastic vector q is a stationary distribution.

The proof of Lemma 3 is shown elsewhere. Taking the sum of both sides of (12) with respect to j,

$$\sum_j [q]_i [P]_{ij} = \sum_j [q]_j [P]_{ji}$$

then using $\Sigma_j [P]_{ij} = 1$, we have $[q]_i = \Sigma_j [q]_j [P]_{ji}$, in Lemma 2. The relation in (12) is the so-called detailed balance equation.

B. Proof of Theorem 1

First, we show that given a finite homogeneous Markov chain with transition probability matrix $\mathbf{P}$ in (10) with (7) and (8) is irreducible and aperiodic. Irreducibility can be shown as follows.

From the definition of transition probability $\mathbf{P}$ in (10), for any pair $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$, there exists a finite n (≥ 1), such that

$$\begin{aligned}[\mathbf{P}^n]_{ij} &= \sum_{j_1 \in \Omega} \sum_{j_2 \in \Omega} \cdots \sum_{j_{n-1} \in \Omega} [\mathbf{P}]_{ij_1} [\mathbf{P}]_{j_1 j_2} \cdots [\mathbf{P}]_{j_{n-1} j} \\ &\geq [\mathbf{G}]_{ik_1} [\mathbf{A}]_{ik_1} [\mathbf{G}]_{k_1 k_2} [\mathbf{A}]_{k_1 k_2} \cdots [\mathbf{G}]_{k_{n-1} j} [\mathbf{A}]_{k_{n-1} j} \\ &> 0\end{aligned}$$

namely we can choose states $\mathbf{x}^{(k_1)}, \ldots, \mathbf{x}^{(k_{n-1})}$ such that from $\mathbf{x}^{(i)}$ to $\mathbf{x}^{(k_1)}$, from $\mathbf{x}^{(k_1)}$ to $\mathbf{x}^{(k_2)}$, and so on, with positive transition probabilities and reaching finally to the state $\mathbf{x}^{(j)}$. Also, aperiodicity is easily shown by Lemma 1, since $[\mathbf{P}]_{ii} > 0$ for any $\mathbf{x}^{(i)}$ from the definition of the acceptance probability in (8).

*A vector is called stochastic if its elements are nonnegative and the sum of all elements is one.

Secondly we show the stochastic vector for (10) satisfies the detailed balance equation in Lemma 3 such that the stationary probability exists and coincides with (1) as follows.

Using the definition of the transition probability matrix in (10), we have, for $\mathbf{x}^{(j)} \in \Omega_i$

$$
\begin{aligned}
& p(\mathbf{x}^{(i)})[\mathbf{P}]_{ij} \\
&= p(\mathbf{x}^{(i)})[\mathbf{G}]_{ij}[\mathbf{A}]_{ij} \\
&= p(\mathbf{x}^{(i)}) \frac{f_g(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))}{\alpha^{(j)}} \\
&\quad \times \min\left(1, \frac{p(\mathbf{x}^{(j)}) f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))\,\alpha^{(j)}}{p(\mathbf{x}^{(i)}) f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))\,\alpha^{(i)}}\right) \\
&= p(\mathbf{x}^{(i)}) \frac{f_g(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))}{\alpha^{(j)}} \frac{p(\mathbf{x}^{(j)}) f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))\,\alpha^{(j)}}{p(\mathbf{x}^{(i)}) f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))\,\alpha^{(i)}} \\
&\quad \times \min\left(1, \frac{p(\mathbf{x}^{(i)}) f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))\,\alpha^{(i)}}{p(\mathbf{x}^{(j)}) f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))\,\alpha^{(j)}}\right) \\
&= p(\mathbf{x}^{(j)}) \frac{f_g(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(j)}))}{\alpha^{(i)}} \\
&\quad \times \min\left(1, \frac{p(\mathbf{x}^{(i)}) f_a(\pi_o(x_s^{(i)}) - \pi(x_s^{(i)}; \mathbf{x}^{(i)}))\,\alpha^{(i)}}{p(\mathbf{x}^{(j)}) f_a(\pi_o(x_s^{(j)}) - \pi(x_s^{(j)}; \mathbf{x}^{(j)}))\,\alpha^{(j)}}\right) \\
&= p(\mathbf{x}^{(j)})[\mathbf{G}]_{ji}[\mathbf{A}]_{ji} \\
&= p(\mathbf{x}^{(j)})[\mathbf{P}]_{ji}
\end{aligned}
$$

where we used the identity $\min(1, b/a) = (b/a)\min(1, a/b)$ for a positive pair a and b, and the relation in (9). Hence by Lemma 3, the stochastic vector $\mathbf{p}$ is a stationary distribution of the Markov chain with transition probability matrix $\mathbf{P}$, namely,

$$\mathbf{p} = \mathbf{p}\mathbf{P}; \qquad \mathbf{p} \equiv [p_1, \ldots, p_{K^{MN}}], \qquad p_i \equiv p(\mathbf{x}^{(i)})$$

which completes the proof of Theorem 1.

IV. EXPERIMENTAL RESULTS

In this section we show experimental results of synthesizing images by the proposed sampling algorithm as well as the Metropolis algorithm from the aspect of observing the phase transition phenomenon. Throughout our

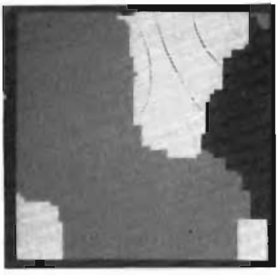


Figure 1 IMAGE-1 ($N_L = 200$).

experiments, we used a GRF model (the so-called first-order multilevel logistic (MLL) model in [14]) for the nearest neighborhood system $\eta^{(1)}$, with the potential, for $|k|+|l| \leq 1$,

$$V_c(x_{(i,j)}|\Theta) = \begin{cases} -\gamma_{k,l} & x_{(i,j)} = x_{(i-k,j-l)} \\ \gamma_{k,l} & x_{(i,j)} \neq x_{(i-k,j-l)} \end{cases} \tag{13}$$

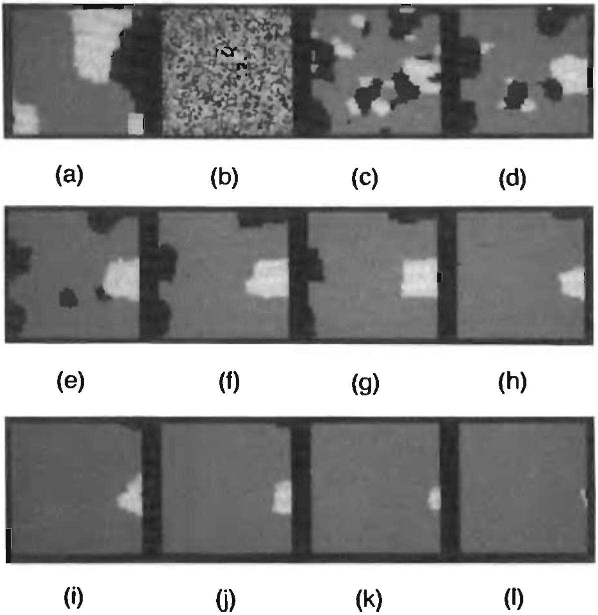
where we denote $s \equiv (i,j)$ and assume $\gamma_{k,l} = \gamma_{-k,-l}$.

Experiment 1. First we generate a 64×64 original image ($M = N = 64$) with 3 gray levels, $\lambda_1 = 40$, $\lambda_2 = 80$, $\lambda_3 = 120$ ($K = 3$) by the Metropolis sampling algorithm for the MLL model with potential parameters, $\gamma_{0,1} = 1$, $\gamma_{1,0} = 1$, $\gamma_{0,0} = 0$. Beginning initially from an independently identical distributed random image, we have the image in Fig. 1 (we call it "IMAGE-1"), after 200 raster scans ($N_L = 200$). Note that one raster scan contains $MN(= 64^2)$ visits from one site to the next site.

Observing the image in Fig. 1, we cannot clearly say whether or not the phase transition phenomena occur. Using IMAGE-1, therefore, we apply the pseudo-likelihood method (shown in the appendix) to obtain estimates of $\Theta \equiv [\gamma_{0,1}\gamma_{1,0}]^T$, whose results are listed in Table 1. Then, using these estimated potential parameters in GRF models, we again generated the images by both the Metropolis and the proposed sampling algorithms. Figures 2 and 3 show snapshots of the synthesized images at $N_L = 0, 10, 40, 60, \ldots, 160, 200, 240, 280, 300$ by the Metropolis and the proposed sampling algorithms, respectively. In the proposed sampling algorithm, relative frequencies $\pi_o(\cdot)$ were determined by counting frequencies of gray levels in IMAGE-1. In Fig. 2 by the Metropolis algorithm, the phase transition phenomenon clearly occurs, i.e., the images in Fig. 2 gradually tend to be dominated by one or two gray levels. However such a phenomenon does not occur in the textured images in Fig. 3.

Table 1 Estimated Potential Parameters

	Parameters	True values	Estimates
IMAGE-1	$\gamma_{0,-1}$	1.231	1.440
	$\gamma_{-1,0}$	1.212	1.126
IMAGE-2	$\gamma_{0,-1}$	1.0	1.16
	$\gamma_{-1,0}$	1.0	0.826
IMAGE-3	$\gamma_{0,-1}$	–	0.649
	$\gamma_{-1,0}$	–	0.902


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 2 Snapshots for IMAGE-1 (Metropolis algorithm): (a) original; (b) initial image; (c) $N_L = 10$; (d) $N_L = 40$; (e) $N_L = 60$; (f) $N_L = 80$; (g) $N_L = 100$; (h) $N_L = 160$; (i) $N_L = 200$; (j) $N_L = 240$; (k) $N_L = 280$; (l) $N_L = 300$.

Experiment 2. Similar experiments were carried out for another 5 gray level image; ($\lambda_k = 40, 80, 120, 160, 200$); namely, we generate the textured image (which is denoted as "IMAGE-2") with 64×64 size from the MLL models with respect to $\eta^{(1)}$ by using the Metropolis sampling algorithm with $N_L = 200$ iterations. The relative frequencies $\pi_o(\cdot)$ are obtained again

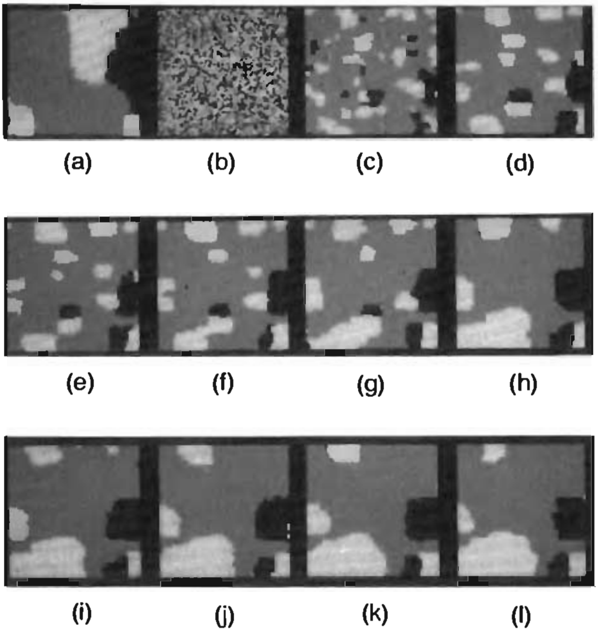


Figure 3 Snapshots for IMAGE-1 (proposed algorithm): (a) original; (b) initial image; (c) $N_L = 10$; (d) $N_L = 40$; (e) $N_L = 60$; (f) $N_L = 80$; (g) $N_L = 100$; (h) $N_L = 160$; (i) $N_L = 200$; (j) $N_L = 240$; (k) $N_L = 280$; (l) $N_L = 300$.

by counting frequencies of gray levels in IMAGE-2. By applying the PLL parameter estimation method, we obtained estimates of Θ. Both the parameters and their estimates are listed in Table 1.

Then from the GRF model with these estimated parameters, synthesized images are generated by the Metropolis as well as the proposed algorithms. Figures 4 and 5 show the images synthesized by the Metropolis and the proposed algorithms, respectively. We can again observe that the phase transition phenomenon occurs in Fig. 4 but not in Fig. 5.

Experiment 3. As a final experiment, we used a real texture image, called "mica" in *Brodaz's texture data* [31]. The image data of "mica" were transformed into a binary image (called "IMAGE-3") with 128 × 128 size and $\lambda_1 = 0$, $\lambda_2 = 255$, by applying the so-called K-means method [32]. After estimating potential parameters by the PLL method, whose results are listed in Table 1, by both the Metropolis and the proposed sampling

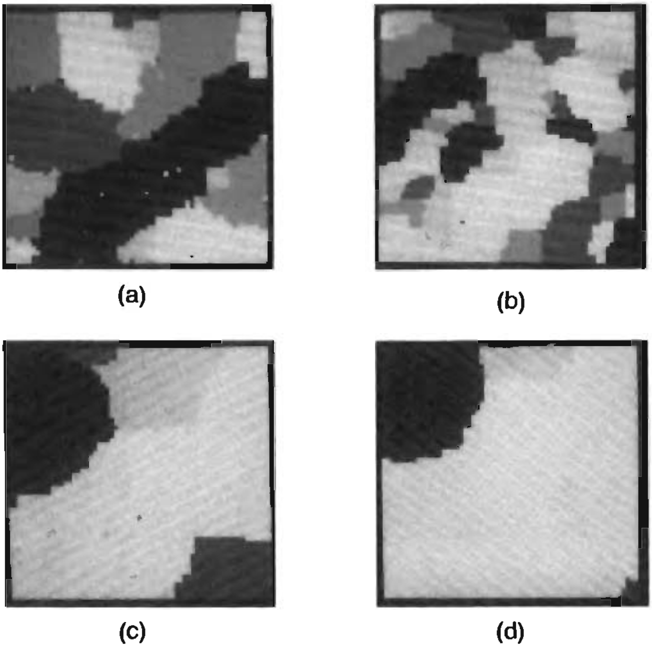


Figure 4 Snapshots for IMAGE-2 (Metropolis algorithm): (a) original; (b) $N_L = 100$; (c) $N_L = 1000$; (d) $N_L = 2000$.

algorithms, synthesized images were generated from the estimated GRF model. These images are shown in Fig. 6c (by the Metropolis algorithm), and Fig. 6d (by the proposed algorithm), respectively.

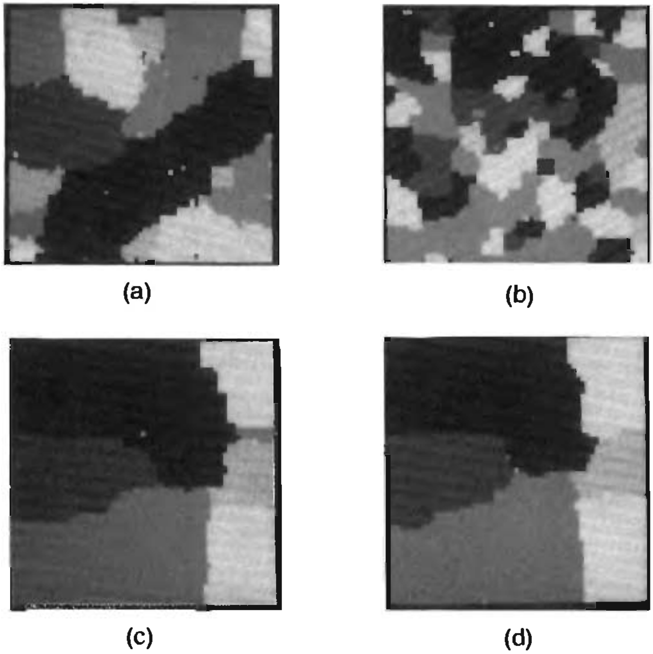
In Table 2, we also list relative frequencies of all images where we can observe that relative frequencies of the images synthesized by the proposed sampling algorithm are very close to those of the original images.

Further in our experiments, the functions f_g and f_a are chosen empirically as follows:

$$f_g(y) = \exp\{ay\}$$
$$f_a(y) = \exp\{-ay\}$$

where we set $a = 2$ in Experiments 1 and 3 for IMAGE-1 and IMAGE-3, and $a = 1.2$ in Experiment 2 for IMAGE-2.

From these experiments, we could confirm that the proposed sampling method can hold the visual resemblance as well as the relative frequencies of the gray level better than the Metropolis algorithm (see Table 2).


(a)

(b)

(c)

(d)

Figure 5 Snapshots for IMAGE-2 (proposed algorithm): (a) original; (b) $N_L = 100$; (c) $N_L = 1000$; (d) $N_L = 2000$.

V. CONCLUSIONS

In this chapter, we discussed MRF image models, especially with respect to generating texture images based on Gibbs random image models. The well-known relaxation-type sampling algorithms such as the Metropolis algorithm and Gibbs sampler often exhibit the so-called phase transition phenomenon. To avoid the phase transition in relaxation-type sampling algorithms, therefore, we proposed a new Metropolis-like sampling algorithm containing *a priori* information on relative frequencies of gray levels in the original image or on an invariant probability measure of the GRF model. *A priori* information on relative frequencies was implemented in the generation and acceptance probabilities as the forms of nonlinear functions $f_a(\cdot), f_g(\cdot)$ such that the generated image holds the same frequency properties as the original image. Also it was shown that the stationary distribution of a finite homogeneous Markov chain associated with a proposed sampling algorithm is identical to the original Gibbs distribution according to Markov chain theory.

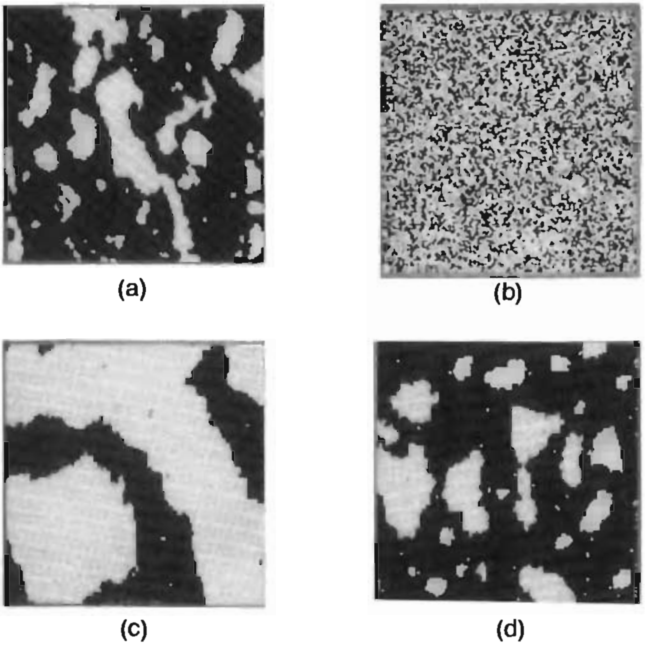


Figure 6 IMAGE-3 and sampling images by Metropolis and proposed algorithms: (a) mica; (b) initial image; (c) by Metropolis algorithm ($N_L = 50$); (d) by proposed algorithm ($N_L = 50$).

As in Section IV, through the computer experiments we have shown that in comparison to the Metropolis sampling algorithm our proposed sampling algorithm could avoid the phase transition phenomenon. The theoretical questions on the existence of the invariant probability measure for the GRF model discussed in Section III and its evaluation method, if such a measure exists, should be investigated in the future.

APPENDIX

We show an iterative estimation algorithm of Gibbsian parameters Θ based on the pseudo-log-likelihood (PLL) method [8,33]. The PLL function $L_p(\mathbf{x};\Theta)$ is defined as follows:

$$\log p(\mathbf{x};\Theta) \simeq L_p(\mathbf{x};\Theta) \equiv \sum_{s\in S} \{-U(x_s; x_{\eta_s^{(m)}}, \Theta) - \log Z_s(\Theta)\}$$

Table 2 Relative Frequencies

(a) IMAGE-1

	Gray levels		
Sampling	$\lambda_1 = 40$	$\lambda_2 = 120$	$\lambda_3 = 200$
Metropolis algorithm	0.008	0.970	0.022
Proposed algorithm	0.100	0.651	0.248
Relative frequencies of IMAGE-1	0.137	0.602	0.261

(b) IMAGE-2

	Gray levels				
Sampling	$\lambda_1 = 40$	$\lambda_2 = 80$	$\lambda_3 = 120$	$\lambda_4 = 160$	$\lambda_5 = 200$
Metropolis algorithm	0.162	0.012	0.007	0.053	0.766
Proposed algorithm	0.311	0.140	0.300	0.056	0.192
Relative frequencies of IMAGE-2	0.305	0.146	0.267	0.073	0.208

(c) IMAGE-3

	Gray levels	
Sampling	$\lambda_1 = 0$	$\lambda_2 = 255$
Metropolis algorithm	0.329	0.671
Proposed algorithm	0.734	0.266
Relative frequencies of IMAGE-3	0.750	0.250

where

$$Z_s(\Theta) \equiv \sum_{x_s=\lambda_1}^{\lambda_K} \exp\{-U(x_s; x_{\eta_s^{(m)}}, \Theta)\}$$

The energy function $U(x_s; x_{\eta_s^{(m)}}, \Theta)$ is assumed to be given as the following function

$$U(x_s; x_{\eta_s^{(m)}}, \Theta) \equiv \mathbf{B}(x_{(i,j)}; x_{\eta_s^{(m)}})\,\Theta \tag{A.1}$$

where

$$\begin{aligned} \mathbf{B}(x_{(i,j)}; x_{\eta_s^{(m)}}) &= \{B_{k,l}(x_{(i,j)}; x_{\eta_s^{(m)}})\} \\ B_{k,l}(x_{(i,j)}; x_{\eta_s^{(m)}}) &\equiv R_{k,l}(\rho\Delta_{k,l}(x_{(i,j)})) + R_{-k,-l}(\rho\Delta_{-k,-l}(x_{(i,j)})) \\ \Theta &= \{\gamma_{k,l}\}; \qquad |k| + |l| \leq m, \end{aligned}$$

and $\Delta_{k,l}(x_{(i,j)})$ is the difference of gray levels between a clique pair $\{x_{(i,j)}, x_{(i-k,j-l)}\}$, i.e., $\Delta_{k,l}(x_{(i,j)} \equiv x_{(i,j)} - x_{(i-k,j-l)}$. $R_{k,l}$ are adequate functions whose domains are given by $[-\rho(\lambda_K - \lambda_1 + 1), \rho(\lambda_K - \lambda_1 + 1)]$; ρ is a positive normalization constant for changing the domain into $[-1,1]$; let $\rho = 1/255$ for an 8-bit pixel image. In the case of the MLL model with respect to $\eta^{(1)}$

$$R_{k,l}(\rho\Delta_{k,l}(x_{(i,j)})) = 2\delta(\rho\Delta_{k,l}(x_{(i,j)}), 0) - 1; \qquad |k| + |l| \leq 1$$

where $\delta(\cdot, \cdot)$ is Kronecker's delta.

Then the estimates of Gibbsian parameters are obtained by solving the following nonlinear equations:

$$\frac{\partial L_p(\mathbf{x}; \Theta)}{\partial \Theta} = 0 \tag{A.2}$$

Here we show an iterative parameter estimation algorithm. Firstly, we evaluate the left-hand side in (A.2)

$$\begin{aligned} \frac{\partial L_p(\mathbf{x}; \Theta)}{\partial \gamma_{k,l}} &= \frac{\partial}{\partial \gamma_{k,l}} \sum_{(i,j)\in S} \{-\mathbf{B}(x_{(i,j)}; x_{\eta_s^{(m)}})\,\Theta^T - \log Z_{(\mathrm{i},j)}(\Theta)\} \\ &= \sum_{(i,j)\in S} \left[-B_{k,l}(x_{(i,j)}; x_{\eta_s^{(m)}}) + \sum_{x_{(i,j)}=\lambda_1}^{\lambda_K} B_{k,l}(x_{(i,j)}; x_{\eta_s^{(m)}})\, p(x_{(i,j)}; x_{\eta_s^{(m)}}, \Theta) \right] \end{aligned} \tag{A.3}$$

where

$$p(x_{(i,j)}; x_{\eta_s^{(m)}}, \Theta) \equiv \frac{1}{Z_{i,j}(\Theta)} \exp\{-U(x_{(i,j)}; x_{\eta_{(i,j)}^{(m)}}, \Theta)\}$$

Assuming the parameter estimates $\{\gamma_{k,l}^{(n)}\}$ are given after *n*th iteration, then we approximate that $Z_{(i,j)}(\Theta) \simeq Z_{(i,j)}(\Theta^{(n)})$ and a first-order Taylor series approximation yields

$$\exp\{-U(x_{(i,j)}; x_{\eta_{(i,j)}^{(m)}}, \Theta)\} \simeq \sum_{(k,l)} D_{k,l}^{(n)}(x_{(i,j)}; \Theta^{(n)})\,\gamma_{k,l} + E^{(n)}(x_{(i,j)}; \Theta^{(n)})$$

where

$$\begin{cases} D^{(n)}_{k,l}(x_{(i,j)}) \equiv -B_{k,l}(x_{(i,j)}) \exp\{-U(x_{(i,j)}; x_{\eta_S^{(m)}}, \Theta^{(n)}\} \\ E^{(n)}(x_{(i,j)}) \equiv \{1 + U(x_{(i,j)}; x_{\eta_S^{(m)}}, \Theta^{(n)})\} \exp\{-U(x_{(i,j)}; x_{\eta_S^{(m)}}, \Theta^{(n)})\} \end{cases} \tag{A.4}$$

Substituting (A.4) into (A.2), we have

$$\begin{aligned} \frac{\partial \log L(\mathbf{x};\Theta)}{\partial \gamma_{k,l}} \simeq & \sum_{(u,v)} \left[\sum_{(i,j)\in S} \frac{1}{Z_{(i,j)}(\Theta^{(n)})} \sum_{x_{(i,j)}=\lambda_1}^{\lambda_K} B_{k,l}(x_{(i,j)}) D^{(n)}_{u,v}(x_{(i,j)}) \right] \gamma_{k,l} \\ & - \sum_{(i,j)\in S} \left\{ B_{k,l}(x_{(i,j)}) - \frac{1}{Z_{(i,j)}(\Theta^{(n)})} \right. \\ & \left. \times \sum_{x_{(i,j)}=\lambda_1}^{\lambda_K} B_{k,l}(x_{(i,j)}) E^{(n)}(x_{(i,j)}) \right\} \\ = & 0 \end{aligned} \tag{A.5}$$

Moreover, we define

$$\begin{aligned} J^{(n)}_{(k,l),(u,v)} &\equiv \sum_{(i,j)\in S} \frac{1}{Z_{(i,j)}(\Theta^{(n)})} \sum_{x_{(i,j)}=\lambda_1}^{\lambda_K} B_{k,l}(x_{(i,j)}) D^{(n)}_{(u,v)}(x_{(i,j)}) \\ K^{(n)}_{(k,l)} &\equiv - \sum_{(i,j)\in S} \left\{ B_{k,l}(x_{(i,j)}) - \frac{1}{Z_{(i,j)}(\Theta^{(n)})} \right. \\ & \left. \times \sum_{x_{(i,j)}=\lambda_1}^{\lambda_K} B_{k,l}(x_{(i,j)}) E^{(n)}(x_{(i,j)}) \right\} \end{aligned}$$

Then from (A.5) we have the following equation

$$\mathbf{J}^{(n)}\Theta = \mathbf{K}^{(n)}$$

where the components of matrices $\mathbf{J}^{(n)}$ and $\mathbf{K}^{(n)}$ are given respectively by

$$\begin{aligned} \mathbf{J}^{(n)} &= \{J^{(n)}_{(k,l),(u,v)}\} \\ \mathbf{K}^{(n)} &= \{K^{(n)}_{k,l}\} \end{aligned}$$

In the case of the nearest neighborhood system $\eta^{(1)}$, for example, $\mathbf{J}^{(n)}$ is given by

$$\mathbf{J}^{(n)} \equiv \begin{bmatrix} J^{(n)}_{(0,1),(0,1)} & J^{(n)}_{(0,1),(1,0)} \\ J^{(n)}_{(1,0),(0,1)} & J^{(n)}_{(1,0),(1,0)} \end{bmatrix}$$

and we can derive the following iterative estimation equation

$$\Theta^{(n+1)} = \mathbf{Q}(\Theta^{(n)}) \equiv {;}\mathbf{J}^{(n)})^{-1}\mathbf{K}^{(n)} \tag{A.6}$$

In our computer experiments in Section IV, we use the following iterative algorithm for estimating the potential parameters,

$$\begin{aligned}\gamma_{0,1}^{(n+1)} &= \frac{J^{(n)}_{(0,1),(1,0)}K^{(n)}_{1,0} - J^{(n)}_{(1,0),(1,0)}K^{(n)}_{0,1}}{J^{(n)}_{(0,1),(0,1)}J^{(n)}_{(1,0),(1,0)} - J^{(n)}_{(0,1),(1,0)}J^{(n)}_{(1,0),(0,1)}}\\ \gamma_{1,0}^{(n+1)} &= \frac{J^{(n)}_{(1,0),(0,1)}K^{(n)}_{0,1} - J^{(n)}_{(0,1),(0,1)}K^{(n)}_{1,0}}{J^{(n)}_{(0,1),(0,1)}J^{(n)}_{(1,0),(1,0)} - J^{(n)}_{(0,1),(1,0)}J^{(n)}_{(1,0),(0,1)}}\end{aligned} \tag{A.7}$$

REFERENCES

1. JW Wood. IEEE Trans Inf Theory IT-18:232–240, 1972.
2. J Besag. J Roy Statis Soc B6:192–236, 1974.
3. AK Jain. Proc IEEE 69:502–528, 1981.
4. RC Dubes, AK Jain. J Appl Stat 16:131–164, 1989.
5. T Katayama, S Sugimoto. Parameter identification for two-dimensional image enhancement. In: SG Tzafestas, ed, Multidimensional Systems. New York: Marcel Dekker, 1986.
6. R Chellappa, A Jain. Markov Random Fields, Theory and Application. New York: Academic Press, 1993.
7. S Geman, D Geman. IEEE Trans Pattern Anal Machine Intell 6:721–741, 1984.
8. JE Besag. J Roy Statist Soc 48:259–302, 1986.
9. D Geman, G Reynolds. IEEE Trans Pattern Anal Machine Intell 14:367–383, 1992.
10. TJ Hebert, R Leahy. IEEE Trans Signal Proc 40:2290–2303, 1992.
11. C Bouman, K Sauer. IEEE Trans Image Proc 2:296–310, 1993.
12. L Bedini, I Gerace, A Tonazzini. CVGIP, Graphical Models and Image Processing, 56:109–123, 1994.
13. TJ Hebert, K Lu. IEEE Trans Image Proc 4:1084–1095, 1995.
14. H Derin, WS Cole. Comput Vision Graphics & Image Proc 35:72–98, 1986.
15. FS Cohen, DB Cooper. IEEE Trans Pattern Anal Machine Intell 9:195–219, 1987.
16. WA Wright. Image Vision Comput 7:144–150, 1989.
17. D Geiger, F Yuille. Int J Comput Vision 6:227–243, 1991.
18. CS Won, H Derin. Comput Vision Graphics Image Proc 54:308–328, 1992.
19. CA Bouman, M Shapiro. IEEE Trans Image Proc 3:162–177, 1994.
20. MG Bello. IEEE Trans Image Proc 3:834–846, 1994.
21. M Hassner, J Sklansky. Computer Graphics and Image Processing 12:357–370, 1980.
22. GR Cross, AK Jain. IEEE Trans Pattern Anal Machine Intell 5:25–39, 1983.

23. FS Cohen, Z Fan, MA Patel. IEEE Trans Pattern Anal Machine Intell 13:192–202, 1991.
24. CO Acuna. CVGIP, Graphical Models and Image Processing 53:240–257, 1991.
25. MT Johnson, MT Desai. Proc ICIP-94, vol. 3, 1994, pp 388–391.
26. H Derin, PA Kelly. Proc IEEE 77:1485–1510, 1989.
27. R Kinderman, JL Snell. Markov Random Fields and Their Applications. Providence, RI: Amer Math Soc, 1980.
28. N Metropolis, A Rosenbluth, M Rosenbluth, A Teller, E Teller. J Chem Phys 21:1087–1092, 1953.
29. E Aarts, J Korst. Simulated Annealing and Boltzmann Machines. New York: John Wiley, 1989.
30. E Seneta. Non-negative Matrices and Markov Chains. New York: Springer-Verlag, 1981.
31. P Brodatz. Textures: A Photographic Album for Artists and Designers. New York: Dover, 1966.
32. GH Ball, DJ Hall. ISODATA – Novel Method for Data Analysis and Pattern Classification. Menlo Park: Stanford Research Institute, 1965.
33. T Yamazaki, MN Shirazi, H Noda. IEICE Trans Inf E76-D:259–268, 1993.

This Page Intentionally Left Blank

Index

Abrupt discontinuities, 59
Acceptance probability, 527–529
Adaptation, 448
Adaptive fuzzy system, 468
AIC (Akaike information criterion), 38–39, 46–47, 60, 87, 98–99, 114, 162, 166–167, 179
  of composite model, 114
  small sample correction, 99
Algebraic Riccati inequality, 3
Algorithm:
  alternating conditional expectation (ACE), 108
  parallel, 483
  relaxation-type, 525
  sequential, 482
  subspace, 99, 403
  supervised learning, 490–491
Amplitude modulation, quadrature, 377
Amplitude process, 331
Antenna array, 400
A priori information, 446–447, 526, 536
AR model or process, 43–47, 337, 335, 404, 443
  time-varying (TVAR), 49, 66
ARMA model or process, 125, 128, 179, 180–182, 190, 464
  multivariable, 182, 194, 360
ARMA order estimation or test, 179–180, 186–187
ARMAX model, 482, 466
ARMAX parametrization, 86, 97
Array error parametrizations, 406
Array response, 402–404
Array signal processing, robust, 399
  simulation results, 421
Array steering vectors, 404
Asymptotically normal, 128, 187, 203, 205–206
Attractor, 502, 507, 515
  in discrete random dynamics, 284
Aumann-like integral, 308, 324

Backward shift operator, 122
Bandwidth expansion, 442
Bayesian approach, 399
  to time series analysis, 41
Beamformer:
  MAP, 413
  MVDR, 412
Beamforming, 404, 412, 425
  adaptive, 409

Bessel function, 440
Beta function, 281
Bias errors, 161
BIC (Bayes information criteria), 86
Binary random variable, 488–489
Binary tree:
- biased-search, 488
- fair-search, 488
- structure, 487

Birkhoff theorem, 385
Birth-and-death process, 290
Bispectrum, 360
- integrated, 362

Blind deconvolution, 375, 378, 379
Bose–Einstein statistics, 276, 293
Boundary preserving mapping, 511
Boundary value problem, 503
Brodaz's texture data, 535
Brownian motion process, 217

Calibration, 433
- array, 402

Canonical variable, 92, 112
Canonical variate analysis (CVA), 83, 89, 91
- applications, 108–116
- computational aspects, 99
- extension to nonlinear system, 107
- model identification, 113
- of past and future, 93
- theorem, 91–92

Cayley–Hamilton theorem, 184, 202
Central limit theorem, 187, 193, 205
Change detection, 454, 471, 482
- robust, 453
  - KDI-based, 462
- scheme, 454, 471
- system, 453–455
- with model uncertainty, 457

Chi-square random variable, 63
Conditional expectation, 11, 107
Conditional likelihood, 91
Conditional probability, 484–485, 504
Consistency, 369
Consistent estimate, 187
Contraction mappings or process, 506–507, 509
- boundary preserving, 508

Controllability, 184
- matrix, extended, 185

Controller form, 136
Covariance matrix, 7, 29, 90–91, 182, 184, 187, 201, 205
- Hankel, 180, 185

Cramer–Rao lower bound, 86, 418
Cross-spectrum, 360–362
- estimators, 364

Cross validation, 86
Cumulant (function or statistics), 357, 376, 379
- fourth-order, 361, 378
- higher-order, 357
- *n*th-order, 376
- second-order, 389
- third-order, 357, 360

Cyclostationary process, 327–328
- discrete-time (DTCS), 327, 334
- spectral theory of, 328

Decimator, 343
de Finetti's theorem, 288
Detailed balance equation, 531
Difference constraints, 40, 43
Direction finding (DF), 399, 402
Direction of arrival (DOA), 400–401, 403, 405, 413
- estimation, 411, 421

Discrete Fourier Transform (DFT), 334, 363, 364, 400
Discrete random dynamics, 284
Discrete robust mapping, 482, 484
Distributed parameter system, 503, 511
Distribution:
- binomial, 69
- chi-square, 187, 198, 207
- conditional, 54, 74, 75, 502, 517
- Dirichlet, 275, 276, 280–281, 297
- Gibbs, 526–527,
- logarithmic series, 290
- multinomial, 275, 277

[Distribution:]
Poisson-Dirichlet, 275, 277, 282, 286, 296–297
Rician, 440
stationary, 275, 530
Zipf's, 290
Dynamic programming, 224

Eigenvalue decomposition (EVD), 188, 191, 195, 203, 207, 410
Elevation angle, 435
Emergent market, shares in, 273
Emitter localization, 399
Emitter waveforms, 405
Energy function, 527
Equalization, 375, 378
Equalizer, 377
Equation error (EE), 126
criterion, 127
formulation, 368, 369
Ergodic, 7, 483, 490
ESPRIT algorithm, 401, 402, 423
Estimation, 434
function, 439
nonstationary trend, 61
of abruptly changing spectrum, 66
of integrated polyspectrum, 363
of MA models, 132, 137
of smoothly changing spectrum, 47
order, 179, 182
parameter, 434
state or state vector, 41, 74, 75, 85, 435
Estimator inconsistency, 400
Euler constant, 63, 296
Evolutionary spectrum, 51
Ewens sampling formula, 276, 278, 282–283, 292
Exchange algorithm, 526
Explorative expansion, 501
Exponential quadratic error criterion, 243

False alarm, 187, 492
curve, 494
rate, 457
Fault detection, 112, 453
Feature pattern, 506, 517
detector, 508
Feature points, 502
invariant subsets of, 502
Feedback-free processes, 12
Filter (*see also* Kalman filter):
analysis, 343
Butterworth, 168
Kautz, 131, 144
Laguerre, 131
Monte Carlo, 61, 75
Poisson, 133, 138
prediction, 443
state variable, 140
synthesis, 343
Wiener, 20, 412
Filter banks, 343–344
alias-free, 328, 346–347
analysis, 345
conjugate quadrature, 351
optimization of, 349
perfect reconstruction (PR), 349, 351
synthesis, 346
Filter divergence, 438
Filtering, 41, 54, 56, 57
minimax, 243
reduced-order, 257, 266–270
Final prediction error (FPE), 179
Finite impulse response (FIR) filter, 351, 443
Fisher information matrix, 101, 371, 445
Fractal dimension, 507
Fractal patterns, 499, 516
Frequency spectrum, 283–285
F-test, 213, 215
Fuchs test, 189
multivariable, 194
Fuzzy:
perception, 304
random data (FRD), 299
random matrix, 306
set, 300–302, 468
vector or matrix, 300–303

Fuzzy random vector (FRV), 299, 303–305
expectation of, 307
second moments of, 310–313
set representation, 305
statistical properties of, 307

GA (genetic algorithm), 163, 460
Gain and phase errors, 407, 421
Gaussian random process, complex, 405
Gaussian sum approximation, 37, 55, 61
General state space modeling, 37
applications of, 59
Generation probability, 527–528
Gibbs random field (GRF), 525–527, 529, 533, 535, 536
Gibbs random image model, 526
Gibbs sampler, 525, 528
Girsanov's transformation, 220
Gladyshev's formula, 327–328, 334
Goodness of fit, 39
Gram–Schmidt method, 165
Gray levels, 526

H_∞ filter, 239
stochastic property of, 243
H_∞ filtering problem, 239
finite horizon, 240
H_∞ Riccati difference equation (RDE), 241, 248
Hamilton–Jacobi equation, 223
numerical procedure, 227–229
Hamiltonian, 262
Hammersley and Clifford theorem, 526
Hankel matrix, 134–135, 188
block, 31, 183
Hard bound, 454
Hausdorff metric, 304
Higher-order statistics, 355, 357
Hilbert space, 9
of the observed data, 6
of nonlinear functions, 107
relative, 15
Homotopy, 511
maximality of, 512
Hyperparameters, 41, 51
Hypergeometric function, confluent, 442

Identifiability, 368
parameter, 358
Identification, 153 (*see also* System identification)
automated and objective, 87
of continuous-time systems, 121
of (system) impulse response, 149
of parametric model, 123
robust, 454
under feedback, 103–107
Ill-conditioned problem, 150
Imaging processes, 501
Impulse response, 377, 392
continuous-time (CTIR), 152–153
of ideal low-pass filter, 330
Inhomogeneous discrete process, 67
Innovation representation, 13, 182
Instantaneous spectrum, 51–52
Interpolator, 343
Intersymbol interference, multichannel 392
Invariant feature, 499, 507–508, 515, 519
existence of, 508
Invariant measure, 507
support of, 507
Invariant probability, 529
Invariant set, 507
Ito's formula, 220

K-distribution, 439–440
K-means methods, 535
Kallianpur-Striebel formula, 219
Kalman filter, 41, 245, 247, 257, 260, 433, 438 (*see also* H_∞ filter)
extended, 75, 436, 440
fading memory, 448
full-order, 264
Kolmogorov–Gabor polynomial, 465
Kronecker product, 198, 366, 416

Kullback discrimination information (KDI), 98, 453–454, 456–457, 460, 469, 471

Kullback(-Leibler) information or number, 38, 98, 101–103, 482–483, 492 (*see also* KDI)
- expected, 101–102

Kullback–Leibler matching criterion, 489

Lagragian multiplier, 262

Laplace operator, 503

Large deviation probabilities, 483

LDU-decomposition, 196–197

Least-squares (LS):
- estimate or estimation, 126, 128, 134
- method, 86, 152, 442
- orthogonal, 152, 163–164
- recursive (RLS), 86, 126
- solution, 367

Least-squares (LS) problem:
- Bayesian interpretation of, 50
- constrained, 50

Level set, 300

Likelihood function, 456, 460

Likelihood functional, 229

Linear matrix inequality (LMI), 3

Linear process, 357–358

Log-likelihood, 42, 55–56, 414, 459

Low-pass approximation, 162

Lyapunov equation, 3

Lyapunov's convexity theorem, 309

Majority voting rule, 472

Maneuvering target tracking, 448

MAP (maximum *a posteriori*):
- criterion, 415
- estimate or estimator, 218, 404–405, 414–416
- nonlinear recursive filtering, 217
- parameter estimation, 413
- technique, 403

MAP–NSF (noise subspace fitting) method, 416–421, 423–424

Markov chain, 286–287
- aperiodic, 530–531
- homogeneous, 526–527, 530
- irreducible, 530

Markov parameters, 131, 132

Markov–Poisson parameters, 133, 135, 140

Markov process, 1, 94, 274, 286
- controlled, 94
- Gaussian, 485

Markov random field (MRF), 525–526 (*see also* Gibbs random field)

Matrix minimum principle, 262

Matrix perturbation theory, 191

Maximal correlation, 108

Maximization criterion:
- multistage, 379, 383
- single-stage, 384

Maximum entropy method (MEM), 442, 444

Maximum likelihood (ML), 86, 91
- criterion, regularized, 415
- estimate or estimator, 42, 65, 367, 371
- estimation, 86
- method, 459
- solution, 401
- structured, 413
- type II, 40

Membership function, 300, 315

Memory, 95–96

Methods of moments, 30

Metropolis algorithm, 525, 527–529

Metropolis-like sampling algorithm, 528

M-function, 441

Minimax filtering problem, 243–244

Minimum entropy deconvolution, 379

Minimum variance distortionless response (MVDR), 412–413

Model:
- abrupt seasonal, 61
- abrupt trend, 61
- black-box polynomial, 125

[Model:]
- discrete-time impulse response (DTIRM), 150
- errors-in-variables, 355, 357
- input–output, 11
- linear-in-parameters, 128
- Markov parameter, 131–132
- parametric form of continuous-time, 125
- random perturbation, 406
- reduction, 85
- structure selection, objective, 97
- time-moment, 137–140
- trend, 43, 59
- urn, Polya-like, 279

Modeling:
- Bayesian, 37
- continuous-time, 122
- gray-box, 129
- hierarchical, 464
- non-Gaussian, 65
- state space, 37, 41
- symmetric, 355
- time-series, 43
- vibrating system, 109–112

Modulated autocovariance, 329

Monte Carlo:
- approach, 38, 57
- methods, 56, 57
- filtering methods, 59, 61
- simulation, 101–102, 516

Morphological computation process, 503

Moving-average (MA) model, 132
- generalized (GMAM), 121, 129

Multichannel blind deconvolution, 375–376, 379
- necessary and sufficient condition for, 388

Multilevel logistic model (MLL), 533–534, 541

Multirate systems, 341

Multiresolution analysis or approximation, 154

Multitarget tracker, 447

MUSIC (MUltiple SIgnal Classification), 401–402, 410–411, 442, 444–446

Narrowband modeling assumptions, 404

Neighborhood system, 526

Neural networks, 481
- multiresolution (MRNN), 152, 159–161
- stochastic binary, 481

Neutrality, 280

Noble identities, 342

Noise subspace fitting (NSF), 411, 416 (*see also* MAP-NSF)

Nonbound approach, 462, 464

Noncausal system, 376

Nonminimum phase system, 375

Nonstationarity in the mean, 42

Nonstationary:
- binary process, 68
- time-series, 43

Nuisance parameters, 403, 405

Null hypothesis, 187–190, 197, 202

Numerical approximation, 55

Oblique projection, 20

Observability, 436
- Gramian, 437–438
- matrix, extended 185

Onsager–Machlup functional, 218–219

Order statistics, 277, 296
- of the market shares, 275

Orthogonal projection, .7, 410

Output error (OE), 126

Parallel distributed scheme or system, 502, 508

PARCOR (partial autocorrelation) coefficients, 49, 52, 66, 67

Partition function, 527

Periodic AR process, 335
- spectral formula of, 336

Persistently exciting, 368

Phase transition phenomenon, 525–526, 535
Polyphase representation, 343, 347
Polyspectrum, 360
  integrated, 356–360, 363
Positive real, 31
Positivity, 29, 31
Potential, 527
Prediction, 41, 54, 56–57
  nonlinear, 108
Prediction error method, 459, 469
Predictive mapping layer, 486–491
Predictor spaces, 18
  oblique, 21
Process monitoring, 112
Pseudo-inverse, 11, 410
Pseudo-log-likelihood (PLL) method, 533, 535, 540

Quasi-ARMAX model, 465–466
  hybrid, 467–468, 474
Quasi-periodic model or process, 70

Radar:
  antenna, 399
  cross-section (RCS), 442
  Doppler weather, 433–434 (*see also* TDWR)
  imaging, 433
  parameter, 445
  tracking, 433
Radial basis function (RBF) network, 151, 158
Random partitions, 274, 282
  exchangeable, 296
Random patterns, 502
Random variable, i.i.d. 277, 357, 378
Rank test(s), 187–188, 196, 200, 208
  based on EVD, 188
  based on LDU decomposition, 200
  comparison and evaluation of, 208–215
  Gramian-based (GRAB), 201, 207, 212
    multivariable, 208
Rational extension, 29–30
Realization, 4, 14
  algorithm, eigensystem (ERA) 109–111, 149
  ARMA model, 134, 138
  deterministic, 19, 26
  finite-interval, 24
  minimal, 3, 19
  minimal partial, 29
  minimum state-space, 182
  state space, 1
  stationary, 18
Reduced-order functional estimator, 259
Redundant parameter vector, 464
Relative frequency, 528–529
Resolution enhancement, 442
Riccati difference equation, 241, 245
Robbins-Monro method, 481
Robustness issues, 402

Sampling algorithm or method, 535
Scaling function, 155, 159
  Haar, 159
  Meyer, 155–156
Scattering theory, Lax–Phillips, 15
Seasonal adjustment, 46
Seasonal component, 43–47
  non-Gaussian, 61
Seismic data, 52, 63
Self-similar pattern, 499, 501, 507, 520
  experiments, 516–521
Self-similarity, 499
  mappings, 501
Sensor position errors, 407
Sequential estimation, statistical, 481
Shalvi–Weinstein approach, 375, 379
Shierpinski gasket, 501
Signal:
  estimation, 412
  processing in radar system, 431
  separation, blind, 375
  subspace fitting (SSF) algorithm, 420

Similarity (*see also* Self-similarity):
  contour, 504
  function, 503
Singular value, 437
  Hankel, 149
Singular value decomposition (SVD), 87, 92, 203, 444, 446
  generalized, 92–93
Smoother, fixed-lag, 58, 418
Smoothing, 42, 54, 56–58
  error, 27
  nonlinear, 73
Smoothness criterion, 40
Smoothness priors, 39–41, 50
Soft bound, 454
  approach, 458
Spectral density matrix, 2, 181
  cyclic, 327, 333, 346
Spectral factor, 2
Spectral factorization theorem, 181
Splitting subspace:
  Markovian, 14–18, 25
  (minimal) oblique, 21, 23
State space model, 45, 92
  estimation, 95–97
  general, 54
  nonlinear, 73
Stationary process, 1, 359
  M-channel weakly, 331
Statistical decision making, 439
Stochastic approximation, 481
Stochastic binary tree search, 487
Stochastic embedding approach, 458
Stochastic mapping, 485
Stochastic realization, 1
  classical, 2
  geometric, 4
Stopping time, 484
Stopping variable, extended, 483–484
Strong cut, 300
Strongly consistent, 365
Subspace(s):
  corrected (SC) algorithm, 403
  corrected (SC) MVDR, 413
  decomposition algorithm, fast, 100
[Subspace(s):]
  identification, 5, 29
  past and future, 10
Subspace fitting, 402
  signal (SSF), 402, 411
  weighted, 402
Subspace(-based) methods, 32, 409
  statistical properties of, 32
Summability conditions, 359
  absolute, 377
Supervised learning, 481
System analysis, 432
System identification, 85 (*see also* Identification)
  automatic, 84
  in continuous-time domain, 124
  structure of, 84
  subspace-based state space (4SID), 99, 180
  transfer function matrix, 365
  using polyspectra, 355

T^2 statistic, 113
Terminal Doppler weather radar (TDWR), 433, 439, 441, 449
Threshold, 482–483
Time-moment sequence, 137, 139
Time-varying variance, 62
Toeplitz matrix, 8, 30
  block, 183
Trade-off analysis, 432
Transfer function matching, 367
Transition probability, 527, 530–532
Trapezoidal approximation, 122
Trispectrum, 361
Truth function or value, 300, 307–308
Tufts–Kumaresan method, 442

Unbiasedness, 261
Under-modeling, 127
Unmodeled dynamics, 122, 454, 457, 462, 469
  index of, 470
Unobservable space, 437
Unstructured errors, 408, 419

Vague perception, 300, 304
  of random phenomena, 299
Variance error, 161

Wavelet, 155
  neural network (WNN), 157–159
  transformation, 151
Weierstrass theorem, 179
White Gaussian process, 436, 442
White noise or process, 1, 41, 179, 181
  finitely additive, 217–218
  Gaussian, 39, 42, 209, 243, 455
Whitening condition, normalized, 378, 384
Wing flutter, 114–115
Wold decomposition theorem, 15
Wolfer sunspot data, 71

Yule–Walker(-type) equation, 190, 206, 335

Zakai equation, 218, 227
Zipf's law, 293