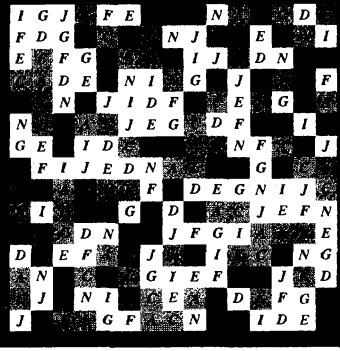
Methods and Models in Statistics

This page intentionally left blank

Methods and Models in Statistics



In Honour of Professor John Nelder, FRS

Editors

Niall Adams
Martin Crowder
David J Hand
David Stephens

Imperial College London, UK

ICP

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: Suite 202, 1060 Main Street, River Edge, NJ 07661
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

METHODS AND MODELS IN STATISTICS
— In Honour of Professor John Nelder, FRS

ISBN 1-86094-463-9

Printed in Singapore by World Scientific Printers (S) Pte Ltd

To John Nelder for his 80th birthday

CONTENTS

PREFACE

Methods and models in statistics: in honour of Professor John Nelder, FRS

This book contains written versions of papers presented at a two-day symposium in March 2004 in honour of Professor John Nelder, FRS. John Nelder is one of today's leading statisticians, having had a fundamental impact on a variety of areas of data analysis. The papers here have been selected to indicate the breadth of that impact. They range from the collection of data (Bailey on experimental design), over the fundamentals of statistics (Cox), through statistical modelling (Cullis, Smith, and Thompson, Wakefield, and Ripley), via statistical computing (Payne), to generalisations of that core statistical concept of likelihood (Pawitan, Lee). Also included is a biographical sketch of John, by Senn, and a bibliography of John's works to date (he is still carrying out and publishing important work).

Readers may not know that John is also a keen ornithologist. In recognition of this, we have included here one of his earliest papers, a paper that combines his statistical and ornithological interests: *A statistical examination of the Hastings Rarities.*

Niall M. Adams, Martin J. Crowder, David J. Hand and David A. Stephens
Imperial College
March 2004

This page intentionally left blank

FOREWORD

I have not been generally in favour of *Festschrifts*, finding that often the contributions did not seem to make any sort of consistent story. Now that I am on the receiving end of one, it is a pleasure to find that the contributions in this volume combine to give such a good view of the present-day practice of statistical science.

Beginning with Finney's definition of our subject as 'making sense of numbers' we can immediately see what a complex business this is going to be. We have only to begin with the thought 'how were the numbers collected and can they answer the questions we wish to pose?' I was brought up in the Rothamsted tradition where matters of design were taken especially seriously. When I first went there I was given some complex experiments on trace elements to analyse, the designs having split rows and columns, with further subdivisions. I was able to construct the form of the anova tables from what I had learnt, correctly I believe, but then I began to wonder what the rules actually were. After a long period of gestation a formal description of the rules for a large class of (generally balanced) designs appeared my 1965 papers in *Proc. Roy. Soc. (A)*. BAILEY has ably summarized what has happened since then, much of the work originating with her, and using tools I did not possess. Given the vital necessity of good design, why are so few statisticians given a thorough grounding in it?

Between design and analysis comes the actual execution of the experiment. All too often this phase is regarded as the sole job of the experimenter, but I believe that every fully-trained statistician must have done an experiment. Experimentation is not as easy as many people think!

After execution comes analysis, an endlessly fascinating exercise. We have to define a statistical model class and then assess how far the patterns that they define express the pattern in the data satisfactorily. There is always the possibility that no member of the class is adequate. This, I believe, is what happened with the famous stackloss data, in which as Dodge showed, all but five of the 21 data values had been found by someone to be an outlier of some sort. By broadening the model class from classical normal models to GLMs, I was able to show that there is a simple model in the wider class in which no outliers whatever appear! RIPLEY shows how difficult this process of extension and selection can be. We still have much to learn here.

After fitting a model and summarizing the results via parameter estimates and associated estimates of uncertainty, we then should be checking

to see if the fit is internally consistent. This process of model checking, whose elements are so clearly described by COX, is, I believe, grossly underused in the current statistical literature. Increasingly powerful software makes it ever easier to fit more and more complex models, but this is of no help if a simple plot then shows clearly that the model cannot be right.

There have been two important extensions of classical normal models in the last 30 years or so; GLMs and linear mixed models. The first extended the distribution class and the scale on which additivity was defined, while the second allowed random effects as well as fixed effects, the errors remaining normal. It was obvious to ask if these two extensions could be combined, with random effects occurring in the linear predictor of a GLM; they can, of course, and the resulting models are known as generalised linear mixed models (GLMMs). It is also worth asking if the distribution class of the random effects can itself be extended, as in the generalization of normal models in GLMs. The answer to this will be found in LEE's paper, which describes some of our joint work over some 15 years. He shows that a class where the random effects have a GLM conjugate distribution can be fitted with a single algorithm. Other major aspects of the work that he writes about are the joint fitting of models for mean and dispersion and the introduction of random effects into the model for the dispersion. This last extension, which has become popular in the modelling of financial data, will, I believe, be found to have applications in many other fields. By using the idea of h-likelihood as a criterion for fitting the models. The analysis of the whole class reduces to the fitting of an interlinked set of GLMs. The idea of h-likelihood is described from a different perspective by PAWITAN, together with an acute analysis of the place of likelihood in both modelling and inference.

Although we have code for fitting our new model class (double hierarchical generalized linear models (DHGLMs)) it is simple-minded and does not take any short cuts. The paper by CULLIS et al. shows how efficient algorithms can be written for GLMMs. I look forward to an extension of their ideas to our wider class.

The class of DHGLMs, described by LEE, though in many ways of considerable generality, relies on linearity through the linear predictor. It does not include non-linear models of the type dealt with by WAKEFIELD. I find it hard to visualize how random-effect models can be developed from general non-linear models. It cannot, or should not, be simply a matter of adding random effects on the end of a complex non-linear function. This looks like a fruitful area of future research.

Underlying all the papers is the absolute necessity of having the computer as a working tool, to organize complex data, to fit models, to simulate sampling distributions etc. PAYNE gives a good history of the development of statistical computing in Britain. Those of us concerned learnt a lot, though it is disappointing to find writers of software 30 years on who seem to have learnt nothing from our mistakes. (It is not wrong to make mistakes, but it is wrong not to learn from them.)

Finally a self-reference: I am not sure if such is rated as good form, but I could not resist the temptation to reprint a paper of mine from 1962, of which few statisticians are aware. It concerns an analysis of extraordinary numbers of bird rarities occurring in the Hastings area at the beginning of the last century. Many readers of the Witherby Handbook had noticed these records, but, with Hilary Fry's help, I decided to analyse them. I then found that two eminent ornithologists, Ferguson-Lees and Nicholson, were making their own enquiries on how the deceptions might have been carried out. Our papers constituted an entire issue of the magazine *British Birds* on this one topic, resulting in enormous interest among ornithologists. The result was the deletion of several hundred records from the official British list. This occurred just after one author had completed his five-volume work on British birds, in which all the Hastings records were accepted as genuine. Here the statistical techniques verged on the trivial, but the consequences were anything but trivial.

I recommend the papers in this book to anyone wanting a view of the way that statistical science is developing, and I thank the authors for their kind remarks about my contributions.

John Nelder, Imperial College, 2004

This page intentionally left blank

JOHN NELDER: FROM GENERAL BALANCE TO GENERALISED MODELS (BOTH LINEAR AND HIERARCHICAL)

Stephen Senn
Department of Statistics
University of Glasgow

A personal portrait of some aspects of John Nelder's life, personality and work is given.

Introduction: some personal remarks

Some personal remarks are in order, since the reader will be baffled as to how a medical statistician such as myself, a practitioner rather than a theoretician, can be a suitable person to give an overview of John Nelder's work in statistics, which has been powerful and deep, and whose importance is only 'applied', rather than 'theoretical', in the sense that it is of great utility to many practitioners in many fields. In fact, this introduction will provide more of a plausible excuse than a valid reason but at least the reader will have some understanding as to why I am writing this.

I first met John Nelder when we were both lecturers at the annual one-day meeting of the Swiss Statistical Association in the early 1990s. John's lecture made more of an impression on me than my own since I can remember that his was about 'significant sameness' but cannot recall what I lectured on (either baseline balance or cross-over trials). There was little opportunity for us to talk to each other and I doubt that we exchanged more than a few words. We next met at the annual meeting of the PSI (Statisticians in the Pharmaceutical Industry) at Bournemouth in 1996. John lectured on Hierarchical Generalised Linear Models (of which more anon) and I on portfolio management but again we exchanged very few words, although I remember the meeting for another contribution of John's and that was that it was the first time I heard him play the piano. (This will

also be covered anon.) However, shortly after the Bournemouth meeting I saw John standing on Harpenden station. I think all who know John will agree that he has an unmistakable appearance: that of a tall, commanding and striking figure who would satisfy any schoolboy's prejudice as to what a scientist should look like. (For some reason the old-fashioned English slang word 'boffin' springs to mind.) There was no mistaking the man and I plucked up courage and introduced myself properly to John this time and have counted myself as a friend, almost from that instant.

John lives with his wife Mary in the village of Redbourn, which is only three miles from Harpenden, in which is situated Rothamsted research station, the statistical department of which has counted Fisher, Yates and also, of course, Nelder, amongst its heads. It is a curious fact that interests me that the rudest letter (with possible exception of one addressed to John Tukey) in the collected statistical correspondence of the first head of statistics at Harpenden[1] is addressed to the man who was destined to become the third. From that you can judge that John Nelder is not deterred easily. Harpenden also has the nearest and most convenient station to Redbourn for getting into London. Although John took early retirement from Rothamsted aged only 60, this made no difference to the interest he has had from an early age in statistics and he has been travelling in by train to Imperial College ever since. Since we are gathered here at Imperial to celebrate John's 80th Birthday, it is a simple calculation to determine that he has been doing this for 20 years. For seven of those years I was in the fortunate position of being a frequent travelling companion of John's, since from 1995-2003, while working at University College London, I lived in Harpenden and after our 1996 introduction frequently found myself on the same train as John.

In fact, John's train-travelling habits have another relevance here since they exhibit many features that also find a parallel in his justifiably famed problem- solving and algorithmic abilities. First of all there is the choice of train. If you wish to get into London of a morning from Harpenden without paying peak time rates, you must wait until after 09:30. The first train is then a slow train, which stops at every station (of which there are eight) between Harpenden and King's Cross. The second is a fast train that only stops at St Albans. However, the second is scheduled to arrive after the first but it only has four carriages as opposed to eight and you frequently have to stand on your way to town. I think you are beginning to perceive that this is a problem with many variables. John's default choice is to take the earlier train but he always pays keen attention to the leader board and will

adjust his plans accordingly. (As a Swiss, I can't resist making the comment that such adjustments are more frequently necessary for travellers in John's country than in mine.) Then there is the choice of seat. John always makes for the same seat, which is facing the direction of travel at the back of the train and next to an exit. The reasons for this choice are not declared by John but I note for the benefit of the curious that this seat a) allows adequate leg room for a tall man b) Permits the traveller to see what is happening on the line c) Permits quick exit from the carriage and d) Deposits the traveller directly opposite the entrance to the underground, which facility he will wish to use if proceeding to Imperial College from King's Cross. John is always *very* disappointed if his seat is taken.

A frequent companion on our travels in to London had been Michael Healy, who lives in Harpenden and it is entirely appropriate, and a very great pleasure to me, and I am sure also to John, that Michael should be giving the after-dinner speech at this meeting. Travelling in with these two wise old men has often made me feel very foolish although, occasionally, despite the fact that they are both very sprightly, it has had the side-effect of making me feel young.

When travelling with John you will need to be warned of a few things. First he is very interested in statistics, second, he believes you are too and third he pays you the compliment of assuming that you know more or less everything about statistics that he does, except, perhaps, what he has just recently discovered. In my case, he is wrong about the last of these by a very big margin and, although I have received a considerable and valuable education travelling in with John over the years and had many enjoyable discussion with him on the subject, I have also frequently emerged from a journey feeling a bit of a fraud. John likes talking about likelihood, be it conditional, joint or marginal, penalised, pseudo or quasi or even extended quasi, partial, restricted or profile or, latterly, hierarchical. In particular, John takes a poor view of those who confuse pseudo and quasi-likelihood[2] and I fear that I do not even know enough about these two to confuse them. To claim I confused them would be as misleading as to say that I found it hard to tell the difference between Czech and Slovak. As well as likelihood, what John doesn't know about deviance is not worth knowing. (This is a dangerous statement to take out of its statistical context and must be interpreted intelligently.)

Childhood

The Child is Father of the Man

Wordsworth

I want also to say something about John's childhood, as I think that this gives a clue to his personality and also to some qualities of his work. John has described his childhood as idyllic and apologised for the term, because he is aware that it is a cliché but knows that it was so. He was born the son of the son of the son of a hotel-keeper[3] in Brushford near Dulverton, a small town in Exmoor and in the beautiful valley of the Barle, close to its confluence with the Exe. He has had this to say about himself as a child

"We swam in the river. We flooded a piece of a field in the winter in case there was enough ice to skate on, which was not very often. I went everywhere on my bicycle, up to the edge of the Moor, and into the woodlands that bordered the rivers Exe and Barle. I think it's hard to imagine a nicer place with no restrictions on where we could go. We collected plants and mounted them in books. I loved the long names of the families - caryophyllaceae, scrophulariaceae and so on. We collected birds' eggs, something that of course would be absolutely forbidden today. We collected butterflies. We learned a lot about natural history from what we simply discovered in our movements around."[3]

This conjures up a delightful image in my mind of John as a child of nature in some latter-day Garden of Eden looking on everything in delight and wonder but consumed with the sort of overwhelming curiosity that got mankind into trouble in the first place. John is, in fact, a great gardener, or perhaps it would be more accurate to say that Mary is a great gardener and so is John (they have a most beautiful and extensive garden at Redbourn). John is also a keen amateur ornithologist and has applied statistical reasoning to this. Also, of course, all of his working life he has worked at an agricultural research station, first at Wellesbourne and then at Rothamsted. I think that a love of nature has been a constant of his life. But I think that the relevance of his childhood goes beyond this. I am convinced that John sees statistics as one great big garden and he is determined to go about in it and discover all the possible varieties of likelihood that there are, encourage them to flourish, determine which one grows best in which soil, and show how they are connected to each other.

I also find it revealing and interesting that John refers to 'we' in this piece, a reference to his brother. Even in his early scientific investigations there was a collaborator. A key feature of his scientific career has been

his many successful collaborations with one other statistician, so we have, Hammersley and Nelder on simulation[4], Nelder and Mead[5] on a simplex method of optimisation, Nelder and Wilkinson creating Genstat[a], Nelder and Wedderburn on generalised linear models[6] and then an important book on the same subject by McCullagh and Nelder[7] and finally (for the moment at least) Lee and Nelder on hierarchical generalised linear models[8].

The section quotation is from Wordsworth's famous poem, which begins,

'My heart leaps up when I behold
A Rainbow in the sky'.

I am sure that John's does too. As I have already stated, his love of Nature is very deep, but he has the advantage over Wordsworth that it also leaps up when he beholds data. Or to adapt the language of *Apocalypse Now,* he could say, 'I love the smell of data in the morning'. To give an example, a few years ago I drew John's attention to a witty article by Yadolah Dodge[9] looking at the history of attempts to analyse Brownlee's famous stack-loss data[10]. Dodge identifies 60 analyses of the data by various authors and points out that of 21 data points given by Brownlee only five have been identified as an outlier by nobody. Now something like this is a challenge to a man of John's mettle and, sure enough, we now have a further paper in the statistical literature by John entitled, 'There are no outliers in the stack-loss data'[11].

General balance

The first of his many important contributions to statistics that I wish to discuss did not, however, start as a collaboration, although it later led to a collaboration with Graham Wilkinson in Adelaide. Let John speak for himself:

"During my first employment at Rothamsted, I was given the job of analysing some relatively complex structured experiments on trace elements. There were crossed and nested classifications with confounding and all the rest of it, and I could produce analyses of variance for these designs. I then began to wonder how I knew what the proper analyses were and I thought that there must be some general principles that would allow one to deduce the form of the analysis from the structure of the design. The idea went underground for about 10 years. I finally resurrected it and constructed the theory of generally balanced designs, which took in virtually

[a]Gentstat®– trademark information supressed hereafter

all the work of Fisher and Yates and Finney and put them into a single framework so that any design could be described in terms of two formulas. The first was for the block structure, which was the structure of the experimental units before you inserted the treatments. The second was the treatment structure - the treatments that were put on these units. The specification was completed by the data matrix showing which treatments went on to which unit."[3]

In fact, this extremely powerful general framework of John's[12,13] is the basis of the analysis of variance capabilities of Genstat, a result of his further collaboration with Wilkinson already alluded to. A very wide class of designs, including completely randomised designs, randomised blocks, split plots, Latin and Graeco-Latin squares, split-split plots, balanced incomplete blocks, balanced lattices, Youden squares and many more[14], in fact all designs possessing the property of 'first-order balance', can be analysed using this approach. As far as I am aware, Genstat is the only package that does this and although I am not going to attempt to explain the property of first-order balance I am going to draw attention to an explicit feature of this whole philosophy of analysis of variance that is lost in many modern approaches to data-analysis.

The feature is that a clear, and to my mind fundamental, distinction is drawn between blocking and treatment structures. Let me give an example from my own field, that of clinical trials. You could have a clinical trial in an indication in which you believed that the outcome would, other things being equal, differ strongly by sex. That being so you could decide to make sex a blocking factor by running two randomisation lists, one for men and one for women. Since, of course, you will have many patients of each sex under each treatment you have the structure of a two-way analysis of variance with replication. In a linear model you could have 'sex' as a main effect and 'treatment' as a main effect and also investigate the interaction between the two. Such a model makes no distinction of type between sex and treatment and in nearly all statistical packages there will be no way of distinguishing them. Not so with Genstat in which you can declare 'sex' as a blocking factor and 'treatment' (appropriately) as a treatment factor. The point is that you have allocated the patients their treatments and these could have been different but you haven't allocated them their sexes and these could not and once you have declared one as a block and the other as a treatment Genstat will go on to encourage you to make different sorts of inferences about them.

Generalised Linear Models

Skipping over much important other work of John's, not least his citation hit with Roger Mead[5] on 'A simplex method for function minimization', we come to another stellar contribution of John's to modelling data, namely his paper with Wedderburn on generalised linear models. Appropriately enough this was published in that annus mirabilis of statistics 1972, a year that gave us not only GLMs but also proportional hazards[15], Bayesian approaches to linear models[16] and the log-rank test[17]. The statistical world changed and if it is now puzzling for medical statisticians to try and imagine what survival analysis looked like before David Cox's seminal paper, it is also difficult for applied statisticians generally to imagine what modelling was like before we had GLMs. Nelder and Wedderburn was a paper that changed the statistical landscape for ever and it is simply impossible now to envisage the modelling world without it.

That this is so is not without its irony and, although John may not be pleased to hear me say this, his later work on modelling has tended to have the effect of making his earlier work on general balance less relevant to us now. The very flexibility of GLMs has encouraged us to fit more things and differently. The randomised experiment with distinction between block and treatment factors, Normal error terms, correct and inevitable partitionings of variances determined by design, and close parallels between randomisation and modelling approaches, seems to us now less like a commanding fortress of excellence, set somewhat apart in the city of inference, but more like a single apartment (albeit, perhaps, the penthouse suite) in the tower-block of data-modelling we all now occupy.

Statistical computing

John has been a major force in statistical computing but I think that his efforts in this direction have not always been crowned with the success that they deserve and, in this context, I am going to permit myself two critical remarks.

First, Genstat, which is a magnificent package, with whose origins and development John has been intimately involved, has failed to make the impact it deserved. This, I believe, is partly traceable to an early decision by the developers to make it a powerful and flexible tool for the expert statistician but user-unfriendly to the amateur. This latter feature was always pointless, since there was no possibility of the Genstat developers dictating what would happen in the world at large. Sure enough, other packages

concentrated on being user-friendly to the statistically inept and as a consequence huge quantities of misleading analyses are produced by all and sundry. But would things have been worse if the Genstat developers had also played this game? It is an irony that Genstat is now one of the most user-friendly packages there is. It has the best menu to command mode integration of any package I know and superb spreadsheet capabilities as well, in addition, of course, to all-round statistical capabilities. (As regards that, I can really only fault it on survival analysis. This reflects its agricultural origins since little work has been done on modelling the life-times of vegetables.)

Second, the algorithms that *were* developed to exploit Nelder and Wedderburn's wonderful synthesis of 1972, were placed in a new package, GLIM[b], rather than being immediately offered via Genstat. So the opportunity was missed to make Genstat a package that everybody needed to have. GLIM and the GLIM community took off rapidly, especially in Britain, but in the long-term the sort of specialist package that GLIM was, was doomed to pass from being cutting-edge to superfluous. Now you can do your generalised linear modelling within SAS®, SPlus® and, of course, best of all, within Genstat.

Collaboration with Youngjo Lee

In recent years, John has been extending our modelling capabilities by combining two different developments in data-analysis. The first of these, due to his work with Wedderburn[6], was to throw off the shackles of Normality. The second, associated with the work of people such as Patterson and Thompson[18], permits general handling of models for data with more than one error term provided that these errors are Normally distributed. Together with Youngjo Lee, John has been developing hierarchical linear models, that is to say models that, 'allow extra error components in the linear predictors of generalized linear models'.[8] Their approach to fitting these avoids the integration necessary in order to marginalise the likelihood over the unmeasured random effect and instead relies on a generalisation of the joint likelihood of Henderson, the hierarchical or h-likelihood. I think that it is fair to say that this work has met with some hostility, but this has not deterred John and Youngjo who bit by bit have revealed the supposed counterexamples to be false without, thereby, convincing all their critics.

[b]GLIM®– trademark information supressed hereafter

Only the future will tell whether this work of theirs is eventually revealed as a misleading diversion or a true advance but either way it seems to me that two points are undeniable. First, that whatever repugnance some may feel for the philosophical basis of h-likelihood their modelling approach performs extremely well in practice. Second that their espousal of h-likelihood is causing many to think more deeply about the issues involved.

I can only describe John's and Youngjo's collaboration with each other as marvellous. Despite considerable differences of age, culture and education and a formidable obstacle, even in this age of electronic communication, of distance, they have achieved a prodigious output of papers and clearly love working together. As the younger partner, it is perhaps understandable that Youngjo has been the more regular traveller and if you visit John either at Imperial or Redbourn you can often come across Youngjo.

I have been racking my brains to think of similar bivariate statistical partnerships but can think of hardly any: Neyman and Pearson, perhaps, but I do not think that they published as often together as Nelder and Lee. Fisher and Yates are two names that go together but in fact they have very little joint published work. Mather was a frequent published collaborator of Fisher but that was in genetics. The most appropriate equivalent partnership, I believe, comes from mathematics and is that of Littlewood and Hardy.

In summary: some personal remarks

Where I live and work now, in Glasgow, I can walk to work in the time it previously took me to walk to the station in Harpenden. I do not miss travelling on Thameslink from Harpenden to London but despite that it is no contradiction to say that I *do* miss my Journeys with John. In addition to discussing statistics, John had the habit of showing me from time to time the latest offer from his wine club and asking me whether I would be interested in going halves on a crate. The answer was always, 'yes,' and John always delivered my share from Redbourn to Harpenden by car. When I have been asked by guests, for example, where I got a particularly curious and fine bottle of Lebanese wine, I have taken great pleasure in replying, "Oh I get them from my wine-merchant. His name is John Nelder. He is also quite well-known as a statistician."

I shall also miss the famous musical matinees at Crown Street, with the entertainment provided by John at the piano assisted by friends and colleagues and, at the interval, the table groaning with cakes made by Mary.

John is a very fine pianist and love of music must count, together with that of statistics and nature, as a major theme of his life.

I wish John and Mary many more musical matinees and John and Youngjo many more papers together.

References

1. Bennett JH. *Statistical Inference and Analysis Selected Correspondence of R.A. Fisher.* Oxford: Oxford University Press, 1990.
2. Nelder JA. Quasi-likelihood and pseudo-likelihood are not the same thing. *Journal of Applied Statistics* 2000;27(8):1007–1011.
3. Senn S. A conversation with John Nelder. *Statistical Science* 2003;18(1):118–131.
4. Hammersley JM, Nelder JA. Sampling from an isotropic Gaussian process. *Proceedings of the Cambridge Philosophical Society* 1955;51:652–666.
5. Nelder JA, Mead R. A simplex method for function minimization. *Computer Journal* 1965;7:308–333.
6. Nelder JA, Wedderburn RWM. Generalized Linear Models. *Journal of the Royal Statistical Society A* 1972;132:107–120.
7. McCullagh P, Nelder JA. *Generalized Linear Models.* Second ed. London: Chapman and Hall, 1989.
8. Lee Y, Nelder JA. Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B- Methodological* 1996;58(4):619–656.
9. Dodge Y. The guinea-pig of multiple regression. In: Rieder H, editor. *Robust Statistics, Data Analysis and Computer Intensive Methods: in honor of Peter J. Huber's 60th birthday. Springer Lecture Notes in Statistics 109.* New York: Springer, 1996.
10. Brownlee KA. *Statistical Theory and Methodology in Science and Engineering.* New York: Wiley, 1960.
11. Nelder JA. There are No Outliers in the Stack-loss Data. *Student* 2000;3(3):211–218.
12. Nelder JA. The analysis of randomised experiments with orthogonal block structure I. Block structure and the null analysis of variance. *Proceedings of the Royal Society of London. Series A* 1965;283:147–162.
13. Nelder JA. The analysis of randomised experiments with orthogonal block structure II. Treatment structure and the general analysis of variance. *Proceedings of the Royal Society of London. Series A* 1965;283:163–178.

14. Payne RW, Lane PW, Digby PGN, Harding SA, Leech PK, Morgan GW, et al. *Genstat 5, Release 3 Reference Manual.* Oxford: Oxford Science Publications, 1993.
15. Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972;34:187–220. 16. Lindley DV, Smith AFM. Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society B* 1972;55:1–41.
16. Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society A* 1972;135(2):185–207.
17. Patterson SD, Thompson R. Recovery of inter-block information when block sizes are unequal. *Biometrika* 1971;58(3):545–554.

This page intentionally left blank

SOME REMARKS ON MODEL CRITICISM

D. R. Cox
Nuffield College and Department of Statistics
Oxford

To John Nelder, on his 80th birthday

Some general comments are made about the role of model criticism in statistical analysis and the relation of the corresponding methods to statistical theory more broadly. Types of procedure are reviewed with examples.

1. Introduction

Box (1980) introduced the term *model criticism* although the notion is much older going back at least to the informal use of least squares residuals in the 19th century and more formally to Karl Pearson's development of the chi-squared test of goodness-of-fit (Pearson, 1901). Box also coined the aphorism *all models are false but some models are useful*, raising immediately the questions as to what makes a false model useful and as to why should we bother to criticize models that we know are false anyway.

The organizers of this meeting asked for a paper on model criticism so that the second and easier question has to be answered, the answer presumably being that the objective is to point the way, if appropriate, to more useful models, and so requiring at least a partial answer to the first question about model usefulness. There are various criteria that may be relevant; see, for example, Cox and Wermuth (1996, pp 18-20). These include providing a link with underlying theory or background knowledge, providing a comparison with previous work in the field, suggesting a potential data-generating process, having parameters with individual clear subject-matter interpretations, specifying an adequate but not overelaborate error structure, including parameters representing key features of the research questions of interest regardless of whether these effects are "significant" and finally giving an adequate fit.

A broad distinction, even if a bit fuzzy at the edges, may be drawn between on the one hand purely empirical statistical models and on the other hand substantive models that contain some specific-subject matter basis. In the former some combination of adequacy of fit, interpretability of parameters and simplicity are appropriate for judging model suitability. In the latter type, it may be unwise to abandon totally a type of model solely on the basis of lack of fit, especially at the edges of the region of investigation.

There is the further implication that a procedure that at least points in the direction of potential improvement in the model is to be preferred over one that merely signals fit or lack of fit.

The remainder of this paper focuses on the issue of fit.

The detailed literature on examining model adequacy is vast and this short review centres around the following themes. First how does model criticism fit into some formal theory of statistical inference? Secondly what broad types of procedure are available and finally how do these procedures fit into statistical practice?

For some purposes it may be helpful to distinguish model criticism from data criticism, the latter but not the former concentrating largely on the detection of relatively small numbers of anomalous observations.

2. Formal theory

We consider a parametric family of models $f_Y(y;\theta)$ with $\theta \in \Omega_\theta$ for observed random vector Y, the density depending also on data z not regarded as random and regarded as fixed. If there is a minimal sufficient statistic S of relatively small dimension the likelihood can be written in the form

$$f_S(s;\theta) f_{Y|S}(y \mid s),$$

so that the second factor is available for examining model fit. Are the data in some sense suspiciously extreme in their conditional distribution given $S = s$?

While discussions of sufficiency typically focus on the role of the first factor, for the present purposes it is the second that is needed.

For the normal theory linear model with unknown variance, S consists of the least squares estimates and the residual sum of squares and the conditional distribution is essentially that of the standardized residuals and any chosen function of them has in principle a distribution not depending on θ or on s, the latter being clear on grounds of invariance. An explicit use of this kind of argument in a different context is by Fisher (1950) in his

examination of goodness of fit with the Poisson distribution. Given a series of observations $y_1, \ldots, y_n$ and the model that they are a random sample of a Poisson distribution of unknown mean θ, Fisher suggested as possible tests of model adequacy the sample variance and the number of zeros. He obtained the exact conditional distribution of these given the sample mean. The former statistic gives essentially the chi-squared dispersion test which Fisher had set out much earlier. In principle similar arguments are possible when the parameter θ is the canonical parameter of an exponential family distribution as in generalized linear models with canonical link (McCullagh and Nelder, 1989). These arguments leave open the question of the choice of the function of the data to be used in the assessment of fit. This is clearly a crucial issue in that all data are extreme in some respect, as Neyman in particular pointed out.

Box (1980), in effect following Fisher, suggested essentially an absolute test of significance in which the sample points are ordered in decreasing value of $f_{Y|S}(y \mid s)$ and the probability calculated of a deviation as or more extreme than that observed. That is, a p-value is found using the ordinate of the conditional density as test statistic, small values being evidence against the model. There seem a number of objections to this as a general procedure. First there are obvious difficulties with continuous variables and while it can be argued that all real problems are discrete there would be some dependence on the mode of discretization adopted. More seriously, perhaps, it will not be true in general that the test statistic points to deviations in the direction or directions of substantive importance. This bears, in particular on the issue of whether a suspiciously small value of, say, chi-squared, is as strong evidence against a model as is a large value of chi-squared.

For problems with no simple reduction by sufficiency one might regard the maximum likelihood estimate and the observed information matrix as approximately sufficient and consider the conditional distribution of Y given those two statistics which in well-behaved problems will have only slight dependence on the unknown θ. One could regard some of the procedures for checking nonlinear models via simulation in this light. A more intuitive approach is to define residuals for each observation, or sometimes groups of observations and to plot these possibly computing summarizing statistics from them. A rather general definition of residuals and discussion of test statistics based on them was given by Cox and Snell (1968, 1969) but has rarely been used outside survival analysis for which there is a quite extensive special literature.

So far it has been implicitly assumed that some procedure formally or

informally akin to a test of significance is involved. That is, we make some assessment of what the data ought to be like if the model were true, for example that the number of zeros in a Poisson series should be simply related to the mean. We then examine the data for consistency with that assessment and if the discrepancy is too great regard the model with suspicion.

A different approach starts from the position that the model is at best a good approximation and that the problem should be regarded as one of estimating the distance between the assumed model and the "true" one. Let us skate over the question as to what the notion of a "true" model might mean. Then a distance has to be defined and in some contexts at least, for example the classical chi-squared goodness of fit with a special multinomial distribution, that is not difficult via the asymptotic expectation of the test statistic under the "true" model. Another appealing possibility with a long history for discrete data is to use a misclassification rate; the fitted presumed false model can be used to classify individuals and the proportion misclassified can be used as a descriptor of model fit and also (Firth and Kuha, 2004) be the basis of formal procedures. The broad approach of estimating a suitable distance parameter is explored systematically by in as yet unpublished work by B. Lindsay. A difficulty with this approach is that the measure of distance may be difficult to define in an interpretable way and more seriously that the estimate may give no clear indication of how the model might be improved if the distance is somehow judged too large.

In a fully personalistic Bayesian discussion, in so far as I understand it, probability assessments cannot be wrong: they are what they are and provided they are coherent, i.e. internally consistent, that is all that is required. At a more pragmatic level there could be difficulties if either the data are in conflict with *any* model of the proposed form or if the data and the prior are discrepant. Thus a sample possibly from a Poisson distribution might have its variance approximately equal to its mean but a mean far in the tails of the proposed prior distribution. That would mean that either the data have a systematic error or that the prior is based on misconceptions (or both). If the prior is supposedly innocuous, or even chosen in the light of a surreptitious look at the data, this difficulty will not arise. If, however, as is an attraction of the Bayesian formalism, the prior is a serious attempt to include additional information, the issue of consistency of prior and data seems in principle important and rarely discussed. In so far as the prior can be regarded as pseudo-data, checks are certainly possible, although typically outside the Bayesian formulation.

3. Choice of criterion

For the rest of the discussion it is assumed that model criticism consists of choosing some aspects of the model that can be predicted under the model and checking whether the prediction is satisfied, as judged possibly by a graphical procedure or by calculating the p- value of one or more statistics. These two methods of assessment are not in principle essentially different in that the graphical procedures cannot be interpreted except via substantial experience or via something akin to a simulation envelope and the relation of the latter to a p-value is clear. The notion of deducing some testable consequence of the model and then examining the consistency of the data with that calculation is no different in principle from that of testing consistency in a deterministic context.

How should the criteria used for assessment be chosen? There is now a major difficulty. First in most applications one has some idea not just of the kinds of departure likely to arise but more importantly of the kinds of departure of substantive interest, the most interesting of which may throw into question the formulation of the research question under investigation. On this count one would like criteria that focus on aspects of importance and also, although this is not essential, are diagnostic of the kind of modification required, ranging from checking and possibly discarding some observations to radical reformulation of the model or even of the research question being studied. To some extent the model failures likely to be important can be foreseen *a priori*. On the other hand there is always the possibility of some important effect of totally unforeseen kind.

This leads to a classification of procedures ranging from the highly focused to those that M.S.Bartlett termed *omnibus tests*. The former, considered as tests, will typically involve a small number of degrees of freedom, for example if converted into chi-squared form, and may indeed essentially be based on estimates of either one or at least a small number of parameters in some model different from that under test. The omnibus tests will be based on a large number of degrees of freedom with a corresponding loss of sensitivity against specific alternatives. Somewhat equivalently an omnibus procedure may be based on inspection of residuals for any one of a large and possibly ill-specified range of anomalous features. There are a large number of such procedures. If they are intended to detect anomalous observations or small numbers of observations, smoothing of residuals is undesirable. Otherwise to detect "smooth" departures from an initially specified model implicit or explicit smoothing is needed. Thus Lin, Wei and Ying (2002)

have discussed a systematic family of procedures based on cumulation of residuals after arranging them in order systematic with respect to features such as explanatory variables.

A different classification distinguishes checks of the assumed form of the systematic variation from those of the error structure, the former being often but not invariably the more important.

Here are some simple examples. An omnibus test of normality could be based on the chi-squared test of grouped data, having an appreciable number of degrees of freedom. A focused test could be based on estimates of the standardized third and fourth cumulants. These could be combined essentially into a chi-squared statistic with two degrees of freedom or, probably preferably, the more significant of the two component statistics used to assess fit; an advantage of the latter is its stronger immediate diagnostic power (Cox and Hinkley, 1974, p. 72). These procedures give no check of possible internal dependencies in the data which for a number of purposes may be more important than changes in distributional shape. Of course there are a large number of other approaches to this problem.

Tests with good diagnostic features are often best derived by supplementing the initial model with one or more additional parameters and then deriving a test, for example based on the additional log likelihood achieved. Such models need not be those to be used as a new basis for interpretation if the need for model change is established. For example inclusion of a quadratic term in a linear model may often be the best base for testing linearity. On the other hand polynomials are often not the best basis for the analysis of clearly nonlinear relations; see, for example, McCullagh and Nelder (1989, sec. 14.3).

A particular instance arises in the study of relatively complicated multivariate dependencies, often based on somewhat limited data. Here it is unreasonable to assume total linearity of, say ordinary linear regressions and logistic regressions and, at the same time, proceed with arbitrary shapes of dependence including interactions. Cox and Wermuth (1993) suggested using probability plots of all possible squared and cross-product terms of explanatory variables as a guide to the detection of interaction and nonlinearity and gave further examples of effective use (Cox and Wermuth, 1996, Chapter 6). Software for implementing these procedures is available on the World Wide Web.

A check of the proportionality of hazard functions can be based on inspecting Kaplan-Meier curves obtained from separate sections of data or on some kind of generalized residual plot; note that the latter can be hard

to interpret whenever there is an appreciable proportion of censored data. A more focused approach is to include in the fitted model one or more terms representing interactions of say treatment effects with time (Cox, 1972). The purpose of such a model is, as with polynomial models used to assess linearity of regression, more to provide a simple easily implemented detection device with probability properties known under the initial model than to serve as a replacement model in its own right.

In regarding model criticism as in some way linked to significance testing it may be objected that we are testing a hypothesis that we virtually know must be false. One answer is that until a departure at least approaching some interesting level of significance has been achieved, it is unclear in what direction to amend the original model. In that sense the initial model is a dividing hypothesis (Cox, 1977).

4. Influence and breakdown

A substantial literature has developed around the notion of the influence on estimates of single observations or small groups of observations and on the related notion of the proportion of observations that may be massively defective without undue effect on an estimate. This raises somewhat different issues from the ones discussed in the present paper, essentially of data criticism rather than of model criticism. It may indeed be very helpful to know that the conclusions from some data depend critically on a fairly small proportion of the data. This is an aspect of the broader issue of the transparency of methods of analysis, i.e. of the ability to follow pathways from the raw data to the conclusion. Nevertheless that is not a basis for model criticism; to have a few very informative observations and a lot of rather uninformative observations is not a good situation but is not in general on its own a basis for rejecting the informative values!

For further discussion, see Atkinson (1985) and Cook and Weisberg (1999).

5. Role in applications

The discussion so far has concentrated on the analysis of in effect a single set of data. In many contexts, however, there are several or indeed many sets of data expected to be broadly similar. A formal or informal formulation via a model in which different sets of data have key parameters governed by a probability distribution is an old idea which modern computational developments make increasingly popular and appropriate especially when

the number of sets of data is appreciable. (Such formulations are occasionally termed Bayesian, but this usage is misleading unless either the variance components involved are given a hyper-prior or interest focuses on parameter values in particular sets of data.) It is important in such analyses that unless there are compelling reasons otherwise models of the same form and with the same parameter space are used for all sets of data; for example, the practice of fitting straight lines to some sets and nonlinear functions to others on the basis either of some formal procedure or from inspection of residuals may be very misleading.

The most suitable checking procedures depend on the amount and complexity of the data involved. Often the fitting of one or more expanded models, preferably supplemented by some graphical representation, will be the simplest and most effective route. It seems in principle, however, that especially in complex problems the possibility of failure of the initial model of some unanticipated kind must be considered and for this graphical analysis of, for example, residuals will often be the best approach. Indeed any account of statistical theory must be seriously defective if in principle it offers no route for criticism of the whole formulation used for analysis. In terms of the practice of statistical analysis the issue raises big challenges, especially with large sets of data. Overzealousness in the search for discrepancies with a model is virtually certain to find something apparently unusual, but search only within an initially specified narrow range of possibilities may overlook crucial mistakes of formulation. Of course within fields where new data can be obtained relatively quickly, independent check of unexpected features with new data will be both desirable and feasible.

Reference

Atkinson, A.C. (1985). *Plots, transformations and regression.* Oxford University Press.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. R. Statist. Soc.* A **143**, 383-430.

Cook, R.D. and Weisberg, S. (1999). *Applied regression and analysis including computing and graphics.* New York: Wiley.

Cox, D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187-220.

Cox, D.R. (1977). The role of significance tests (with discussion). *Scand. J. Statist. 4*, 49-70.

Cox, D.R. and Hinkley, D. V. (1974). *Theoretical statistics.* London: Chapman and Hall.

Cox, D.R. and Snell, E.J. (1968). A general definition of residuals (with discussion). *J. R. Statist. Soc.* B **30**, 248-275.

Cox, D.R. and Snell, E.J. (1971). On test statistics calculated from residuals. *J. R. Statist. Soc.* B **58**, 589-594.

Cox, D.R. and Wermuth, N. (1994). Tests of linearity, multivariate normality and the adequacy of linear scores. *Appl. Statist.* **43**, 347-355.

Cox, D.R. and Wermuth, N. (1996). *Multivariate dependencies.* London: Chapman and Hall.

Firth, D. and Kuha, J. (2004). On the index of dissimilarity for lack of fit in log linear models. Under review.

Fisher, R.A. (1950). The significance of deviations from expectation in the Poisson series. *Biometrics* **6**, 17-24.

Lin, D.Y., Wei, L.J. and Ying, Z. (2002). Model-checking procedures based on cumulative residuals. *Biometrics* **58**, 1-12.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models.* 2nd ed. London: Chapman and Hall.

Pearson, K. (1901). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Phil. Mag.* Series 5 **50**, 157-175.

This page intentionally left blank

LIKELIHOOD PERSPECTIVES IN THE CONSENSUS AND CONTROVERSIES OF STATISTICAL MODELLING AND INFERENCE

Yudi Pawitan
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
Email: yudi.pawitan@meb.ki.se

Likelihood ideas have driven most of the progress in statistical modelling in the last 3 decades, in particular in the area of generalized linear models and survival analysis. Various extensions of the likelihood concept, to deal with observation- and parameter-driven complexities in modelling, are reviewed. There is now an extremely rich modelling framework capable for dealing with complex datasets. We can say with reasonable confidence that we have a general consensus about the utility of the likelihood function for modelling. Unfortunately, the same cannot be said for statistical inference, where consensus is lacking even for something as simple as interpretation of the confidence intervals or computation of 2-sided P-values, and a direct use of likelihood is still controversial. I will review the consensus and use this opportunity to indicate where the likelihood perspective, in particular the Fisherian idea of 'ladder of uncertainty', throws some lights on the controversies. The key idea is that uncertainty can be expressed by both likelihood and probability, where likelihood is a weaker measure of uncertainty and only probability allows objective verification in terms of long-term frequencies.

1. Introduction

"The problems of theoretical statistics," wrote Fisher in 1921, "fall into two main classes: a) To discover what quantities are required for the adequate description of a population, which involves the discussion of mathematical expressions by which frequency distributions may be represented; b) To determine how much information, and of what kind, respecting these population-values is afforded by a random sample, or a series of random samples." It is clear that these two points refer to statistical modelling and

inference.

In the same paper, for the first time, Fisher coined the term 'likelihood' and contrasted it with 'probability', two "radically distinct concepts [that] have been confused under the name of 'probability'...". It is a significant fact that likelihood is a key concept in both modelling and inference. In 1912 Fisher's first insight was to use likelihood (not explicitly called so then), which he called 'an absolute criterion', for estimation (modelling). His last insights on likelihood in his last book (edited posthumously, Fisher 1973) were mostly on the use of likelihood for inference.

The likelihood-based methodology is of course not the only approach one can take in statistics. The least-squares method and the method of moments were around in 1912. Later, other methods of estimation or inference appeared in the statistical scene, such as the method of optimal unbiased estimation, rank tests, admissible estimation, robust estimation, etc. Fisher (1912) dismissed the least-squares approach as 'obviously inapplicable to frequency curves', while the method of moments was deemed 'arbitrary'.

We will discuss in more detail later in what ways the non-likelihood based methodologies are lacking compared to the likelihood ones. For now, we need only refer to Fisher's point (a) above, where "discussion of mathematical espressions by which frequency distributions may be represented" clearly indicates data modelling. Most, if not all, non-likelihood-based methods of *estimation* are simply mathematical or numerical techniques, which are somewhat removed from the statistical modelling of the observed data.

By 'modelling' we usually mean a deeper interaction between a statistician and the data, where features of the data influence the statistician's decisions on what to do. If one is determined to use, say, the least-squares method, one will find it hard to deal with discrete or heavy-tailed outcomes. By 'modelling' a statistician first assesses and decides what 'frequency curve' is appropriate for the observed data, and the likelihood methodology automatically takes the decision into account.

Futhermore, most non-likelihood estimation methods are typically empty of inferential consequences. For example, if one gets estimates from the least-squares optimization, other principles are required to perform inference on the results. Large-sample results, usually in the form of asymptotic normality of the estimates, are the main theoretical basis for inference. In contrast to the likelihood approach, such results are usually not invariant with respect to transformation of the parameters.

The richness of likelihood-based modelling is manifest when one searches, for example, the Current Index to Statistics for 'likelihood'. The

other side of likelihood, its direct role in inference, however, is still controversial. In this paper I will review the consensus of the power of likelihood for modelling, and discuss the lack of consensus in inference.

2. Likelihood in modelling

Modelling involves fitting mathematical structures to real observations so we can perceive order or patterns more clearly; we then use the models, for example, for description or prediction. There are two aspects of modelling: 'fitting' and 'structures'. 'Fitting' requires techniques or methods, and 'structures' require models for data distribution to capture the randomness as well as models for the systematic structure, such as linear or nonlinear models.

Interesting models and methods were born in data-rich environments: astronomy data led to the method of least-squares and Laplacian calculus of probability (later known as Bayesian statistics), anthropometric data to regression and correlation methods, industrial data to the t-test, and agricultural experiments to the analysis of variance and design methodologies. Note, however, that the most common distributional model associated with those techniques was the Gaussian model. While these classical methods have served us well, newer applications of statistics, such as in the biomedical areas, have provided us with more complex problems. It is for dealing with these problems that the likelihood becomes indispensable.

Complexity in modelling comes from two sources: (a) observation-driven and (b) parameter-driven. In observation-driven complexity, the outcome of interest might be non-Gaussian (discrete or heavy-tailed), censored, missing, irregularly observed, or measured from an unbalanced design. Discrete data make model identification and model checking via residual analysis harder compared to continuous data. Heavy-tailed data can simply throw off many classical Gaussian-based procedures, and produce misleading answers. Analysis of censored or missing data requires specialized assumptions and techniques. Missing/unbalanced data may complicate an otherwise simple orthogonal structure from designed experiments.

Parameter-driven complexity arises from a large number of parameters needed to represent the data distribution. These types of problems includes, for example, the nonparametric function estimation (smoothing or inverse problems), semi-parametric models, repeated/dependent observations and mixed models.

Dependent outcomes, even Gaussian ones, such as time series or spa-

tial data, have a complexity that is somewhat on the border between the two types. Complexity arises if we are willing (or forced) to model the dependence structure. For an analysis of a stationary time series, if the dependence structure is allowed to be free or 'nonparametric', then the complexity is parameter-driven. But if we impose a parametric structure, such as an autoregressive integrated moving average (ARIMA) model, then the number of parameters is low, but the analysis is still more complicated than the analysis of non-time-series data. In some cases it is possible to 'ignore' the dependence structure, and thus simplify the analysis, and still come up with a valid analysis (e.g., the generalized estimating equation technique; see Liang and Zeger, 1986, or Diggle et al, 2002).

The development of generalized linear models (GLM) (Nelder and Wedderburn, 1972), and Cox's proportional hazard regression (Cox, 1972) addressed much of the complexity for modelling non-Gaussian and censored data. These likelihood-based methods are practically unchallenged as the standard methodology for such data. No general-purpose technique is yet recognized for missing data problems, but the likelihood approach is almost the only viable one.

As we would expect, the slowest to develop methodologically are the methods to deal simultaneously with observation-driven and parameter-driven complexities. In these problems we might have clustered binary outcomes, with a large number of clusters, and we are interested in the estimates of cluster effects and, possibly, other fixed predictors. Or, we might have a time series of count data, and be interested in forecasting or concise descriptions of the dependencies. Specialized techniques may be developed for particular problems, but the likelihood-based generalized linear mixed models (GLMM) of Lee and Nelder (1996, 2001), hierarchical GLM (HGLM) and Double hierarchical GLM (DHGLM), provide the richest modelling framework to deal with these problems. Here the likelihood concept needs to be extended to deal with random parameters. It seems clear to me that current and future work in likelihood-based general-purpose models will be directed in this area.

3. Extensions of likelihood

The coverage of the likelihood approach is greatly extended by various modifications, or in some cases, compromises, to deal with data too complex for an exact likelihood. There is a bewildering array of likelihoods of various kinds, which seems intimidating to those new to likelihood methodology.

Similar to factors that drive the problem complexity in the previous section, it is instructive to categorize likelihood extensions as observation-driven or parameter-driven.

3.1. *Profile, marginal, conditional, modified profile likelihoods*

The first type of parameter-driven extension occurs as modifications to deal with nuisance parameters. Much theoretical effort is directed to this problem, and the benefits can be applied to both modelling and inference. The main objective is to remove the nuisance parameters and arrive at a 'sensible' likelihood of some parameters of interest. Among these extensions we find the idea of *profile likelihood*, *marginal likelihood*, *conditional likelihood*, and various *modified* or *adjusted profile likelihoods*. One might also mention the *pseudo-likelihood* and *estimated likelihood*, although these terms are much less standardized.

Suppose that (θ, η) are the parameters neeeded to specify the distribution of the observed data x, where θ is the parameter of interest and η is the nuisance parameter. Suppose that $L(\theta, \eta)$ is the joint likelihood of (θ, η). Then the profile likelihood is

$$L_p(\theta) = \sup_{\eta} L(\theta, \eta)$$

where the supremum is taken over η at fixed θ. As a contrast, one might quickly mention the estimated likelihood or pseudo-likelihood (Gong and Samaniego, 1981)

$$L_e(\theta) = L(\theta, \widehat{\eta}),$$

where $\widehat{\eta}$ is some estimate of η from the data. A special theory must be developed to account for the variability due to estimation of η.

In many applications, the profile likelihood can be interpreted directly as the usual likelihood, although it is not a 'true likelihood' since in general it does not correspond to a probability of an observed event. It known that the profile likelihood leads to biased estimates and over-precision; the amount of bias is partly determined by the number of nuisance parameters. The problem becomes serious when the number of nuisance parameters is of the same order of magnitude as the sample size.

The marginal and conditional likelihoods alleviate the problems associated with profile likelihood. Suppose that we have some statistics $u = u(x)$ and $v = v(x)$ such that x has a one-to-one map with (u, v). If the marginal

distribution of u depends only on θ,

$$p_{\theta,\eta}(u,v) = p_\theta(u)p_{\theta,\eta}(v|u)$$

and then we have the marginal likelihood $L(\theta) = p_\theta(u)$. If the conditional distribution of u given v depends only on θ, then the conditional likelihood is $L(\theta) = p_\theta(u|v)$. These likelihoods are 'real' likelihoods, since they are based on probability of some observed events, so they obey the usual properties of likelihood. Some loss of efficiency might occur from ignoring the information contained in v, but in applications of these likelihoods we usually add an extra argument that such loss is not substantial.

Even when available, for example in the exponential family models, exact conditional or marginal likelihoods, may be difficult to derive. The modified or adjusted profile likelihood can be thought of as an approximate marginal or conditional likelihood. Suppose $\widehat{\eta}_\theta$ is the maximum likelihood estimate of η for fixed θ, and $I(\widehat{\eta}_\theta)$ is the associated observed Fisher information. Then the modified profile likelihood (Barndorff-Nielsen, 1983) is

$$\log L_m(\theta) = \log L_p(\theta) - \frac{1}{2}\log|I(\widehat{\eta}_\theta)| + \log\left|\frac{\partial\widehat{\eta}}{\partial\widehat{\eta}_\theta}\right|, \tag{1}$$

see Pawitan (2001, Section 10.6) for an elementary heuristic derivation. The last term on the right hand side makes this version of modified profile likelihood invariant with respect to transformation of the parameters, but it is usually hard to derive; see Barndorff-Nielsen and Cox (1994) for estimation of the last term.

Suppose that the parameters of interest and the nuisance parameters are almost orthogonal such that $\partial\widehat{\eta}/\partial\widehat{\eta}_\theta \approx 1$, or that $\widehat{\eta}_\theta$ slowly varying over θ. Hence, we get Cox and Reid's (1987) adjusted profile likelihood

$$\log L_a(\theta) = \log L_p(\theta) - \frac{1}{2}\log|I(\widehat{\eta}_\theta)|. \tag{2}$$

The modified and adjusted profile likelihood formulas apply even to problems where $(\widehat{\theta},\widehat{\eta})$ is not sufficient. If there is an ancillary statistic $a(x)$, these modifications of profile likelihoods are approximate conditional likelihoods given the ancillary statistic.

3.2. *Partial and empirical likelihoods*

Partial and empirical likelihoods are also parameter-driven extensions, to deal with infinite-dimensional nuisance parameters. Cox's partial likelihood was introduced in 1972 and soon became the key analytical tool for survival

analysis. It can be derived in many ways, but one that is theoretically illuminating is its derivation as the profile likelihood for the proportional hazard model, by profiling out the unknown nonparametric baseline hazard (e.g., Whitehead, 1980). It is a remarkable result that profiling out the infinite-dimensional nuisance parameter still produces a sensible likelihood.

The empirical likelihood is also a remarkable and relatively recent discovery (Owen, 1988). In the simplest case, given an iid sample $(x_1, \ldots, x_n)$ from a distribution with mean θ, the empirical likelihood of θ is defined as the profile likelihood obtained by profiling out the space of all distributions with mean θ. Specifically,

$$L(\theta) = \sup_{F_\theta} \prod_{i=1}^{n} p_\theta(x_i)$$

where the supremum is taken over all distributions supported on $x_1, \ldots, x_n$, such that the mean $\int x dF_\theta(x) = \theta$. Empirical likelihood for more complex models are covered in Owen (1990, 1991) and Kolaczyk (1994), and in recent years it has been extended to cover a great number of applications (Owen, 2002).

There is a close connection between the empirical likelihood and the nonparametric bootstrap (Davison et al, 1992; Pawitan 2000). In some sense the empirical likelihood is the 'proper likelihood' associated with the bootstrap . Some deep theories have been developed to show the advantage of empirical likelihood over the bootstrap (DiCiccio et al, 1993).

3.3. *Quasi- and pseudo-likelihoods and estimating equations*

Observation-driven extensions of the likelihood are usually motivated by the need to deal with complex data types, for example, dependent non-Gaussian outcomes including repeated measures, time series or spatial data. Among these extensions we find the *quasi-likelihood*, *extended quasi-likelihood*, *pseudo-likelihood* and so-called *Gaussian estimation* in time series. It is useful to mention here the closely related estimating equation approach.

The quasi-likelihood (Wedderburn, 1974) is associated with the adoption of the exponential family likelihood regardless of the true distribution of the data. The approach is suited for estimation of model parameters for many non-Gaussian outcomes beyond the usual binomial or Poisson models. Wedderburn considered the quasi-likelihood approach as an extension

of Nelder and Wedderburn's (1972) GLM class of models, since the quasi-likelihood does not have to be a real likelihood. He cited Fisher (1949), who analysed *continuous* data, where the variance was proportional to the mean, as if the data were Poisson.

Suppose we have independent outcomes $y_1, \ldots, y_n$ with mean $Ey_i = \mu_i$ and variance $v(\mu_i)$, where μ_i is a function of unknown regression parameters $(\beta_1, \ldots, \beta_p)$. Wedderburn's original definition of quasi-likelihood (strictly it is quasi-log-likelihood) is a function $K(y_i, \mu_i)$ satisfying

$$\frac{\partial K(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{v(\mu_i)}$$

Wedderburn showed that the quasi-likelihood is a true log-likelihood if and only if the outcome y_i comes from the exponential family model with density $\exp\{\theta_i y_i - A(\theta_i) + c(y_i)\}$. Furthermore, as far as first-order inference is concerned (e.g., up to asymptotic normality of the regression estimates), the quasi-likelihood implied by a mean-variance relationship behaves largely like a true likelihood, provided an exponential family exists for such a relationship. To be clear, such a result assumes that the mean-variance relationship is correctly specified. McCullagh (1983) reviewed the large-sample theory of quasi-likelihood; in one key difference, the quasi-likelihood ratio statistic has a $n^{-1/2}$-rate of convergence (to the χ^2-distribution), while the true likelihood ratio statistic has a faster n^{-1}-rate of convergence.

With little effort, Wedderburn's one-parameter exponential family can be extended to the exponential dispersion family (Jørgensen, 1987), where the likelihood contribution of the outcome y_i is

$$\log L(\mu_i, \phi; y_i) = \{y_i\theta_i - A(\theta_i)\}/\phi + c(y_i, \phi), \tag{3}$$

with a known function $A()$, but unknown dispersion ϕ. In particular we can now fit binomial or Poisson data with overdispersion. The function $c(y_i, \phi)$ is implicitly determined, since the density must integrate to one, and is only available in a few special cases such as the normal and gamma models.

The unknown ϕ and implicitly-defined $c(y_i, \phi)$ do not affect the estimation of the regression parameters $(\beta_1, \ldots, \beta_p)$. However, a likelihood-based estimation of ϕ needs an explicit $c(y_i, \phi)$. Nelder and Pregibon (1987) defined an extended quasi-likelihood, where the contribution of y_i is

$$\log L_i = -\frac{1}{2}\log(\phi v(y_i)) - \frac{1}{2\phi}D(y_i, \widehat{\mu}_i)$$

where $D(y_i, \widehat{\mu}_i)$ is the so-called deviance function defined by

$$D(y_i, \mu_i) = 2\log\frac{L(y_i, \phi = 1; y_i)}{L(\mu_i, \phi = 1; y_i)}.$$

The extended quasi-likelihood allows likelihood-based modelling of the dispersion parameter using the deviance as 'data'.

Even though quasi-likelihood was originally defined for regression parameters, it is clear that the idea of using an assumed likelihood is useful more generally in situations where natural probability models may not be available. First-order inference, based on asymptotic normality of the estimates, can be developed quite easily. There is a general theory for inference based on an assumed or 'possibly wrong' working model (e.g., Serfling, 1980, Chapter 7), where the asymptotic distribution of the estimates depends only on the first and second derivatives of the assumed log-likelihood. This is, for example, the motivation behind the Gaussian estimation or Whittle likelihood in time series analysis, where the Gaussian likelihood is applied regardless of whether the observed series is Gaussian or not. It is a useful approach since the scope for non-Gaussian time series models is limited.

By starting at the score equation of the quasi-likelihood one gets the estimating-equation or M-estimation approach. The advantage is that it is possible to use an estimating equation which is not a derivative of a log-likelihood. An important example is the generalized estimating equation (GEE) method to deal with non-Gaussian repeated measures (e.g., Diggle et al, 2002). One key modelling strategy in GEE, that is different from the original quasi-likelihood approach, is the emphasis that the mean-variance relationship does not have to be correct. A robust variance formula is used to get a valid inference.

In image analysis, exact likelihoods, for example based on the Gibbs distribution, are also much too complicated to use for statistical purpose. Here, the idea of pseudo-likelihood, which is based on the product of conditional probabilities, has been shown to have reasonable properties (e.g., Besag, 1974; Strauss, 1982). Note the term 'pseudo' here, pointing to the need to regularize the names we assign to different likelihoods.

Inference from the estimating-equation approach typically relies on Wald-type statistics, i.e., asymptotic normality of the estimates. For example, confidence intervals would be of the form $\widehat{\theta} \pm 1.96\text{se}(\widehat{\theta})$. Better inference can be achieved, for example, via the bootstrap. Extension to likelihood-based inference would have to be based on the empirical likelihood idea, although this is not routine practice. In principle, we can construct the empirical likelihood from the bootstrap (e.g., Davison and Hinkley, 1988; Pawitan, 2000).

3.4. *Predictive and hierarchical likelihoods*

The last extensions of likelihood that we consider here are motivated by the desire to use the likelihood-based method to deal with unobserved random variables, such as those in prediction or random-effects estimation problems. Suppose that y is the observed number of successes from a binomial experiment with n trials and unknown probability θ. What is wanted is a prediction for z, the number of successes for the next m independent trials. The classical definition of likelihood does not work: treating the unknown z as a realized and fixed value, we find that the conditional distribution of y given z is free of z (because of independence). It is interesting that such a problem does not occur with the Bayesian approach, where we would simply find the conditional density $p(z|y)$, by integrating out the unknown parameter θ.

The effort to define *predictive likelihood* began with Lauritzen (1974) and Hinkley (1979). Suppose (y, z) has a joint density $p_\theta(y, z)$, and $R(y, z)$ is a sufficient statistic for θ, so that the conditional distribution of (y, z) given $R = r$ is free of θ. One can then derive the conditional distribution of of z given y, which also free of θ; thus the 'predictive likelihood' of z is

$$L(z) = \frac{p_\theta(y, z)}{p_\theta(r(y, z))}$$

The need to have a sufficient statistic, for getting rid of the nuisance parameter θ, restricts the application of this predictive likelihood to more general problems.

Butler (1986) and, later, Bjørnstad (1990, 1996) developed and investigated a more general concept of joint likelihood of the unknown parameters (θ, z) defined by

$$L(\theta, z) = p_\theta(y, z). \tag{4}$$

This definition is clearly an extension of Fisher's original definition of likelihood for fixed parameters, so it deserves to be called 'extended likelihood'. As in the case of classical likelihood for fixed parameters, the problem of the nuisance parameter θ leads to many different versions of predictive likelihood. By far the simplest is the profile likelihood

$$L_p(z) = \sup_\theta L(\theta, z) = L(\widehat{\theta}_z, z),$$

where $\widehat{\theta}_z$ is the MLE of θ assuming that z is observed. The same modifications of profile likelihood as discussed in Section 3.1 apply to $L(z)$. For

example, one might use the adjusted profile likelihood

$$\log L_a(z) = \log L_p(z) - \frac{1}{2}\log|I(\widehat{\theta}_z)|,$$

where $I(\widehat{\theta}_z)$ is the observed Fisher information in the estimation of θ for fixed z. As for the fixed-effects estimation, one might consider the adjusted profile likelihood as an approximate conditional-distribution-based predictive likelihood as defined by Hinkley (1979).

What is the 'proper' likelihood for a random parameter? Bayarri et al (1987) concluded that no general definition could be found. Bjørnstad (1996) extended the definition of sufficiency, conditionality and likelihood principles to include the random parameters, and proved Birnbaum's (1962) theorem that sufficiency plus conditionality principles are equivalent to the likelihood principle, assuming that we use the definition (4). This implies that the extended likelihood (4) contains all the information about (θ, z) in the data y. As we shall see, however, this result does not 'solve' our problem yet, since the likelihood principle does not tell us how to deal with inference with individual parameters θ and z, but it least it does confirm that one must start with (4).

Example 1: Suppose that we have a one-way random effects model

$$y_{ij} = \mu + b_i + e_{ij}, \quad i = 1, \ldots, n; \; j = 1, \ldots, m$$

where conditional on b_i, y_{ij} is $N(\mu + b_i, 1)$, and the $b_1, \ldots, b_n$ are iid $N(0, 1)$. Using (4), with μ and b_i's as the unknown parameters, one gets

$$\log L(\mu, b_1, \ldots, b_n) = -\frac{1}{2}\sum_{ij}(y_{ij} - \mu - b_i)^2 - \frac{1}{2}\sum_i b_i^2.$$

By jointly maximizing this likelihood one gets the 'maximum likelihood estimate' (MLE) $\widehat{\mu} = \bar{y}$. But suppose we reparameterize the random effects, by assuming that $b_i = \log a_i$. Allowing the Jacobian, we now get

$$\log L(\mu, a_1, \ldots, a_n) = -\frac{1}{2}\sum_{ij}(y_{ij} - \mu - \log a_i)^2 - \frac{1}{2}\sum_i(\log a_i)^2 - \sum_i \log a_i,$$

and, by maximizing, obtain $\widehat{\mu} = \bar{y} + 1$. Hence the estimate of the fixed-effects is not invariant with respect to reparameterization of the random effects.

It is worth repeating the usual caveat that the likelihood principle does not tell us 'what to do', for example how to estimate individual component parameters or how to deal with parameter transformation. The invariance of MLE in classical likelihood theory is not a consequence of the likelihood principle or any frequentist consideration such as the repeated sampling

principle; the invariance of MLE is implied by the axiom of invariance of likelihood ratio (e.g., Pawitan, 2001, Chapter 2). The fact that we do not have invariance of MLE with respect to transformation of the random parameters does not invalidate the extended likelihood (4). It does mean, however, that additional principles are needed to deal with the component parameters and parameter transformation.

It is obvious that the proper likelihood for the fixed parameter θ should be based on the marginal density $p_\theta(y)$, so

$$L(\theta) = \int_z L(\theta, z)dz.$$

Inference based on $L(\theta)$ is invariant with respect to arbitrary transformation of z. The profile likelihood $L_p(z)$ for the random parameter satisfies the invariance with respect to transformation of the fixed parameter. Furthermore, the estimate of z derived as a maximizer of $L(\theta, z)$ coincides with the maximizer of the profile likelihood $L_p(z)$. Since in general $Eg(Z) \neq g(EZ)$, except if $g()$ is linear, it seems sensible to require only that the random parameter estimate is invariant with respect to linear transformation of the random parameter itself.

It is interesting that Bjørnstad (1990) only listed 'time series and forecasting' as the major area of application of predictive likelihood. It seems clear now that random-effects estimation in generalized linear mixed models constitutes a substantial extension. It is in this setting that Lee and Nelder (1996) defined the *hierarchical (h-) likelihood* specifically for extending GLM class of models to include random effects.

Suppose that, conditional on random effects b, the outcome y_i follows the exponential family model (3) with mean μ_i. Let μ be the vector of μ_i's, and assume that

$$h(\mu) = X\beta + Zb,$$

where $h()$ is a known link function applied element-by-element, X and Z are design matrices, β is a fixed parameter and b is random with density $p_\theta(b)$. The fixed parameter θ contains the dispersion parameters. The h-likelihood is the extended likelihood

$$L(\beta, \theta, b) = p_{\beta,\theta}(y, b),$$

where the random effects b *appear additively* as Zb in the linear predictor. Let us call this the additivity condition. For fixed θ, the estimates of β and b are computed as the maximizer of the h-likelihood (MHL).

Under the additivity condition, Lee and Nelder (1996) showed that in several important models (e.g., normal y – normal b, Poisson y – gamma b) the MHL estimate of β coincides with the MLE from the marginal likelihood, and more generally the distance between the MHL estimate and MLE are asymptotically of order $O_p(n^{-1})$. Lee and Nelder (2001) showed that the h-likelihood is the only extended likelihood that gives invariant MHL with respect to linear transformation of the random effects b. This means that the additivity condition provides a natural scale for the random effects, leading to convenient estimation of the fixed regression parameter β.

The use of extended likelihood and the associated likelihood principle borders on inference, so it is not surprising that controversies should occur. The likelihood principle does not tell us how to get parameter estimates and how to deal with invariance. We know that jointly maximizing the extended likelihood with respect to all unknown parameters will lead to contradictions or lack of invariance, and no one ever suggests that we should do that. The h-likelihood suggests a *particular scale* of the random parameters that allows joint estimation of the fixed and random-effects *mean parameters*. It does not allow joint estimation of the dispersion parameters; extra principles, such using the adjusted profile likelihood, are needed for this purpose. The price for this restriction is that one is not free to transform the scale arbitrarily. As with the original definition of quasi-likelihood, the value of the particular definition of likelihood is judged by the class of models that it covers.

4. Various comments

4.1. *A case for likelihood-based inference*

Many students leave a theory of statistics course with the wrong impression that Wald-statistic-based inference is equivalent to likelihood-based inference. Confidence intervals are based on the asymptotic distribution of the likelihood ratio statistic (Wilk statistic) or of the maximum likelihood estimate (Wald statistic). For example, in the scalar case, the likelihood-based 95% confidence interval is

$$\{\theta, 2\log\frac{L(\widehat{\theta})}{L(\theta)} \leq 3.84\}$$

and the Wald interval is

$$\widehat{\theta} \pm 1.96\ \mathrm{se}(\widehat{\theta}).$$

While both are based on asymptotic theory of the same first order, the likelihood-based interval is better in the following sense.

Suppose that

$$\frac{\widehat{\theta}-\theta}{\mathrm{se}(\widehat{\theta})} \sim N(0,1)$$

is true, then the Wald interval is exact. Otherwise it is first-order accurate, in the sense that its true coverage is $95\% + O(1/\sqrt{n})$. In contrast, the likelihood-based interval is exact as long as *there exists* a transformation $g(\cdot)$, which we do not need to know, so that

$$\frac{g(\widehat{\theta})-g(\theta)}{\mathrm{se}(g(\widehat{\theta}))} \sim N(0,1).$$

Otherwise, it is second-order accurate, in the sense that its true coverage is $95\% + \mathrm{O}(1/\mathrm{n})$.

This means that the applicability of the likelihood interval is much wider and, consequently, it is much safer to use than the Wald interval. The main source of problems with the Wald interval is that $\widehat{\theta}$ may be far from normal, and if we want to transform it to improve the normality we need to know which transformation to use. All that is done automatically by the likelihood interval.

4.2. *Computing*

Likelihood methods are inherently computational: given data and model they can proceed quite automatically with little input from the statistician. Thus, their progress has coincided with the advent of cheap computing, where now there is little need for compromise in model complexity. Nelder and Wedderburn's GLM was immediately associated with the Gauss-Newton or iteratively-weighted-least-squares (IWLS) algorithm. Computer-intensive methods, such as the bootstrap and Gibbs sampling or Monte-Carlo Markov Chain (MCMC) simulation, have also helped in reducing or avoiding the analytical problems associated with likelihood inference in complex models.

There is an obvious computational problem in producing parameter estimates for a model. Traditionally, we rely on analytical work to get our inference; e.g., getting the Fisher information and using the Delta Method. More complex problems now also require computational methods to produce inference. The simplest approach is to estimate the standard error

using the bootstrap method (Efron and Tibshirani, 1993). A more elaborate scheme is to produce an empirical likelihood using the bootstrap. The Bayesian perspective leads to Gibbs sampling or Monte Carlo Markov Chain (MCMC) method to produce the posterior likelihood.

4.3. *Non-likelihood methods*

While the likelihood method is not the only approach for many of the problems, it is usually the one that

- can adapt/extend most naturally to different types of data, and
- specifies explicit, and usually testable, assumptions about the different elements under study.

A commonly-cited weakness of the likelihood method is that it is not robust with respect to distributional assumption. We can view this positively as an encouragement to look at the data carefully and to perform sufficient model checking; in the end we shall be rewarded as the likelihood method is theoretically optimal, and modelling data properly contributes to better understanding of the phenomena under study.

As an example, nonparametric regression estimation or smoothing can be easily done using the non-likelihood-based kernel method, but

- it is not natural for non-Gaussian data,
- the choice of smoothing kernel is arbitrary,
- a special kernel is needed on the boundaries, and
- the choice of smoothing parameter requires special techniques or additional principles.

Similar comments may apply to other general non-likelihood techniques such as the methods of unbiased estimation, rank tests, admissible estimation, robust estimation, etc. I am here talking in general terms; it is of course possible to offer specific cases for which a non-likelihood method is just what we need. From a theoretical point of view we can vouch for the likelihood method for its large-sample optimality properties, though its application does not rely on large samples.

5. Consensus

Except for experimental design, every area of statistics has had a large input of likelihood thinking. (In this context I also include the whole area of Bayesian methods as likelihood-based in the sense that likelihood is a key

component in the methodology.) There is a general consensus in modelling in the sense that

- the class of datasets that can be addressed by routine methods, using off-the-shelf software, has become very large,
- when faced with a certain type of data, we expect sensible statisticians to agree on an appropriate methodology.

Another marker of consensus is that there is now very little debate about the merits or otherwise of the Bayesian approach. With the advent of MCMC simulation methods the Bayesian approach has proven viable for solving complex modelling problems, so it can now be justified almost by the utilitarian principle alone, rather than by orthodox philosophical reasons.

I believe that consensus is reached in modelling because the models we have developed actually serve our purpose quite well. Another reason is that there is an evolutionary process where useful models get used and improved, while less useful ones get forgotten. There are also specific techniques developed to compare models with the reality they are meant to represent. In their evolution models interact and are confronted with reality, so it is no surprise that we converge to a set of powerful models and techniques.

Example 2: DuMouchel (1999) reported a data-mining exercise for unusual adverse events in a large frequency table from the Federal Drug Administration Spontaneous Reporting System. There were 1398 drugs considered and 952 types of adverse events (a total of around $1398 \times 952 = 1.3$ million cells). Given such a large number of cells we can guess that they are quite sparse, and it is quite easy to get spurious results.

Let N_{ij} the frequency of drug i and adverse effect type j. Assuming that a person contributes to a single cell only (e.g., only a major side effect is reported), then we can model N_{ij} as Poisson(μ_{ij}), where in obvious notation

$$\log \mu_{ij} = \mu + \text{drug}_i + \text{type}_j + r_{ij},$$

where r_{ij} is the interaction between drug i and side effect j. A high value of r_{ij} indicates that drug i has an unusually large side effect of type j, i.e., it is larger than what one would expect from an independence assumption. Since there is a large number of r_{ij}'s it makes sense to make a random-effects assumption that they are iid with some distribution G_θ. The random-effects assumption would automatically produce a shrinkage estimate, thus avoiding spurious estimates.

Estimation of the r_{ij}'s can proceed quite straightforwardly as a problem in GLMM. This is what I meant by the consensus: the relatively straightforward way of thinking and the availability of rich models to deal with the problem. One may still disagree with some of the details; e.g., the exact form of G_θ, but the bulk of the modelling strategy is quite settled.

Example 3: Recent advances in genomic technology allow simultaneous measurements of the expression of thousands of genes in a sample. In a breast cancer study in the Karolinska Hospital in Stockholm (Pawitan et al, 2004), we obtained the expression of 5549 genes from each sample/patient, and in all there were 62 patients. The objective in the study is to use the gene expression values as prognostic indicators. For each patient we ascertained the age, size of tumours, and cancer relapse or death were considered as the endpoint during followup. We fitted a Cox proportional-hazards regression model

$$\log \lambda_i(t) = \log \lambda_0(t) + x_i'\beta + z_i'b \tag{5}$$

where $\lambda_0(t)$ is the baseline hazard function, β is a fixed regression parameter and b is a random-effects parameter for the gene expression; z_i is the vector of gene expression data and x_i is a covariate vector of other information that will be used to provide an adjustment of the gene expression data. While the large number of variables involved is non-standard, the analysis is relatively straightforward via a re-expression of the hazard regression in terms of Poisson regression, so the problem reduces to GLMM.

6. Controversies in Inference

We can view statistical methods and models as a firm upper structure in the statistical edifice. But how is the foundation? It appears that the sturdy upper structure of statistics has supported a frail foundation by suspension. While we have developed highly sophisticated tools we still have not agreed on how to interpret a confidence interval! (By confidence interval I mean any form of interval we form to express our uncertain or incomplete knowledge about a parameter, so it includes Bayesian intervals.)

Those not convinced that controversies still exist in Statistics should read Lindsey (1999) and the discussions in it. Those who feel that a discussion about the foundation of Statistics is not worth the time and effort should reflect on how they would feel if biologists, for example, did not care strongly whether *homo sapiens* arrived by evolution or by creation using the same principle that 'in practice it matters little.'

Example 4: When we say that the 95% 'confidence' interval for θ is 1.2 to 5.4, Bayesians would argue that the degree of uncertainty of 95% does apply to the specific interval, while frequentists argue passionately otherwise.

This is a symptom of a deeper malaise: there is a fundamental and unsolved difficulty of satisfying both the obvious psychological need to attach uncertainty to a unique event and the long-term objective meaning of probability statements. I would put this as the foremost challenge in inference.

Example 5: The *Forsooth!* Section in the October 1999 issue of the *RSS News* published the following:

> The Met Office give a 40 per cent chance of clear skies for the eclipse, but they admit a 30 per cent chance that they may be wrong. (Daily Telegraph, 5 August 1999)

This is such a 'standard' statement of uncertainty that it is surprising to see the subsequent controversy on how one should react rationally to such a statement. The issue is whether we can provide a measure of uncertainty associated with an estimated probability of a specific event. There is apparently a similar and continuing controversy in genetic counseling.

There is still no generally-accepted method to compute two-sided P-values for asymmetric sampling distributions. For example, on observing a Poisson variate $x = 4$, what is the two-sided P-value for the hypothesis that the mean $\theta_0 = 1$? The answer varies from 0.019 to 0.08 depending on what definition we use; see Dunne, Pawitan and Doody (1998). The issue is both embarrassing, as we cannot answer a client convincingly, and exasperating, as there is no indication that statisticians will eventually agree on a sensible definition.

A recent informal poll (in the Allstat email distribution list) on the 'exact' analysis of 2-by-2 tables still indicates general disagreement among statisticians. For example, should we condition on the margins? There is a related question: should we use Yates's corrected χ^2? (Apparently, not everyone is aware that Yates's correction makes the two-sided P-value from the χ^2 test very close to doubling the one-sided P-value from the hypergeometric distribution obtained by conditioning on the margins. Notwithstanding, there is still no consensus on two-sided P-values.) See Haviland (1990) for an extensive discussion.

Deeper foundational issues arise in the analyses of sequential experiments. Here P-values depend on unobserved future data, which may or may not be collected, thus leading to paradoxes where a statement of evidence depends on the experimenter's intention. Since any experiment can be imbedded in a sequential experiment, the issue has general relevance. The first example below is from Berger and Wolpert (1988) and the second arose in a consulting session.

Example 6: An experimenter finished a study (say, with a sample size of 100, to be definite) and found $z = 2.0$ for a one-sample test of the mean. This is 'of course' a significant result at the 0.05 level. But wait, what if it weren't? Suppose the experimenter would have continued to collect more data (another 100 subjects), then $z = 2.0$ is no longer a significant result! This is because it is now part of a sequential test, so the critical value needs to be adjusted to 2.18 to achieve the 0.05 level. So the intention of the experimenter becomes relevant in the statement of evidence from the experiment.

Example 7: If we use the current data to calculate how much more data we should collect, can we then use all the data for the final analysis? If yes, should we somehow 'pay a penalty' in our analysis?

7. Views of probability

The main purpose of statistical inference is to produce measures and statements of uncertainty as we proceed from the known to the unknown. It is well known that our view of probability determines how statements of uncertainty are produced and interpreted. It is here we still see, as deep as ever, the rift between the Bayesian view of probability as a subjective measure of uncertainty and the frequentist view that probability is a measure of long-term frequency.

We can recognize two kinds of uncertainty. One is stochastic, which is an uncertainty due to randomness; e.g., we do not know what the outcome of a coin toss will be. There is no controversy that probability theory is the main tool for dealing with this sort of uncertainty. The other is uncertainty due to some or total ignorance about an unknown fixed parameter. For example, while riding a train we might not know whether the restaurant is at the front or at the back of the train. Bayesians would also use the same probability theory to deal with the second uncertainty, while frequentists do not deal with it directly. Note also that the second uncertainty is associated with a specific instance.

8. Paradoxes

There are well-known paradoxes under both Bayesian and frequentist views. A paradox is an absurd conclusion drawn from a seemingly reasonable argument. The existence of a paradox is a warning that there is something incomplete, if not wrong, in our reasoning. Paradoxes have always been important in the history of mathematics. The foundation of mathematics had had several major reconstructions in response to paradoxes; for example, Russell's paradox in 1901 – about the set of all sets that are not members of themselves – created a (creative) crisis that was not fully solved until Gödel's startling theorem in 1931 about the incompleteness of arithmetic.

8.1. *Exchange paradox*

Example 8: A swami puts an unknown amount, θ dollars, in one envelope and 2θ dollars in another. He asks you to pick one envelope at random, open

it and then decide if you would exchange it with the other envelope. You pick one (randomly), open it and see the outcome $x = 100$ dollars. You reason that, suppose Y is the content of the other envelope, then Y is either 50 or 200 with 50-50 chance; if you exchange it you expect to get $(50 + 200)/2 = 125$, which is bigger than your current 100. So, you would exchange the envelope, wouldn't you?

Here is the wonderful paradox: the reasoning above holds for any value of x, which means that you actually *do not need to open the envelope* in the first place and you would still want to exchange it! The exchange paradox has been analyzed from the Bayesian perspective, e.g., Christensen and Utts (1992), who found that our intuitive reasoning above corresponds to using a uniform prior on $\log\theta$. If such a choice leads to a paradox, not knowing θ, what proper prior can we use?

First note that, if we do not exchange or always exchange, our expected winning is $3\theta/2$. It is instructive to discuss a frequentist strategy (Ross, 1994) that leads to better returns:

(1) Generate a random variate u with *any* strictly positive density $f(u)$ on the positive real axis.
(2) Take one envelope at random and observe the amount x.
(3) Compare x and u. If $x < u$ then exchange the envelope, otherwise keep it.

It is a nice exercise in a probability course to show that the expected return from such a randomized strategy is *strictly greater* than $3\theta/2$

Note, however, how the frequentist strategy can be applied in practice. To follow a frequentist reasoning, *first assume that θ is fixed.* Then the above strategy is *not* applicable to a sensible person playing a repeat game. To see this, suppose that at the first draw the player sees the amount x. Then he should ask for an exchange so he knows what θ is exactly; after that it will be absurd to have any randomized strategy.

This means that the strategy cannot be applied to one person acting sensibly; it can be applied to a group of individuals acting independently, each playing once and *there is no sharing of any information,* but at the end the winnings are divided out equally. If such a group does not actually exist, another interpretation is for a person to play once and to imagine that such a group exists only *hypothetically* and be happy with the result of his randomized procedure; now it is not clear if the hypothetical group-based winning is relevant for this single player.

If θ is random (but with an unknown distribution), then one person can play a repeat game and use the randomized strategy to improve his winning.

A Bayesian strategy can also be devised to take advantage of randomness in θ; see Christensen and Utts (1992)

8.2. *Saint Petersburg Paradox*

Bayesians stress individual decision making, while frequentists emphasize long-term properties (insisting that what is good in the long term is good individually). As richly illustrated by the celebrated Saint Petersburg paradox, these two goals are not always compatible. Probability theory grew as a method to settle fair prices in games of chance. The infinite expectation of the Saint Petersburg game, first stated in a letter by Nicolas Bernoulli in 1713, bewildered generations of probabilists until Feller solved it in 1937 (see Feller, 1968, page 251–253).

Example 9: A *single play* of the Petersburg game consists of a series of tosses of a fair coin until it turns out heads. If this occurs at the r'th throw, the player receives $x = 2^r$ pounds. The expected value of the payment is

$$\begin{aligned} EX &= \sum_{r=1}^{\infty} 2^r P(X = r) \\ &= \sum_{r=1}^{\infty} 2^r 2^{-r} = \infty. \end{aligned}$$

What is the fair entrance fee to play the game?

An informal frequentist interpretation of the infinite expected payment means that a player should be willing to pay a large amount of money in exchange of potentially large payment. Here is the paradox: we know that it only takes an average of 2 tosses to throw heads, so nobody in his right mind would want to pay anything but a small amount. It is almost certain that the game will end by the tenth toss (with probability $1-2^{-10} = 0.999$), at which point the average payment is only 10. So here we have a situation where a sensible individual will not follow a decision – e.g., pay £100 to play the game – that is guaranteed profitable by a long-term argument.

Feller's solution was ingenious. Let $x_1, \ldots, x_n$ be the payments for n independent plays of the game and $S_n = \sum_i x_i$. Feller showed that

$$\frac{S_n}{n \log_2 n} \to 1$$

in probability as n goes to infinity. This means, informally, for n large S_n becomes close to $n \log_2 n$. So, if there are n players (each playing once), then each should be willing to pay $\log_2 n$ to make it a fair game for the house,

i.e., there is a variable entrance fee depending on the number of players. For example, if there are $2^{10} = 1024$ players, each should pay £10. This is supposed to 'solve' the paradox.

Actually, the paradox is not really solved: the theory merely confirms that Saint Petersburg game is an impossible game. For starters, when the house opens for business it does not know how many players will play and how many games will be played, so it cannot set a price; an individual who does not know how many games he will play cannot even decide for himself what is a fair price. Furthermore, fundamentally, *what is fair for the house is not appealing to the individual player*, and vice versa. There is an irreconcilable duality between the individual's and the house's (or group) points of view which is impossible to satisfy simultaneously. To see this, suppose that there are 1024 players, each wanting to play 1024 games. Using Feller's result, from a player's point of view a fee of £10 per game is fair, but not for the house, since it has to deal 2^{20} games, so it has to charge £20 per game. It is important to note that each point of view is valid.

To put this parable in statistical context, frequentist long-term properties, such as the overall type-I error probability (or α level) or confidence level, are group properties in line with the house's perspective, while the Bayesian view of individual decisions is more sympathetic to the player's. The Bayesian rejection of frequentist methods is in the latter's insistence that a player should take the house's point of view in his decision making. On the other hand, in stressing the individual decision makers, the Bayesians get criticized for ignoring the house's perspective.

8.3. *Prisoner's dilemma*

This is not a paradox, but a closely related dilemma. Statisticians are not the only professionals that think carefully about the problem of rational decision. There is a large literature in political, social or moral philosophy on the so-called prisoner's dilemma , (see, for example, Campbell and Sowden, 1985), which concerns the rational basis of moral behaviour. There are many versions of this dilemma, but here is one that explains the name (Campbell, 1985).

Example 10: Imagine that you and an accomplice commit a crime and are waiting for a trial. The prosecutor offers you a deal: "there is enough evidence to convict both of you, so even if both of you remain silent you will be sentenced to one year. But, if you confess and your friend remain silent, you will go free and your friend will get 10 years. The converse is true if you remain silent, but your friend confesses. However, if both of you confess you will both get 9 years." What

should you do?

You as a 'rational' person would think:

- Either the other prisoner will confess or he will not
- If he does, it is better for you to confess
- If he does not, it is also better for you to confess
- So, it is better for you to confess.

You know, however, that the other prisoner would obviously go through the same reasoning and end up confessing, so you would both get 9 years. Had both of you remained silent you would only get one year. How can it be rational to confess?

The crux of the prisoner's dilemma is that what is rational for one person (i.e., maximizing utility) is not sensible if everyone behaves the same way. There are many situations that resemble the prisoner's dilemma, for example; peace negotiation between two warring parties, everyday trade negotiation, voting, vaccination schemes, etc.

The interesting and highly relevant fact for us statisticians is that the philosophers and political scientists clearly recognize several *distinct* problems:

- one-off game, two players
- n-games, two players
- one-off game, n players
- many games, many players

What we learn from the literature is that there is some agreement that for a one-off game

- strategy is not meaningful
- 'be rational' is not meaningful advice as far as maximizing utility is concerned
- players are doomed to frustration

The only cases where it is possible to have a strategy (e.g., tit-for-tat) are when many games are involved. Can we not agree on something similar for our inference paradoxes?

9. Ladder of uncertainty

Settling the foundation of inference does not mean we want a new set of principles that will change practical statistics. As an analogy we can look

at the early history of calculus, from the time of Newton and Leibniz in the late 1600s, where mathematicians had to rely on the idea of infinitesimal. It was a 'number next to zero,' needed in the steps to compute derivatives and integrals. Obviously there is no such number, so its presence in calculus sticks out like a sore thumb. It took more than a century for mathematics to settle with the idea of epsilon-delta limits by the mid-1800s and get rid of the infinitesimals. Here limits serve as a rigorous foundation, but do not change 'practical' calculus – in fact the infinitesimal made a late come-back in 1960s in the idea of 'hyper-reals'; see later. We are in search of something similar in inference.

The main feature of the previous and many other paradoxes in statistical inference is the fundamental duality between the individual (or unique case) and group reasoning. The controversy is in the use of probability in individual reasoning; this is precisely where we face the problem of inference. The problem is that while we can take the Bayesian and frequentist views as the extremes of a spectrum there is no articulated middle ground on offer. It seems that what we must look closely at is our view of uncertainty, and, in particular, we must accept a 'ladder of uncertainty', a Fisherian idea contained in his last book *Statistical Methods and Scientific Inference* (1973). The proposal can be summarized as follows:

- whenever possible we should base inference on probability statements, otherwise it should be based on the likelihood
- the likelihood can be interpreted subjectively as a rational degree of belief, but it is weaker than probability, since it does not allow an objective verification, and
- in large samples there is a strengthening of likelihood statements where it becomes possible to attach some probabilistic properties.

The distinguishing view is that inference is possible directly from the likelihood function; this is neither Bayesian nor frequentist, in fact both schools would reject such a view as they allow only probability-based inference. This Fisherian view, however, also differs from the so-called 'pure likelihood view' that considers the likelihood as the sole carrier of evidence in statistical inference (e.g. Royall, 1997).

To emphasize, in this proposal we recognize two 'well-defined levels of logical status' for uncertainty about unknown quantities, one supplied by probability and the other by likelihood; a likelihood-only statement is used to 'analyze, summarize and communicate statistical evidence of types too weak to supply true probability statements' (Fisher, 1973, page 75); fur-

thermore, when available, a probability statement must allow an objective verification. In 1921 Fisher already recognized likelihood and probability as 'two radically distinct concepts', but it seems only in his last book – first published in 1956 – he explicitly considered a direct use of likelihood for a weaker form of inference.

Putting this proposal into practice is not going to be simple. It would be convenient to be able to say that (i) unique-case reasoning and statements should be based on likelihood, and group-based reasoning and statements should be based on probability; and (ii) when a probability statement is provided it is in general not attached to a unique result (such as a specific confidence interval), but to a collection of results, and that in this case there is no need to satisfy the psychological sense of uncertainty. But Fisher would immediately be against that, since for him the (fiducial) probability applies to unique-case reasoning.

Other ladders in mathematics

The idea of a ladder of uncertainty can be compared with

- the ladder of infinity in number theory
- the ladder of truth in the theory of logic
- the ladder of real numbers.

Georg Cantor's investigation of the infinite led to a surprising discovery that there were layers of infinity. It is now 'standard' knowledge, for example, that there are more reals than integers. This idea solves many ancient paradoxes, such as Zeno's paradoxes.

Alfred Tarski's ladder of truth is more exotic. Statements about the objective world can be true_0 or false_0, but statements about those statements maybe true_1 or false_1. Thus "the statement that 'the snow is yellow' is false_0" is true_1. Such a hierarchy of truth gets rid of self-referential paradoxes, as in the statement 'this sentence is false', as meaningless statements.

The ladder of reals, consisting of the standard-real and the hyper-real, is an effort to revive the infinitesimal. There is now a respectable branch of mathematics called 'non-standard analysis', based on the hyper-real, which is a real number plus a 'cloud' of infinitesimals around it. It is possible using non-standard analysis to develop calculus without limiting arguments.

Actually it is a common trick of mathematics to expand the concepts when there are things that cannot be handled. For example, the complex number is introduced as there is no real solution to $x^2 = -1$, and now

it is impossible to imagine the awkwardness of mathematics without the complex numbers. Rules governing the complex numbers are not the same as those for the reals.

9.1. *Likelihood versus probability*

There are several ways in which likelihood is weaker than probability, e.g., it cannot function as weight for averaging purposes and it is not calibrated.

Example 11: For the previous exchange paradox: if we know absolutely nothing about θ, then the other envelope is either 50 or 200 with equal likelihood (not probability). It simply says there is no rational way of preferring one envelope over the other. We cannot then use the likelihood as weight for averaging, thus avoiding the paradox. Here the likelihood satisfies the psychological need to attach some uncertainty in a unique case, something denied by the frequentists.

The likelihood is not calibrated in the following sense. A probability of 0.02 has an objective meaning in repeated-sampling terms, but a likelihood ratio of 50 does not have a universal meaning; in particular, the meaning can depend on the experiment and the size of the parameter space. This must be accepted as part of the 'weakness' of likelihood as opposed to probability.

Example 12: We pick a card at random out of a deck of 52 cards and, say, we observe the ace of clubs. Then consider two hypotheses: H: it is a standard deck of cards or A: it is a deck of 52 aces of clubs. The likelihood ratio of A against H is 52. If the hypothesis A is formed after seeing the result then we feel it is spurious; but it can also be formed prior to observing a card, where it is one of a large collection of hypotheses of all possible decks of cards. Likelihood by itself cannot tell if the evidence is spurious or not.

Finally, the likelihood does not follow the rules of probability. In particular, rules regarding transformation of parameters are handled using the invariance axiom of the likelihood ratio. This implies that we are equally ignorant regardless of how we parameterize our model, and it avoids the well-known difficulty of the Bayesian methods regarding the invariance and choice of the prior distribution.

10. Settling the controversies?

I believe that the ladder of uncertainty will go some distance towards settling the controversies in statistical inference. Total settlement, however, seems unlikely at the moment, because

- a probability-based statement is not actually ruled out for unique-case inference, but at the moment there is no complete guidance as to where

it is possible. Fisher's condition on recognizable subsets may apply, i.e., a probability statement for a confidence interval is applicable if we cannot find a subset of the sample space under which condition a different probability statement is true.

- there is still a theoretical difficulty in dealing with nuisance parameters.
- it is not clear if there is a canonical way to calibrate the likelihood. At the moment the Akaike information criterion is the main tool for continuous parameter models, but its relevance is not clear if the parameter space is discrete (as in a collection of K models, where K maybe large).

In conclusion, we have discussed a Fisherian proposal on the ladder of uncertainty that occupies a truly middle-ground position between the Bayesian and frequentist extremes. It satisfies, via likelihood, the psychological need to attach a degree of uncertainty to a unique event and, via probability, the scientific requirements of objectivity. From this new perspective both Bayesian and frequentist methods achieve stronger results than likelihood, but at the price of more assumptions in their applications. A deep theoretical result was shown by Birnbaum (1962) that all evidence in an experiment is contained in the likelihood function, so reporting the likelihood should be routine. What needs to be done is to work out exactly where probability statements are possible and how routine reporting of applied results should be done.

References

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994). *Inference and asymptotics.* London: Chapman & Hall.

Bayarri, M.J., DeGroot, M.H., and Kadane, J.B. (1987). What is the likelihood function? In *Statistical Decision Theory and Related Topics IV,* Vol. 1, S.S. Gupta and J. Berger (Eds.). New York: Springer Verlag.

Berger, J.O. and Wolpert, R.L. (1988). *The likelihood principle.* Hayward: Lecture Notes - Monograph Series, Vol. 6, Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundation of statistical inference (with discussion). *Journal of the American Statistical Association,* 53, 259-326.

Bjørnstad, J.F. (1990). Predictive likelihood: a review. *Statistical Science*, **5**, 242–265.

Bjørnstad, J.F. (1996). On the generalization of the likelihood function and likelihood principle. *Journal of the American Statistical Association*, **91**, 791–806.

Butler, R.W. (1986). Predictive likelihood inference with applications (with discussion). *Journal of the Royal Statistical Society*, **48**, 1–38.

Campbell, R. and Sowden, L. (1985). *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem.* Vancouver: University of British Columbia Press.

Christensen, R. and Utts, J. (1992). Bayesian resolution of the 'exchange paradox.' *American Statistician* **46**, 274–276. Correction in the same Journal in 1994, p. 98.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, Series B, **34**, 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society*, Series B, **49**, 1–39.

Davison, A.C., Hinkley, D.V., Worton, B.J. (1992). Bootstrap likelihoods. *Biometrika* **79**, 113–30.

DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Annals of Statistics*, **19**, 1053–1061.

Diggle, P.J., Heagerty, P.J., Liang, K-Y. and Zeger, S. (2002). *Analysis of longitudinal data.* Oxford: Oxford University Press.

DuMouchel, W. (1999). Bayesian data mining in large frequency tables (with discussion). *American Statistician* **53**, 177–202.

Dunne A., Pawitan, Y. and Doody L: Two-sided P-values from discrete asymmetric distributions. *Statistician*, **45**, 397–405.

Efron, B., Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Feller, W. (1968). *An introduction of probability theory and its applications.* Volume I. New York: Wiley.

Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, **41**, 155–160.

Fisher, R.A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, **1**, 3–32.

Fisher, R.A. (1949). A biological assay of tuberculin. *Biometrics*, **5**, 300-316.

Fisher, R.A. (1973). *Statistical Methods and Scientific Inference.* 2nd Edition. Oxford: Oxford University Press.

Haviland M.G. (1990). Yates correction for continuity and the analysis of 2x2 contingency-tables. *Statistics in Medicine* **9**, 363–367.

Hinkley, D.V. (1979). Predictive likelihood. *Annals of Statistics* **7**, 718–728.

Kolaczyk, E.D. (1994). Empirical likelihood for generalized linear models. *Statistica Sinica*, **4**, 199–218.

Lauritzen, S.L. (1974). Sufficiency, prediction and extreme models. *Scandinavian Journal of Statistics*, **1**, 128–134.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society*, Series B, **58**, 619–678.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: a synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika*, **88**, 897–1006.

Liang, K-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lindsey, J.K. (1999) Some statistical heresies (with discussion). *Statistician* **48**, 1–40.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, **11**, 59–67

Nelder J. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, **A 135**, 370–384.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–49.

Owen, A.B. (1991). Empirical likelihood for linear models. *Annals of Statistics*, **19**, 1725–1747.

Owen, A.B. (2002). *Empirical likelihood.* Boca Raton: Chapman & Hall/CRC.

Pawitan, Y. (2000). Computing empirical likelihood from the bootstrap. *Statistics and Probability Letters* **47** 337–345.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood.* Oxford: Oxford University Press.

Pawitan Y, Bjohle J, Wedren S, Humphreys K, Skoog L, Huang F, Amler L, Shaw P, Hall P, Bergh J. (2004). Gene expression profiling using Cox regression. To appear in *Statistics in Medicine.*

Ross, S. (1994). Comment to Christensen, R. and Utts, J. (1992). *American Statistician*, **48**, 267.

Royall, R.M. (1997). *Statistical Evidence.* London: Chapman & Hall.

Serfling, R.J. (1980). Approximation theorems of mathematical statistics. New York: Wiley.

Strauss, D. (1992). The many faces of logistic regression. *American Statistician*, **46**, 321–327.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Whitehead, J. (1980). Fitting Cox's regression model to survival data using GLIM. *Applied Statistics*, **29**, 268–275.

PERSPECTIVES OF ANOVA, REML AND A GENERAL LINEAR MIXED MODEL

B. R. CULLIS and A. B. SMITH[a]
Wagga Wagga Agricultural Institute,
Private Bag,
Wagga Wagga, NSW, Australia
E-mail: brian.cullis@agric.nsw.gov.au

R. THOMPSON
Biomathematics,
Rothamsted Research,
Harpenden, Herts, AL5 4JQ, England
E-mail: robin.thompson@bbsrc.ac.uk

1. Introduction

Since the original paper by Patterson and Thompson[1] there has been substantial interest and development of the technique known as Residual Maximum Likelihood (REML) estimation. REML is now the method of choice for estimating variance components (or more generally variance parameters) in linear mixed models. REML was devised as a variant of the Maximum Likelihood (ML) estimation of variance components of Hartley and Rao[2] for the problem of estimating intra-block and inter-block weights in the analysis of incomplete block designs with block sizes not necessarily equal. Nelder[3] had earlier devised an efficient estimation of weights in generally balanced designs (GBDs) in which the blocks are usually, though not always, of equal size. When block sizes are equal REML estimation and Nelder's approach yield identical results for GBDs.

[a]Work partially supported by Grains Research and Development Corporation of Australia.

Simplistically, in REML, the relevant likelihood function is the likelihood for a set of "error" contrasts, rather than that for the full likelihood function. The difference is analogous to the well known difference between the two methods of estimating the variance (σ^2) of a normal distribution given a sample of size n. Both ML and REML use the same sum of squares of deviations about the sample mean, but ML equates the sum of squares to $n\sigma^2$, whereas REML equates the sum of squares to $(n-1)\sigma^2$. That is, REML accounts for the "loss in degrees of freedom" attributable to estimation of the mean (more generally the model's fixed effects). Recently, Verbyla[4] presented an illuminating derivation of REML using the fact that the Residual Likelihood can be regarded as a marginal likelihood (in the sense described, for example, by Cox and Hinkley[5](pp 16-18.)). Other justifications for REML have been given, such as a marginal posterior density, as an iterative MIVQUE or MINQUE procedure, or as an adjusted profile likelihood (Cox and Reid[6]; Lee and Nelder[7]).

The rapid expansion of interest in REML over the past 15 years has been largely a result of its availability in both commercially and public domain software such as SAS[8], S-PLUS[9], GENSTAT 5[10](pp413-503), ASReml[11] and samm[12]. These implementations have adopted more efficient computing strategies, including sparse matrix methods[13, 14] and algorithms that are not as computer intensive as the Fisher Scoring (F-S) algorithm suggested by Patterson and Thompson[1]. These algorithms include derivative free methods utilised by Smith and Graser[14], which became the basis of the DFREML package[15], first-order schemes such as the Expectation-Maximisation (EM) scheme[16] and the computationally-efficient second-order scheme known as the Average Information (AI) algorithm[17], which has become the basis for the ASReml GENSTAT 5 and samm implementations. This algorithm has the second-order convergence properties of the F-S or Newton-Raphson (N-R) algorithms but removes the burden of computing the trace terms in the expected and observed information matrices required by the F-S and N-R algorithms, respectively.

Paralleling these developments there has also been an expansion in linear mixed-effects models to what we term a general (not to be confused with generalised) linear mixed model, in which correlation may exist within and between random effects. REML is now widely used (or recommended) for the analysis of longitudinal and repeated-measures data[18, 19], spatial and geostatistics data[20, 21], animal and plant breeding applications[22, 23], non-parametric or semi-parametric regression[24] and bioinformatics data [25, 26].

In this paper we present a review of REML. Our review is centred on what we see as the four major themes that constitute the REML "story" thus far. As such, the structure of the paper is as follows. We begin in Sec. 2 with a description of the origin of REML through its links to the analysis of variance methods for designed experiments proposed in several key papers by Nelder[27, 28, 3]. In Sec. 3 we present the formulation of a general linear mixed model and discuss estimation in Sec. 4. Sec. 5 deals with computational strategies, reviewing computing algorithms, presenting several new schemes and providing a limited comparison of these algorithms using some published data-sets. In Sec. 6 we consider inference for linear mixed-effects models with particular emphasis on the approach of Kenward and Roger[29], who considered inference concerning fixed effects within the REML framework, and illustrate how their adjustments for Wald tests and standard errors can be readily implemented within the AI algorithm. The analysis of data using linear mixed models usually requires forming predictions of a linear combination of fixed and random effects as a summary to explore the relationships established in the analysis. Lane and Nelder[30] described a general approach for forming predicted values in general(ised) linear models, which was extended by Lane[31] to lessen the computational burden of obtaining standard errors of predicted values. In Sec. 7 we consider the problem of forming predictions in a general linear mixed model and present the approach of Gilmour *et al.*[32], which builds on the work of Lane and Nelder.

2. REML and the analysis of Generally Balanced Designs

2.1. *Preliminaries*

The class of generally balanced designs (GBD) was introduced by Nelder[27, 28]. This class includes many experiment designs that have been widely used in practice. In a GBD the n experimental units have an orthogonal block structure, which is defined by the sampling or randomisation employed in their selection. This block structure defines the null (i.e. in the absence of treatment effects) analysis of variance. It can be shown that this results in a variance structure for the data given by

$$\text{var}\,(\boldsymbol{y}) = \boldsymbol{V} = \sum_{i=0}^{q+1} \xi_i \boldsymbol{P}_i$$

where $\boldsymbol{y}$ is the vector of data, the $\boldsymbol{P}_i$ are mutually orthogonal idempotent $(n \times n)$ matrices summing to the identity matrix of order n, which define

the *strata* of the analysis ($\boldsymbol{P}_0$ defining the stratum that corresponds to the general mean) and the ξ_i are known as the stratum variances. We also note that the null analysis of variance follows from the decomposition of the total sum of squares, $\boldsymbol{y}'\boldsymbol{y}$, into $q+2$ components, $\boldsymbol{y}'\boldsymbol{P}_i\boldsymbol{y}$, which is termed the sum of squares for the ith stratum and the rank of $\boldsymbol{P}_i$, denoted ν_i, is the degrees of freedom associated with this stratum sum of squares.

If we consider the application of, say, p treatments, then we denote the vector of fixed effects for these treatments by $\boldsymbol{\tau}$. Following Nelder[3], we further assume that the complete set of treatment effects can be represented in terms of a linear (treatment) model of terms involving crossed, nested, or a mixture of both, factors that relate to the treatment structure. It is therefore possible to write a linear model for the vector $\boldsymbol{\tau}$ given by

$$\boldsymbol{\tau} = \boldsymbol{T}\boldsymbol{\tau} = \sum_{j=1}^{l} \boldsymbol{T}_j \boldsymbol{\tau}$$

where, again, the $\boldsymbol{T}_j$ are mutually-orthogonal, idempotent $(p \times p)$ matrices summing to the identity matrix of order p.

The model for $\boldsymbol{y}$ in terms of the first and second moments can therefore be equivalently written as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \sum_{i=1}^{q+1} \boldsymbol{Z}_i \boldsymbol{u}_i \\ &= \boldsymbol{X}\boldsymbol{\tau} + \sum_{i=1}^{q} \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{e}\,, \end{aligned}$$

say, where $\boldsymbol{Z}_{q+1} = \boldsymbol{I}_n$ and $\boldsymbol{u}_{q+1} = \boldsymbol{e}$, $\boldsymbol{X}$ is a known fixed-effects design matrix assigning treatments to experimental units, $\boldsymbol{Z}_i$ is a $(n \times b_i)$ known design matrix, and $\mathrm{E}(\boldsymbol{u}_i) = \boldsymbol{0}, \mathrm{var}(\boldsymbol{u}_i) = \sigma_i^2 \boldsymbol{I}_{b_i}$ and $\mathrm{cov}(\boldsymbol{u}_i, \boldsymbol{u}_j) = \boldsymbol{0}$. Note that we use the conventional notation so that the residual variance, $\sigma_{q+1}^2 = \sigma^2$. The vectors $\boldsymbol{u}_i$ are called random effects, hence the linear mixed-effects model. The variance parameters (σ_i^2) are called variance components. There is a unique relationship between the stratum variances (ξ_i) and the variance components that depends on the blocking structure and hence the strata. We consider two examples in Secs. 2.2 and 2.3 to illustrate this.

The full analysis of variance is best constructed via a non-singular transformation of the data vector $\boldsymbol{y}$ to $\boldsymbol{K}'\boldsymbol{y}$, where $\boldsymbol{K} = [\boldsymbol{K}_0, \boldsymbol{K}_1, \ldots, \boldsymbol{K}_{q+1}]$. The $\boldsymbol{K}_i$ are full-rank matrices of size $n \times \nu_i$ chosen so that

$$\boldsymbol{P}_i = \boldsymbol{K}_i \boldsymbol{K}_i'$$

We note that $\boldsymbol{K}_i'\boldsymbol{K}_j = \boldsymbol{0}$ and $\boldsymbol{K}_i'\boldsymbol{K}_i = \boldsymbol{I}_{\nu_i}$. Hence the sub-vectors of $\boldsymbol{K}'\boldsymbol{y}, \boldsymbol{K}_i'\boldsymbol{y}$ are mutually independent with means $\boldsymbol{K}_i'\boldsymbol{X}\boldsymbol{\tau}$ and variances $\xi_i\boldsymbol{I}_{\nu_i}$.

If we now consider estimation of $\boldsymbol{\tau}$ in the ith stratum, then the least squares (ML under the assumption of normality for the vectors of random effects) estimate satisfies

$$\boldsymbol{T}\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}\hat{\boldsymbol{\tau}}_{[i]} = \boldsymbol{T}\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}\,. \tag{1}$$

The matrix $\boldsymbol{T}\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}$ is the information matrix for the fixed effects in stratum i. It may not be of full rank and hence obtaining a unique solution to Eq. [1] depends on finding an appropriate generalised inverse. A GBD is defined to be one in which

$$\boldsymbol{T}\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T} = \sum_{j=1}^{l}\lambda_{ij}\boldsymbol{T}_j\,, \tag{2}$$

for all i, j.

The λ_{ij} in Eq. [2] are the effective replication factors for the jth treatment term in the ith stratum. Therefore a generalised inverse of $\boldsymbol{T}\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{X}\boldsymbol{T}$ is $\sum_{j=1}^{l}\lambda_{ij}^{-1}\boldsymbol{T}_j$ where we set $\lambda_{ij}^{-1} = 0$ if $\lambda_{ij} = 0$. When $\lambda_{ij} \neq 0$ then the estimate of $\boldsymbol{T}_j\boldsymbol{\tau}$ in stratum i is given by

$$\boldsymbol{T}_j\hat{\boldsymbol{\tau}}_{[i]} = \lambda_{ij}^{-1}\boldsymbol{T}_j\boldsymbol{X}'\boldsymbol{P}_i\boldsymbol{y}\,,$$

with variance $(\xi_i/\lambda_{ij})\boldsymbol{T}_j$. The treatment sum of squares for stratum i is $\lambda_{ij}(\boldsymbol{T}_j\hat{\boldsymbol{\tau}}_{[i]})'(\boldsymbol{T}_j\hat{\boldsymbol{\tau}}_{[i]})$ with degrees of freedom equal to the rank of $\boldsymbol{T}_j$.

If we consider the λ_{ij} as elements of a $(q+2)\times l$ matrix, say $\boldsymbol{\Lambda}$, termed the effective replication matrix after Nelder[3], with rows corresponding to strata and columns corresponding to treatment terms, then orthogonal designs have the property that each column contains only one non-zero element, i.e. all the information for each treatment term is contained in only one stratum. For these designs it is then simple to complete the full analysis of variance by subdividing the sum of squares in each stratum into a (total) treatment sum of squares and an error sum of squares (obtained by difference from the total sum of squares for that stratum).

When there is more than one non-zero element in any column j, of $\boldsymbol{\Lambda}$ then there exist independent estimates of the treatment effects, $\boldsymbol{T}_j\boldsymbol{\tau}$, with variances $(\xi_i/\lambda_{ij})\boldsymbol{T}_j$. The λ_{ij} are known and so the problem of combining information is one of estimation of the ξ_i.

Using the approach outlined in the preceding developments, we now use two simple examples to illustrate the link between the ANOVA estimates of

stratum variances and REML for designs with orthogonal block and treatment structure; we also extend this, using an idea due to Thompson[33], to demonstrate the link between REML and the approach of Yates[34] for a balanced incomplete block design.

2.2. *Split Plot Design*

Consider a split plot design split plot design with b blocks, w whole-plots in each block, s sub-plots in each whole-plot and $n = bws$. Two treatment factors, given by A and B, are assigned to the whole-plots and sub-plots respectively. There are four strata, corresponding to the mean, blocks, whole-plots within blocks and sub-plots within whole-plots, with degrees of freedom $1, b-1, b(w-1)$ and $bw(s-1)$ respectively. If we denote the three variance components for blocks, whole-plots within blocks and error by σ_1^2, σ_2^2 and σ^2 then, noting that ξ_0 is confounded with the overall mean,

$$\xi_1 = ws\sigma_1^2 + s\sigma_2^2 + \sigma^2,$$

$$\xi_2 = s\sigma_2^2 + \sigma^2,$$

$$\xi_3 = \sigma^2.$$

The treatment model may be written as

$$\boldsymbol{T}\boldsymbol{\tau} = \sum_{i=1}^{4} \boldsymbol{T}_j \boldsymbol{\tau}$$

and the matrix $\boldsymbol{\Lambda}$ is given by

$$\boldsymbol{\Lambda} = \begin{bmatrix} b & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & b & 0 & 0 \\ 0 & 0 & b & b \end{bmatrix},$$

with rows indexed by mean, block, block.wplot and block.wplot.splot and columns by mean, A, B and A.B.

The analysis of variance table can be constructed by decomposing the total sum of squares for each of the four strata, noting that there is no residual sum of squares for the mean stratum, there is no treatment sum of squares for the block stratum, hence the residual sum of squares equals the total sum of squares, and the residual sum of squares for the block.wplot and block.wplot.splot strata are given by

$$s_2 = \boldsymbol{y}'\boldsymbol{P}_2\boldsymbol{y} - s_A$$

$$= \boldsymbol{y}'\boldsymbol{P}_2^*\boldsymbol{y}$$

$$s_3 = \boldsymbol{y}'\boldsymbol{P}_3\boldsymbol{y} - s_B - s_{AB}$$
$$= \boldsymbol{y}'\boldsymbol{P}_3^*\boldsymbol{y}\,,$$

where s_A, s_B and s_{AB} are the sum of squares for A, B and A.B respectively.

The ANOVA estimates of the stratum variances are then given by $\hat{\xi}_i = s_i/\nu_i^*, i = 1,2,3$, where ν_i^* is the rank of $\boldsymbol{P}_i^*$ (note that $\boldsymbol{P}_1 = \boldsymbol{P}_1^*$). If we assume normality for the random effects, then these estimates are equivalent to the estimates that maximise the residual likelihood (ie, the likelihood of the error constrasts); for this example the residual likelihood can be shown to be equal to the likelihood of $s_i, i = 1,2,3$, the log of which is, ignoring constants, given by,

$$\begin{aligned}\ell_R &= \ell(\xi_1; s_1) + \ell(\xi_2; s_2) + \ell(\xi_3; s_3)\\ &= -\tfrac{1}{2}\left[(b-1)\log\xi_1 + s_1/\xi_1 + (b-1)(w-1)\log\xi_2 + s_2/\xi_2\right.\\ &\quad \left. + (b-1)w(s-1)\log\xi_3 + s_3/\xi_3\right]. \qquad (3)\end{aligned}$$

The log-likelihood of $\boldsymbol{y}$, after replacing the treatment effects by their least squares estimates, is

$$\begin{aligned}\ell = -\tfrac{1}{2}&\left[\log\xi_0 + (b-1)\log\xi_1 + s_1/\xi_1 + b(w-1)\log\xi_2 + s_2/\xi_2\right.\\ &\left. + bw(s-1)\log\xi_3 + s_3/\xi_3\right] \qquad (4)\end{aligned}$$

The terms that depend on the data are the same in both Eq. [3] and Eq. [4], but the coefficients of the $\log\xi_i$ differ, one being the degrees of freedom for the stratum residual sum of squares (Eq. [3]: REML) the other the total degrees of freedom for the stratum (Eq. [4]: ML).

2.3. *Balanced Incomplete Block Design*

The results presented in Sec. 2.2, linking REML estimates to ANOVA estimates of variance components for designs with orthogonal block and treatment structures, are well known but cannot be readily extended when there are departures from orthogonality. It is possible, however, for the class of GBD with equal block sizes, to consider an idea proposed by Thompson[33], which intuitively links REML estimation to the approach of Yates[34] and gives the same estimates as Nelder[3].

We consider the simple example of a balanced incomplete block design based on the design in Cochran and Cox[35] (pp 444). We assume that

there are t treatments replicated r times in b blocks in each replicate, each comprising p plots. The block structure is the same as the split-plot design and hence we consider the same decomposition. We consider the simple treatment model containing a general mean and deviations from the general mean, i.e. $\boldsymbol{T}\boldsymbol{\tau} = \boldsymbol{T}_1\boldsymbol{\tau} + \boldsymbol{T}_2\boldsymbol{\tau}$.

The matrix $\boldsymbol{\Lambda}$ is given by

$$\boldsymbol{\Lambda} = \begin{bmatrix} r & 0 \\ 0 & 0 \\ 0 & r(1-E) \\ 0 & rE \end{bmatrix},$$

where $E = t(p-1)/[p(t-1)] = 0.6$ for the example we consider, with $r = 5, p = 2, t = 6$ and $b = 3$.

Thompson[33] proposes a two-stage approach similar to the approach used for the ANOVA decomposition for orthogonal designs in which we compute the residual sum of squares for each stratum given by

$$\begin{aligned} s_1 &= \boldsymbol{y}'\boldsymbol{P}_1\boldsymbol{y} \\ s_2 &= \boldsymbol{y}'\boldsymbol{P}_2\boldsymbol{y} - \lambda_{22}(\boldsymbol{T}_2\hat{\boldsymbol{\tau}}_{[2]})'(\boldsymbol{T}_2\hat{\boldsymbol{\tau}}_{[2]}) \\ &= \boldsymbol{y}'\boldsymbol{P}_2^*\boldsymbol{y} \\ s_3 &= \boldsymbol{y}'\boldsymbol{P}_3\boldsymbol{y} - \lambda_{32}(\boldsymbol{T}_2\hat{\boldsymbol{\tau}}_{[3]})'(\boldsymbol{T}_2\hat{\boldsymbol{\tau}}_{[3]}) \\ &= \boldsymbol{y}'\boldsymbol{P}_3^*\boldsymbol{y}, \end{aligned}$$

with degrees of freedom $r-1, r(b-1)-t+1$ and $rb(p-1)-t+1$ respectively. It is clear, however, that there is information on the stratum variances that is being lost by estimation of $\boldsymbol{T}_2\boldsymbol{\tau}$ in two strata. To recover this we form the sum of squares of the difference $\boldsymbol{T}_2(\boldsymbol{\tau}_{[3]} - \boldsymbol{\tau}_{[2]})$, which is given by

$$s_t = (\hat{\boldsymbol{\tau}}_{[3]} - \hat{\boldsymbol{\tau}}_{[2]})'\boldsymbol{T}_2'\boldsymbol{T}_2(\hat{\boldsymbol{\tau}}_{[3]} - \hat{\boldsymbol{\tau}}_{[2]}),$$

which has $t-1$ degrees of freedom and expectation $\lambda_{22}^{-1}\xi_2 + \lambda_{32}^{-1}\xi_3$. It is clear that s_1, s_2, s_3 and s_t are mutually independent and that maximisation of the likelihood of s_1, s_2, s_3, s_t for $\sigma_1^2, \sigma_2^2, \sigma^2$ is equivalent to maximisation of the likelihood of $\boldsymbol{K}'\boldsymbol{y}$ for $\boldsymbol{K}$ being a full rank matrix of size $n \times (n-t)$ such that $\boldsymbol{K}'\boldsymbol{X} = \boldsymbol{0}$. The likelihood of $\boldsymbol{K}'\boldsymbol{y}$ is the residual likelihood.

Maximisation of the former likelihood can be easily achieved using a generalised linear model, after McCullagh and Nelder[36], with a Gamma distribution, identity link and weight vector $\nu_i/2$. For the example from

Cochran and Cox we have $(s_1, s_2, s_3, s_t) = (298.467, 174.833, 77.33, 32.139)$, weights equal to $(2, 2.5, 5, 2.5)$ and design matrix

$$\begin{bmatrix} 6 & 2 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0.833 \end{bmatrix}$$

yielding REML estimates $\hat{\sigma}_1^2 = 8.44, \hat{\sigma}_2^2 = 8.28$ and $\hat{\sigma}^2 = 7.38$.

Thompson[33] shows that this can be extended to the case when treatment estimates for $\boldsymbol{T}_j\boldsymbol{\tau}$ are available on $m_j + 1$ strata, by constructing differences between the first m_j estimates and the last estimate, and hence calculating an $m_j \times m_j$ matrix representing mean squares and products of differences (see Thompson[33] for details).

3. A general linear mixed model

The general linear mixed model we present extends the linear mixed-effects model used in Sec. 2. This model accommodates all of the analyses that were described in Sec. 1. The particular feature is the extended generality for variance models for the random effects, $\boldsymbol{u}$ and the residual $\boldsymbol{e}$. We denote these as G-structures and R-structures. We also consider both crossed and nested variance models and, where sensible, exploit the assumption of separability without loss of generality to reduce the computational burden.

3.1. *The Model*

If $\boldsymbol{y}$ is the $(n \times 1)$ vector of observations, the general linear mixed model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{5}$$

where $\boldsymbol{\tau}$ is the $(p \times 1)$ vector of fixed effects, $\boldsymbol{X}$ is the $(n \times p)$ design matrix (assumed to be of full rank) for fixed effects, $\boldsymbol{u}$ is the $(b \times 1)$ vector of random effects, $\boldsymbol{Z}$ is the $(n \times b)$ design matrix for random effects, and $\boldsymbol{e}$ is the $(n \times 1)$ vector of residual errors. It is assumed that

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \sigma_H^2 \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \right),$$

where $\boldsymbol{G} = \boldsymbol{G}(\gamma)$ and $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\phi)$. The variance parameters, σ_H^2 and σ^2, are included to change parameterisations from variance components

to variance component ratios. Their inclusion depends on the application and the form of the $G-$ and R-structures.

The distribution of $\boldsymbol{y}$ is then $N(\boldsymbol{X\tau}, \boldsymbol{V})$, where $\boldsymbol{V} = \sigma_H^2 \boldsymbol{H}$ and $\boldsymbol{H} = \boldsymbol{ZGZ}' + \boldsymbol{R}$.

3.2. *R- and G-structures*

To allow for the analysis of data from different experiments or from distinct sections, such as in glasshouse or microarray experiments, we allow a very general structure for the variance of the residual vector. We assume that there exists an indexing factor that delineates the sections of the data and we partition $\boldsymbol{e}$ conformably with this indexing. Thus, $\boldsymbol{e} = [\boldsymbol{e}_1', \boldsymbol{e}_2', \ldots, \boldsymbol{e}_s']'$. The variance matrices for the sections may differ, but generally we assume that the errors from different sections are independent. Thus,

$$\boldsymbol{R} = \oplus_{j=1}^{s} \boldsymbol{R}_j.$$

In the simplest case the matrix $\boldsymbol{R}$ is an identity matrix. More complex variance structures arise in many applications in which $\boldsymbol{R}_j$ may be the kronecker product of one or more component matrices. The component matrices are related to the underlying structure of the data (see for example Smith *et al.*[37].)

We assume that the vector of random effects is given by $\boldsymbol{u} = [\boldsymbol{u}_1' \ \boldsymbol{u}_2' \ \ldots \ \boldsymbol{u}_q']'$, where $\boldsymbol{u}_i$ is a $b_i \times 1$ vector. In most applications the $\boldsymbol{u}_i$ relate to separate terms in the linear mixed model and are assumed to be mutually independent. In some applications, for example random regressions, separate terms may be correlated. In either case we can assume that

$$\boldsymbol{G} = \oplus_{i=1}^{q} \boldsymbol{G}_i,$$

where separate terms are grouped according to the nature of the covariance model. Correspondingly, we have $\boldsymbol{Z} = [\boldsymbol{Z}_1 \ \boldsymbol{Z}_2 \ \ldots \ \boldsymbol{Z}_q]$. As for $\boldsymbol{R}_j$, $\boldsymbol{G}_i$ is assumed to be the Kronecker product of one or more component matrices. These matrices are usually indexed by the factors that are used in the construction of the relevant term in the linear mixed model.

3.3. *Identifiability of variance models*

The generality of the mixed model can often result in problems of identifiability of variance models. The cause of the non-identifiability can be

sometimes hard to diagnose. The problem is similar in principle to ensuring that the fixed-effects model is not over-parameterised. That is, variance models may not be identifiable as they are over-parameterised. Also, there may be insufficient data to estimate the parameters of the chosen variance model. General principles exist for combining variance models that can be implemented to avoid most problems in practice (see Gilmour *et al.*[11]).

4. Estimation

Estimation in a linear mixed model encompasses both the estimation of the variance parameters $\boldsymbol{\sigma} = \left(\sigma_H^2, \boldsymbol{\kappa}'\right)'$, where $\boldsymbol{\kappa} = \left(\boldsymbol{\gamma}', \sigma^2, \boldsymbol{\phi}'\right)'$, and estimation (or prediction) of the effects ($\boldsymbol{\tau}$ and $\boldsymbol{u}$) for given values of the variance parameters. The two processes are very closely linked and are combined in the computing algorithm used to carry out the estimation.

4.1. *REML estimation of variance parameters*

Verbyla[4] presented an illuminating derivation of the residual likelihood. He partitioned the full likelihood for the mixed model in Eq. [5] into two independent parts: one relates to the treatment (fixed-effect) contrasts and the other to the residual contrasts (i.e. $\boldsymbol{Zu}+\boldsymbol{e}$). Maximization of the former provides estimates of the fixed effects, whereas maximization of the residual likelihood provides REML estimates of the variance parameters.

Briefly, Verbyla[4] considers a non-singular matrix $\boldsymbol{L} = [\boldsymbol{L}_1\ \boldsymbol{L}_2]$, where $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ are $(n \times p)$ and $(n \times (n-p))$ matrices chosen to satisfy $\boldsymbol{L}_1'\boldsymbol{X} = \boldsymbol{I}_p$ and $\boldsymbol{L}_2'\boldsymbol{X} = \boldsymbol{0}$. The distribution of the transformed data, $\boldsymbol{L}'\boldsymbol{y} = (\boldsymbol{y}_1'\ \boldsymbol{y}_2')'$, say, is given by

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim N\left\{ \begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{0} \end{bmatrix}, \sigma_H^2 \begin{bmatrix} \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_1 & \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_2 \\ \boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_1 & \boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2 \end{bmatrix} \right\}.$$

We express the likelihood of $\boldsymbol{L}'\boldsymbol{y}$ as the product of the conditional likelihood of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ and the marginal likelihood of $\boldsymbol{y}_2$. The marginal distribution of $\boldsymbol{y}_2$ is

$$\boldsymbol{y}_2 \sim N\left(\boldsymbol{0}, \sigma_H^2\, \boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2\right),$$

and the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is

$$\boldsymbol{y}_1 | \boldsymbol{y}_2 \sim N\left(\boldsymbol{\tau} + \boldsymbol{y}_2^*, \sigma_H^2\, \left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}\right)$$

where $\boldsymbol{y}_2^* = \boldsymbol{L}_1'\boldsymbol{H}\boldsymbol{L}_2\left(\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2\right)^{-1}\boldsymbol{y}_2$. The associated log-likelihood functions (excluding constant terms) are given by

$$\begin{aligned}\ell_R &= \ell(\sigma_H^2, \boldsymbol{\kappa}; \boldsymbol{y}_2) \\ &= -\tfrac{1}{2}\left\{(n-p)\log\sigma_H^2 + \log|\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2| + \boldsymbol{y}_2'\left(\boldsymbol{L}_2'\boldsymbol{H}\boldsymbol{L}_2\right)^{-1}\boldsymbol{y}_2/\sigma_H^2\right\} \quad (6)\end{aligned}$$

and

$$\begin{aligned}\ell_1 &= \ell(\boldsymbol{\tau}, \sigma_H^2, \boldsymbol{\kappa}; \boldsymbol{y}_1|\boldsymbol{y}_2) \\ &= -\tfrac{1}{2}\left\{p\log\sigma_H^2 + \log|\left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}| \right. \\ &\qquad \left. + (\boldsymbol{y}_1 - \boldsymbol{\tau} - \boldsymbol{y}_2^*)'\left(\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}\right)(\boldsymbol{y}_1 - \boldsymbol{\tau} - \boldsymbol{y}_2^*)/\sigma_H^2\right\}.\end{aligned}$$

Clearly, the likelihood of $\boldsymbol{y}_2$ contains no information on $\boldsymbol{\tau}$ so that $\boldsymbol{\tau}$ must be estimated from the conditional distribution of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$. The MLE of $\boldsymbol{\tau}$ is

$$\hat{\boldsymbol{\tau}} = \left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y}.$$

The likelihood of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is a function of $\boldsymbol{\tau}, \sigma_H^2$ and $\boldsymbol{\kappa}$, but, since $\boldsymbol{\tau}$ and $\boldsymbol{y}_1$ are both vectors of length p, once $\boldsymbol{\tau}$ has been estimated there is no information left to estimate σ_H^2 and $\boldsymbol{\kappa}$. The variance parameters are therefore estimated using the marginal likelihood of $\boldsymbol{y}_2$, that is, the residual likelihood. This is given in Eq. [6], though a better known form is given by

$$\begin{aligned}\ell_R = -\tfrac{1}{2}\{&(n-p)\log\sigma_H^2 + \log|\boldsymbol{H}| + \log|\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{X}| - \\ &\log|\boldsymbol{X}'\boldsymbol{X}| + \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma_H^2\}, \end{aligned} \quad (7)$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}\left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}$.

The REML estimates of σ_H^2 and $\boldsymbol{\kappa}$ are obtained by solving the REML score equations:

$$\begin{aligned}U_R(\sigma_H^2) &= -\tfrac{1}{2}\left\{(n-p)/\sigma_H^2 - \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/\sigma_H^4\right\}, \\ U_R(\kappa_i) &= -\tfrac{1}{2}\left\{\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) - \boldsymbol{y}'\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}/\sigma_H^2\right\}, \end{aligned} \quad (8)$$

where $\dot{\boldsymbol{H}}_i = \partial\boldsymbol{H}/\partial\kappa_i$. Hence it follows that the REML estimate of σ_H^2, given $\boldsymbol{\kappa}$, is

$$\hat{\sigma}_H^2 = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}/(n-p)$$

4.2. *Prediction and the mixed model equations*

In this section we consider the prediction of the linear combination $\boldsymbol{c}_1'\boldsymbol{\tau}+\boldsymbol{c}_2'\boldsymbol{u}$ of fixed and random effects, where $\boldsymbol{c}_1$ is a known $(p \times 1)$ vector and $\boldsymbol{c}_2$ is a known $(b \times 1)$ vector. If σ_H^2 and $\boldsymbol{H}$ are known, implying that σ_H^2 and $\boldsymbol{\kappa}$ are known, the predictor that has the minimum mean-square error among the class of linear unbiased predictors is given by $\boldsymbol{c}_1'\hat{\boldsymbol{\tau}} + \boldsymbol{c}_2'\tilde{\boldsymbol{u}}$, where

$$\hat{\boldsymbol{\tau}} = \left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{y},$$

$$\tilde{\boldsymbol{u}} = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{P}\boldsymbol{y}.$$

A simple extension of this yields $\hat{\boldsymbol{\tau}}$ as the best linear unbiased estimator (BLUE) of $\boldsymbol{\tau}$ and $\tilde{\boldsymbol{u}}$ as the best linear unbiased predictor (BLUP) of $\boldsymbol{u}$.

Robinson[38] provides an excellent account of BLUP and many other aspects concerning the prediction of random effects. In his paper Robinson presents several derivations of BLUP, however the original derivation of BLUP presented in Henderson[39] provides an important link with the estimation of variance parameters. Henderson described the BLUP estimates as being "joint maximum likelihood estimates". Later, Henderson[40] retracted this statement and suggested that this terminology should not be used, as the function being maximised is not a likelihood. However, Henderson's derivation was based on maximising a function derived from the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}$. The log-density function for $(\boldsymbol{y}, \boldsymbol{u})$ can be written as

$$\log f_Y(\boldsymbol{y} \mid \boldsymbol{u}\ ;\ \boldsymbol{\tau}, \sigma_H^2, \sigma^2, \boldsymbol{\phi}) + \log f_U(\boldsymbol{u}\ ;\ \sigma_H^2, \boldsymbol{\gamma})$$

This is the log-joint distribution of $(\boldsymbol{y}, \boldsymbol{u})$. *It is not a log-likelihood as* $\boldsymbol{u}$ *is not observed.* The vectors of fixed and random effects ($\boldsymbol{\tau}$ and $\boldsymbol{u}$) can be "estimated" by maximising this function. Differentiation with respect to $\boldsymbol{\tau}$ and $\boldsymbol{u}$ and setting the result to zero leads to the system of equations known as the mixed model equations (MME), as proposed by Henderson[39]. This can be written in matrix-vector notation as

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}$$

A more abbreviated representation of the MMEs is

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y}, \tag{9}$$

where $\boldsymbol{W} = [\boldsymbol{X}\ \boldsymbol{Z}]$, $\boldsymbol{\beta}' = (\boldsymbol{\tau}'\ \boldsymbol{u}')$ and

$$\boldsymbol{C} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^*,$$

$$\boldsymbol{G}^* = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}^{-1} \end{bmatrix}.$$

The solution to Eq. [9] requires an estimate of $\boldsymbol{\kappa}$. In practice these are replaced by their REML estimates. The resulting solutions are termed Empirical BLUEs (E-BLUEs) and Empirical BLUPs (E-BLUPs).

5. Iterative schemes for REML estimation of variance parameters

In general, maximisation of the residual likelihood in Eq. [7] requires an iterative scheme. Patterson and Thompson[1] used a F-S algorithm, which requires calculation of the expected information matrix for the variance parameters. The associated computational burden is prohibitive for large data-sets or complex variance models. This was the motivation behind the development of the AI algorithm (Gilmour *et al.* [17]), which employs an information matrix that is an approximate average of the observed and expected information matrices. The AI algorithm has proved to be a computationally efficient algorithm for variance-parameter estimation in a wide range of applications, including variance-component models in large unbalanced data-sets (see Hofer[41] for a recent computational comparison of methods) and factor-analytic or unstructured models for multi-environment plant variety trials. In complex models such as the latter, however, second-order methods, including AI, are sensitive to the choice of starting values. The convergence sequence is not guaranteed to be monotonic, nor are the variance parameters ensured to remain within the parameter space, so poor start values may cause convergence difficulties. For this reason first-order schemes, the most popular of which is the Expectation-Maximisation (EM) algorithm (Dempster *et al.* [16]), have been widely used for REML estimation of variance parameters. The EM algorithm is stable (insensitive to choice of starting values), has a monotonic convergence sequence and parameters are guaranteed to remain within the parameter space. However, it may be very slow, in the sense of requiring a large number of iterations for convergence, and may not be applicable for some variance models. The Parameter Expanded EM (PXEM) algorithm (Liu *et al.* [42]) is a variant of the EM algorithm that was formulated in order to reduce the number of iterations. In general, this algorithm still requires more iterations than a second-order method.

In this section we review the AI, EM and PXEM algorithms. We then describe a series of hybrid algorithms that involve a combination of AI

and EM (or PXEM) updates. The aim is to exploit the desirable properties of both types of algorithm, namely the speed of the AI algorithm and the stability of the EM (or PXEM) algorithm. Finally, we investigate the performance of the various schemes using two published data-sets. Further work will use these results to suggest robust procedures for generating better initial values for variance parameters and suggest efficient schemes to keep parameters within the required parameter space. This could involve a Choleski parameterization (Lindstrom and Bates[43]) or log transformation (see Lee and Nelder[44], for example).

5.1. *The Average Information algorithm*

Gradient methods for variance parameter estimation are based on the linearisation of the score equations using the first term in a Taylor's expansion. Expanding the score equations in Eq. [8] about the value $\boldsymbol{\sigma}^{(m)}$, then equating to zero yields an updated value of

$$\boldsymbol{\sigma}^{(m+1)} = \boldsymbol{\sigma}^{(m)} + \left[\boldsymbol{I}_o{}^{(m)}\right]^{-1} \boldsymbol{U}_R(\boldsymbol{\sigma}^{(m)}),$$

where $\boldsymbol{I}_o^{(m)}$ is the observed information matrix for $\boldsymbol{\sigma}$ evaluated at $\boldsymbol{\sigma}^{(m)}$. We note that, since an algebraic form for $\boldsymbol{\sigma}_H^{2(m)}$ exists, we only need an update for $\boldsymbol{\kappa}$. This is given by

$$\boldsymbol{\kappa}^{(m+1)} = \boldsymbol{\kappa}^{(m)} + \left[\boldsymbol{I}_o{}^{(m)\kappa\kappa}\right] \boldsymbol{U}_R(\boldsymbol{\kappa}^{(m)}), \tag{10}$$

where $\boldsymbol{I}_o{}^{(m)\kappa\kappa}$ is the portion of $\left[\boldsymbol{I}_o{}^{(m)}\right]^{-1}$ relating to $\boldsymbol{\kappa}$. This scheme is known as the N-R algorithm. Closely related to this is the F-S algorithm in which the expected information matrix (denoted $\boldsymbol{I}_e$) is used instead of observed information. The AI algorithm is obtained by using the so-called average information matrix (denoted $\boldsymbol{I}_A$). The matrix $\boldsymbol{I}_A$ is a scaled residual sums of squares matrix given by

$$\boldsymbol{I}_A = \tfrac{1}{2}\boldsymbol{Q}'(\boldsymbol{P}/\sigma_H^2)\boldsymbol{Q},$$

where the columns of $\boldsymbol{Q}$ are working variables corresponding to σ_H^2 and $\boldsymbol{\kappa}$ and are given by

$$\boldsymbol{q}_{\sigma_H^2} = \boldsymbol{HPy}/\sigma_H^2,$$
$$\boldsymbol{q}_{\kappa_i} = \dot{\boldsymbol{H}}_i\boldsymbol{Py}.$$

Note that for models in which the variance structure is linear in the parameters (for example in variance component models), elements in $\boldsymbol{I}_A$ are

exact averages of the corresponding elements in $\boldsymbol{I}_o$ and $\boldsymbol{I}_e$. In other models they are approximate averages in which $\boldsymbol{y}'\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y}$ is approximated by its expectation $\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right)$ (where $\ddot{\boldsymbol{H}}_{ij} = \partial^2\boldsymbol{H}/\partial\kappa_i\partial\kappa_j$).

Thus, the $\boldsymbol{I}_A$ matrix does not involve the computationally-intensive trace terms needed for $\boldsymbol{I}_o$ and $\boldsymbol{I}_e$. Additionally, it can be computed efficiently via absorption of the mixed model coefficient matrix $\boldsymbol{C}$ on the working variable matrix. That is, by commencing with the augmented coefficient matrix

$$\begin{bmatrix} \boldsymbol{Q}'\boldsymbol{R}^{-1}\boldsymbol{Q} & \boldsymbol{Q}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Q}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Q} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Q} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1} \end{bmatrix}$$

and then sequentially absorbing the rows and columns of this matrix into the first element, $\boldsymbol{Q}'\boldsymbol{R}^{-1}\boldsymbol{Q}$. This approach forms the basis of the implementations of the AI algorithm in ASReml, GENSTAT 5 and samm.

5.2. *The Expectation-Maximisation algorithm*

The EM algorithm (Dempster *et al.* [16]) is a widely-used technique for calculating parameter estimates via maximum likelihood. It is particularly well suited to variance parameter estimation in linear mixed models since the random-effects vector $\boldsymbol{u}$ is a natural choice for the "missing" data. In this section we present a REML-EM algorithm for estimation of variance parameters in the linear mixed model. Each iteration of the algorithm comprises two steps: the expectation (E) step and the maximisation (M) step. For the E-step we evaluate the conditional expectation of the joint likelihood of the so-called complete data $(\boldsymbol{u}', \boldsymbol{y}')'$ given the observed part of the data that relates to estimation of the variance parameters, i.e. $\boldsymbol{y}_2$. Note that it is the conditioning on $\boldsymbol{y}_2$ rather than $\boldsymbol{y}$ that provides the basis of a REML-EM algorithm in contrast to the standard ML algorithm. The conditional expectation is evaluated at the current updates for σ^2_H and $\boldsymbol{\kappa}$. The M-step involves maximisation of the resultant expectation with respect to $\boldsymbol{\kappa}$. As in the AI algorithm an update for σ^2_H given $\boldsymbol{\kappa}$ can be obtained algebraically so we need only consider updates for $\boldsymbol{\kappa}$.

The joint likelihood (though we note it is not strictly a likelihood since $\boldsymbol{u}$ cannot be observed) of $(\boldsymbol{u}', \boldsymbol{y}')'$ is, ignoring constants, given by

$$\ell_c = -\tfrac{1}{2}\left\{(n+b)\log\sigma^2_H + n\log\sigma^2 + \log|\boldsymbol{\Sigma}| + \boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e}/(\sigma^2\sigma^2_H) + \log|\boldsymbol{G}| + \boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u}/\sigma^2_H\right\}.$$

The expected value of this joint likelihood, conditional on $\boldsymbol{y}_2$ and evaluated at the current iterate $\boldsymbol{\sigma}^{(m)} = \left[\sigma_H^{2(m)}, \boldsymbol{\kappa}^{(m)\prime}\right]'$, is given by

$$\ell_{ce}(\boldsymbol{\kappa})^{(m)} = \ell_{ce}(\boldsymbol{\gamma})^{(m)} + \ell_{ce}(\sigma^2, \boldsymbol{\phi})^{(m)},$$

say, where

$$\begin{aligned}
\ell_{ce}(\boldsymbol{\gamma})^{(m)} &= -\tfrac{1}{2}\mathrm{E}\left(b\log\sigma_H^2 + \log|\boldsymbol{G}| + \boldsymbol{u}'\boldsymbol{G}^{-1}\boldsymbol{u}/\sigma_H^2 \mid \boldsymbol{y}_2,\ \boldsymbol{\sigma} = \boldsymbol{\sigma}^{(m)}\right) \\
\ell_{ce}(\sigma^2, \boldsymbol{\phi})^{(m)} &= -\tfrac{1}{2}\mathrm{E}\big(n\log\sigma_H^2 + n\log\sigma^2 + \log|\boldsymbol{\Sigma}| + \\
&\qquad \boldsymbol{e}'\boldsymbol{\Sigma}^{-1}\boldsymbol{e}/(\sigma^2\sigma_H^2) \mid \boldsymbol{y}_2,\ \boldsymbol{\sigma} = \boldsymbol{\sigma}^{(m)}\Big).
\end{aligned}$$

Evaluation of these expectations requires use of the conditional distributions of $\boldsymbol{u} \mid \boldsymbol{y}_2$ and $\boldsymbol{e} \mid \boldsymbol{y}_2$ (evaluated at the current iterate). Thence we obtain

$$\begin{aligned}
\ell_{ce}(\boldsymbol{\gamma})^{(m)} = -\tfrac{1}{2}\Big\{&b\log\sigma_H^{2(m)} + \log|\boldsymbol{G}| + \mathrm{tr}\left(\boldsymbol{G}^{-1}\boldsymbol{C}^{ZZ(m)}\right) + \\
&\tilde{\boldsymbol{u}}^{(m)\prime}\boldsymbol{G}^{-1}\tilde{\boldsymbol{u}}^{(m)}/\sigma_H^{2(m)}\Big\}
\end{aligned}$$

where $\boldsymbol{C}^{ZZ}$ is the portion of the inverse of $\boldsymbol{C}$ corresponding to $\boldsymbol{u}$ and

$$\begin{aligned}
\ell_{ce}(\sigma^2, \boldsymbol{\phi})^{(m)} = -\tfrac{1}{2}\Big\{&n\log\sigma_H^{2(m)} + n\log\sigma^2 + \log|\boldsymbol{\Sigma}| + \\
&\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{C}^{-1(m)}\boldsymbol{W}'\right)/\sigma^2 + \tilde{\boldsymbol{e}}^{(m)\prime}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{e}}^{(m)}/(\sigma^2\sigma_H^{2(m)})\Big\}.
\end{aligned}$$

This is the E-step. The M-step involves maximisation of ℓ_{ce} with respect to $\boldsymbol{\kappa}$. The $(m+1)^{th}$ update for a parameter γ_{ij} associated with $\boldsymbol{G}_i$ is obtained by equating the following derivative to zero:

$$\begin{aligned}
\frac{\partial\ell_{ce}(\boldsymbol{\gamma})^{(m)}}{\partial\gamma_{ij}} = -\tfrac{1}{2}\Big\{&\mathrm{tr}\left(\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_{ij}\right) - \mathrm{tr}\left(\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_{ij}\boldsymbol{G}_i^{-1}\boldsymbol{C}^{Z_iZ_i(m)}\right) - \\
&\tilde{\boldsymbol{u}}_i^{(m)'}\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_{ij}\boldsymbol{G}_i^{-1}\tilde{\boldsymbol{u}}_i^{(m)}/\sigma_H^{2(m)}\Big\},
\end{aligned}$$

This usually provides an explicit solution for the updated parameter γ_{ij}.

The $(m+1)^{th}$ update for a parameter ϕ_i associated with $\boldsymbol{\Sigma}$ is obtained by equating the following derivative to zero

$$\begin{aligned}
\frac{\partial\ell_{ce}(\sigma^2, \boldsymbol{\phi})^{(m)}}{\partial\phi_i} = -\tfrac{1}{2}\Big\{&\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\boldsymbol{C}^{(m)-1}\boldsymbol{W}'\right)/\sigma^2 - \\
&\tilde{\boldsymbol{e}}^{(m)'}\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_i\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{e}}^{(m)}/(\sigma^2\sigma_H^{2(m)})\Big\}
\end{aligned}$$

which may still require an iterative solution itself (see Foulley *et al.*[22]).

For σ^2 we have

$$\frac{\partial \ell_{ce}(\sigma^2, \boldsymbol{\phi})^{(m)}}{\partial \sigma^2} = -\tfrac{1}{2}\left\{\text{tr}\left(\boldsymbol{I}_n/\sigma^2\right) - \text{tr}\left(\boldsymbol{W}\boldsymbol{C}^{(m)-1}\boldsymbol{W}'\boldsymbol{\Sigma}^{-1}\right)/\sigma^4 - \tilde{\boldsymbol{e}}^{(m)'}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{e}}^{(m)}/(\sigma^4)\right\}$$

$$\Rightarrow \sigma^{2(m+1)} = \left\{(p+b)\sigma^{2(m)} - \sigma^{2(m)}\text{tr}\left(\boldsymbol{C}^{(m)-1}\boldsymbol{G}^{(m)*}\right) + \tilde{\boldsymbol{e}}^{(m)'}\boldsymbol{\Sigma}^{(m)-1}\tilde{\boldsymbol{e}}^{(m)}\right\}/n.$$

It is interesting to note that EM updates can be expressed in an analogous form to gradient methods, that is, as in Eq. [10], but with an information matrix that corresponds to the complete data. Specifically, the elements of this matrix are given by $-\partial^2 \ell_{ce}/\partial\kappa_i \partial\kappa_j$. (Also see Jensen *et al.* [45].)

5.2.1. *Example: EM updates for an unstructured G-matrix*

We illustrate the ideas developed above for a linear mixed model with an unstructured variance model for the random effects. This model often arises in multivariate, multi-environment trial or random regression applications. For simplicity we assume only a single random effect (i.e. the one of interest) and only present the updating scheme for the variance parameters associated with the G-structure. We assume a model of the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}, \tag{11}$$

where $\boldsymbol{u}^{gt\times 1}$ is a vector of random effects corresponding to sires (genotypes) within traits (trials) with associated design matrix $\boldsymbol{Z}^{n\times gt}$. We assume that the random effects and residuals have a joint Gaussian distribution with variance matrix

$$\text{var}\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{pmatrix} = \sigma_H^2\begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{I}_n\end{bmatrix},$$

where $\boldsymbol{G} = \boldsymbol{G}_t \otimes \boldsymbol{I}_g$, the subscript t denoting traits (trials).

Following the approach outlined previously, the EM update for $\boldsymbol{G}_t$ is obtained as

$$\boldsymbol{G}_t^{(m+1)} = \left(\tilde{\boldsymbol{U}}^{(m)'}\tilde{\boldsymbol{U}}^{(m)}/\sigma_H^{2(m)} + \boldsymbol{\Psi}^{(m)}\right)/g,$$

where $\boldsymbol{U}$ is the $g \times t$ matrix such that $\text{vec}(\boldsymbol{U}) = \boldsymbol{u}$. Also, $\boldsymbol{\Psi}^{t\times t}$ is a matrix of average prediction errors for each trait (trial), that is, with elements given by $\psi_{ij} = \text{tr}\left(\boldsymbol{C}_{ij}^{ZZ}\right)$ and $\boldsymbol{C}^{ZZ} = \left\{\boldsymbol{C}_{ij}^{ZZ}\right\}$ is partitioned accordingly into $g \times g$ matrices for each pair of traits (trials).

5.3. *The Parameter Expanded EM algorithm*

Although the EM algorithm has been widely used for the estimation of variance parameters in the linear mixed model it can be slow to converge. This is particularly a problem when the estimates of the variance parameters are on or near the boundary of the parameter space (Laird and Ware[18]). Furthermore, Foulley and van Dyk[46] suggest that biometricians working in animal breeding have been among the largest users of the EM algorithm, but note that the EM algorithm can be very slow to converge in these applications due to the relative magnitude of some of the variance components. To improve the rate of convergence of the EM algorithm, Liu *et al.* [42] introduced the parameter expanded EM or PXEM algorithm. In the case of linear mixed models, the algorithm involves the re-scaling of the random effects for simple variance components models or a rotation of the random effects for unstructured G-matrices. In this section we briefly review the PXEM algorithm and illustrate its application in a simple example. For a more thorough review the reader is referred to Foulley and van Dyk[46].

The PXEM algorithm assumes that the parameter vector $\boldsymbol{\kappa}$ can be expanded to a larger set of parameters $\boldsymbol{\Gamma}' = (\boldsymbol{\kappa}^{*\prime}\ \boldsymbol{\lambda}')$ where $\boldsymbol{\lambda}$ is a "working" parameter. The expanded parameterisation must satisfy the following two conditions

(1) it can be reduced to the original parameterisation, $\boldsymbol{\kappa}$, maintaining the same data model via a many-to-one reduction, $\boldsymbol{\kappa} = F(\boldsymbol{\Gamma})$;
(2) when $\boldsymbol{\lambda}$ is set to its "null" value, say $\boldsymbol{\lambda}_0$ this induces the same *complete* data model as with $\boldsymbol{\kappa} = \boldsymbol{\kappa}^*$.

Once we have set up the expanded parameter set the PXEM algorithm proceeds in a fashion simliar to the EM algorithm, in which there is an E-step and an M-step. The PX-E step computes the conditional expectation of the joint density of the complete data given $\boldsymbol{y_2}$, the so-called observed data, with $\boldsymbol{\Gamma}^{(m)}$ set to $\left(\kappa^{*\prime(m)}, \lambda' = \lambda_0'\right)'$. The PX-M step then maximises this conditional expectation with respect to the expanded parameters and $\boldsymbol{\kappa}^{(m)}$ is updated via the reduction $\boldsymbol{\kappa}^{(m+1)} = F(\boldsymbol{\Gamma}^{(m+1)})$.

5.3.1. *Example: PXEM updates for an unstructured G-matrix*

We consider the model in Eq. [11], which can be expanded to

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z\Lambda f} + \boldsymbol{e}, \tag{12}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_t \otimes \boldsymbol{I}_g)$, $\boldsymbol{u} = (\boldsymbol{\Lambda}_t \otimes \boldsymbol{I}_g)\boldsymbol{f}$ and $\text{var}(\boldsymbol{u}) = \sigma^2_{\scriptscriptstyle H}\boldsymbol{G} = \sigma^2_{\scriptscriptstyle H}(\boldsymbol{G}_t \otimes \boldsymbol{I}_g)$,and $\text{var}(\boldsymbol{f}) = \sigma^2_{\scriptscriptstyle H}\boldsymbol{D} = \sigma^2_{\scriptscriptstyle H}(\boldsymbol{D}_t \otimes \boldsymbol{I}_g)$; thus, $\boldsymbol{G}_t = \boldsymbol{\Lambda}_t\boldsymbol{D}_t\boldsymbol{\Lambda}_t'$ and $\boldsymbol{G} = \boldsymbol{\Lambda}\boldsymbol{D}\boldsymbol{\Lambda}'$. The matrix $\boldsymbol{\Lambda}_t^{t\times t}$ is assumed to be invertible while the matrix $\boldsymbol{D}_t^{t\times t}$ is symmetric positive definite. Further we let $\boldsymbol{\gamma} = \text{vech}(\boldsymbol{G}_t)$, $\boldsymbol{\lambda} = \text{vec}(\boldsymbol{\Lambda}_t)$ and $\boldsymbol{d} = \text{vech}(\boldsymbol{D}_t)$. The reduced parameter vector is $\boldsymbol{\kappa}' = (\boldsymbol{\gamma}', \boldsymbol{\phi}')$, where in this case $\boldsymbol{\phi}$ is the null vector, while the expanded variance parameter vector is $\boldsymbol{\Gamma}' = \left[\boldsymbol{d}'\ \boldsymbol{\lambda}'\right]$. The role of the extra parameter matrix $\boldsymbol{\Lambda}_t$, is simply to rotate the random effects. Note that the null value of $\boldsymbol{\lambda}_0 = \text{vec}\,(\boldsymbol{\Lambda}_{t0}) = \text{vec}\,(\boldsymbol{I}_t)$ results in the same variance model parameterisation as the reduced variance parameter model with $\boldsymbol{G}_t = \boldsymbol{D}_t$.

We then define the complete data to be $(\boldsymbol{f}', \boldsymbol{y}')'$ and consider the joint likelihood, which is given by

$$\ell_c(\boldsymbol{\Gamma}; \boldsymbol{f}, \boldsymbol{y}) = -\tfrac{1}{2}\left\{(n+gt)\log\sigma^2_{\scriptscriptstyle H} + \boldsymbol{e}'\boldsymbol{e}/\sigma^2_{\scriptscriptstyle H} + \log|\boldsymbol{D}| + \boldsymbol{f}'\boldsymbol{D}^{-1}\boldsymbol{f}/\sigma^2_{\scriptscriptstyle H}\right\}.$$

The expected value of this joint likelihood conditional on $\boldsymbol{y}_2$ and evaluated at $\boldsymbol{\Gamma} = \left[\boldsymbol{\kappa}^{*(m)\prime}, \boldsymbol{\lambda}_0'\right]'$, is given by

$$\ell_{ce}(\boldsymbol{\Gamma})^{(m)} = \ell_{ce}(\boldsymbol{d})^{(m)} + \ell_{ce}(\boldsymbol{\lambda})^{(m)},$$

say, where

$$\ell_{ce}(\boldsymbol{d})^{(m)} = -\tfrac{1}{2}\text{E}\left(gt\log\sigma^2_{\scriptscriptstyle H} + \log|\boldsymbol{D}| + \boldsymbol{f}'\boldsymbol{D}^{-1}\boldsymbol{f}/\sigma^2_{\scriptscriptstyle H} \mid \boldsymbol{y}_2,\ \boldsymbol{\Gamma} = [\boldsymbol{\kappa}^{*(m)\prime}, \boldsymbol{\lambda}_0']'\right),$$

$$\ell_{ce}(\boldsymbol{\lambda})^{(m)} = -\tfrac{1}{2}\text{E}\left(n\log\sigma^2_{\scriptscriptstyle H} + \boldsymbol{e}'\boldsymbol{e}/\sigma^2_{\scriptscriptstyle H} \mid \boldsymbol{y}_2,\ \boldsymbol{\Gamma} = [\boldsymbol{\kappa}^{*(m)\prime}, \boldsymbol{\lambda}_0']'\right).$$

We maximise ℓ_{ce} with respect to both $\boldsymbol{d}$ and $\boldsymbol{\lambda}$ to obtain updates from which $\boldsymbol{\gamma}$ is then updated via

$$\boldsymbol{G}^{(m+1)} = \boldsymbol{\Lambda}^{(m+1)}\boldsymbol{D}^{(m+1)}\boldsymbol{\Lambda}^{(m+1)\prime}.$$

We first consider $\boldsymbol{\lambda}$. Maximising ℓ_{ce} with respect to $\boldsymbol{\lambda}$ leads to updates of the form

$$\boldsymbol{\lambda}^{(m+1)} = \boldsymbol{A}^{-1(m)}\boldsymbol{b}^{(m)},$$

where $\boldsymbol{b}$ is a vector of length t^2 whose $(i,j)th$ element is b_{ij} and $\boldsymbol{A}$ is a $t^2 \times t^2$ matrix whose rows and columns are indexed by (i,j) and (k,l) respectively with elements given by $a_{ij;kl}$. The elements of $\boldsymbol{b}$ and $\boldsymbol{A}$ are given by

$$b_{ij} = \tilde{\boldsymbol{u}}^{(m)\prime}\dot{\boldsymbol{\Lambda}}_{ij}'\boldsymbol{Z}'\hat{\boldsymbol{y}}^{*(m)} - \sigma^{2(m)}_{\scriptscriptstyle H}\text{tr}\left(\dot{\boldsymbol{\Lambda}}_{ij}'\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{C}^{XZ(m)}\right),$$

$$a_{ij;kl} = \tilde{\boldsymbol{u}}^{(m)\prime}\dot{\boldsymbol{\Lambda}}_{ij}'\boldsymbol{Z}'\boldsymbol{Z}\dot{\boldsymbol{\Lambda}}_{kl}\tilde{\boldsymbol{u}}^{(m)} + \sigma^{2(m)}_{\scriptscriptstyle H}\text{tr}\left(\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Lambda}_{kl}\boldsymbol{C}^{ZZ(m)}\dot{\boldsymbol{\Lambda}}_{ij}{}'\right),$$

where $\boldsymbol{y}^* = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau}$ and $\dot{\boldsymbol{\Lambda}}_{ij} = \partial\boldsymbol{\Lambda}/\partial\lambda_{ij}$ and $\boldsymbol{C}^{XZ}$ is the off-diagonal portion of the inverse of $\boldsymbol{C}$ corresponding to $\boldsymbol{\tau}$ and $\boldsymbol{u}$.

The updating formula for $\boldsymbol{d}$ is obtained using the same approach as for the EM updates for unstructured G-matrices presented earlier.

5.4. *Improved iterative schemes*

When REML estimates of variance parameters are within the parameter space the most common cause of convergence difficulties for second-order methods, including AI, is the use of poor starting values. Given that second-order methods converge much faster than first-order methods when updates are "close to" the solution, we propose hybrid schemes that use EM (or PXEM) updates until close enough to the solution, then invoke AI updates. Operationally, we compute the score vector and AI matrix at each iteration, and then check that the AI matrix is positive definite and calculate a global measure of proximity of the current estimates to the REML solution. The proximity measure for the m^{th} iteration is calculated as $\boldsymbol{U}_R(\boldsymbol{\sigma}^{(m)})'\boldsymbol{I}_A(\boldsymbol{\sigma}^{(m)})^{-1}\boldsymbol{U}_R(\boldsymbol{\sigma}^{(m)})$. In the hybrid schemes we only use AI updates if the AI matrix is positive definite and the proximity measure has a p-value greater than some pre-determined tolerance when compared with a chi-square reference distribution with degrees of freedom equal to the total number of variance parameters. We choose to always use AI updates for variance parameters associated with $\boldsymbol{R}$ since we believe most convergence difficulties are in relation to parameters in $\boldsymbol{G}$. In the remainder of this paper these schemes will be referred to as the "EM/AI" and "PXEM/AI" schemes.

5.4.1. *Local schemes*

A computationally "cheaper" alternative to using full EM (or PXEM) iterations in the hybrid schemes is to use an internal or so-called local EM (or PXEM) scheme that is invoked within the external iterations. These hybrid schemes will be referred to as the "local EM/AI" and "local PXEM/AI" schemes. The motivation for these approaches was the use of augmented dispersion models for the estimation of dispersion parameters in hierarchical generalised linear models (Lee and Nelder[44]). We describe our hybrid schemes in the context of a single random factor and let $\sigma_H^2 = \sigma_H^{2(0)}$, and $\boldsymbol{\phi} = \boldsymbol{\phi}^{(0)}$ denote the values for these parameters from the external loop. After absorption of fixed effects in the mixed model equations (of the external loop) we have

$$(\boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z} + \boldsymbol{G}^{(0)-1})\tilde{\boldsymbol{u}} = \boldsymbol{t}$$

where

$$\boldsymbol{t} = \boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{y}$$

and

$$\boldsymbol{S}^{(0)} = \boldsymbol{R}^{(0)-1} - \boldsymbol{R}^{(0)-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{R}^{(0)-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}^{(0)-1}.$$

Thus, we have that

$$\begin{aligned}\mathrm{E}(\boldsymbol{t}) &= \boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{X}\boldsymbol{\tau} = \boldsymbol{0}\\ \mathrm{var}(\boldsymbol{t}) &= \sigma_H^{2(0)}(\boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z} + \boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z})\end{aligned}$$

The matrix $\boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z}$ will in general be singular and not readily available within the framework of mixed-models software such as ASRemlHence we propose in the following to approximate this matrix by another matrix that we denote by $\boldsymbol{\Omega}^{(0)}$, which is non-singular and whose elements are accessible from the (current) external loop. The method of approximation will in general depend on the application and the form of the model for $\boldsymbol{G}$.

Define the modified working effects vector by

$$\boldsymbol{z} = \boldsymbol{\Omega}^{(0)-1}\boldsymbol{t}/\sqrt{\sigma_H^{2(0)}}$$

and hence

$$\begin{aligned}\mathrm{E}(\boldsymbol{z}) &= \boldsymbol{0}\\ \mathrm{var}(\boldsymbol{z}) &= \boldsymbol{\Omega}^{(0)-1}(\boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z} + \boldsymbol{Z}'\boldsymbol{S}^{(0)}\boldsymbol{Z})\boldsymbol{\Omega}^{(0)-1}\\ &\simeq \boldsymbol{G} + \boldsymbol{\Omega}^{(0)-1}\end{aligned}$$

Hence we have approximately,

$$\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{G} + \boldsymbol{\Omega}^{(0)-1})$$

Since $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\gamma})$ the opportunity exists to update $\boldsymbol{\gamma}$ from an internal updating scheme, in which we have implicitly fixed the estimates of the fixed effects and the other variance parameters from the current external loop. The approach is illustrated in the following section for an unstructured variance matrix.

Example: Local scheme for an unstructured G-matrix

In this section we consider the example considered in sections 5.2.1 and 5.3.1. The "local" linear mixed model is given by

$$\begin{aligned}\boldsymbol{z} &= \boldsymbol{u}_z + \boldsymbol{e}_z\\ &= \boldsymbol{\Lambda}\boldsymbol{f}_z + \boldsymbol{e}_z\end{aligned} \tag{13}$$

and, as before, $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_t \otimes \boldsymbol{I}_g$, $\boldsymbol{u}_z = \boldsymbol{\Lambda} \boldsymbol{f}_z$ and $\mathrm{var}(\boldsymbol{u}_z) = \boldsymbol{G} = \boldsymbol{G}_t \otimes \boldsymbol{I}_g, \mathrm{var}(\boldsymbol{f}_z) = \boldsymbol{D} = \boldsymbol{D}_t \otimes \boldsymbol{I}_g$; thus, $\boldsymbol{G}_t = \boldsymbol{\Lambda}_t \boldsymbol{D}_t \boldsymbol{\Lambda}_t'$. As before the matrix $\boldsymbol{\Lambda}_t^{t \times t}$ (and hence $\boldsymbol{D}_t$) is assumed to be invertible while the matrix $\boldsymbol{D}$ is symmetric positive definite. Lastly, we have that $\mathrm{var}(\boldsymbol{e}_z) = \boldsymbol{\Omega}^{(0)-1}$, which is assumed to be known (as is σ_H^2).

To implement the local scheme we first consider an appropriate full rank approximation to the matrix $\boldsymbol{Z}'\boldsymbol{S}\boldsymbol{Z}$. If $\boldsymbol{C}^{ZZ(0)} = \{\boldsymbol{C}_{ij}^{ZZ(0)}\}$ ($i, j = 1 \ldots t$) is the coefficient matrix from the mixed model equations associated with $\boldsymbol{u}$ evaluated at the current parameter values, then, in order to obtain a $\boldsymbol{\Omega}^{(0)}$, we suggest approximating $\boldsymbol{C}^{ZZ(0)}$ by

$$\boldsymbol{C}^{(0)*} = \left\{ \mathrm{diag}\left(\boldsymbol{C}_{ij}^{ZZ(0)} \right) \right\}$$

and hence we choose

$$\boldsymbol{\Omega} = \boldsymbol{C}^{(0)*-1} - \boldsymbol{G}^{(0)-1}$$

In terms of the data vector $\boldsymbol{z}$ we have

$$\boldsymbol{z} = \boldsymbol{\Omega}^{(0)-1} \boldsymbol{C}^{ZZ(0)-1} \tilde{\boldsymbol{u}}^{(0)} / \sqrt{\sigma_H^{2(0)}},$$

where

$$\tilde{\boldsymbol{u}}^{(0)} = \boldsymbol{C}^{ZZ(0)} \boldsymbol{Z}' \boldsymbol{S}^{(0)} \boldsymbol{y}.$$

In order to implement the scheme we must apply the same approximation for the calculation of $\boldsymbol{z}$. That is, we must use

$$\boldsymbol{z}^* = \boldsymbol{\Omega}^{(0)-1} \boldsymbol{C}^{(0)*-1} \tilde{\boldsymbol{u}}^{(0)} / \sqrt{\sigma_H^{2(0)}}.$$

Since Eq. [13] is a (trivial) linear mixed model in which the fixed-effects design matrix is the null matrix, we can implement either EM or PXEM schemes quite simply.

5.5. *Analysis of data-sets*

Here we examine the performance of a range of iterative schemes on two published data-sets, namely the lamb weight data presented in Callanan and Harville[47] and the ultrafiltration data presented in Foulley and van Dyk[46]. We have used models as given in these papers. Note in particular that the variance parameterisation used in both cases corresponds to the scale parameter σ_H^2 being fixed at the value 1. Convergence was defined as being achieved when the norm $\sqrt{(\sum(\sigma_i^{(m)} - \sigma_i^{(m-1)})^2)/(\sum \sigma_i^{(m)2})}$ of all variance parameters was smaller than 1^{-8}. For the hybrid schemes, AI was

invoked when the p-value for the proximity measure was greater than 0.5. The local schemes require a choice for the number of internal iterations. For the PXEM scheme we used a single internal iteration and for EM we used 3 iterations. The reasoning behind our choices was that the updates from a single internal iteration of local EM are equivalent to those from a single global iteration of EM so, in order to take advantage of the economies of the local scheme, it is desirable to perform more than a single iteration (we chose 3). In contrast, updates from a single iteration of local PXEM are different from the global PXEM updates since the prediction error covariance terms between fixed and random effects (ie. $\boldsymbol{C}^{XZ}$) are ignored. One of the principal motivations for the local PXEM scheme was the difficulty in implementing global PXEM in ASReml (due to the need for terms involving $\boldsymbol{C}^{XZ}$). Thus we have an interest in assessing the performance of a scheme that is essentially PXEM but ignores these terms. The local PXEM scheme with a single internal iteration provides such an algorithm. Determination of an optimum number of internal iterations for either EM or PXEM is an issue that requires further research.

5.5.1. *Lamb weight data*

The data consist of the weights at birth of 62 single-birth male lambs. Each lamb was the progeny of one of 23 sires and the sires belonged to one of 5 population lines. Each lamb had a different dam and the age groups of the dams were recorded (1=1-2 years; 2=2-3 years; 3=over 3 years). The model fitted to the data can be written in symbolic form as

$$\text{weight} \sim \text{mu} + \text{age} + \text{line} + \boldsymbol{sire}$$

where age, line and sire are factors with 3, 5 and 23 levels, respectively, and mu is the intercept term. The effects of sire (within line) are fitted as random effects; all other effects are fitted as fixed. In matrix notation we have

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{\tau}$ is the vector of fixed effects comprising the intercept and the main effects of age and line, and $\boldsymbol{u}$ is the vector of random sire effects. The associated design matrices are $\boldsymbol{X}^{62\times 7}$ (assumed to be of full column rank) and $\boldsymbol{Z}^{62\times 23}$. The vector of residual effects is given by $\boldsymbol{e}$. We assume that $\boldsymbol{u}$ and $\boldsymbol{e}$ are independent with

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_s^2 \boldsymbol{I}_{23})$$

$$\text{and} \quad \boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_{62}).$$

We used nine different iterative schemes to obtain REML estimates of the variance components σ_s^2 and σ_e^2. These schemes included standard AI, EM and PXEM and the hybrid schemes EM/AI, PXEM/AI, local EM/AI and local PXEM/AI. In addition, we used a linearised AI and a linearised method of successive approximations (MSA) scheme in order to compare our results with those presented in Callanan and Harville[47]. The standard and linearised AI schemes are analogous to the standard and linearised N-R schemes labelled "NR2" and "LN1" respectively in Callanan and Harville[47]; the linearised MSA scheme is "LMSA1" as per Callanan and Harville[47] (and note that our EM scheme is equivalent to Callanan and Harville's "EM1"). Two sets of starting values were used, namely $(\sigma_s^2 = 0.01, \sigma_e^2 = 1)$ and $(\sigma_s^2 = 5, \sigma_e^2 = 1)$.

The REML estimates of the variance components were $\hat{\sigma}_s^2 = 0.5171$ and $\hat{\sigma}_e^2 = 2.9616$. The numbers of iterations to convergence for all schemes (other than the linearised schemes) are given in table 1. The linearised AI scheme required 12 iterations (from both sets of start values) and the linearised MSA scheme required 12 and 14 iterations (for the start values of $\sigma_s^2 = 0.01$ and $\sigma_s^2 = 5$ respectively). Note that in the AI scheme the first update for σ_s^2, when started from the value 5, was negative. We have implemented a strategy in ASReml for such cases, namely to change the inadmissable estimate to an admissable value (chosen to be 0.0001) then re-update other parameters subject to this alteration. Without this remedial action the AI algorithm failed for this data-set when the start value 5 was used for σ_s^2.

We have graphed the iteration sequences for σ_s^2 for the AI algorithm and all hybrid schemes (see figure 1). Only the first 10 iterations are shown.

5.5.2. *Ultrafiltration data*

The data consist of the ultrafiltration response of 20 membrane dialysers measured at 7 different transmembrane pressures with an evaluation made at two blood flow rates. The model fitted to the data can be written in symbolic form as

$$\text{ufr} \sim \text{mu} + \text{bfr} + \text{tmp} + \text{tmp2} + \text{tmp3} + \text{tmp4} + \boldsymbol{subject} + \boldsymbol{subject.tmp} + \boldsymbol{subject.tmp2},$$

where bfr (blood flow rate) is a factor with 2 levels, tmp (transmembrane pressure) is a covariate (as are tmp2=tmp^2, tmp3=tmp^3 and tmp4=tmp^4),

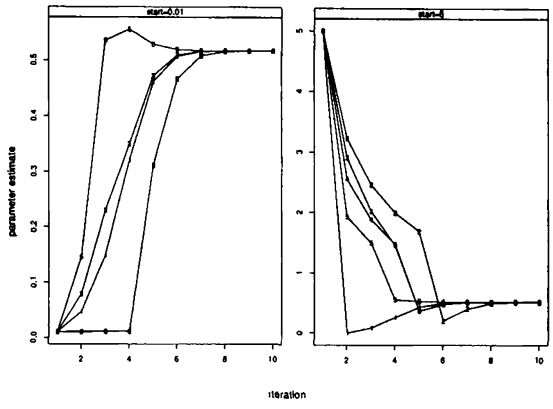


Fig. 1. Lamb weight data: estimates for sire variance component for 5 algorithms: 1=standard AI; 2=EM/AI; 3=PXEM/AI; 4=local EM/AI; 5=local PXEM/AI. First 10 iterations only. (a) start values $\sigma_s^2 = .01$, $\sigma_e^2 = 1$ and (b) start values $\sigma_s^2 = 5$, $\sigma_e^2 = 1$

Table 1. Number of iterations to convergence for 2 data-sets and a range of algorithms. For hybrid schemes, the break-down into non-AI and AI iterations is given in parentheses. Notes: lamb[1] start values $\sigma_s^2 = 0.01, \sigma_e^2 = 1$, lamb[2] start values $\sigma_s^2 = 5, \sigma_e^2 = 1$ and [3] updates for σ_s^2 constrained to be positive

Algorithm	Lamb[1]	Lamb[2]	Ultrafiltration
AI	14	14[3]	Failed
EM	1296	342	249
PXEM	83	78	42
EM/AI	15 (3/12)	16 (4/12)	13 (5/8)
PXEM/AI	13 (3/10)	14 (3/11)	8 (2/6)
local EM/AI	15 (3/12)	14 (3/11)	11 (3/11)
local PXEM/AI	13 (2/11)	13 (2/11)	7 (1/6)

and subject (dialyser) is a factor with 20 levels. The effects in bold are fitted as random effects; all other effects are fitted as fixed. In matrix notation we have

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

where $\boldsymbol{\tau}$ is the vector of fixed effects and $\boldsymbol{u} = (\boldsymbol{u}_0', \boldsymbol{u}_1', \boldsymbol{u}_2')'$ is the vector of random effects, partitioned as the subject intercept, linear and quadratic coefficients respectively. The associated design matrices are $\boldsymbol{X}^{140\times 6}$ (as-

sumed to be of full column rank) and $\boldsymbol{Z}^{140\times 60}$. The vector of residual effects is given by $\boldsymbol{e}$. We assume that $\boldsymbol{u}$ and $\boldsymbol{e}$ are independent with

$$\begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \\ \boldsymbol{u}_3 \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} \sigma_{00} & & \\ \sigma_{01} & \sigma_{11} & \\ \sigma_{02} & \sigma_{12} & \sigma_{22} \end{bmatrix} \otimes \boldsymbol{I}_{20} \right)$$

$$\text{and} \quad \boldsymbol{e} \sim N(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_{140})$$

We used seven different iterative schemes to obtain REML estimates of the variance parameters. These schemes comprised standard AI, EM and PXEM and all hybrid schemes. The starting values used were as given in Foulley and van Dyk[46], namely $\sigma_{00} = \sigma_{11} = \sigma_{22} = 4$, $\sigma_{01} = 2$, $\sigma_{02} = -1.2$, $\sigma_{12} = -2.4$ and $\sigma_e^2 = 4$. The use of these values revealed the instability of second-order methods when starting values are poor, with standard AI failing to converge, whereas standard EM and PXEM converged. The REML estimates of the variance parameters were $\hat{\sigma}_{00} = 2.25, \hat{\sigma}_{11} = 24.08, \hat{\sigma}_{22} = 2.17$, $\hat{\sigma}_{01} = -3.73$, $\hat{\sigma}_{02} = 0.69$, $\hat{\sigma}_{12} = -6.83$ and $\hat{\sigma}_e^2 = 3.32$. The numbers of iterations for convergence are given in table 1.

We have graphed the iteration sequences for the random regression variance parameters for the AI algorithm and all hybrid schemes (see figure 2). Only the first 10 iterations are shown.

5.6. *Discussion of Results*

Foulley and van Dyk[46] state that the EM algorithm is widely used for REML estimation of variance parameters in a linear mixed model. They suggest that a major reason for its popularity is the stable convergence property but that a disadvantage is its slow convergence. They therefore focus on "EM procedures and ways to improve them". In particular, they advocate use of the PXEM algorithm. We have adopted the converse strategy, that is, our focus is on a fast, computationally-efficient algorithm, namely the second-order method of AI, and ways to improve its stability. At present, attention has been restricted to cases in which the REML solution is within the parameter space and our discussion here is in that context. The problem of iterative schemes when the REML solution is outside the parameter space is the subject of current research.

Our analysis of the lamb weight data clearly shows that when initial values for variance parameters are reasonable the AI algorithm converges in far fewer iterations (14) than either PXEM (83) or EM (1296). The

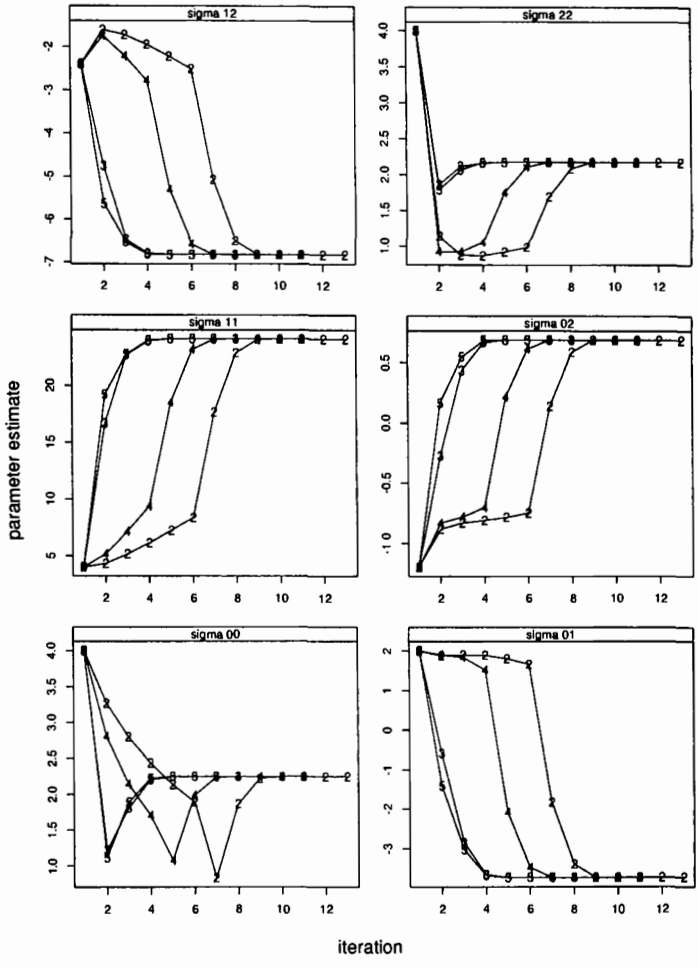


Fig. 2. Ultrafiltration data: estimates for random regression variance parameters for 4 algorithms: 2=EM/AI; 3=PXEM/AI; 4=local EM/AI; 5=local PXEM/AI. First 10 iterations only.

superiority over EM reflects the well-known difference between the speed of first- and second-order methods. The direct comparison of AI and PXEM suggests that PXEM, although an improvement over EM, still falls well short of AI in terms of speed of convergence. However, with poor initial values, second-order methods, including AI, may be unstable. In simple

models, in particular variance component models, this may not be an issue provided that non-negativity constraints are imposed and properly dealt with during the iteration sequence. This is illustrated in our analysis of the lamb weight data. Instability is most likely to arise in models with a more complex variance structure such as the unstructured model used in the random regression context. Our analysis of the ultrafiltration data reveals a potential solution in the form of hybrid schemes that employ a sequence of EM or PXEM updates followed by AI. All of the proposed hybrid schemes converged in far fewer iterations (ranging from 7 for the local PXEM/AI scheme to 13 for the EM/AI scheme) than either PXEM (42) or EM (249). The initial EM-type updates in the hybrid schemes are merely a tool to move close enough to the REML solution to enable AI to be successfully invoked. An obvious alternative is to ensure that initial values are sufficiently close to the solution. General approaches for generating good initial values for a range of variance models is the subject of current research.

Although illustrated in terms of small examples, the algorithms presented here are all amenable for use on large data-sets (i.e. with large numbers of random effects or many variance parameters). Harville[48] dismisses the use of second order methods in such settings, suggesting that they are so computer intensive as to be infeasible. This may be the case for the N-R and F-S schemes but the AI algorithm is routinely implemented on large data-sets, particulary in animal and plant breeding applications and requires little more computation than first-order schemes[17]. For example, we routinely use the AI algorithm to fit complex variance models to multi-environment plant variety trial data where the number of random effects often exceeds 60,000.

6. Inference in linear mixed models

6.1. *Hypothesis tests for variance models*

REML likelihood ratio tests (abbreviated to the acronym REMLRT) can be used to compare nested models. That is, for a comparison of (nested) models M_0 and M_1 with the same fixed model, where M_1 contains k extra variance parameters, the REMLRT statistic is given by

$$D = -2\left(\ell_{R0} - \ell_{R1}\right),$$

where ℓ_{Ri} is the residual log-likelihood for model i.

The statistic D is asymptotically distributed as a chi-squared variable with k degrees of freedom. The exception is when the test involves a null hypothesis with the parameter on the boundary of the parameter space (see Stram and Lee[49] for further discussion). REMLRT cannot be used to compare non-nested models. In these situations the Akaike Information Criterion AIC[50] has been proposed as a model selection criterion. This is given by $AIC = -2\ell_R + 2k$, where ℓ_R is the value of the maximised REML log-likelihood and k is the number of variance parameters being estimated. Various other criteria have been suggested to improve the performance of the AIC criteria in different settings. Some examples of these include the Corrected Akaike Information Criterion $AICC$[51], the Bayesian Information criterion BIC[52] and Residual Information Criterion[53].

6.2. *Inference for fixed effects*

For many applications of the general linear mixed model given in Eq. [5], interest centres on the vector of fixed effects. In such cases it is well known that specification of the variance model, either via a design-based or model-based route, significantly affects inference for such fixed effects. The distribution of the E-BLUE of $\boldsymbol{\tau}$ is in general not known. For certain situations, particularly in small samples or in designs with multiple strata or complex variance models, this can have a significant impact on the distribution of the E-BLUE (see, for example, Welham and Thompson[54]).

Wald-type test procedures are generally favoured for conducting tests concerning $\boldsymbol{\tau}$, though these are based on asymptotic χ^2 approximations and hence also ignore the additional variability in the estimated variance parameters.

Kenward and Roger[29] considered this general problem in detail. They pursue the concept of construction of Wald-type test statistics through an adjusted variance matrix of $\hat{\boldsymbol{\tau}}$. Note that in the following, rather than introduce additional notation, and for consistency with Kenward and Roger, we shall use $\hat{\boldsymbol{\tau}}$ to denote the E-BLUE, obtained as the solution to Eq. [9] after replacing $\boldsymbol{\kappa}$ by its REML estimate. Kenward and Roger argue that it is initially useful to consider an improved estimator of the variance matrix of $\hat{\boldsymbol{\tau}}$ which has less bias and accounts for the variability in estimation of the variance parameters. There are two reasons for this. Firstly, the small sample distribution of Wald tests is simplified when the adjusted variance matrix is used. Secondly, if measures of precision are required for $\hat{\boldsymbol{\tau}}$, or effects therein, those obtained from the adjusted variance matrix will gen-

erally be preferred. The adjusted variance matrix builds on the work of Harville and colleagues (see for example[55]) and is given by

$$\boldsymbol{\Phi}_A = \boldsymbol{\Phi} + 2\boldsymbol{\Phi}\left[\sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij}\left(\dot{\boldsymbol{D}}_{ij} - \dot{\boldsymbol{D}}_i\boldsymbol{\Phi}\dot{\boldsymbol{D}}_j - \frac{1}{4}\ddot{\boldsymbol{D}}_{ij}\right)\right]\boldsymbol{\Phi}, \tag{14}$$

where r is the dimension of $\boldsymbol{\sigma}$, and

$$\dot{\boldsymbol{D}}_i = \boldsymbol{X}'\frac{\partial \boldsymbol{V}^{-1}}{\partial \sigma_i}\boldsymbol{X},$$

$$\dot{\boldsymbol{D}}_{ij} = \boldsymbol{X}'\frac{\partial \boldsymbol{V}^{-1}}{\partial \sigma_i}\boldsymbol{V}\frac{\partial \boldsymbol{V}^{-1}}{\partial \sigma_j}\boldsymbol{X},$$

$$\ddot{\boldsymbol{D}}_{ij} = \boldsymbol{X}'\boldsymbol{V}^{-1}\frac{\partial^2 \boldsymbol{V}}{\partial \sigma_i \partial \sigma_j}\boldsymbol{V}^{-1}\boldsymbol{X},$$

$$\boldsymbol{\Phi} = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1},$$

and w_{ij} is $(i,j)th$ element of the inverse of the AI matrix. In the following we have omitted the "hat" unless there is ambiguity. All matrices that are functions of the variance parameter vector $\boldsymbol{\sigma}$ are replaced by the equivalent matrices with $\boldsymbol{\sigma}$ replaced by its REML estimate.

For inferences concerning $\boldsymbol{L}'\boldsymbol{\tau}$, where $\boldsymbol{L}$ is a $p \times v$ known matrix of full rank, Kenward and Roger [29] propose the use of the adjusted variance matrix presented in Eq. [14] in a Wald-type statistic given by

$$F_e = (\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})'\boldsymbol{L}\left(\boldsymbol{L}'\boldsymbol{\Phi}_A\boldsymbol{L}\right)^{-1}\boldsymbol{L}'(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})/v.$$

An appropriate F-approximation is achieved by consideration of a scaled form of F_e, say F^*, where and $F^* = \lambda F_e$, so that $F^* \sim F_{v,m}$. The formulae for m and λ are

$$m = 4 + (v+2)/(v\rho - 1),$$

$$\lambda = m/(\mathrm{E}(F)(m-2)),$$

where

$$\rho = \mathrm{var}(F)/2\mathrm{E}(F)^2.$$

We replace $\mathrm{E}(F)$ and $\mathrm{var}(F)$ by

$$E^* = (1 - A_2/v)^{-1},$$

$$V^* = 2(1 + c_1 B)/\{v(1 - c_2 B)^2(1 - c_3 B)\},$$

where

$$c_1 = g/\{3v + 2(1-g)\},$$
$$c_2 = (v-g)/\{3q + 2(1-g)\},$$
$$c_3 = (v+2-g)/\{3v + 2(1-g)\},$$
$$g = \{(v+1)A_1 - (v+4)A_2\}/\{(v+2)A_2\},$$
$$B = (A_1 + 6A_2/(2v).$$

All of these are simple to compute once we have A_1 and A_2. These can be written as

$$A_1 = \sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij}\mathrm{tr}\left(\boldsymbol{\Theta}\boldsymbol{A}_{1i}\right)\mathrm{tr}\left(\boldsymbol{\Theta}\boldsymbol{A}_{1j}\right),$$

$$A_2 = \sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij}\mathrm{tr}\left(\boldsymbol{\Theta}\boldsymbol{A}_{1i}\boldsymbol{\Theta}\boldsymbol{A}_{1j}\right),$$

for

$$\boldsymbol{\Theta} = \boldsymbol{L}\left(\boldsymbol{L}'\boldsymbol{\Phi}\boldsymbol{L}\right)^{-1}\boldsymbol{L}',$$
$$\boldsymbol{A}_{1i} = \boldsymbol{\Phi}\dot{\boldsymbol{D}}_i\boldsymbol{\Phi}.$$

6.3. *Computing the scaled F and adjusted variance matrix*

Computation of the scale and residual degrees of freedom for F^*, as well as the adjusted variance matrix of $\hat{\boldsymbol{\tau}}$, can be challenging in many applications of the general linear mixed model. For problems with either a large number of fixed effects or a large number of variance parameters the computational burden is substantial. In this section we present an efficient approach to compute the quantities. The approach fits into the AI algorithm and, at the time of writing, a *beta*-version has been implemented for testing in ASReml.

Analogous to the concept of working variables within the AI algorithm, we define working effects matrix for each variance parameter σ_i, given by

$$\boldsymbol{Q}_i^x = \dot{\boldsymbol{V}}_i\boldsymbol{V}^{-1}\boldsymbol{X}\boldsymbol{\Phi} = \dot{\boldsymbol{V}}_i\boldsymbol{H}^{-1}\boldsymbol{X}\left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}, \quad i = 1,\ldots,r$$

where

$$\dot{\boldsymbol{V}}_i = \frac{\partial \boldsymbol{V}}{\partial \sigma_i}$$

then it can be shown that

$$\boldsymbol{\Phi}\left(\dot{\boldsymbol{D}}_{ij} - \dot{\boldsymbol{D}}_i\boldsymbol{\Phi}\dot{\boldsymbol{D}}_j\right)\boldsymbol{\Phi} = \boldsymbol{Q}_i^{x\prime}\boldsymbol{P}_v\boldsymbol{Q}_j^x,$$

where $\boldsymbol{P}_v = \boldsymbol{P}/\sigma_H^2$.

Thus, excluding the last term, the adjusted variance matrix is a weighted sum of residual sums-of-squares and cross-products matrices for the working effects matrices. The final term can be computed if necessary by noting that

$$\boldsymbol{\Phi}\ddot{\boldsymbol{D}}_{ij}\boldsymbol{\Phi} = \left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\ddot{\boldsymbol{Q}}_{ij}^x,$$

where $\ddot{\boldsymbol{Q}}_{ij}^x = \ddot{\boldsymbol{V}}_{ij}\boldsymbol{H}^{-1}\boldsymbol{X}\left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}$ and

$$\ddot{\boldsymbol{V}}_{ij} = \frac{\partial^2\boldsymbol{V}}{\partial\sigma_i\partial\sigma_j}.$$

This quantity is available as an intermediate term in the absorption of $\boldsymbol{C}$ on the working effects matrix $\ddot{\boldsymbol{Q}}_{ij}^x$. This term is zero for a large class of variance models, otherwise our experience suggests this term can be ignored.

The matrix $\boldsymbol{A}_{1i}$ required for the scale and residual degrees of freedom can also be computed in a similar manner by noting that

$$\boldsymbol{A}_{1i} = -\left(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{H}^{-1}\boldsymbol{Q}_i^x.$$

A computationally cheap alternative, using numerical differentiation of $\boldsymbol{\Phi}$ with respect to σ_i, can be used by noting that

$$\boldsymbol{A}_{1i} = -\frac{\partial\boldsymbol{\Phi}}{\partial\sigma_i}.$$

This approximation appears adequate for the range of problems we have considered thus far.

6.4. *Kenward adjustments in ANOVA settings*

Consider the linear mixed-effects model with variance components (setting $\sigma_H^2 = 1$), where $\boldsymbol{G} = \oplus_{j=1}^q\gamma_j\boldsymbol{I}_{b_j}$. Then, if it is possible to consider a natural ordering of the variance component parameters, including σ^2, we can use an idea due to Thompson[33] to further reduce the computational load. That is, consider diagonalisation of the expected (or average) information matrix. If $\boldsymbol{F}$ is the average information matrix for $\boldsymbol{\sigma}$, let $\boldsymbol{U}$ be an upper triangular matrix such that $\boldsymbol{F} = \boldsymbol{U}'\boldsymbol{U}$. Further, we define

$$\boldsymbol{U}_c = \boldsymbol{D}_c\boldsymbol{U},$$

where $\boldsymbol{D}_c$ is a diagonal matrix whose elements are given by the inverse elements of the last column of U, i.e. $d_{cii} = 1/u_{ir}, i = 1, \ldots, r$. The matrix $\boldsymbol{U}_c$ is therefore upper triangular with the elements in the last column equal to one. If the vector $\boldsymbol{\sigma}$ is ordered in the "natural" way, with σ^2 being the last element, then we can define the vector of so called pseudo stratum-variance components by

$$\boldsymbol{\xi} = \boldsymbol{U}_c \boldsymbol{\sigma}.$$

Thence,

$$\text{var}\,(\boldsymbol{\xi}) = \boldsymbol{D}_c^2.$$

The diagonal elements can be manipulated to produce effective stratum degrees of freedom (Thompson[33]) viz

$$\nu_i = 2\xi_i^2 / d_{cii}^2$$

To compute the adjusted variance matrix and the terms for the scaled F-test we need to calculate the working effects matrices for the pseudo stratum-variance parameters ξ_i from the working effects matrices for σ_i. That is,

$$\begin{aligned} \boldsymbol{Q}_i^{x(\xi)} &= \frac{\partial \boldsymbol{V}}{\partial \xi_i} \boldsymbol{V}^{-1} \boldsymbol{X} \boldsymbol{\Phi} \\ &= \sum_{j=1}^{i} u_c^{ji} \boldsymbol{Q}_j^{x(\sigma)}. \end{aligned}$$

It follows immediately that

$$\boldsymbol{\Phi}_A = \boldsymbol{\Phi} + \sum_{i=1}^{r} d_{cii}^2 \boldsymbol{Q}_i^{x(\xi)\prime} \boldsymbol{P} \boldsymbol{Q}_i^{x(\xi)}$$

and similarly

$$A_1 = \sum_{i=1}^{r} d_{cii}^2 \{\text{tr}\,(\boldsymbol{\Theta} \boldsymbol{A}_{1i})\}^2,$$

$$A_2 = \sum_{i=1}^{r} d_{cii}^2 \text{tr}\,(\boldsymbol{\Theta} \boldsymbol{A}_{1i} \boldsymbol{\Theta} \boldsymbol{A}_{1i}),$$

where

$$\boldsymbol{A}_{1i} = \sum_{j=1}^{i} u_c^{ji} \frac{\partial \boldsymbol{\Phi}}{\partial \sigma_j}.$$

A similar approach can be used to reduce the computational burden for a general model by suitable diagonalisation of the AI matrix, thus avoiding the need for the double summation in Eq. [14].

7. Prediction for the general linear mixed model

Lane and Nelder[30] describe a general approach for forming predictions in general(ised) linear models. Briefly, their approach involves forming the fitted values for all combinations of the variables in the model, then taking marginal means across the variables not relevant to the current prediction. Their approach has been implemented in GENSTAT 5. Some computational limitations with the calculation of the standard errors of predicted values have been recently removed (Lane[31]). This algorithm, however, is not generally suitable for use in linear mixed models. An alternative approach, suited to the class of balanced linear mixed models with several random terms that can be analysed by ANOVA, is to replace predictions by treatment means. This approach may not be suitable for unbalanced or non-orthogonal data sets. Where random effects are present in the model, a decision must be made about how to treat these terms in prediction, and this might differ according to the purpose of a particular prediction. For correlated random effects, information on effects present in the data may be used to predict effects not present in the data set, with prediction standard errors allowing for the extra uncertainty associated with the effects not being observed. The application of this principle to the residual error gives the kriging predictions used in geostatistics.

In the following sections we briefly review the algorithm described by Gilmour *et al.*[32], which has been implemented into both ASReml and the REML directive of GENSTAT 5.

7.1. *The Prediction Model*

We define a prediction to be a linear function of the (empirical) BLUP of random effects with the (empirical) BLUE of fixed effects. A prediction is typically formed as the predicted response from an experiment for a subset of explanatory variables at given values, with the remaining explanatory variables in the model being either averaged over, ignored, or taking a specified value. Welham *et al.*[56] consider the possible roles of fixed and random model-terms in prediction and conclude that, while fixed model-terms can never be ignored, random terms may be either included (for a conditional prediction) or ignored (to obtain a marginal prediction). They

also illustrate that flexibility in the averaging process is required to allow for different weighting schemes over factors, or combinations of factors. Aliasing and nesting must be determined to ensure invariance of predictions to the parameterisation used.

7.2. *Steps in the prediction process*

Gilmour *et al.*[32] consider four main steps, which are

(1) Choosing the explanatory variable(s) and their respective values for which predictive margins are required; the variables involved are called the *classify* set.
(2) Determining which variables should be averaged over to form predictions. The values to be averaged over must also be defined for each variable; the variables involved are called the *averaging* set. The combination of the classify set with the averaging set defines a multi-way *hyper-table*. Formally, variables to be evaluated at a single specified value within the prediction, e.g. a covariate evaluated at its mean value, can be equivalently included as a member of either the classify or averaging sets.
    At this point, there may be some explanatory variables in the model that do not classify the hyper-table. These variables will normally only occur in random terms that are ignored when forming the fitted values.
(3) Determining which terms from the linear mixed model are to be used in forming predictions for each cell in the multi-way hyper-table.
(4) Choosing the weighting for forming means over each dimension (or combination of dimensions) of the hyper-table.

7.3. *Prediction process*

If we denote the vector of predictive margins by $\tilde{\boldsymbol{\pi}}$, then

$$\tilde{\boldsymbol{\pi}} = \boldsymbol{D}\tilde{\boldsymbol{\beta}},$$

where $\boldsymbol{D}$ is a $d \times (p + b)$ matrix . We require both $\tilde{\boldsymbol{\pi}}$ and the matrix of prediction errors, $\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}'$. The matrices $\boldsymbol{D}$ and $\boldsymbol{C}$ are often very large, so that it is not practical to directly compute the matrix products involved. Gilmour *et al.*[32] decomposed $\boldsymbol{D}$ into matrices which relate to the four steps described in the prediction process. That is,

$$\boldsymbol{D} = \boldsymbol{A}\boldsymbol{W}_{M}\boldsymbol{M}\boldsymbol{S}, \tag{15}$$

where $\boldsymbol{S}$ is a $r \times (p+b)$ binary matrix that selects the elements of $\boldsymbol{\beta}$ which are used to form the predictions, $\boldsymbol{M}$ is a $c \times r$ "design" matrix that forms (a portion of) the multiway hyper-table for the specified combinations of the classify set plus the averaging set, $\boldsymbol{W}_M$ is a $c \times c$ diagonal matrix of weights and $\boldsymbol{A}$ is a $d \times c$ matrix, that, when combined with $\boldsymbol{W}_M$, averages the multiway table to produce the predictive margins.

As pointed out by Lane[31], aliasing is an important problem to be aware of and keeping $\boldsymbol{A}$ and $\boldsymbol{W}_M$ separate helps to control the type of averaging of the multi-way hyper-table, which is important for problems in which aliasing has occurred. Care must be taken whenever aliasing occurs to ensure that sensible averaging occurs.

The main difference between the Gilmour *et al.* algorithm and the algorithm proposed by Lane and Nelder[30] is the presence of the matrix $\boldsymbol{S}$.

7.4. *Computing Strategy*

One of the major obstacles with the implementation of the Lane and Nelder algorithm in GENSTAT 5 has been limits on the size of the model and or the data-set for which predictions and associated standard errors can be readily obtained. Use of sparse-matrix methods, and judicious formation of $\boldsymbol{D}$ and the matrix of prediction error variances, has to a large degree ameliorated this problem whereby over 20,000 predictions can be obtained without recourse to inordinately large computers.

Ignoring most of the details of the algorithm, such as initialisation of component matrices, checks on aliasing and forming $\boldsymbol{D}$, we simply illustrate how the algorithm fits easily within the AI framework. $\boldsymbol{M}$ is the augmented mixed model matrix, given by

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{0} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{W} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D} \\ \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{D}' & \boldsymbol{C} \end{bmatrix}.$$

Absorption of $\boldsymbol{C}$ gives

$$\boldsymbol{M}^* = \begin{bmatrix} \boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} & -\tilde{\boldsymbol{\pi}}' \\ -\tilde{\boldsymbol{\pi}} & -\boldsymbol{D}\boldsymbol{C}^{-1}\boldsymbol{D}' \end{bmatrix}.$$

The absorption is performed using a reordering of the mixed model matrix, retaining a high degree of sparsity (Gilmour *et al.*[17]).

Acknowledgments

We would like to acknowledge the inspirational support of our colleague Arthur Gilmour for years of collaboration and lively debate. He is also thanked for implementation of most of what has been presented into ASReml, which has shown its utility in many circumstances. Similarly, we would like to also acknowledge Sue Welham for her collaboration and the implementation into GENSTAT 5. Some parts of Sec. 6 form the masters thesis of Sharon Neilsen and parts of Sec. 5 are from the PhD thesis of Emma Knight. The long-term financial support of the Grains Research and Development Corporation of Australia is gratefully acknowledged. Lastly, we would like to respectfully acknowledge the influence that John Nelder has had in much of what we have presented here.

References

1. Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **31**, 545–554.
2. Hartley, H. O. & Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108.
3. Nelder, J. A. (1968). The combination of information in generally balanced designs. *Journal of the Royal Statistical Society, Series B* **30**, 303–311.
4. Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* **32**, 227–230.
5. Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics,* London: Chapman and Hall.
6. Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B* **49**, 1-39.
7. Lee, Y. & Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika* **88**, 987-1006.
8. Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996). *SAS System for Mixed Models,* SAS Institute Inc.: Cary, NC.
9. Pinheiro, J. & Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*, Springer-Verlag: New York.
10. Welham, S. J. & Thompson, R. (2000). *The Guide to Genstat, Part 2: Statistics, REML analysis of mixed models,* VSN International Ltd, Wilkinson House, Jordan Hill Road, Oxford, UK.

11. Gilmour, A. R., Cullis, B. R., Welham, S. J., Gogel, B. J. & Thompson, R. (2003). ASREML, reference manual. Technical report, VSN International.
12. Butler, D. G., Cullis, B. R., Gilmour, A. R. & Gogel, B. J. (2003). SAMM, reference manual. Technical report, QPDI: Brisbane.
13. Mitzal, I. & Perez-Enciso, M. (1993). Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation maximisation. *Journal of Dairy Science* **76**, 1479–1483.
14. Smith, S. P. & Graser, K-U. (1986). Estimating variance components in a class of mixed models by restricted maximum likelihood. *Journal of Dairy Science* **69**, 1156–1165.
15. Meyer, K. (1989). Restricted maximum likelihood to estimate covariance components of animal models with several random effects using a derivative free algorithm. *Genetics, Selection and Evolution* **21**, 318–340.
16. Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.
17. Gilmour, A. R., Thompson, R. & Cullis, B. R. (1995). AI, an efficient algorithm for REML estimation in linear mixed models. *Biometrics* **51**, 1440–1450.
18. Laird, N. & Ware, J. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–975.
19. Diggle, P. J., Liang, K.-Y. & Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford: Clarendon Press.
20. Stein, M. L. (1999). *Interpolation of Spatial Data.* New York: Springer.
21. Gilmour, A. R., Cullis, B. R. & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics* **2**, 269–293.
22. Foulley, J-L., Jaffrezic, F. & Robert-Granie, C. (2000). EM-REML estimation of covariance parameters in Gaussian mixed models for longitudinal data analysis. *Genetic Selection and Evolution.* **32**, 129–141.
23. Smith, A. B., Cullis, B. R., Luckett, D. J., Hollamby, G. & Thompson, R. (2002). Exploring variety-environment data using random effects AMMI models with adjustments for spatial field trend. II Application. *Proceedings of the International Genetics and Plant Breeding Symposium, ed. Kang, M. S. Commonwealth Agricultural Bureau - International: New York.*

24. Verbyla, A. P., Cullis, B. R., Kenward, M. G. & Welham, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines (with discussion). *Applied Statistics* **48**, 269–311.
25. Rodriguez-Zas, S. I. & Southey, B. R. (2002). Linear mixed effects models for microarray gene expression data. *7th World Congress on Genetics Applied to Livestock Production.* 16:04.
26. Seaton, G. R., Haley, C. S., Knott, S. A. & Visscher, P. M. (2002). QTL Express: Rapid and user-friendly mapping of quantitative trait loci in livestock. *7th World Congress on Genetics Applied to Livestock Production.* 28:11.
27. Nelder, J. A. (1965a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and null analysis of vaiance. *Proceedings of the Royal Society Series A* **283**, 147–162.
28. Nelder, J. A. (1965b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proceedings of the Royal Society Series A* **283**, 163–178.
29. Kenward, M. G. & Roger, J. H. (1997). The precision of fixed effects estimates from restricted maximum likelihood. *Biometrics* **53**, 983–997.
30. Lane, P. W. & Nelder, J. A. (1982). Analysis of covariance and standardisation as instances of prediction. *Biometrics* **82**, 613–621.
31. Lane, P. W. (1998). Predicting from unbalanced linear or generalised linear models. In *COMPSTAT98 Proceedings in Computational Statistics.* Heidelburg: Physica-Verlag.
32. Gilmour, A. R., Cullis, B. R., Welham, S. J., Gogel, B. J. & Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis* **44**, 571-586.
33. Thompson, R. (1980). Maximum likelihood estimation of variance components. *Math. Operationsforsch Statistics, Series, Statistics* **11**, 545–561.
34. Yates, F. (1940). The recovery of inter-block information in balanced incomplete block designs. *Annuls of Eugenics* **10**, 317–325.
35. Cochran, W. G. & Cox, G. M. (1957). *Experimental Designs (2nd ed.).* New York: Wiley.
36. McCullagh, P. & Nelder, J. A. (1994). *Generalized Linear Models (2nd Ed.).* London: Chapman and Hall.
37. Smith, A., Cullis, B. R. & Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57**, 1138–1147.

38. Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15–51.
39. Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* **21**, 309–310.
40. Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the animal breeding and genetics symposium in honour of Dr. Jay L. Lush, Champaigne*, pages 10–41.
41. Hofer, A. (1998). Variance component estimation in animal breeding: a review *Journal of Animal Breeding and Genetics* **115**, 247–265.
42. Liu, C., Rubin, D. B. & Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755-770.
43. Lindstrom, M. J. & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed effects models for repeated measures data. *Journal of American Statistical Association* **83**, 1014–1022.
44. Lee, Y. & Nelder, J. A. (1996). Hierarchical generalized linear models (with Discussion). *Journal of the Royal Statistical Society, Series B* **58**, 619-678.
45. Jensen, J., Mntysaari, E. A., Madsen, P. & Thompson, R. (1997). Residual maximum likelihood estimation of (Co) variance components in multivariate mixed linear models using average information. *Indian Journal Agricultural Statistics* **49**, 215-236.
46. Foulley, J-L. & van Dyk, D. A. (2000). The PX-EM algorithm for fast stable fitting of Henderson's mixed model. *Genetic Selection and Evolution*, **32**, 143-163.
47. Callanan, T. P. & Harville, D. A. (1991). Some new algorithms for computing restricted maximum likelihood estimates of variance components. *Journal of Statistical Computation and Simulation*, **38**, 239-259.
48. Harville, D. A. (2004). Making REML computationally feasible for large data-sets: use of the Gibbs sampler. *Journal of Statistical Computation and Simulation* **74**, 135-154 .
49. Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects setting. *Biometrics* **50**, 1171–1177.
50. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. & Csaki, F., editors, *Proceedings 2nd International Symposium on Information Science, Budapest*, pages 267–281. Akademai Kiado.
51. Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 271–293.

52. Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
53. Shi, Pedde. & Tsai, C-L. (2002). Regression model selection - a residual likelihood approach. *Journal of the Royal Statistical Society, Series B* **64**, 237-252.
54. Welham, S. J. & Thompson, R. (1997). Adjusted likelihood ratio tests for fixed effects. *Journal of the Royal Statistical Society, Series B* **59**, 701-714.
55. Kackar, A. N. & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, **79**, 853-862.
56. Welham, S. J., Cullis, B. R., Gogel, B. J., Gilmour, A. R. & Thompson, R. (2004). Prediction in mixed linear models. *Australian and New Zealand Journal of Statistics* **46**, to appear.

ALGORITHMS, DATA STRUCTURES AND LANGUAGES – THE COMPUTATIONAL INGREDIENTS FOR INNOVATIVE ANALYSIS

Roger Payne
Rothamsted Research,
Harpenden,
Herts,AL5 2JQ,
United Kingdom.

John Nelder's career in statistical computing is traced with particular reference to his ideas and contributions to statistical algorithms, data structures, languages and software. Key concepts were (1) that algorithms should be comprehensive to avoid constraining the users within a restricted set of facilities constructed in a piecemeal way, (2) that output from algorithms should be able to be saved in suitable data structures so that it can be used as input to other algorithms, and (3) that the command language of any statistical system should be usable as a programming language in its own right to facilitate the efficient development of new ideas. The development of these concepts is described with particular reference to their implementation in Genstat, GLIM and GLIMPSE.

1. Context and history

Practical statistics looked very different in October 1949 when John Nelder received his first appointment as a statistician, at the newly set-up Vegetable Research Station of the Agricultural Research Council (ARC). Later this was renamed as the National Vegetable Research Station (NVRS), and it has now become Horticultural Research International, Wellesbourne. Analyses were performed using mechanical calculators, operated by hand or perhaps (if you were lucky) with the help of an electric motor – the sprained wrist must have been the precursor of Repetitive Strain Injury! Nevertheless, some fairly sophisticated analyses were possible. For example, Yates (1937) described methods for some of the experimental designs routinely analysed in this way, including 2^n designs with confounding be-

tween treatment interactions and blocks, split plots, and quasi-factorials. Similarly, Finney (1947) showed how to fit probit lines and probit planes by hand. However, data sets were relatively small, models were fairly simple, and the pressure of routine analysis must have severely limited the opportunities for innovation. As Frank Yates noted in the Rothamsted Annual Report for 1952:

> "The analytical work has again involved a very considerable computing effort. " (sic).

1.1. *Rothamsted*

The buildings at Wellesbourne had not been completed in 1949. So John worked initially at Rothamsted (another ARC station), in the Statistics Department then headed by Frank Yates. This began John's association with Rothamsted, which continued after his arrival at Wellesbourne. Some of his collaborations are recorded by Yates in the Rothamsted Annual Reports. For example, in 1957 John was investigating the use of the Newton-Raphson iterative procedure for solving least-square equations with S. Lipton (see 1957 Rothamsted Annual Report). The context was the fitting of logistic curves, foreshadowing John's development of his generalized logistic curve (Nelder 1962). In 1959 Howard Simpson programmed the Jackson method for analysis of capture-recapture data for John – foreshadowing their long collaboration on Genstat. In 1960 G.W. Bonsall used John's method to estimate components of variance in genetical data and, in 1961, collaborated with him and with members of Birmingham University in the application of improved techniques to estimate genetical parameters in experiments on *Drosiphila* and tomatoes. In 1962 John collaborated with Rothamsted on the automatic recording of data using the "Port-a-punch" system, and the 1964 Rothamsted Annual Report noted that NVRS statisticians

> "do much of their own programming and tape preparation and cooperate with us in the writing of general programs"

The Rothamsted association became more formal in 1968, when John took over as Head of the Statistics Department following Yates' retirement. At that point, Rothamsted was anticipating the arrival of its fourth computer, an ICL 4-70. The first Rothamsted computer, an Elliot 401, had been obtained in 1954 and can be claimed to be the first computer to be associated primarily with agricultural research and with statistics. An Elliot 402 was added in 1961, before both were replaced by a Ferranti Orion

in December 1963. These computers were available to other ARC stations, but this must have been very inconvenient, even after the addition of a telex link to NVRS in 1967. The 401 and 402 supported only one user at a time, programs were written in machine code, and considerable ingenuity had to be used to make efficient use of the rotating disk that stored data and programs during execution. The aim was to ensure that successive instructions passed the reading head at just the right moment to avoid a wasted rotation of the disk. Nevertheless, many analysis programs were developed. The 1964 Rothamsted Annual Report noted that, in their final year of operation, the 401 and 402 analysed 14,357 variates (or analysis variables). However, this took a total of 4,731 hours! The Orion was much more efficient, and provided an early computing language called Extended Mercury Autocode. Facilities provided by the programs developed in EMA included analysis of 2^n or 3^3 experiments and of factorial designs, multiple regression, and analysis of surveys. The Orion could run more than one program at a time but, in practice, this was of limited benefit as most programs required all the available disk space.

The ICL 4-70 was delivered to Rothamsted in July 1970 and became usable by the Statistics Department in November. In the interim, programs were developed on an IBM 360 computer at Edinburgh Regional Computing Centre, accessed through a telex link. The 4-70 was a major step forward, and provided many of the computing facilities that we now take for granted. It supported a high-level programming language (Fortran), allowed programs and data to be stored on disk, and even provided a rather slow interactive mode of use (known as RIRO, standing for roll-in-roll-out). The operating system, Multijob, also enabled larger programs and data sets to be handled by "overlaying" subroutines and common blocks that would not all be needed at once, although it was not until 1971 that this was operating reliably.

So John arrived at Rothamsted at a time when many new opportunities were opening up for statistical computing. In particular, it was now feasible to develop general-purpose statistical systems. John had prepared his ideas for just such a system during his time at NVRS and also, importantly, on a visit to the Waite Institute of the University of Adelaide in 1965-1966 where he was able to work with Graham Wilkinson of the CSIRO Division of Mathematical Statistics. An additional bonus of this trip was that, on the boat to Australia, he was able to write a program to implement the Nelder-Mead simplex algorithm for optimization, which is now a citation classic (Nelder & Mead 1965). Furthermore, within the Rothamsted Department,

there was considerable expertise that could be deployed, and several existing statistical programs to provide an additional source of algorithms and ideas (see Gower 1986). The result eventually became the system now known as GenStat *for Windows*; see Section 4.1.

John's other statistical computing interests during his time at Rothamsted included the Royal Statistical Society's Working Party on Statistical Computing (Section 1.2), and the development of the program GLIM (Section 4.2). He retired from Rothamsted in October 1984, but has remained a regular and welcome visitor, and has retained a continuing interest in Genstat.

1.2. *Working Party on Statistical Computing*

Another important thread in John's career began on 15th December 1966 when, with Brian Cooper, he organised a meeting at the Atlas Computer Laboratory, Chilton, on *Statistical Programming – the Present Situation and Future Prospects.* The organisers believed that many good ideas had been developed amongst a diverse set of systems, but there had been much duplication of effort and the systems were increasingly incompatible with each other. With the arrival of a new generation of computers which, for example, might facilitate conversational use, they felt that it might be opportune to look closely at progress and prospects for statistical computing. John circulated a memorandum setting out some thoughts on the possibilities of a statistical language using standardised data structures and instruction formats. This received an encouraging response, which led Brian Cooper to suggest the meeting. The participants, 41 in all, included statisticians, programmers and representatives of computer manufacturers, and the talks described current work at Rothamsted (Gower, Simpson & Martin 1967), the Atlas Computer Laboratory (Cooper 1967), University of Lancaster (Colin 1967), the Meteorological Office (Craddock & Freeman 1967), and in the USA (Chambers 1967). In the discussion John revealed that he had not volunteered to discuss Genstat as he had recently spoken about the system to the Biometric Society.

The discussion contains many threads that would not be out of place today, for example concern about potential misuses of statistical systems and methods e.g. multiple regression ("one of the most used, and misused of standard statistical programs": Page 1967). Others have been overtaken by time. For example, it was felt by several contributors that conversational modes of use could be impracticable because pressure of use would preclude

the necessary time for thought. However, Vickers (1967) pointed out that

> "Clearly if those of us around this table were presented with a console most of us would spend too much time racking our brains, and activity would be pretty slight. But let us not forget that the sixth-form boy who sees a console today and then goes on to a degree in statistics and computation will be a very fine console-operator in five years time. Such people, who can react to consoles quickly, will be with us soon."

The graduates of 1971 would agree, although console-operation may have proved more of a means within their careers than an end in itself! More tellingly, Meier (1967) predicted that consoles would soon become sufficiently cheap for them to be left unused (if necessary) until needed.

At the end of the meeting Ewan Page proposed, and Michael Healy seconded, a proposal that a working party be set up. David Finney, who had chaired the discussion, called for nominations and as a result the Working Party on Statistical Computing was formed with the following members: John Nelder (Chairman), Brian Cooper (Secretary), James Craddock, Ewan Page, John Gower and Michael Healy.

The first activity of the Working Party was to establish and maintain a statistical algorithms section in the RSS journal *Applied Statistics*. The initial aim seems to have been to support good computing practice by providing implementations of the basic building blocks of a statistical program. Later very much more complicated techniques were added, and the publication of an algorithm for a new piece of methodology became an equally valid (and perhaps more effective) way of registering a new idea. The section was instigated in 1968, when John was one of the two joint editors of Applied Statistics. An editorial announcement *Statistical algorithms* appeared in issue 1 of the 1968 volume, the first seven algorithms appeared in issue 2, and a further five in issue 3. By 1971, the volume of work justified the establishment of an explicit Algorithms Editor on the Applied Statistics Editorial Board. John took the title for that year (giving up his joint-editorship), before passing the role to David Hill who continued until 1976. Subsequent editors were Howard Simpson (1977–1979), Paul Griffiths (1979–1984), Patrick Royston (1985–1987), Janet Webb (1985–1988), David Muxworthy (1988–1991), Peter Fayers (1988–1991), Carl O'Brien (1991–1997), and Nick MacLaren (1992–1994).

John was very keen to support the algorithms section, personally and through his NVRS and Rothamsted colleagues. By the time the activity

ceased, in 1997, 321 algorithms had been published, over 9% of which originated from John and his immediate colleagues. The other major activity of the Working Party was the program GLIM , which is described in Section 4.2. John resigned as Chairman of the Working Party in 1984, at the time of his retirement from Rothamsted.

1.3. *Imperial College*

John was first appointed as a Visiting Professor at Imperial College (or, to use its formal title: Imperial College of Science, Technology, & Medicine, London) in 1972. Following his retirement from Rothamsted, this became his main centre of operations. The first of these was the GLIMPSE project, described in Section 4.3. Later projects included his Genstat-based *K*, *MD* and *HG* systems (Section 4.3), and his associated and continuing research with Youngjo Lee on hierarchical generalized linear models (Section 3.3).

2. Data structures

Right from the outset, John regarded the existence of suitable data structures as fundamental to the design of a successful statistical system. He introduced Genstat in the 1968 Rothamsted Annual Report as follows:

> "The system under construction (for GENeral STATistical work) is based on the idea that the easiest way to ensure compatibility of different programming modules is to define the basic data structures on which they operate. Some of these structures can be seen in raw experimental data, whereas others are created during analysis. At any stage, the analysing program will have assembled a set of these structures in the core-store of the machine."

The initial Genstat system supported the data matrix (as a collection of columns with equal lengths), three types of matrix for matrix arithmetic (rectangular, symmetric and diagonal), the multi-way table, the scalar, the textual string, a special structure to store latent roots and vectors, and a structure to refer to sets of other structures which became known as the *pointer*. In the 1969 Report he noted that

> "All GENSTAT programs use these standard structures as input and produce others of the same kind as output; their use thus gives the automatic compatibility between programs that is so important if the system is to be easy to use.".

A key requirement is the ability to save output from algorithms in data structures so that it can be used as input to other algorithms. Nelder (1974) explained why this was essential for multi-stage analyses.

> "Almost everyone accepts the idea that the analysis of a substantial body of data must be a multi-stage process of trial and error. Our models can only be tentative, may be contradicted by the data themselves, and will usually need revision. But although everyone may accept this principle, far too many people accept computer programs that, far from encouraging a multi-stage approach, actively foster the single-stage procedure
>
> input
> analyse
> output
> stop.
>
> Those who do not wish to think may welcome the magic black box that tabulates their data in all possible ways, does all possible regressions, calculates all possible statistics, prints the result on several kilometres of paper and then stops. The intelligent user will reject this approach, but he needs to recognise the symptoms that a program is based implicitly on a single-stage view of analysis. The critical question is – does the system allow the output from a procedure to be named as instances of suitable data structures, and used subsequently as input to other procedures? If the answer is no then, real multi-stage analysis is going to be at best difficult and at worst impossible. Thus facilities for naming output and the definition of a common set of structures for both input and output are key features of a system supporting multi-stage analysis. A further essential feature is of course the ability to save, and subsequently to retrieve, not just the raw data but data structures derived from them."

Payne & Nelder (1976) extended this campaign by examining ten popular statistics packages to see what types of data structure were provided, and how these could be used. Again the focus was on the ability to link algorithms together through their data structures. In their introduction they stated

> "The availability of a particular structure, although not absolutely essential for the feedback of information from an algorithm into

> the system, very much affects the extent to which this is provided. For example packages without scalar structures seem reluctant to allow the user internal access to scalar quantities like a regression coefficient (see 1.3). This internal access to output structures is crucial if the user is not to be constrained within the (relatively small) set of problems that have been programmed by the package designers. For example: the regression algorithm in a package may simply take sets of x- and y-variates, perform the regression, print the results and plot the regression lines and residuals. This would provide a rather restricted analysis. If however the algorithm, allows the regression coefficients, residuals and fitted values to be accessed by the graphical and transformation algorithms, different sets of lines can be plotted in the same frame, residuals can be plotted against other possible x-variates, or as a histogram etc. Furthermore, if weighted regression is available, the way is open to exploit techniques based on iterative weighted least squares, using the fitted values from one iteration to define the weights used in the next. This procedure is, of course, greatly simplified if macro and looping facilities are provided."

It was perhaps no surprise to find that Genstat came out top in the comparison, but the aim of the paper was more to raise standards than to promote Genstat (which, at that stage was not being marketed commercially). The conclusion stated

> "The purpose of this paper has not been to recommend any particular package as a 'best buy' (although some packages are clearly more limited than others) but to suggest other ways of evaluating packages apart from examining the algorithms they provide, and to highlight areas for improvement. Two points seem worth repeating here. First, the use of the data matrix as the basic data structure is rather inflexible – it forces all variates to be of the same length, precludes the use of scalar structures and the provision of general calculation facilities involving structures of different types.
> Secondly, if a package is to be useful not only to the routine user, but also to the user who wishes to develop new forms of analysis – to use the package as a research tool rather than as a standard black box – it is important to allow output from algorithms to be copied into named data structures of a form recognised by other algorithms in the package. There has certainly been much misuse of

> packages by unthinking users, and it is important that originators should not encourage rigidity by making their packages difficult to use flexibly."

Now, over 25 years later, it should seem surprising that this was ever a controversial view, but a repeat of the same survey would no doubt find that some packages are still far from flexible in this respect. However, many users demand flexibility, and they have a range of systems such as Splus and, of course, Genstat itself able to satisfy their needs (see Section 4.1).

3. Algorithms

Implicit in John's work in Genstat and GLIM, and stated explicitly by Nelder & Payne (1991), is the aim that algorithms should be comprehensive. In their discussion on statistical features, Nelder & Payne (1991) state that

> "One way in which a package designer can encourage flexibility is to select algorithms that are as general and comprehensive as possible. For example, analysis of designed experiments is often broken up into separate programs for a one-way classification, two-way classification, Latin square, and so on; this piecemeal approach tends to foster the attitude whereby users think of designing an experiment as being merely a process of selecting from a limited set of alternatives, instead of the more challenging – and rewarding – task of devising the arrangement that most effectively meets the requirements of the investigation. Thus, the analysis-of-variance algorithm in Genstat [9, 13] {i.e. Payne & Wilkinson 1977, Wilkinson 1970} handles the complete class of generally balanced designs [4, 5] {Nelder 1965}, and these cover virtually all the standard experimental designs: Genstat is distributed with a data file that shows how it can analyse all the worked examples in the standard text of Cochran and Cox [1] {1957}."

In fact, in Genstat *for Windows* you can enjoy the best aspects of both the general and the piecemeal approaches, because the general algorithms are used by special-purpose procedures – and now menus – that provide custom interfaces to the simple analyses, such as one-way anova. However, there are also menus that allow the full power of the algorithms to be invoked by those with wider needs and knowledge.

3.1. *Analysis of variance*

Design and analysis of experiments was the first area that John studied. His concept of general balance (Nelder 1965) brought virtually all of the traditional experimental designs into a single framework, and facilitated their analysis by a single algorithm. In this case, the algorithm was developed by Graham Wilkinson, initially at the Waite Institute and at CSIRO in Adelaide Australia; then later (from 1971–1975) it was revised extensively at Rothamsted in collaboration first with Charlie Rogers, and then with Roger Payne (i.e. this author) – who took over the subsequent development after Wilkinson left Rothamsted (see e.g. Payne & Wilkinson 1977, Payne & Tobias 1992, Payne 1998). Wilkinson (1970) noted that the algorithm was derived partly from the theory of general balance of Nelder (1965), and its ability to produce a correct and complete analysis when there is more than one *block* (or *error*) term is an aspect that, even now, is a strong advantage of Genstat compared to other statistical systems.

When there are several block terms, the total sum of squares can be partitioned up into components known as *strata*, one for each block term. Each stratum contains the sum of squares for the treatment terms estimated between the units of that stratum, and a residual representing the random variability of those units. The properties of a generally balanced design are that (i) the block terms are mutually orthogonal, (ii) the treatment terms are also mutually orthogonal, and (iii) the contrasts of each treatment term all have equal efficiency factors in each of the strata where they are estimated. The mathematics underlying the Genstat analysis algorithm (Wilkinson 1970, James & Wilkinson 1971) are not straightforward to relate to those of general balance. In fact, the class of designs that the Genstat algorithm can analyse (known as designs with *first-order balance*) include some that are not generally balanced. The precise connection between the two classes was made by Payne & Tobias (1992). First-order balance does not require an orthogonal block structure (condition i), but Payne & Tobias (1992) showed that a first-order balanced design that does have an orthogonal block structure will also be generally balanced.

Implicit in the definition of general balance is the idea that treatment effects may be estimated in more than one stratum. The simplest example is the balanced incomplete block design, where there are treatment estimates both between blocks, and between the plots within blocks. Under these circumstances, it is advantageous to present treatment estimates that combine the information from each of the strata where the treatment is

estimated. Nelder (1969) showed how to do this in the analysis of variance of a generally-balanced design. Payne & Tobias (1992) extended this to analysis of covariance, and Payne added this facility to the Genstat anova algorithm (Payne *et al.* 1993). Payne & Tobias (1992) noted that it was not known how to calculate effective degrees of freedom for the standard errors of the combined effects. This problem has now been solved. Effective degrees of freedom are now provided by Release 7.1 of Genstat (Payne *et al.* 2003), and studies are under way to see how well the combined effects can be assessed by t-distributions with that number of degrees of freedom.

General balance, together with John's other contributions to experimental design, is described in more detail in Chapter 9.

3.2. *Generalized linear models*

The development of generalized linear models , by Nelder and Wedderburn (1972), now provides one of the most useful techniques in any data-analyst's tool kit. The discussion here is concerned more with its implications for statistical computing.

Iteratively reweighted least squares was in use prior to that paper, notably for probit analysis (Finney 1947). However, the focus on bioassay (and, originally, on hand calculation) seems to have meant that only a few types of model were considered: those with a single dose variate i.e. probit lines; those with a variate (dose) and a factor (type of drug) where the aim is to compare new drugs with a standard; and those with two variates (e.g. quantity and concentration of a drug) i.e. probit planes. Essentially, this provided another piecemeal approach. Nelder & Wedderburn (1972) not only expanded the range of models covered by the framework. They also emphasised that the algorithm was fundamentally a regression algorithm, and should thus be able to fit any of the models that would be considered in that, more familiar situation. The regression analogy allows non-statisticians to use the methodology with confidence, building on their regression experience. For example, the Genstat Guide (Payne *et al.* 2003, Part 2 page 218) states that:

> Generalized linear models extend the ordinary regression framework to situations where the data do not follow a Normal distribution, or where a transformation (known as the *link function*) needs to be applied before a linear model can be fitted.

So, the potential user needs to be aware that data may not follow a normal

distribution and that, for example, proportions may arise from binomial distributions and counts from Poisson distributions. This should not be a problem: for example, both of these distributions are now taught in schools in Britain, and no doubt in many other countries. The user also needs to realise that relationships may be nonlinear, again a relatively straightforward idea to communicate – and an essential one when the response variable is constrained to be non-negative (as with counts).

The implementation of the generalized linear models algorithm, initially (from 1974) in GLIM and then (from 1976) in Genstat, led to dramatic improvement in the quality of statistical analysis allowing unsatisfactory approximate analyses, such as those involving the angular transformation of percentage data, to be discarded.

3.3. *Hierarchical generalized linear models*

During the 1980's interest grew about ways of incorporating extra error terms into generalized linear models. The paper by Schall (1991) established *generalized linear mixed models*, in which there were additional normally-distributed terms in the linear part of the generalized linear model, as a standard. (Note, though, that there had been earlier, related, work by Thompson 1979.) By that time, John had completed his work on GLIMPSE (Section 4.3), and was again researching into generalized linear models with Genstat as his computing environment (Section 4.4). At the Genstat Conference, at University of Kent in Canterbury in July 1993, he spoke on *Extending the frontiers of generalized linear models.* The final section of the talk described *conjugate GLMMs*, in which the random components are assumed to follow the conjugate distribution to that of the GLM errors (Nelder 1993). By the time of the Genstat conference at Wagga Wagga in November 1994, these had become *double generalized linear models* (Nelder & Lee 1994). The theory allowed for two-stage models where both between- and within-group components came from exponential families, and included both generalized linear mixed models and conjugate double generalized linear models. Finally, by the time of the Genstat Conference at University College, Dublin in July 1995 (Nelder 1995) and at the Royal Statistical Society in December 1995 (Lee & Nelder 1996), the full theory of *hierarchical generalized linear models* had emerged.

From the point of view of statistical computing, the models offer the huge advantage of a unifying approach, bringing together a wide range of models within a single framework. As with the other general algorithms

described in this section, this empowers users, giving them the flexibility to select a good model, rather than the one that is least inappropriate. The models are fitted by a computationally efficient algorithm. So users are encouraged to analyse their data interactively, and to find the right model rather than simply stopping with relief when they find one that converges. Finally, the fact that a hierarchical generalized linear model can be defined from two interlinked generalized linear models (Lee & Nelder 2001) means that we have access to a familiar repertoire of model checking techniques, and can thus base our choice of error distributions on the data rather than on prejudice or software limitations. The theory of hierarchical generalized linear models is described in Chapter 10, and their implementation in Genstat (Payne *et al.* 2003) is described in Section 4.4.

4. Statistical software

From the earlier sections of this chapter, it will be apparent that John's interests in statistical methodology have always been aimed at generating implementations of the new theory, to enable the methods to be used in practice. John's work has resulted in two very widely used systems, Genstat and GLIM. This section traces his statistical computing career, starting with Genstat, continuing with GLIM, then investigating expert systems through GLIMPSE, and finally returning to Genstat as his environment for implementing, investigating and distributing new algorithms and systems.

4.1. *Genstat*

The ideas underlying Genstat took shape during John's visit to the Waite Institute of the University of Adelaide in 1965–1966. The first program called Genstat was produced there during 1966 by John with Graham Wilkinson, who was then the acting Reader of Biometry at the Waite on secondment from CSIRO Division of Mathematical Statistics. The main features were an early version of Wilkinson's analysis-of-variance algorithm (Section 3.1), and some facilities for multiple regression. However, development stalled following John's return to Wellesbourne, and Graham's return to CSIRO – who were unwilling to support the project further. John's appointment at Rothamsted in 1968 rescued the project, and provided access to the many additional ideas and the extensive expertise of the Rothamsted Statistics Department. In 1970 Graham Wilkinson joined Rothamsted too, and the first release of the new Genstat appeared. Additional facilities included principal component and principal coordinate analysis, single-

linkage cluster analysis and facilities for general calculations on vectors, matrices and tables. So Genstat was already well on the way to becoming an authentically general-purpose statistical system.

By the second major release, in 1972, the Genstat language was sufficiently general to implement new methods, and was used for example by Colin Banfield at Rothamsted to program canonical correlation analysis. In the third major release (1973), the distribution of these programs was simplified by the inclusion of a macro facility. A standard Library of macros was collected and released in 1975. The initial Library had just six macros. However, by 1983 it contained 35 macros, and there were a further 16 macros in an additional library contributed by CSIRO (see Alvey *et al.* 1983). To facilitate the efficient execution of these programs, and to allow for the use of labels and jumps, Genstat programs ran through two phases: first a block of commands was compiled into an internal form, then it was executed. Some of the inefficiencies of interpreted languages were thus avoided. However, as we shall see later, this greatly complicated interactive use. Two other major landmarks, associated especially with John, were the extension of the regression algorithm in 1976 (Release 3.09) to cover generalized linear models, and the inclusion in 1982 (Release 4.04) of facilities to generate tables of predictions from generalized linear models (see Lane & Nelder 1982).

The first interactive use of Genstat took place at Edinburgh in 1975 on an ICL 4-75 with a paged operating system. This prompted John to note that

> "interactive use, even of such a large program as Genstat, is possible and efficient, given a suitable operating system and software"

(Nelder 1976). The ICL 4-70 at Rothamsted was too limited to support widespread use of an interactive Genstat, although a version was prepared that required less than 200 "store units" on the 4-70 (about 100k bytes) to run. Reasonable response times could be obtained outside the core Rothamsted working day, but in the middle of the working day there was plenty of time to think between command prompts (c.f. Section 1.2!). However, interactive use was complicated by the two phases of execution: for example, to execute commands one at a time, each one had to be accompanied by a 'RUN' command. Also output had been designed to exploit the spaciousness of a printer, and was all in upper case. So, in 1983, with the opportunity for true interactivity becoming available though the replacement of the ICL-470 by DEC VAX computers, it was time to consider a

redesign of Genstat.

Further incentives for change were the complications and incompatibilities that had built up as the system had expanded, resulting in a plethora of different concepts being used across the language. Furthermore, the macro facility was proving very complicated to program and use. Information was passed into each macro using global data structures (and, in fact, all the data structures involved in the macro program were taken to be global unless they were explicitly defined as local). So, in effect, the "parameters" of the macro had fixed names, and there were no defaults. The macros had to be recovered from Genstat's backing-store files explicitly, and complex programming was done using labels and jumps rather than the more recent and more reliable mechanisms of structured programming (e.g. if-then-else and case).

So, during 1983 with John's advice and support, a specification was prepared for a radically revised version, Genstat 5 (Payne 1983, Payne & Lane 1986, Payne *et al.* 1987). This would have a redesigned, simpler and consistent, syntax which would be easier to learn and remember. It would support structured programming and the compile/run distinction would be abolished; the only time that a command would not be executed immediately would be when a loop of commands was being defined. The macro would be replaced by a new *procedure* structure. Procedures would have an identical syntax for use to that of the standard Genstat commands, and they would be accessed automatically from libraries as required. The facilities for pointers (see Section 2) would be extended to support the ideas of Lamacraft & Payne (1980) that statistical data often need to be represented in compound and hierarchical structures. There would be a general text vector, and facilities for manipulation of text (editing, concatenation &c.). Finally facilities were added to plot graphs in high resolution, and to fit a range of standard curves.

The first release of Genstat 5 took place in 1985, in time for the Genstat Conference at University of York, and in the ensuing years it has fulfilled its promise to be accessible to novice users while still providing an environment for statistical experts to develop their own ideas. John's HG-system (Section 4.4) provides a good example of Genstat's use in research, and there are plenty of others in the manual that describes the wide variety of procedures now in the official Genstat Library (Payne & Arnold 2003). These days, statistical analysis seems to take place mainly on PC's running MS Windows, and there use has become even more straightforward with the development of seven "Editions" of "GenStat *for Windows*" during

1996 – 2003 (see e.g. Baird *et al.* 2003). Now, as Section 4.4 illustrates, new research can be devised, implemented and evaluated using Genstat commands, distributed in procedures in the Genstat Library, and then made accessible even to those with little programming expertise by extensions to Genstat's menu interface.

Initially Genstat was distributed by Rothamsted, but in 1979 (when there were about 100 licenses in 20 countries) distribution was taken over by the Numerical Algorithms Group, Oxford. Now, in 2004, Genstat is developed and marketed by VSN International, a company set up for that purpose by Rothamsted and the Numerical Algorithms Group.

4.2. *GLIM*

By 1972 the Algorithms Section of Applied Statistics was running smoothly, and the Working Party on Statistical Computing was ready for another challenge. In meetings in January of that year, John suggested that they should collaborate on an interactive program to implement the class of generalized linear models about to be published by Nelder & Wedderburn (1972). The system drew on John's Genstat experiences, for example in its use of the same syntax for specifying models (see Wilkinson & Rogers 1973). However, it had a simplified language, that aimed to support an interactive style of use. The commands (again known as *directives*) had fewer operands, and did not for example have the dichotomy of "'options" and "parameters" of the Genstat syntax. GLIM also supported a much more limited set of data structures. In fact the first release provided only a single data matrix, represented by individual columns of factors or variates. So all vectors had to be of the same length. A further restriction was that the factor levels could be represented only by the integers 1 upwards.

GLIM 1 appeared in 1974 (see Nelder 1975), and was distributed on behalf of the Royal Statistical Society by the Numerical Algorithms Group to about 50 sites on about 10 different ranges of computer (see Richardson & Baker 1980). This was followed in 1975 by GLIM 2, which reached about 130 sites. Notable improvements were the macro facility, and the inclusion of 26 scalar structures named %A, %B,... %Z, and the inclusion of offsets into the models. In addition to John himself, the main contributors to these first two releases, were Michael Clarke, David Hill, Charlie Rogers and Robert Wedderburn up to the time of his tragic death in June 1975.

GLIM 3 was a major step forward, which took until 1978 to appear. A lesson was (re)learned about data structures in that vectors were now

no longer constrained all to have the same length. Also suffices were introduced, to allow subsets of vectors to be defined. Model fitting was simplified by removal of the requirement to specify the maximal model (i.e. the largest model that might be considered) before any fitting took place. Other improvements in the modelling facilities included the ability for users to specify their own links and error distributions, thus enhancing its use as a research tool. There were also additional functions and enhancements to the programming facilities. The main contributors were Bob Baker, Michael Clarke and John himself. By 1980 GLIM 3 had been distributed to nearly 300 sites, and it had had an immense influence on the new generation of practical statisticians. For many it provided their first experience of analysing data interactively. It encouraged them to think about each data set, instead of directing it at a black box with a request for "statistics all". It provided opportunities to investigate a rich set of models, and good diagnostics (including plots) to assess which one would be most appropriate. The macro library also provided opportunities to develop and distribute new techniques, although considerable ingenuity (and very convoluted coding) was often needed to overcome GLIM 3's limited set of data structures and operations (e.g. no matrices nor tables, and no matrix nor table arithmetic).

The popularity of the interactive approach led to an investigation from 1978 onwards of the possibility of putting the Genstat analysis of variance algorithm into the GLIM framework (see Payne 1982). Other potential new modules for GLIM included a tabulator and calculator for arrays (Green 1982), and facilities for displaying graphics on high-resolution devices (Slater 1982). An unsuccessful attempt was made to combine these all into a single system, to be known as PRISM (Baker 1982). However, the name PRISM had to be abandoned because of trademark clashes, the project itself failed to gel, and the collaboration between Rothamsted and the Working Party on Statistical Computing came to a close in 1984 when John retired as Head of the Rothamsted Statistics Department and Chairman of the Working Party. So instead GLIM 3.77 appeared in 1985, and the Rothamsted focus returned to Genstat and its reincarnation as a fully interactive system in Genstat 5 (Section 4.1).

4.3. *GLIMPSE*

In 1984, when John retired from Rothamsted, it seemed clear that Expert Systems would be the next big idea in statistical computing. For example,

no topic attracted more papers in the proceedings of the COMPSTAT 1986 conference (see Antoni, Lauro & Rizzi 1986), or generated more challenging aspirations. The British Government was willing to support research in this area, through its *Alvey* programme – and one of these grants was to Imperial College and to the Numerical Algorithms Group to develop the GLIMPSE system (O'Brien 1989, Nelder 1991).

GLIMPSE was developed by John, with Carl O'Brien at the Numerical Algorithms Group and David Wolstenholme at Imperial College, to provide a knowledge-based front-end for GLIM. GLIM was a good candidate for such a project. It was a rather minimalist program, offering little on-line help on the syntax and no assistance on modelling strategy other than diagnostics when fitting was unsuccessful. However, it did focus on a clearly defined set of models, and provided language facilities such as macros that the front-end could use for its communications. The project set itself the realistic goal of building a system able to advise on data validation, data exploration, model selection and model selection within GLIM. It was also realistic about the potential users, realising that they might sometimes want to ignore the advice.

The system had three main components, all written in the computing language Prolog: the User Interface, the Translator and the Abstract Statistician. The Interface managed the communications between GLIMPSE and the user, and included the facility for them to ask each other questions. The questions from GLIMPSE included not only the obvious categories such as a query-the-user facility to obtain the necessary information about the dataset. It might also ask the user's own opinion, for example, as to whether a plotted relationship could be taken to be linear. Of course, the user could then reply "what do you think?", but this still leads to a less authoritarian approach, and one in which the user might learn as well as observe a strategy. The user questions also supported this approach, and included: "What answers are possible?" (e.g. is "don't-know" an option..); "What does that mean?" (connecting to explanations of terminology in the help system); "What is the background to that question?" (again linking to explanations in the on-line help); "Why is the question being asked?" (resulting in an explanation of how the question relates to the Prolog rules that are being followed); and "What was the question again?" (to repeat the question after all this help and explanation!).

The key statistical component was the Abstract Statistician, which aimed to embody the statistical expertise of the team as a set of Prolog rules. To execute the rules the user was asked for information, through the

User Interface. Once sufficient information had been obtained, tasks were generated by the Translator for execution by GLIM. The GLIM output was absorbed back into the front-end, so that it could be presented to the user, or used to draw conclusions to decide on the next stage of the analysis. The conclusions could be drawn by GLIMPSE, but users also had the ability to depart from the Abstract Statistician's strategy and pursue their own approach instead. Of course the price of liberty was the ability to generate an analysis that might be statistical nonsense. However, it also meant that there would be no barriers to the investigation of new research.

In practice GLIMPSE became more a tool for experts than a tool to provide expert help to novices. Its long-term effect on statistical computing has been perhaps more through the influence of its ideas of modelling strategy and model checking, for example in John's later modelling systems (Section 4.4) and within the menus of GenStat *for Windows*, than through its own wide-spread use. However, it contained many very interesting and far-sighted ideas and, when it was released in 1989, it was one of the first statistical expert systems to be made available commercially – and perhaps one of the few to deliver what the originators had promised.

4.4. *The K-, MD- and HG-systems*

By the end of the 1980's, the transformation of Genstat into the fully-interactive and easily extendable system, Genstat 5, was complete (see Payne *et al.* 1987). Following the GLIMPSE project, John was keen to find a powerful framework, with sympathetic developers and a reliable future, in which to investigate and implement his future research. Genstat seemed to fulfil all these criteria. Furthermore, the ability that it provided for users to customize their environment, through procedure libraries and commands to control aspects like the mode of execution and the style of graphics, enabled him to adapt it to his own requirements. In particular, it allowed him to achieve the best of both worlds, by building a system within Genstat known as the K-system (Nelder 1993) to provide the features that he most missed from GLIM.

The first requirement was to provide the ability to set the individual aspects of the model without having to (re)specify a complete Genstat MODEL command. The second requirement was to duplicate the GLIM feature whereby output components are saved automatically in data structures with standard names (such as %FV for the fitted values or %DV for the deviance). Within Genstat, users are required to save these components

explicitly and define their own names, using commands such as RKEEP and AKEEP. This gives the user more freedom, and avoids confusion amongst the richer set of models that Genstat supports. However, the GLIM convention saves time for users who are interested only in generalized linear models and who are willing to accept standardization. A final requirement was to provide a very succinct set of commands to minimize typing.

Tools to assist with the first two requirements were provided in Genstat 5 Release 3 by adding a new command called WORKSPACE. This allowed *workspaces* of data structures to be defined and then accessed anywhere within a program. Another extension allowed the workspaces defined within a procedure to refer to data structures within the main program. So the K-system could define a workspace to record all the aspects of a model, and use this to fill in a complete Genstat MODEL command every time an aspect needed to be changed. Other workspaces were defined to contain the various output components (with their standard names). Workspaces have proved to be a very powerful concept particularly for problems that require suites of procedures (including the more recent MD- and HG-systems). They can be used to store and transmit working variables and status information, safely out of sight of the user.

The K-system, with its 71 procedures and 6 workspaces, supports a very interactive approach to model fitting with a strong focus on model validation and checking. It has also provided a foundation upon which John Nelder and Youngjo Lee have built two additional systems: the MD-system for joint modelling of mean and dispersion, and the HG-system for fitting hierarchical generalized linear models (see Section 3.3 and Chapter 10). The HG-system has provided a test base for the theory as it has evolved. Thus the power and flexibility of Genstat as a programming environment has played a key role to enable the theory to be developed. In the 6th Edition of Genstat *for Windows* additional procedures were implemented (Lee, Nelder & Payne 2002) to provide access to the hglm algorithms, independently to the K-system. At the same time, a menu interface was included. So hierarchical generalized linear models have now become part of the standard set of analyses available to any user.

5. Conclusion

In the course of one career, practical statistics has gone from an onerous and expert task to a task that can be handled by novices – for whom the expert can quickly develop and disseminate new ideas. Many of the

necessary developments have been due to John himself, or to the example and inspiration that he has provided to those of us that have had the good fortune to work with him.

Acknowledgments

Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom

References

Alvey, N.G., Banfield, C.F., Baxter, R.I., Gower, J.C., Krzanowski, W.J., Lane, P.W., Leech, Penelope K., Nelder, J.A., Payne, R.W., Phelps, Kathleen M., Rogers, C.E., Ross, G.J.S., Simpson, H.R., Todd, A.D., Wedderburn, R.W.M. & Wilkinson, G.N. (1983). *Genstat A General Statistical Program (Release 4.03).* Oxford: Numerical Algorithms Group.

Baird, D.B., Harding, S.A., Lane, P.W., Murray, D.A., Payne, R.W. & Soutar, D.M. (2003). *Genstat for Windows (7th Edition) Introduction.* Oxford: VSN International.

Baker, R.J. (1982). PRISM – an overview. In: *Lecture Notes in Statistics, No. 14. GLIM 82: Proceedings of the International Conference on Generalised Linear Models* (ed. R. Gilchrist), pp. 3–24. New York: Springer-Verlag.

Chambers, J.M. (1967). Some general aspects of statistical computing. Applied Statistics, **16**, 100–110.

Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs (second edition).* Wiley, New York.

Colin, A.J.T. (1967). On-line access systems in statistics. *Applied Statistics*, **16**, 111–119.

Cooper, B.E. (1967). ASCOP - a statistical computing procedure. *Applied Statistics*, **16**, 100–110.

Craddock, J.M. & Freeman, M.H. (1967). The METO computer language. *Applied Statistics*, **16**, 100–110.

Finney, D.J. (1947). *Probit Analysis.* Cambridge: Cambridge University Press.

Green. M. (1982). Array manipulation in PRISM. *Lecture Notes in Statistics, No. 14. GLIM 82: Proceedings of the International Conference on Generalised Linear Models* (ed. R. Gilchrist), pp. 36-42. New York: SpringerVerlag.

James, A.T. & Wilkinson, G.N. (1971). Factorization of the residual operator and canonical decomposition of non-orthogonal factors in the analysis of variance. *Biometrika*, **58**, 279–294.

Lane, P.W. & Nelder, J.A. (1982). Analysis of covariance and standardization as instances of prediction. *Biometrics*, **38**, 613–621.

Lamacraft, R.R. & Payne, R.W. (1980). A new look at data structures for statistical languages. In: *COMPSTAT 1980: Proceedings in Computational Statistics* (ed. Marjorie M. Barritt & D. Wishart), pp. 463–469. Vienna: PhysicaVerlag.

Lee, Y., & Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, **58**, 619–678.

Lee, Y., & Nelder, J.A. (2001). Hierarchical generalized linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y., Nelder, J.A. & Payne, R.W. (2002). Procedures HGANALYSE, HGDISPLAY, HGFIXEDMODEL, HGKEEP, HGPLOT and HGRANDOMMODEL. In: *GenStat Release 6.1 Reference Manual, Part 3 Procedure Library PL14* (Ed. R.W. Payne & G.M. Arnold), pp. 243–252. Oxford: VSN International.

Meier, P. (1967). Discussion of contributions at Meeting on Statistical Programming. *Applied Statistics*, **16**, 142.

Nelder J.A. (1962). An alternative form of generalized logistic equation. *Biometrics*, **18**, 614–616.

Nelder, J.A. (1965a). The analysis of randomized experiments with orthogonal block structure. I Block structure and the null analysis of variance. *Proceedings of the Royal Society, Series A*, **283**, 147–162.

Nelder, J.A. (1965b). II Treatment structure and the general analysis of variance. *Proceedings of the Royal Society, Series A*, **283**, 163–178.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Nelder, J.A. (1974). A user's guide to the evaluation of statistical packages. *Int. Stat. Rev.*, **42**, 291–298.

Nelder, J.A. (1975). GLIM generalized linear interactive modelling program. *Applied Statistics*, **24**, 259–261.

Nelder, J.A. (1991). GLIMPSE, a knowledge-based front end for GLIM. In: *IMA Volume in Mathematics and its Applications 36: Computing and Graphics in Statistics* (Ed. A. Buja & P.A. Tukey), pp. 125–131. New York: Springer Verlag.

Nelder, J.A. & Payne, R.W. (1991). Genstat as a computing environment. In: *IMA Volume in Mathematics and its Applications 36: Computing and Graphics in Statistics* (Ed. A. Buja & P.A. Tukey), pp. 133–138. New York: Springer Verlag.

Nelder, J.A. (1993). The K system for GLMs in Genstat. *Technical Report TRI/93.* Oxford: Numerical Algorithms Group.

Nelder, J.A. (1993). Extending the frontiers of generalized linear models. In: *Genstat Conference Program and Abstracts, 8th International Genstat Conference, University of Kent at Canterbury, 19–23 July 1993.*

Nelder, J.A. & Lee, Y. (1994). Double generalized linear models. In: *A Statistical Conference of Genstat Users, Wagga Wagga, New South Wales Australia, 28–30 November 1994*, p. 20.

Nelder, J.A. (1995). Hierarchical generalized linear models. In: *9th International Genstat Conference, University College Dublin, 10–14 July 1995, Program and Abstracts*, p. 21.

O'Brien, C.M. (1989). *The GLIMPSE System Manual.* Oxford: Numerical Algorithms Group.

Page E.S. (1967). Discussion of contributions at Meeting on Statistical Programming. *Applied Statistics*, **16**, 133.

Payne, R.W. & Nelder, J.A. (1976). Data structures in statistical computing. *Proceedings of the 9th International Biometric Conference*, Vol.II, 191–208.

Payne, R.W. & Wilkinson, G.N. (1977). A general algorithm for analysis of variance. *Applied Statistics*, **26**, 251–260.

Payne, R.W. (1982). AOV: the Prism module for analysing designed experiments. In: *Lecture Notes in Statistics, No. 14. GLIM 82: Proceedings of the International Conference on Generalised Linear Models* (ed. R. Gilchrist), pp. 58–68. New York: SpringerVerlag.

Payne, R.W. (1983). Plans for Genstat 5. *Genstat Newsletter*, **12**, 23–27.

Payne, R.W. & Lane, P.W. (1986). Design criteria for a flexible statistical language. In: *COMPSTAT 86 Proceedings in Computational Statistics* (ed. F De Antoni, N. Lauro & D. Rizzi), pp. 345–350, Heidelberg: PhysicaVerlag.

Payne, R.W., Lane, P.W., Ainsley, A.E., Bicknell, K.E., Digby, P.G.N., Harding, S.A., Leech, P.K., Simpson, H.R., Todd, A.D., Verrier, P.J., White, R.P., Gower, J.C., Tunnicliffe Wilson, G. & Paterson, L.J. (1987). *Genstat 5 Reference Manual.* Oxford: Oxford University Press.

Payne, R.W. & Tobias, R.D. (1992). General balance, combination of information and the analysis of covariance. *Scandinavian Journal of Statistics*, **19**, 3–23.

Payne, R.W., Lane, P.W., Digby, P.G.N., Harding, S.A., Leech, P.K., Morgan, G.W., Todd, A.D., Thompson. R., Tunnicliffe Wilson, G., Welham, S.J. & White, R.P. (1993). *Genstat 5 Reference Manual, Release 3.* Oxford: Oxford University Press.

Payne, R.W. (1998). Detection of partial aliasing and partial confounding in generally balanced designs. *Computational Statistics*, **13**, 213–226.

R.W. Payne & G.M. Arnold (2003). *GenStat Release 7.1 Reference Manual, Part 3 Procedure Library PL15.* Oxford: VSN International.

Payne, R.W., Baird, D.B., Cherry, M., Gilmour, A.R., Harding, S.A., Kane, A.F., Lane, P.W., Murray, D.A., Soutar, D.M., Thompson. R., Todd, A.D., Tunnicliffe Wilson, G., Webster, R. & Welham, S.J. (2003). *The Guide to GenStat Release 7.1, Part 1: Syntax and Data Management. Part 2: Statistics.* Oxford: VSN International.

Richardson, M.G. & Baker, R.J. (1980). The development off the GLIM system up to GLIM-4. In: *COMPSTAT 1980: Proceedings in Computational Statistics* (ed. Marjorie M. Barritt & D. Wishart), pp. 463–469. Vienna: PhysicaVerlag.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.

Slater, M. (1982). The GRAPH module. In: *Lecture Notes in Statistics, No. 14. GLIM 82: Proceedings of the International Conference on Generalised Linear Models* (ed. R. Gilchrist), pp. 43–57. New York: Springer-Verlag.

Thompson, R. (1979). Sire evaluation. *Biometrics*, **35**, 339–353.

Vickers, T. (1967). Discussion of contributions at Meeting on Statistical Programming. *Applied Statistics*, **16**, 142.

Wilkinson, G.N., A general recursive algorithm for analysis of variance. *Biometrika* **57** (1970) 19–46.

Wilkinson, G.N. & Rogers, C.E. (1973). Symbolic description of factorial models for analysis of variance. *Applied Statistics*, **22**, 392–399.

Yates, F. (1937). *The Design and Analysis of Factorial Experiments.* Technical Communication No. 35 of the Commonwealth Bureau of Soils. Farnham Royal: Commonwealth Agricultural Bureaux.

NON-LINEAR REGRESSION MODELLING AND INFERENCE

J.C. WAKEFIELD
Departments of Statistics and Biostatistics,
University of Washington,
Seattle,
WA 98195, United States
Email: jonno@u.washington.edu

In this paper frequentist and Bayesian approaches to non-linear regression modelling are described, critiqued and contrasted. Estimating functions provide a unifying framework for frequentist inference, and sampling-based methods provide a flexible computational technique for carrying out Bayesian analyses. Special interest is focused upon the effects of model misspecification; in this regard the use of the (linear) exponential family is beneficial, and provides one advantage of using the generalized linear model class. A new application of the inverse polynomial models introduced in [36] is presented: the analysis of data from a pharmacokinetic experiment.

1. Introduction

Over recent years, increases in computer power, algorithmic development and the inclusion of such algorithms within statistical software, have unshackled the statistician in his/her ability to fit models of choice, rather than models imposed by mathematical and/or computational convenience. In this paper the analysis of data using non-linear models is considered.

In preparation for analysis the strategy that is stressed is:

1 To formulate an initial model class on the basis of the context.
2 To examine this class with respect to its statistical properties; specifically the behaviour of estimators and posterior distributions (in particular with respect to model misspecification) may be examined from, respectively, frequentist and Bayesian perspectives.
3 To examine computational aspects.

In either 2 or 3 the model may be altered to correct mathematical or computational shortcomings. The approach followed in this paper is rooted firmly in the tradition of attempting to understand structure within data through parametric modelling of the mean, in contrast to the predictive view of statistical inference (see [5] and the ensuing discussion).

There are a number of challenges associated with the ability to fit ever more complex models. First, the statistical properties of complex models are often not fully understood, in particular with respect to model misspecification. A second problem is the potential for loss of information on parameters of interest when the number of nuisance parameters is unnecessarily increased by expanding the model; further discussion of this issue is given in Section 3.2.

A third difficulty is that there now exists great potential for over-fitting in which models become too dataset-specific as they are refined on the basis of the examination of diagnostics. In practice, if refinement is carried out through the fitting of alternative models (e.g. transformation of covariates, choice of distribution for the responses), then interval estimates will often be too narrow since they are produced by conditioning on the final model, and hence do not reflect the mechanism by which the model was selected (see [7], and the accompanying discussion). From a frequentist standpoint estimators and test statistics should be examined via their long-run behaviour *given* the model-fitting process, including refinement. To be more explicit, let P denote the procedure by which a final model M is decided upon. Then suppose it is of interest to examine the bias of a statistic T,

$$\mathrm{E}[T|P] = \mathrm{E}_{M|P}\{\mathrm{E}[T \mid M]\}. \tag{1}$$

In general it will be incorrect to report $\mathrm{E}[T \mid \widehat{M}]$ where $\widehat{M}$ is the final model chosen, since this does not reflect the procedure by which $\widehat{M}$ was chosen, but rather acts as if the final model is the "truth". From a Bayesian standpoint the same problem exists because the posterior distribution should reflect all sources of uncertainty and *a priori* all possible models that may be entertained should be explicitly stated, with prior distributions being placed upon different likelihoods and the parameters of these likelihoods; model averaging should then be carried out across the different possibilities, a process which is fraught with difficulties not least in placing "comparable" priors over what may be fundamentally different objects (see Section 6 for an approach to rectifying this problem).

One solution to this third difficulty is to never refine the model for a given data set. This approach is operationally pure but pragmatically

dubious (unless one is in the context of a randomized experiment) since we may obtain appropriate inference for a model that is a very poor description of the phenomenon under study. The philosophy suggested here is to think as carefully as possible about the initial model class before the analysis proceeds, but after fitting to carry out model checking and refine the model in the face of *clear* model misspecification, with refinement ideally being carried out within distinct *a priori* known classes[a]. Inference then proceeds as if the final model were the one that were chosen initially. This is clearly a subjective procedure but can be informally justified via either philosophical approaches.

Under a frequentist approach inference follows from the behaviour of an estimator under repeated sampling from the true model, and if an initial model is clearly wrong on the basis of a residual plot (say), then it is very unlikely to be close to the "true" model and hence it is more appropriate to obtain properties of estimators under the assumed model. With reference to (1), if a model is chosen because it is clearly superior to the alternatives, then it may be reasonable to assume that $\mathrm{E}[T \mid P] \approx \mathrm{E}[T \mid \widehat{M}]$, because $\widehat{M}$ would be consistently chosen in repeated sampling under these circumstances.

In a similar vein, under a Bayesian approach the above procedure is consistent with model-averaging but with the posterior model weight being concentrated upon the chosen model (since alternative models are only rejected on the basis of clear inadequacy). The aim is to provide probability statements, from either philosophical standpoints that are "honest" representations of uncertainty. The above approach is relevant to analyses that are more confirmatory in their outlook, as opposed to being used for prediction, or for more exploratory purposes (for example, to gain clues to models that may be appropriate for future data analyses).

The structure of this paper is as follows. The frequentist approach to the analysis of non-linear models is considered in Section 2, with an estimating functions approach being emphasized, and specific choices being suggested by likelihood and quasi-likelihood. The Bayesian approach is described in Section 3 with computation via direct sampling from the posterior being described. A critique and comparison of the frequentist and Bayesian approaches is carried out in Section 4; in particular, situations in which one may be preferred over the other are delineated. Specific non-linear model

[a]So that, for example, examining quantile-quantile plots for different t distributions and picking the one that produces the straightest line would not be a good idea.

classes are considered in Section 5; with generalized linear models being described in Section 5.1 and compartmental models in Section 5.2. The approach to modelling followed in the paper is illustrated with the analysis of a set of pharmacokinetic data in Section 6. The paper ends with a concluding discussion in Section 7.

2. Frequentist Inference

Under the frequentist approach to inference procedures are assessed with respect to their long-run properties under hypothetical repeated sampling. If estimation is the objective then the aim is to obtain an estimator whose distribution is "close" to the true value. A fundamental criterion is *consistency*, which heuristically states that the estimator tends to the true value as the sample size increases. Another criterion by which we may compare two competing asymptotically unbiased estimators is via comparison of their asymptotic variances; an asymptotically *efficient* estimator is one that attains the lowest possible asymptotic variance.

Estimating functions have emerged as a unifying approach to much of frequentist inference and in the next section we review the basics before giving specific examples of estimating functions in the following two sections, specifically those arising from likelihood in Section 2.2 and quasi-likelihood in Section 2.3. In Section 2.4 sandwich estimation as a method of obtaining a consistent estimator of the variance of an estimator is described.

2.1. *Estimating Functions*

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$, represent n observations from a distribution indexed by a p-dimensional parameter $\boldsymbol{\theta}$, with $Y_i|\boldsymbol{\theta}$ (conditionally) independent. An *estimating function* is a function

$$\boldsymbol{G}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \sum_{i=1}^{n} \boldsymbol{G}_i(\boldsymbol{\theta}) \tag{2}$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta})] = \mathbf{0}. \tag{3}$$

The estimating function $\boldsymbol{G}(\boldsymbol{\theta})$ is a random variable because it is a function of $\boldsymbol{Y}$. The corresponding *estimating equation* that defines the estimator $\widehat{\boldsymbol{\theta}}$ has the form

$$\boldsymbol{G}(\widehat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \boldsymbol{G}_i(\widehat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{4}$$

For inference, the frequency properties of the estimating function are derived and are then transferred to the resultant estimator. This is an ingenious approach because the estimating function may be constructed to be a simple function of the data, while the estimator of the parameter that solves (4) will often be unavailable in closed form. The estimating function (2) is a sum of random variables which provides the opportunity to evaluate its asymptotic properties via a central limit theorem. The *art* of constructing estimating functions is to make them dependent on distribution-free quantities, for example, the population moments of the data; in Section 5.1 we will see that estimators arising from exponential family models are particularly appealing. We now state a theorem that forms the basis for asymptotic inference.

Theorem: The estimator $\widehat{\boldsymbol{\theta}}_n$ which is the solution to the estimating equation

$$\boldsymbol{G}(\widehat{\boldsymbol{\theta}}_n) = \sum_{i=1}^{n} \boldsymbol{G}_i(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0},$$

has asymptotic distribution

$$\widehat{\boldsymbol{\theta}}_n \;\dot{\sim}\; N_p\left(\boldsymbol{\theta}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{T-1}\right),$$

where

$$\boldsymbol{A} = \boldsymbol{A}_n(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\theta}}\right] = \sum_{i=1}^{n} \mathrm{E}\left[\frac{\partial \boldsymbol{G}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right],$$

and

$$\boldsymbol{B} = \boldsymbol{B}_n(\boldsymbol{\theta}) = \mathrm{cov}(\boldsymbol{G}) = \sum_{i=1}^{n} \mathrm{cov}\{\boldsymbol{G}_i(\boldsymbol{\theta})\}.$$

The form of the covariance of the estimator here, the covariance of the estimating function, flanked by the inverse of the Jacobian of the transformation from the estimating function to the parameter, is one that will appear again in Section 2.4 in the context of sandwich estimation.

In practice, $\boldsymbol{A} = \boldsymbol{A}_n(\boldsymbol{\theta})$ and $\boldsymbol{B} = \boldsymbol{B}_n(\boldsymbol{\theta})$ are replaced by $\widehat{\boldsymbol{A}} = \boldsymbol{A}_n(\widehat{\boldsymbol{\theta}}_n)$ and $\widehat{\boldsymbol{B}} = \boldsymbol{B}_n(\widehat{\boldsymbol{\theta}}_n)$, respectively. In this case, from a Slutsky Theorem (see for example [17], Chapter 6),

$$\widehat{\boldsymbol{\theta}}_n \;\dot{\sim}\; N_p\left(\boldsymbol{\theta}, \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^{T-1}\right), \tag{5}$$

since $\widehat{\boldsymbol{A}} \rightarrow_p \boldsymbol{A}$ and $\widehat{\boldsymbol{B}} \rightarrow_p \boldsymbol{B}$.

The accuracy of the asymptotic approximation to the sampling distribution of the estimator is dependent on the parameterization adopted. A rule of thumb is to obtain the confidence interval on a reparameterization which takes the parameter onto the real line (for example, the logistic transform for a probability, or the logarithmic transform for a dispersion parameter), and then to transform to the more interpretable scale; examples are presented in Section 6. Estimators for functions of interest, $\phi = g(\boldsymbol{\theta})$, may be obtained via $\widehat{\phi} = g(\widehat{\boldsymbol{\theta}})$, and the asymptotic distribution may be found using the delta method.

2.2. *Likelihood*

We begin by giving the definition of *likelihood* (as given by [19], p. 24).

> *Definition:* Viewing $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ gives the *likelihood function* which we denote by $L(\boldsymbol{\theta})$.

To follow a likelihood approach one must, therefore, specify the probability distribution of the observed data given the model parameters, that is $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. In this paper we consider models that are appropriate when the data are independent and identically distributed conditional on $\boldsymbol{\theta}$, so that we have

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}).$$

The probability model for the full data, which determines the likelihood (and includes the independence assumption), is based upon the context and all relevant accumulated knowledge. The level of belief in this model will clearly be context-specific and in many situations there will be insufficient information available to confidently specify all components of the model. Depending on the confidence in the likelihood, which in turn depends on the sample size (since large n allows examination of the assumptions of the model), the likelihood may either be effectively viewed as "correct" in that inference proceeds as if the true model were known, or may instead be seen as an initial *working* model from which an estimating function is derived, the properties of the subsequent estimator then being determined in a more general setting. For example, in Section 2.4 we describe a method for producing an estimator of the variance of the estimator that does not depend on the full probability model.

The value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, denoted $\widehat{\boldsymbol{\theta}}$, is known as the *Maximum Likelihood Estimator* (MLE); the MLE is therefore that value of $\boldsymbol{\theta}$ that gives the highest probability to the observed data. We now define some functions of the likelihood which will aid in the development of the asymptotic distribution of the MLE.

For both computation and the evaluation of analytical properties, it is convenient to consider the *log likelihood* function which is given by

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(Y_i \mid \boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}),$$

and the *score* function

$$\boldsymbol{G}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, ..., \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p}\right]^T = [\boldsymbol{G}_1(\boldsymbol{\theta}), ..., \boldsymbol{G}_p(\boldsymbol{\theta})]^T,$$

which, as we show below, satisfies the requirements of an estimating function upon which inference may be based.

Definition: Fisher's *expected* information is given by

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\{\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})^T\},$$

a $p \times p$ matrix.

Result: Under regularity conditions:

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta})] = \mathrm{E}\left[\frac{\partial l}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}, \tag{6}$$

and

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\{\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})^T\} = -E\left[\frac{\partial \boldsymbol{G}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right] = -E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} l(\boldsymbol{\theta})\right]. \tag{7}$$

Since $\mathrm{E}[\boldsymbol{G}(\boldsymbol{\theta})] = \mathbf{0}$ we have $\boldsymbol{I}(\boldsymbol{\theta}) = \mathrm{cov}\{\boldsymbol{G}(\boldsymbol{\theta})\}$. From this result we have

$$-\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{B}(\boldsymbol{\theta}) = \boldsymbol{I}(\boldsymbol{\theta})$$

and so, from the theorem of Section 2.1:

$$\widehat{\boldsymbol{\theta}}_n \;\dot{\sim}\; N_p\{\boldsymbol{\theta}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}\}.$$

It can be shown that MLEs are asymptotically efficient, *if* the model from which the score was derived is correct, see for example [45], Chapter 8.

As an alternative to using the asymptotic distribution, resampling methods such as the bootstrap may be used to examine the sampling distribution, see [15] and [10]. We do not discuss the bootstrap further, but acknowledge that a large literature now exists on both its theoretical properties and

its use in practice (though its use in small samples is not recommended). Likelihood ratio tests may be used to obtain confidence intervals (and are invariant to the parameterization adopted), and profile likelihood provides a method of examining the likelihood function for a parameter of interest alone.

In multiparameter situations *adjusted profile likelihood* may be used to create confidence intervals for a parameter of interest while making an attempt to account for the estimation of nuisance parameters. This approach can be computationally intensive and is not always reliable, and a number of modifications have been suggested, see for example [41].

If the model is misspecified then the MLE is that value of the parameter that brings the assumed model closest, in a Kullback-Leibler sense, to the true model ([26], [53]).

2.3. *Quasi-Likelihood*

In this section we describe an estimating function that is, at least on the surface, based upon the mean and variance of the data only. Specifically we assume

$$\mathrm{E}[\boldsymbol{Y}|\boldsymbol{\beta}] = \boldsymbol{\mu}(\boldsymbol{\beta}),$$

$$\mathrm{cov}(\boldsymbol{Y}|\boldsymbol{\beta}) = \phi \boldsymbol{V}\{\boldsymbol{\mu}(\boldsymbol{\beta})\},$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), ..., \mu_n(\boldsymbol{\beta})]^T$ represents a regression function and $\boldsymbol{V}$ is a diagonal matrix (so the observations are uncorrelated), with

$$\mathrm{var}(Y_i|\boldsymbol{\beta}) = \phi V\{\mu_i(\boldsymbol{\beta})\},$$

and $\phi > 0$ is a scalar which is independent of $\boldsymbol{\beta}$. The aim is to obtain the asymptotic properties of an estimator of $\boldsymbol{\beta}$. The specification of the mean function in a parametric regression setting is unavoidable, and least squares would indicate that properties for an estimator may be obtained from the additional specification of the variance.

To motivate an estimating function we follow [33] (see also [18] for an exceptionally clear description of quasi-likelihood) and consider the sum of squares

$$(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\phi, \tag{8}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta})$. To minimize this sum of squares there are two ways to proceed. Perhaps the more obvious route is to acknowledge

that both $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are functions of $\boldsymbol{\beta}$ and differentiate with respect to $\boldsymbol{\beta}$ to give

$$-2\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})/\phi + (\boldsymbol{Y}-\boldsymbol{\mu})^T\frac{\partial \boldsymbol{V}^{-1}}{\partial \boldsymbol{\beta}}(\boldsymbol{Y}-\boldsymbol{\mu})/\phi, \tag{9}$$

where $\boldsymbol{D}$ is the $n \times p$ matrix of derivatives with elements $\partial\mu_i/\partial\beta_j$, $i = 1,...,n; j = 1,...,p$. Unfortunately, if we only assume that $\mathrm{E}[Y_i] = \mu_i(\boldsymbol{\beta})$, the expectation of (9) is not necessarily zero, and so a consistent estimator of $\boldsymbol{\beta}$ will not generally result in this situation if it based on (9). However, if the true variance is equal to $\phi\boldsymbol{V}$, then there is an efficiency loss in ignoring the second term, since it contains information on $\boldsymbol{\beta}$. This illustrates the classic efficiency-robustness trade-off that must be addressed whenever a model (procedure) is chosen for inference.

Alternatively we may suppose that $\boldsymbol{V}$ is not a function of $\boldsymbol{\beta}$ when we differentiate (8), and then solve

$$\boldsymbol{D}(\widehat{\boldsymbol{\beta}})^T\boldsymbol{V}(\widehat{\boldsymbol{\beta}})^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\}/\phi = \boldsymbol{0}.$$

As shorthand we write this function as

$$\boldsymbol{U} = \boldsymbol{D}^T\boldsymbol{V}^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}\}/\phi. \tag{10}$$

This estimating function is linear in the data and so its properties are straightforward to evaluate. In particular:

(1) $\mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta})] = \boldsymbol{0}$.

(2) $\mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}/\phi$.

(3) $-\mathrm{E}\left[\dfrac{\partial \boldsymbol{U}}{\partial \boldsymbol{\beta}}\right] = \mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}/\phi$.

The similarity of these properties with those of the score function (equations (6) and (7)) is apparent and has lead to (10) being referred to as a *quasi-score* function. Note that the derivation of (3) depends only on correct mean specification, while (2) relies on correct variance specification also. We can apply the theorem of Section 2.1 directly to obtain the asymptotic distribution of the maximum quasi-likelihood estimator (MQLE) as

$$\widehat{\boldsymbol{\beta}} \;\dot{\sim}\; N_p\{\boldsymbol{\beta}, (\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\phi\},$$

where we have so far assumed that ϕ is known. Note that $\widehat{\boldsymbol{\beta}}$ does not depend on ϕ, a consequence of assuming that ϕ is a multiplier in the variance function. Since

$$\mathrm{E}[(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})] = n\phi,$$

an unbiased estimator of ϕ would be

$$(\boldsymbol{Y}-\boldsymbol{\mu})^T\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})/n.$$

A "degrees of freedom corrected" (but not in general, unbiased) estimate is therefore given by the Pearson statistic divided by its degrees of freedom

$$\widehat{\phi}=\frac{1}{n-p}\sum_{i=1}^{n}\frac{\{Y_i-\widehat{\mu}_i\}^2}{V(\widehat{\mu}_i)}, \tag{11}$$

where $\widehat{\mu}_i=\widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$. The benefit of this approach, as opposed (say) to constructing an estimating function for ϕ from a likelihood, is again one of robustness, since (11) will in general be more appropriate under a broader range of circumstances (those in which the mean and variance-covariance models are correct) than a model-based alternative. The asymptotic distribution that is used in practice is given by

$$\widehat{\boldsymbol{\beta}} \;\dot{\sim}\; N_p\{\boldsymbol{\beta},(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\widehat{\phi}\},$$

so that the uncertainty in $\widehat{\phi}$ is not accommodated in the uncertainty for $\widehat{\boldsymbol{\beta}}$ (see Section 3.2 for related discussion). [32] and [8] give conditions under which this asymptotic result applies. [9] gives counter-examples in which a linear estimating function such as (10) does not perform well; these examples are mostly of theoretical interest but do indicate that one should not assume that linear estimating functions always perform well.

Integration of the quasi-score (10) gives

$$l(\mu;y)=\int_y^{\mu}\frac{y-t}{\phi V(t)}\mathrm{d}t,$$

which, if it exists, behaves like a log-likelihood, explaining the genesis of the label "quasi-likelihood"; [51] was the first to consider this class. As an example, for the model $\mathrm{E}[Y]=\mu$ and $\mathrm{var}(Y)=\phi\mu$ we have

$$l(\mu;y)=\int_y^{\mu}\frac{y-t}{\phi t}\mathrm{d}t=\frac{1}{\phi}[y\log\mu-\mu+\mathrm{c}],$$

where $\mathrm{c}=-y\log y-y$ and $y\log\mu-\mu$ is the log likelihood of a Poisson random variable. The word "quasi" refers to the fact that the score may or may not equate to a probability function. For example, the variance function $\mu^2(1-\mu)^2$ does not correspond to a probability distribution (but was shown by [34], Example 9.2.4, to be useful in a particular application). If the estimating function (10) corresponds to a score function then the subsequent estimator corresponds to the MLE. Hence, although the mean and variance only are specified in the estimating function, there may be an

implicit model in the sense that the estimating function corresponds to a particular likelihood function. As a trivial example, the estimating function based on $\mathrm{E}[Y] = \mu$, $\mathrm{var}(Y) = \phi$ corresponds to the model $Y \underset{ind}{\sim} N(\mu, \phi)$.

The prediction of *observable* data Y is not possible with quasi-likelihood, since there is no probabilistic mechanism to appeal to.

2.4. *Sandwich Estimation*

A general method of avoiding stringent modelling conditions when the variance of an estimator is calculated is provided by *sandwich estimation*. The basic idea is to estimate the variance of the data empirically with minimum modelling assumptions, and to incorporate this in the estimation of the variance of an estimator. While the idea may be traced at least as far as [26], the paper of [52] implemented the technique for the linear model, and [42] provided a clear and simple account with many examples; [30] and [54] described the technique in the context of longitudinal data by using the replication across individuals to estimate within-person correlations empirically. [6], Appendix A.3 provide a good review.

We have seen that when the estimating function corresponds to a score equation, then *under the model*

$$\boldsymbol{I} = \boldsymbol{A} = -\boldsymbol{B}$$

so that

$$\mathrm{var}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{A}(\boldsymbol{\theta})^{-1}\boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta})^{T-1} = \boldsymbol{I}(\boldsymbol{\theta})^{-1}.$$

If the model is not correct then this equality does not hold, and the variance estimator will be incorrect. An alternative is to evaluate the variance *empirically* via

$$\widehat{\boldsymbol{A}} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{G}(\widehat{\boldsymbol{\theta}}, Y_i),$$

and

$$\widehat{\boldsymbol{B}} = \sum_{i=1}^{n} \boldsymbol{G}(\widehat{\boldsymbol{\theta}}, Y_i)\boldsymbol{G}(\widehat{\boldsymbol{\theta}}, Y_i)^T.$$

This method is general and can be applied to any estimating function, not just those arising from a score equation.

Suppose we assume $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$ and $\mathrm{var}(\boldsymbol{Y}) = \phi \boldsymbol{V}$ with $\mathrm{var}(Y_i) = \phi V(\mu_i)$, and $\mathrm{cov}(Y_i, Y_j) = 0$, $i, j = 1, ..., n$, $i \neq j$, as a *working* covariance model.

Under this specification it is natural to take (10) as an estimating function, in which case $\text{cov}\{U(\boldsymbol{\beta})\} = \boldsymbol{D}^T\boldsymbol{V}^{-1}\text{cov}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}/\phi^2$ to give

$$\text{var}_s(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}\text{cov}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1},$$

and so the appropriate variance is obtained by substituting in the correct form for $\text{cov}(\boldsymbol{Y})$ which is, of course, unknown. However, a simple "sandwich" estimator of the variance is given by

$$\text{var}_s(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1},$$

where $\boldsymbol{R} = (R_1, ..., R_n)^T$ is the $n \times 1$ vector with $R_i = Y_i - \mu_i(\widehat{\boldsymbol{\beta}})$. This estimator is consistent for the variance of $\widehat{\boldsymbol{\beta}}$, under correct specification of the mean, and with uncorrelated data. There is finite sample bias in R_i as an estimate of $Y_i - \mu_i(\boldsymbol{\beta})$ and versions that adjust for the estimation of the parameters $\boldsymbol{\beta}$ are also available, see for example [29].

The great advantage of sandwich estimation is that it provides a consistent estimator of the variance in very broad situations. There are two things to bear in mind when one considers the use of this technique, however. The first is that for small sample sizes, the sandwich estimator may be highly unstable, and in terms of mean squared error model-based estimators may be preferable for small to medium sized n (for small samples one would anyway want to avoid the reliance on the asymptotic distribution). Hence *empirical* is a better description of the estimator than *robust*. The second consideration is that if the model is correct, then the model-based estimators are more efficient.

3. Bayesian Inference

3.1. *Summarising the Posterior Distribution*

In the Bayesian approach, all unknown quantities which are contained in a probability model for the observed data (including, the model parameters, and any missing or censored data) are considered to be random variables. This is in contrast to the frequentist view in which parameters are treated as *constants*.[b] Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ denote all of the unknowns of the model, and $\boldsymbol{y} = (y_1, ..., y_n)^T$ the vector of observed data. Also let $\mathcal{I}$ represent all information relevant to the analysis that is currently available to the individual who is carrying out the analysis, in addition to $\boldsymbol{y}$.

[b]Here, strictly, *fixed* effects parameters are being considered. So-called *random* effects are assumed to arise from a population distribution and are viewed as random.

Inference is made through the posterior probability distribution of $\boldsymbol{\theta}$, after observing $\boldsymbol{y}$:

$$p(\boldsymbol{\theta}|\boldsymbol{y},\mathcal{I}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta},\mathcal{I})\ \pi(\boldsymbol{\theta} \mid \mathcal{I})}{p(\boldsymbol{y} \mid \mathcal{I})}, \tag{12}$$

where $p(\boldsymbol{y} \mid \boldsymbol{\theta},\mathcal{I})$ is the likelihood, and $\pi(\boldsymbol{\theta} \mid \mathcal{I})$ the *prior* distribution representing the probability beliefs for $\boldsymbol{\theta}$ *before* observing the data $\boldsymbol{y}$, based on the current information $\mathcal{I}$. Different individuals will have different information $\mathcal{I}$ and so in general priors, and for that matter likelihoods, may differ. The normalizing constant is given, for continuous $\boldsymbol{\theta}$, by

$$p(\boldsymbol{y} \mid \mathcal{I}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta},\mathcal{I})\ \pi(\boldsymbol{\theta} \mid \mathcal{I})\ \mathrm{d}\boldsymbol{\theta} \tag{13}$$

and is the marginal distribution of the data, given the likelihood and prior. From this point onwards we suppress the dependence on $\mathcal{I}$, for notational convenience.

To summarise the posterior distribution marginal distributions for parameters of interest may be considered. For example, the univariate marginal distribution for a component θ_i is given by

$$p(\theta_i \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}_{(i)}} p(\boldsymbol{\theta} \mid \boldsymbol{y})\ \mathrm{d}\boldsymbol{\theta}_{(i)}, \tag{14}$$

where $\boldsymbol{\theta}_{(i)}$ is the vector of all parameters, $\boldsymbol{\theta}$, excluding θ_i. Posterior moments may be evaluated from the marginal distributions; for example the posterior mean is given by

$$\mathrm{E}[\theta_i \mid \boldsymbol{y}] = \int_{\theta_i} \theta_i\ p(\theta_i \mid \boldsymbol{y})\ \mathrm{d}\theta_i. \tag{15}$$

Posterior means, in contrast to MLEs, are not invariant to transformation, that is, $\mathrm{E}[g(\boldsymbol{\theta})|\boldsymbol{y}] \neq g(\mathrm{E}[\boldsymbol{\theta}|\boldsymbol{y}])$, unless g is a linear function). Further summarisation may be carried out to yield the $100{\times}q\%$ quantile, $\theta_i(q)$ $(0 < q < 1)$ by solving

$$\int_{-\infty}^{\theta_i(q)} p(\theta_i \mid \boldsymbol{y})\ \mathrm{d}\theta_i = q. \tag{16}$$

In particular, the posterior median, $\theta_i(0.5)$, will often provide an adequate summary of the location of the marginal posterior, and a $100{\times}p\%$ equi-tailed *credible interval* is provided by $[\theta_i\{(1-p)/2\}, \theta_i\{(1+p)/2\}]$ for probability p, $0 < p < 1$.

Another useful inferential quantity is the *predictive* distribution for future observations, $\boldsymbol{z}$, which is given, under conditional independence, by

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}. \tag{17}$$

If we wish to compare models M_0 and M_1 then a natural measure is given by the *posterior odds*

$$\frac{\Pr(M_0 \mid \boldsymbol{y})}{\Pr(M_1 \mid \boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid M_0)}{p(\boldsymbol{y} \mid M_1)} \times \frac{\Pr(M_0)}{\Pr(M_1)}, \tag{18}$$

where the *Bayes factor* $p(\boldsymbol{y} \mid M_0)/p(\boldsymbol{y} \mid M_1)$ is the ratio of the marginal distributions of the data under the two models, and $\Pr(M_0)/\Pr(M_1)$ is the *prior odds*. To calculate the former, integrals of the form (13) are required.

Bayesian inference is deceptively simple to describe probabilistically, but there have been two major obstacles to its routine use. The first is how to specify prior distributions and will be considered in Section 3.3. The second is how to evaluate the integrals required for inference, for example (13)–(17), given that for most models (and for all but the most trivial non-linear models always), these are analytically intractable. A general method for non-hierarchical non-linear models is described in Section 3.4.

If the likelihood is correctly specified then, under certain conditions, the posterior distribution is asymptotically normal with mean the true value, and variance-covariance matrix given by the inverse of the expected information, for non-technical derivations see for example [38], p. 75 and [20], Appendix B. An important practical condition is that the prior does not exclude any part of the support of the parameter. More rigorous treatments can be found in, for example, [50] and [27], where it is shown that under suitable regularity conditions the posterior distribution tends to a normal distribution with mean the MLE, and variance-covariance matrix given by the inverse of the observed information, evaluated at the MLE.

While the posterior distribution is asymptotically independent of the prior distribution, so that point and interval estimates are often robust to the prior choice with increasing n, Bayes factors are asymptotically sensitive to the prior, for further discussion see for example [38], p. 195. [28] give a review of Bayes factors, including discussion of computation and prior choice.

3.2. *Model Misspecification*

The behaviour of Bayesian estimators under misspecification of the likelihood has received less attention than frequentist estimators. As discussed

above, under sensible prior distributions the posterior distribution mimics the sampling distribution of the MLE, and so properties of the latter, such as consistency, can be transferred to, for example, the posterior mean or median.

Rather than the effects of misspecification, the emphasis in the Bayesian literature has been on sensitivity analyses ([38], Chapter 7, gives a review of approaches to address sensitivity to prior and likelihood choices), or on embedding a particular likelihood or prior choice within a larger class. If a discrete number of choices are considered then model averaging has been used (for a review see [14]), while others (e.g. [21]) prefer to embed the model within a continuous class and then integrate over this class.

[4] argue that examining the behaviour of an estimator under model misspecification (which they term *criterion robustness*) is inadequate since as the model varies the criterion should change also. While this is certainly true in some situations, it is not true in general and so should not be used as a reason to reject the approach out of hand. Perhaps the reason that such an approach has not been followed is because it is more difficult to apply when no closed form estimator is available. The philosophies behind consideration of misspecification are therefore very different under frequentist and Bayesian approaches.

A major problem with considering model classes with large numbers of unknown parameters is that uncertainty on parameters of interest will be increased if a simple model is closer to the truth, so there will be an efficiency loss associated with considering models that are *too* large. In particular, model expansion may lead to a decrease in precision. The following discussion relates to likelihood inference as well as to Bayesian inference, but we include it here because the emergence of MCMC has encouraged the use of larger and larger models within a Bayesian approach.

We examine the form of the posterior variance. As n increases the prior effect is negligible and the posterior variance is given by the inverse of the observed information; for convenience, we consider the expected information, which is asymptotically equivalent. Suppose that we have a $k \times 1$ vector of parameters, $\boldsymbol{\beta}$, in an original model (and these include the parameters of interest), and $p - k$ additional parameters, $\boldsymbol{\gamma}$, in an expanded model. Then consider

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix}, \tag{19}$$

where $\boldsymbol{I}_{11}$ is a $k \times k$ matrix corresponding to the information for $\boldsymbol{\beta}$, and

$\boldsymbol{I}_{22}$ is the $(p-k)\times(p-k)$ information for γ. In the simpler model the information on the parameters of interest is $\boldsymbol{I}_{11}$, while for the enlarged model it is

$$\boldsymbol{I}_{11}-\boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21},$$

which is never greater than $\boldsymbol{I}_{11}$.

To illustrate in a simple regression setting, consider an observational study in which the covariate of interest in not orthogonal to all other potential confounding variables. As a specific example, the model

$$Y_i=\beta_0^e+\beta_1^e(x_i-\bar{x})+\gamma(z_i-\bar{z})+\epsilon_i,$$

is an expansion of the model

$$Y_i=\beta_0^r+\beta_1^r(x_i-\bar{x})+\epsilon_i,$$

$i=1,...,n$, where $\epsilon_i \underset{iid}{\sim} N(0,\sigma^2)$, with σ^2 known. Here we have distinguished between β_0^e and β_1^e in the expanded model, and β_0^r and β_1^r in the reduced model, because the parameters have different interpretations, and we need to distinguish between them when the posterior variance of each is considered below. Letting $\boldsymbol{x}$ denote the $n\times 3$ matrix with i–th row $[1 \;\; x_i \;\; z_i]$, and $\boldsymbol{\beta}^e=(\beta_0^e \;\; \beta_1^e)^T$, we have

$$\boldsymbol{I}(\boldsymbol{\beta}^e,\gamma)=\sigma^{-2}(\boldsymbol{x}^T\boldsymbol{x})=\sigma^{-2}\begin{bmatrix} n & 0 & 0 \\ 0 & S_{xx} & S_{xz} \\ 0 & S_{xz} & S_{zz}\end{bmatrix},$$

where $S_{xx}=\sum_i(x_i-\bar{x})^2$, $S_{xz}=\sum_i(x_i-\bar{x})(z_i-\bar{z})$, $S_{zz}=\sum_i(z_i-\bar{z})^2$. Hence

$$\boldsymbol{I}(\boldsymbol{\beta},\gamma)^{-1}=\sigma^2\begin{bmatrix} 1/n & 0 & 0 \\ 0 & S_{zz}/D & -S_{xz}/D \\ 0 & -S_{xz}/D & S_{xx}/D\end{bmatrix},$$

where $D=S_{xx}S_{zz}-S_{xz}^2$, giving

$$\text{var}(\beta_1^e|\boldsymbol{y})=\frac{\sigma^2}{S_{xx}-S_{xz}^2/S_{zz}}\geq\frac{\sigma^2}{S_{xx}}=\text{var}(\beta_1^r|\boldsymbol{y}),$$

with equality if $S_{xz}=0$, i.e. if x and z are orthogonal. Intuitively, the posterior variance is increased because when z is present in the model there are competing explanations for the observed association between y and x. Of course, one of the reasons for including additional variables is to reduce bias; however, it is straightforward to phrase the above argument in terms of

mean squared error and reach the same conclusion when the bias reduction due to the inclusion of z is not great.

With unknown σ^2 the situation is more complex since important covariates will reduce the size of the estimate of the variance, $s^2 = \text{RSS/DF}$ (where RSS is the residual sum of squares and DF the degrees of freedom), and (asymptotically, or with flat priors) $\text{var}(\beta_1^e|\boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{x})^{-1}s^2$ and so the posterior variance will be reduced also. However, at some point this variance will also increase since s^2 increases as unimportant covariates are added to the model. So again the overall conclusion is the same: models should not be chosen to be as large as possible, because the variance of quantities of interest will be unnecessarily increased.

We now briefly discuss the general situation in which *estimated* parameters are used in the information matrix. In this situation the variance of quantities of interest is again increased (as just discussed in the normal case). The simplest example is in a generalized linear model with scale parameter ϕ. Assuming ϕ is known corresponds to one family, while ϕ unknown corresponds to another family, as examples the Poisson becomes the negative binomial, and the exponential becomes the gamma. If the data are truly from the simpler model then interval estimates will be unnecessarily widened if the larger model is assumed. This occurs even though the posterior distributions of $\boldsymbol{\beta}$ and ϕ are asymptotically independent (so that in (19) $\boldsymbol{I}_{12}$ and $\boldsymbol{I}_{21}$ are zero); the extra uncertainty is introduced when estimates are substituted into the information. As an aside, the Poisson and exponential scenarios are perhaps not the best illustrations since not allowing excess variation in these two models would be a very dangerous modelling strategy.

The above is a very informal discussion, for a far deeper discussion of the choice between Student's t and normal errors see [25]. An interesting theoretical finding is that even if the errors are truly t, if the degrees of freedom are estimated, for small values of n, it will be more efficient to assume normal errors, because of the extra uncertainty involved in the estimation of the degrees of freedom.

3.3. *The Prior Distribution*

The specification of the prior distribution is clearly a crucial aspect of the Bayesian approach. We distinguish between two situations. In the first an analysis is required in which the prior distribution has minimal impact, so that the likelihood is concentrated upon. Such an analysis may be used

as a comparison with other analyses in which more informative priors are specified, in order to determine the information being provided by the prior. In this situation, the Bayesian formulation may also be seen as a convenient way of carrying out computation for those with a likelihood bent. The second situation is one in which it is desired to incorporate more substantial prior information in the analysis.

For non-linear models care must be taken to ensure that the posterior corresponding to a particular prior choice is proper. In particular the use of an improper uniform prior is not to be universally recommended. Such forms for fixed effects in a generalized linear model will usually lead to a proper posterior ([11]) although not for some pathological cases; for example if a uniform prior is used on $\log\{p/(1-p)\}$ and $y = 0$ or $y = n$.

To illustrate the non-propriety in more general non-linear models consider the model

$$Y_i \underset{ind}{\sim} N\{\exp(-\theta x_i), \sigma^2\},$$

$i = 1, ..., n$, with $\theta > 0$ and σ^2 assumed known. With an improper uniform prior on θ we have the posterior

$$p(\theta \mid \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathrm{e}^{-\theta x_i})^2\right\}.$$

As $\theta \to \infty$

$$p(\theta \mid \boldsymbol{y}) \to \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right\}$$

that is, a constant, so that the posterior is improper. Intuitively, the problem here is that as $\theta \to \infty$ the fitted values do not move increasingly away from the data but to the asymptote $y = 0$. There are no asymptotes in the linear model and so as the parameters increase/decrease to ∞/-∞, the fitted line moves increasingly far from the data, and the likelihood tends to zero.

3.4. *Simulation-Based Inference*

Simulation-based methods have revolutionised the practical applicability of Bayesian methods. Such methods build on the duality between samples and densities ([44]); given a sample we can reconstruct the density, and given an arbitrary density we can generate a sample, given the range of generic random variate generators available (see [12]). With respect to the latter, the ability to obtain *direct* samples from a distribution decreases as the

dimensionality of the parameter space increases and in this case Markov chain Monte Carlo (MCMC) methods may be used as an alternative, the disadvantage being that iteration is needed to produce samples that can be viewed as from the density of interest, and these samples are dependent. It is also not straightforward to calculate marginal densities such as (13) with MCMC, see [13] for a review.

For hierarchical models direct sampling is rarely possible (though feasible if the random effects may be integrated out, as in a linear hierarchical model), and MCMC needs to be considered. This paper concentrates on non-hierarchical models and in this case direct sampling is often feasible. We now describe a rejection algorithm that we will use to carry out Bayesian inference in Section 6.

Let $\boldsymbol{\theta}$ denote the unknown parameters and assume that we can evaluate the maximized likelihood $M = \sup_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta})$. The algorithm then proceeds as follows:

(1) Generate $U \sim U(0,1)$ and, independently, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ (the prior).
(2) Accept $\boldsymbol{\theta}$ if

$$U < \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{M},$$

otherwise return to 1.

The probability that a point is accepted is given by

$$p_a = \frac{1}{M} \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = \frac{p(\boldsymbol{y})}{M},$$

([47]). Hence the empirical rejection rate, $\widehat{p}_a$, can be used to derive the marginal likelihood from (13) as

$$\widetilde{p}(\boldsymbol{y}) = M \times \widehat{p}_a. \tag{20}$$

An alternative importance sampling estimator that is more efficient ([16], [39]) is given by

$$\widehat{p}(\boldsymbol{y}) = \frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(s)}),$$

where $\boldsymbol{\theta}^{(s)} \underset{iid}{\sim} \pi(\boldsymbol{\theta})$. Hence it is straightforward to calculate Bayes factors using the rejection algorithm.

Clearly we need a proper prior distribution to implement the above algorithm, and the efficiency of the algorithm will depend on the correspondence

between the likelihood and the prior, as measured through $p(\boldsymbol{y})$. As n increases, the algorithm will often become less efficient; typically, M increases as n increases. As we will demonstrate in Section 6, it is straightforward to specify the prior distribution in one parameterization, and specify the likelihood in another. The latter is useful since we may be able to specify the prior in terms of a set of model-free parameters, and then compare different likelihoods with an "egalitarian" prior. Another potential advantage is that the above algorithm does not require the functional form of the prior. [47] used a predictive distribution from a Bayesian analysis of a set of data as the prior for the analysis of a separate data set; samples from the predictive distribution could be simply generated, even though no closed form was available for this distribution.

For a generalized linear model (GLM) let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ where $\boldsymbol{\beta}$ is the vector of regression parameters and ϕ is a dispersion parameter. Standard software may be used to find the MLE, $\widehat{\boldsymbol{\theta}}$; however, some software (R for example), by default supplies a method of moments estimator, rather than the MLE, of ϕ. Since the MLE of $\boldsymbol{\theta}$ does not depend on ϕ in a GLM, one simply needs to find the MLE $\widehat{\phi}$ (which is available in R for some families, and in particular for the normal and gamma likelihoods that are used in Section 6).

4. Comparison of Frequentist and Bayesian Methods

In this section we will describe situations in which frequentist and Bayesian methods are likely to agree, and when one is preferable over the other. We concentrate on estimation since point and interval estimation are directly comparable under the two paradigms. For model comparison the objectives of Bayes factors and hypothesis tests are fundamentally different, see for example [1], and so comparison is more difficult.

In terms of interpretation, the Bayesian approach is more straightforward since one can make probability statements, for example, credible intervals are probabilistic. In contrast, frequentist confidence intervals are not so simple to interpretable.

Another appealing characteristic is that the Bayesian approach to inference may be derived via decision theory, see for example [3]. The likelihood principal, [2], also leads one towards a Bayesian approach since all frequentist criteria invalidate this principle, and a true likelihood approach, as followed by for example, [43], is difficult to calibrate. One may of course question the whole endeavor of establishing optimality, given that the sub-

sequent use depends on the specification of likelihoods and priors, both of which are fraught with difficulties.

In contrast the frequentist approach has been justified within a frequentist set of guidelines. For example, there is a Gauss-Markov theorem for linear estimating functions (e.g. [23], [32]), while [9] considers the optimality of quadratic estimating functions (which for implementation unfortunately require assumptions about the third and fourth moments). If one accepts that frequentist criteria are natural, then it would be desirable to find an estimator which minimises the mean squared error with respect to the sampling distribution. Unfortunately, this is not in general possible for finite n, so instead adjusted criteria (such as minimum variance unbiased estimation) become desirable.

A major problem with the frequentist approach is that, in contrast to the Bayesian approach, there is no rigid prescription for carrying out inference. Hence, for example, different types of likelihood (e.g. conditional, marginal, partial, profile, adjusted profile) exist as alternatives when conventional likelihood methods are inadequate (though in such cases the use of Bayesian methods usually requires careful prior specification). Some of these procedures are to deal with nuisance parameters, again the Bayesian approach is theoretically straightforward since posterior distributions for parameters of interest are obtained through marginalisation.

The greatest drawback of the Bayesian approach is the need to specify both a likelihood and a prior distribution. Sensitivity to each of these components can be carried out but the extent of such an investigation is difficult to determine, and one then is faced with the difficulty of how the results are reported. As we have discussed, assessing the behaviour of procedures under model misspecification is far more developed for frequentist methods than for Bayesian methods. For example, although a specific likelihood may be used to define the estimator, the properties of this estimator can be evaluated under more general models.

Bayesian methods are far more amenable to situations in which n is small. In this situation it is not possible to check the likelihood and inference will in general be sensitive to both likelihood and prior choices. When the model is very complex then Bayesian methods are again advantageous since they allow a rigorous treatment of nuisance parameters; MCMC has allowed the consideration of more and more complicated hierarchical models. Spatial models, particularly those that exploit Markov random field second stages, provide a good example of models that are very naturally analysed using MCMC, where the conditional independencies may be ex-

ploited. Unfortunately assessments of the effects of model misspecification are rarely carried out for such complex models, instead sensitivity studies are again typically carried out. Bayesian methods are also a good idea in situations in which the maximum likelihood estimator provides a poor summary of the likelihood, for example in variance components problems where the likelihood may be highly skewed.

If n is sufficiently large for asymptotic normal approximation of the sampling distribution to be accurate, then frequentist methods begin to become preferable. In particular, sandwich estimation can be used to provide a consistent estimator of the variance-covariance matrix of the estimator. Hence, if the estimator of a parameter of interest is consistent also, reliable confidence coverage will be guaranteed. We stress that n needs to be sufficiently large for the sandwich estimator to be stable. A typical Bayesian approach would be to increase model complexity, often through the introduction of random effects. The difficulty with this is that although this allows more flexibility, a specific form needs to be assumed for the mean-variance relationship, whereas sandwich estimation is consistent in more general situations (quasi-likelihood lies between the two, though there is usually an implicit model underlying the quasi-score function).

5. Non-Linear Regression Models

In this section we briefly review two classes of models, in anticipation of their use in Section 6.

5.1. *Generalized Linear Models*

Generalized linear models were introduced by [35], and the most comprehensive and interesting description is still [34]; an excellent review is also given by [18]. A GLM is defined by two components:

(1) The responses y_i follow an exponential family so that the distribution is of the form

$$p(y_i|\theta_i, \phi) = \exp(\{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i, \phi)),$$

where θ_i and ϕ are scalars. This is sometimes referred to as a *linear* or *natural* exponential family. It is straightforward to show (using the results of Section 2.2) that

$$\mathrm{E}[Y_i|\theta_i, \phi] = \mu_i = b'(\theta_i)$$

and

$$\text{var}(Y_i|\theta_i, \phi) = b''(\theta_i)a(\phi),$$

$i = 1, ..., n$, with $\text{cov}(Y_i, Y_j|\theta_i, \theta_j, \phi) = 0$ for $i \neq j$. This describes the stochastic part of the model.

(2) We have $g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is $1 \times p$ and $\boldsymbol{\beta}$ is $p \times 1$ so that we have a linear predictor on a scale determined by the so-called link function $g(\cdot)$. This describes the deterministic part of the model.

While computational advances have unshackled the statistician from the need to use GLMs, they are still an extremely useful class of models. The use of the exponential family is advantageous because the score equation can be written

$$a(\phi)\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\{y_i - \mu_i(\boldsymbol{\beta})\}\frac{\partial \theta_i}{\partial \boldsymbol{\beta}}$$

where $l = l(\boldsymbol{\theta}, \phi)$ is the log-likelihood, and so if the mean is specified correctly the MLE of $\boldsymbol{\beta}$ will be consistent (see the theorem of Section 2.1). Bayes estimators are consistent in this case also (so long as the priors do not exclude a part of the parameter space), due to the asymptotic equivalence between the sampling distribution of the MLE and the posterior distribution (Section 3.1). It is not necessary to have a linear predictor on any particular scale so, for example, the sums of exponentials models of the next section will share this consistency, if the responses arise from the exponential family (so long as regularity conditions are satisfied). So called canonical links in which $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ provide simplifications in terms of computation.

GLMs are also very useful pedagogically since they separate the deterministic and random components of the model; this aspect was emphasized by [35] who wrote in the abstract: "The implications of the approach in designing statistics courses are discussed".

5.2. *Compartmental Models*

Pharmacokinetics is the study of the time course of a drug and its metabolites following introduction into the body. In this section we describe a class of models that has been extensively used in such studies to model individual drug concentrations, $y(x)$, as a function of time x. The drug may be introduced into the body via a variety of routes of administration including

intravenously (directly into the bloodstream via either a bolus or an infusion), subcutaneously (beneath the skin), or orally. After introduction the drug undergoes the processes of absorption, distribution and elimination. These processes may be modelled by assuming the body consists of a series of homogenous pools or compartments, and then considering a set of differential equations that determine the rate of flow of drug between the different compartments, see [22] for a comprehensive account of pharmacokinetic models and principles, and [24] for an account of compartmental modelling in general.

As a simple example consider a model with a single compartment for the distribution and elimination, and an oral dose; we make use of this model in Section 6. We may think, nominally, of the compartment corresponding to the blood; in general pharmacokinetic modelling via a compartmental system is a convenient visualisation but the compartments often have no physiological meaning, rather phyisological parameters such as the time to maximum concentration, maximum concentration, elimination half-life and clearance are of interest. These parameters are defined in Section 6.

Let $w_i(x)$ represent the amount of drug in compartment i, $i = 0, 1$, at time x, with compartment 0 representing the site from which absorption occurs. The differential equations describing the drug flow between the compartments may be assumed to be of the form

$$\begin{aligned} \frac{dw_0}{dx} &= -k_a w_0 \\ \frac{dw_1}{dx} &= k_a w_0 - k_e w_1 \end{aligned} \tag{21}$$

where k_a is the absorption rate constant associated with flow from compartment 0 to compartment 1, and k_e is the elimination rate constant. Assuming that $w(0) = D$ is the dose at time zero and that the (apparent) volume of distribution (which converts total amount of drug into concentration) is V we may solve (21) to obtain the time course of the *concentration*, $\mu(x)$, as

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} \{\exp[-k_e x] - \exp[-k_a x]\}. \tag{22}$$

This model is sometimes known as the flip-flop model because there is a basic identifiability in that the same curve is achieved with the parameter sets (V, k_a, k_e) and $(Vk_e/k_a, k_e, k_a)$, and it is often assumed that $k_a > k_e$ in order to enforce identifiability.

We now consider the stochastic part of the model. In addition to measurement error in the assay technique, errors are introduced due to model

misspecification (particularly at later phases of drug development which are carried out in a poorly controlled environment, and so the reported sampling times may be subject to error, for example). Assay precision is often found to increase with increasing true concentrations and models of the form

$$y(x) = \mu(x) + \delta(x),$$

where $\delta(x) \sim N\{0, \mu(x)^\gamma \sigma_\delta^2\}$ with $\gamma > 0$ have been used. The variance power γ is either fixed, with $\gamma = 2$ being a common choice to produce a constant coefficient of variation, or estimated. A constant coefficient of variation can also be approximately achieved by taking

$$\log y(x) = \log \mu(x) + \epsilon(x),$$

with $\epsilon(x) \sim N(0, \sigma^2)$.

[49] provide a review of pharmacokinetic and pharmacodynamic modelling including more details on both biological and statistical aspects.

6. Pharmacokinetic Data Analysis

An oral dose of 1mg of Theophylline was administered to a new born baby, and concentration time data (x_i, y_i) were subsequently collected for $i = 1, ..., 8$. These data were previously analyzed by [46], and are reproduced in Table 1.

Table 1. Concentration (y) as a function of time (x), obtained from a new-born baby following the administration of a 1mg dose of Theophylline.

i	x_i (hours)	y_i (mg/liter)
1	1.00	60.22
2	1.42	73.41
3	3.58	63.43
4	5.08	56.43
5	6.83	48.81
6	9.08	30.40
7	12.3	20.67
8	23.8	7.28

Traditionally, the so-called one-compartment open model, as described in Section 5.2 would be fitted to these data. Under this model the concentration at time x is given by (22) which we reproduce here, along with

an alternative form, in order to motivate a log-linear inverse polynomial model:

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} \{\exp[-k_e x] - \exp[-k_a x]\}$$

$$= D\exp(\beta_0 + \beta_1 x)\{1 - \exp[-(k_a - k_e)x]\}, \tag{23}$$

where $\beta_0 = \log\{k_a/(V(k_a - k_e)\}$ and $\beta_1 = -k_e$. Typically, interest does not focus upon (V, k_a, k_e), but rather on the following derived parameters:

- The elimination half-life, which is the time it takes for the drug concentration to drop by 50%, when elimination is the dominant process:

$$x_{1/2} = (\log 2)/k_e.$$

- The time to maximum

$$x_{\max} = \frac{1}{k_a - k_e} \log\left(\frac{k_a}{k_e}\right).$$

- The maximum concentration

$$\mu(x_{\max}) = \frac{Dk_a}{V(k_a - k_e)} \{\exp(-k_e x_{\max}) - \exp(-k_a x_{\max})\}$$

$$= \frac{D}{V}\left(\frac{k_a}{k_e}\right)^{k_a/(k_a - k_e)}$$

- The clearance, which is the amount of blood cleared of drug in unit time, is given by $Cl = D/\text{AUC}$ where AUC is the area under the concentration time curve so that

$$Cl = V \times k_e.$$

We assume that

$$\log y_i = \log \mu_i(x_i) + \epsilon_i,$$

where $\epsilon_i \underset{iid}{\sim} N(0, \sigma^2)$.

As an alternative to the above compartmental model, we here fit the log-linear inverse polynomial model, a GLM, given by

$$\mu(x) = D\exp(\beta_0 + \beta_1 x + \beta_2/x).$$

Comparison with (23) shows that β_2 is the parameter that determines the absorption; the model only makes sense if it produces an increasing absorption phase and a decreasing elimination phase which correspond, retrospectively, to $\beta_2 < 0$ and $\beta_1 < 0$. This model is very much in the spirit

of [37] in which the inverse polynomial form was suggested as a model for enzyme-kinetic data. Note that each of the derived parameters are functions of $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2)$. Specifically:

- The elimination half-life

$$x_{1/2} = -(\log 2)/\beta_1.$$

- The time to maximum

$$x_{\max} = (\beta_2/\beta_1)^{1/2}.$$

- The maximum concentration

$$\mu(x_{\max}) = D \exp\{\beta_0 - 2(\beta_1\beta_2)^{1/2}\}.$$

- From the definition as D/AUC the clearance is given by

$$Cl = \frac{\sqrt{\beta_1/\beta_2}}{2\exp(\beta_0)\text{BesselK}\{2(\beta_1\beta_2)^{1/2}\}}, \tag{24}$$

  where BesselK denotes a modified Bessel function of the third kind.

The data are assumed to be gamma distributed, specifically

$$Y_i \underset{ind}{\sim} \text{Ga}\{\phi^{-1}, (\mu_i\phi)^{-1}\},$$

so that $\phi^{1/2}$ is the coefficient of variation. [31] examine various distributional choices for pharmacokinetic data, and found the gamma assumption to be reasonable for their examples. A more extensive discussion of the application of this model to pharmacokinetic data may be found in [48]. A disadvantage of this model compared to compartmental models is that if multiple doses are considered the mean function does not correspond to a GLM.

The lognormal compartmental and gamma log-linear models were fitted in R, with maximum likelihood estimation of $\boldsymbol{\beta}$, and the moment estimator, (11), for the dispersion parameter. Confidence intervals based on the asymptotic distribution of the MLE were calculated for the parameters of interest using the delta method. These parameters are all positive and so the intervals were obtained for the log transforms, and then exponentiated (the delta method for the clearance under the log-linear model was not used because of the intractability of the calculations, the sampling-based Bayesian approach that we describe shortly is straightforward, however). The results are displayed in Table 2, and the fitted curves in Figure 1. Each of these summaries shows a remarkable level of agreement across models.

The maximized log-likelihoods were -21.58 for the gamma model and -20.89 for the lognormal model; these models are not nested and so a likelihood ratio statistic is not available, but the use of AIC is valid and suggests no significant differences between the models.

We now describe a Bayesian implementation of each of these models using the rejection algorithm described in Section 3.4. We place prior distributions on the half-life, $x_{1/2}$, time to maximum, $x_{\max}$, maximum concentration, $\mu(x_{\max})$ and coefficient of variation; this is more natural for each of the models (and in particular for the log-linear model within which β_2 is not straightforward to interpret). Another benefit of specifying the prior in terms of model-free parameters is that models may be compared using Bayes factors on an equal footing, in the sense that the prior input for each model is identical. For more discussion of this issue, see [40]. We choose independent lognormal priors for these four parameters. For a generic parameter, θ, denote the prior by $\theta \sim \text{Lognormal}(\mu, \sigma)$. To obtain the moments of these distributions we specified the prior median, θ_m, and the 95% point of the prior, θ_u, and then solved for the moments via:

$$\mu = \log(\theta_m), \quad \sigma = \{\log(\theta_u) - \mu\}/1.645.$$

The third line of Table 2 gives the illustrative prior choices; samples were simulated from the prior in order to estimate empirically the quantiles of the induced prior for Cl. This prior could be criticised for the assumption of independence; it would be straightforward in principle to specify a multivariate lognormal, however, perhaps with the moments being based on a population pharmacokinetic analysis of a group of patients who are thought to be exchangeable with the specific subject being considered.

To implement the rejection algorithm we sample from the prior on the parameters of interest, and then back-solve for the parameters that describe the likelihood. For the compartmental model we transform back to the original parameters via

$$\begin{aligned} k_e &= (\log 2)/x_{1/2} \\ 0 &= x_{\max}(k_a - k_e) - \log\left(\frac{k_a}{k_e}\right) \\ V &= \frac{D}{\mu(x_{\max})}\left(\frac{k_a}{k_e}\right)^{k_a/(k_a-k_e)} \end{aligned} \tag{26}$$

so that k_a is not directly available, but must be obtained as the root of (26). For the log-linear model the transformation to $\boldsymbol{\beta}$ is via

$$\beta_1 = -\log 2/x_{1/2}, \quad \beta_2 = \beta_1 x_{\max}^2, \quad \beta_0 = \log \mu(x_{\max}) + 2(\beta_1\beta_2)^{1/2}.$$

The rejection algorithm described in Section 3.4 was used, with the MLEs for the β's obtained from the analyses reported earlier (and replacing the method of moments estimators with the MLEs for the dispersion parameters), and 500 samples being generated from the posterior distributions; the acceptance rates were 0.0030 and 0.0015 for the gamma and lognormal models, respectively. Table 2 summarizes inference for the parameters of interest with the interval estimates and medians being obtained as the sample quantiles. Note that inference for the clearance is straightforward since samples can be substituted directly in to the form (24). Figures 2 and 3 show the posteriors for the functions of interest under both models. These figures and Table 2 show that Bayesian inference under both of the models is very similar; frequentist and Bayesian methods are also in close agreement for this example. The posteriors are skewed for all functions of inference apart from the clearance parameter, indicating that the posterior medians are more representative than the MLEs. The clearance parameter is often found to be well-behaved, since it is a function of the area under the curve, which is very stably estimated.

We evaluated the normalizing constants using (20), and calculated the Bayes factor comparing the gamma and lognormal models (denoted M_G and M_L, respectively) on the $\log_2$-base scale (which is suggested by [28]). We obtain $\log_2 p(\boldsymbol{y} \mid M_G) - \log_2 p(\boldsymbol{y} \mid M_L) = -39.56 - (-39.52) = -0.04$, showing no significant difference between the models, in agreement with the AIC conclusion described earlier.

Table 2. Point and 90% interval estimates for the data of Table 1; *Cl* denotes Clearance, CV Coefficient of Variation (expressed as a percentage). The Bayesian point estimates correspond to the posterior medians.

Model	$x_{1/2}$	$x_{\max}$	$\mu(x_{\max})$
Comp MLE	6.27 (5.66,6.95)	1.87 (1.39,2.53)	70.5 (56.3,88.3)
GLM MLE	6.12 (4.46,8.39)	1.72 (1.36,2.17)	68.5 (53.0,88.5)
Prior	5.00 (2.78,9.00)	1.00 (0.333,3.00)	65.0 (52.8,80.0)
Comp Posterior	6.44 (5.74,7.28)	1.69 (0.700,2.23)	70.4 (64.4,78.1)
GLM Posterior	6.37 (5.69,7.18)	1.40 (0.556,2.06)	69.2 (62.6,77.2)

Model	Cl ($\times 10^3$)	CV ($\times 10^2$)
Comp MLE	1.28 (1.19,1.36)	11.3 (7.46,17.0)
GLM MLE	1.27 (–,–)	9.68 (7.89,11.9)
Prior	1.50 (3.29,13.9)	10.0 (2.50,40.0)
Comp Posterior	1.28 (1.19,1.40)	12.3 (7.82,20.3)
GLM Posterior	1.27 (1.16,1.37)	12.5 (7.95,21.2)

Residuals were examined to assess the appropriateness of the mean function,

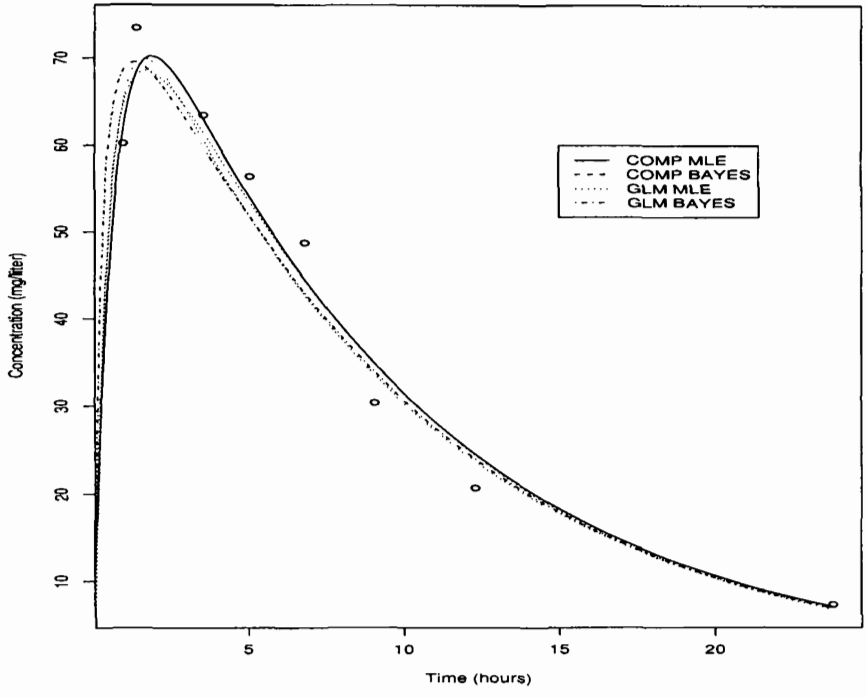


Fig. 1. Fitted curves for Theophylline data.

the mean-variance relationship, and the distribution of the errors. No clear inadequacy was evident for this (admittedly small) data set.

7. Discussion

In this paper a review of parametric non-linear modelling has been presented, with both frequentist and Bayesian approached to inference being described. It has been argued that models should first arise from the context, with mathematical and computational aspects being subsequently examined. The computational convenience of GLMs is a major benefit, and since their introduction in [35] GLMs have been widely used in an array of contexts, a testimony to their flexibility and their continued competitiveness with the increased array of models that are now computationally feasible for the practicing statistician. GLMs also have desirable statistical properties; in particular the use of the linear exponential family yields con-

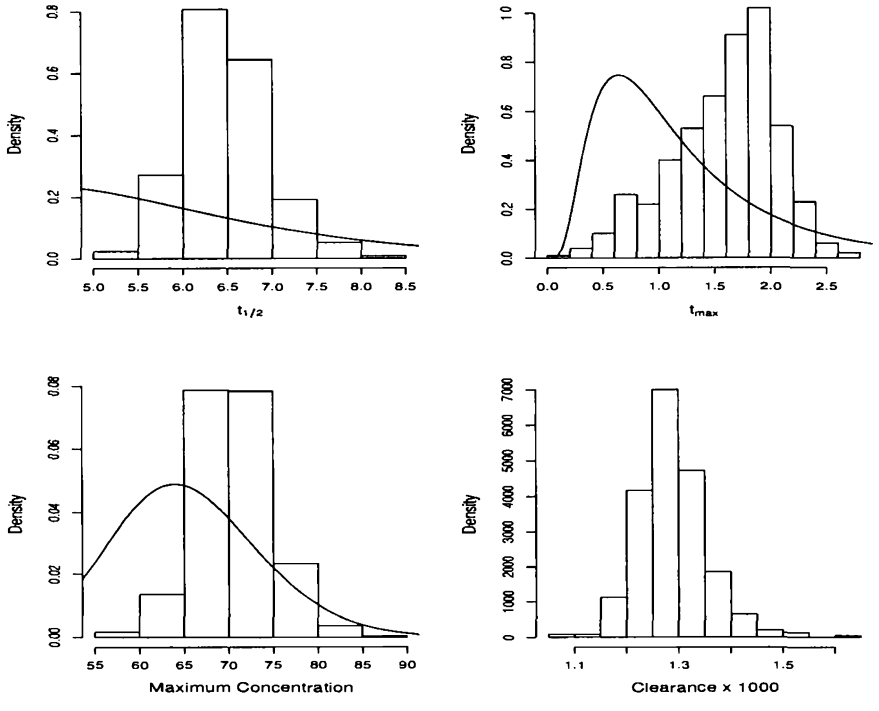


Fig. 2. Histogram representations of posterior distributions for the compartmental model; solid curves denote the lognormal prior distributions.

sistent estimators from likelihood or Bayesian approaches, so long as the mean model is correctly specified.

We have also described a simple rejection algorithm that may be used to produce independent samples from the posterior distribution and is very convenient in situations in which informative prior distributions are available, and the maximised likelihood can be simply calculated. The advantages of such sampling-based approaches have also been illustrated, in particular, inference for functions of interest is straightforward.

Acknowledgments

The author would like to thank John Nelder for discussions that greatly helped in the formulation of the models that were used for the pharmacokinetic data example.

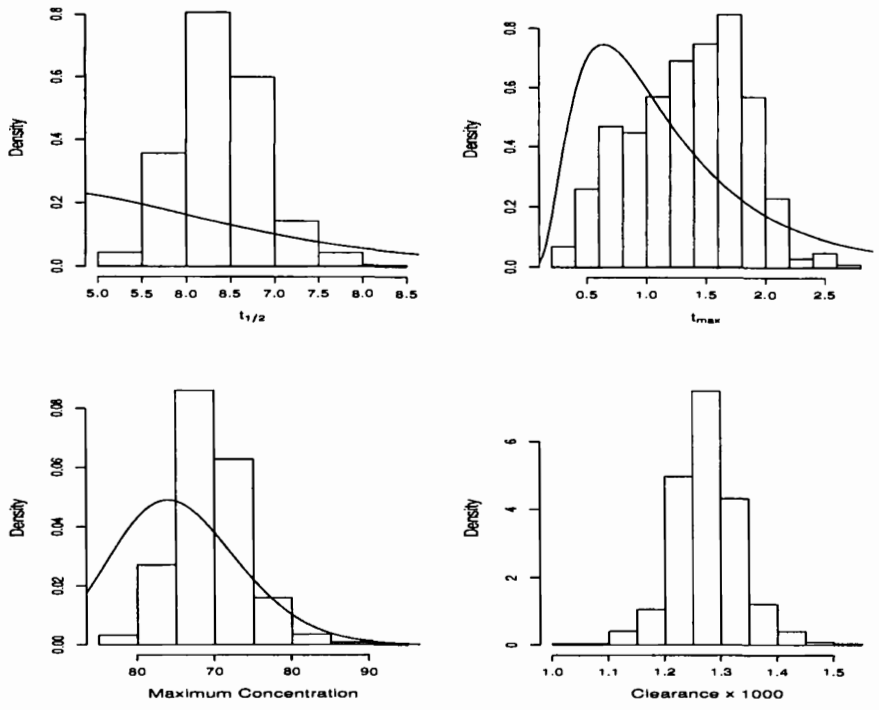


Fig. 3. Histogram representations of posterior distributions for the log-linear inverse polynomial model; solid curves denote the lognormal prior distributions.

References

1. J.O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18:1–32, 2003.
2. J.O. Berger and R.L. Wolpert. *The Likelihood Principle: A Review, Generalizations, and Statistical Implications.* Hayward: Institute of Mathematical Statistics, 1988.
3. J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* John Wiley and Sons, New York, 1994.
4. G. E. P. Box and G. C. Tiao. A note on criterion robustness and inference robustness. *Journal of the Royal Statistical Society, Series B*, 51:169–173, 1964.
5. L. Breiman. Statistical modeling: The two cultures (with discussion). *Statistical Science*, 16:199–231, 2001.

6. R.J. Carroll, D. Ruppert, and L.A. Stefanski. *Measurement Error in Nonlinear Models.* Chapman and Hall/CRC, Boca Raton, 1995.
7. C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society, Series A*, 158:419–466, 1995.
8. M. Crowder. On consistency and inconsistency of estimating equations. *Econometric Theory*, 2:305–330, 1986.
9. M. Crowder. On linear and quadratic estimating functions. *Biometrika*, 74:591–597, 1987.
10. A.C. Davision and D.V. Hinkley. *Bootstrap Methods and their Application.* Cambridge University Press, 1997.
11. P. Dellaportas and A. F. M. Smith. Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, 42:443–459, 1993.
12. L. Devroye. *Non-Uniform Random Variate Generation.* Springer-Verlag, New York, 1986.
13. T.J. DiCiccio, R.E. Kass, A. Raftery, and L. Wasserman. Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92:903–915, 1997.
14. D. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, 57:45–97, 1995.
15. B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman and Hall/CRC, Boca Raton, 1993.
16. M. Evans and T. Swartz. Rejoinder to: Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 11:54–64, 1995.
17. T.S. Ferguson. *A Course in Large Sample Theory.* Chapman and Hall/CRC, 1996.
18. D. Firth. Recent developments in quasi-likelihood methods. In *Proceedings of the ISI 49th Session*, pages 341–358, 1994.
19. R.A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.
20. A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, London, 1995.
21. A. Gelman and X.-L. Meng. Discussion of: Assessment and propagation of model uncertainty, by D. Draper. *Journal of the Royal Statistical Society, Series B*, 57:83, 1995.
22. M. Gibaldi and D. Perrier. *Drugs and the Pharmaceutical Sciences,*

Volume 15: Pharmacokinetics, Second Edition. Marcel Dekker, 1982.
23. V.P. Godambe and C.C. Heyde. Quasi-likelihood and optimal estimation. *International Statistical Review*, 55:231–244, 1987.
24. K.R. Godfrey. *Compartmental Models and their Applications.* Academic Press, London, 1983.
25. N.L. Hjort. The exact amount of t-ness that the normal model can tolerate. *Journal of the American Statistical Association*, 89:665–675, 1994.
26. P.J. Huber. The behavior of maximum likelihood estimators under non-standard conditions. In L.M. LeCam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. University of California Press, 1967.
27. R.A. Johnson. Asymptotic expansions associated with posterior distributions. *The Annals of Mathematical Statistics*, 41:851–864, 1970.
28. R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
29. G. Kauermann and R.J. Carroll. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96:1387–1396, 2001.
30. K.Y. Liang and S.L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.
31. J.K. Lindsey, W.D. Byrom, J. Wang, P. Jarvis, and B. Jones. Generalized nonlinear models for pharmacokinetic data. *Biometrics*, 56:81–88, 2000.
32. P. McCullagh. Quasi-likelihood functions. *The Annals of Statistics*, 11:59–67, 1983.
33. P. McCullagh. Quasi-likelihood and estimating functions. In D.V. Hinkley, N. Reid, and E.J. Snell, editors, *Statistical Theory and Modelling*, pages 265–286. Chapman and Hall, 1991.
34. P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition.* Chapman and Hall/CRC, Boca Raton, 1989.
35. J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society, Series A*, 135:370–384, 1972.
36. J.A. Nelder. Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, 22:128–141, 1966.
37. J.A. Nelder. Generalized linear models for enzyme-kinetic data. *Biometrics*, 47:1605–1615, 1991.
38. A. O'Hagan. *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference.* Arnold, London, 1994.

39. D.K. Pauler, J.C. Wakefield, and R.E. Kass. Bayes factors for variance component models. *Journal of the American Statistical Association*, 94:1242–1253, 1999.
40. J. M. Pérez and J. O. Berger. Expected-posterior prior distributions for model selection. *Biometrika*, 89:491–512, 2002.
41. N. Reid. The roles of conditioning in inference. *Statistical Science*, 10:138–199, 1995.
42. R. Royall. Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, 54:221–226, 1986.
43. R. Royall. *Statistical Evidence – A Likelihood Paradigm.* Chapman and Hall/CRC, Boca Raton, 1997.
44. A.F.M. Smith and A.E. Gelfand. Bayesian statistics without tears: a resampling approach. *The American Statistician*, 46:84–88, 1992.
45. A.W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, 1998.
46. J.C. Wakefield. *The Bayesian Analysis of Pharmacokinetic Models.* PhD thesis, Nottingham University, 1992.
47. J.C. Wakefield. Bayesian individualization via sampling-based methods. *Journal of Pharmacokinetics and Biopharmaceutics*, 24:103–131, 1996.
48. J.C. Wakefield. Gamma generalized linear models for pharmacokinetic data. Technical report, Department of Biostatistics, University of Washington, 2004.
49. J.C. Wakefield, L. Aarons, A., and Racine-Poon. The Bayesian approach to population pharmacokinetic/pharmacodynamic modelling. In Gatsonis C., Kass R.E., Carlin B.P., Carriquiry A.L., Gelman A., Verdinelli I., and West M., editors, *Case Studies in Bayesian Statistics, Volume IV*, pages 205–265. Springer-Verlag, New York, 1999.
50. A.M. Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society, Series B*, 31:80–88, 1970.
51. R.W.M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika*, 61:439–447, 1974.
52. H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48:1721–746, 1980.
53. H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–26, 1982.
54. S.L. Zeger and K.Y. Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42:121–130, 1986.

This page intentionally left blank

SELECTING AMONGST LARGE CLASSES OF MODELS

B. D. Ripley
Department of Statistics,
University of Oxford,
Oxford OX1 3TG, UK
E-mail: ripley@stats.ox.ac.uk

Traditionally a 'model' is a family of probability distributions for the observed data parametrized by a set of parameters (of fixed and finite dimension), but it is often helpful to consider all the models considered as subsets of one model, as well as some even larger models used in 'over-fitting' as part of the validation process. Traditional distinctions between 'parametric' and 'non-parametric' models are often moot, when people now (attempt to) fit neural networks with half a million parameters. We consider how to select in such model classes.

1. Introduction

Statisticians and other users of statistical methods have been choosing models for a long time, but the current availability of large amounts of data and of computational resources means that model choice is now being done on a scale which was not dreamt of 25 years ago. Unfortunately, the practical issues are probably less widely appreciated than they used to be, as statistical software has made it so much easier for the end user to trawl through literally thousands of models (and in some cases many more).

Model choice is a large subject, and this paper deliberately chooses to look at only some aspects of it, most particularly some of the misunderstandings about formal methods such as AIC and cross-validation. Whole books have been written about these and other aspects: two recent ones are Harrell [22] and Burnham and Anderson [10].

2. Why do we want to select a model?

I have slowly come to realize that this is an important question and one that is asked too seldom. First we need to ask *where do our models come from?*

- Sometimes a set of models is provided based on subject-matter theory. In my experience good theory is very rare. Sometimes these are called *mechanistic models*. One example is the Black–Scholes theory of option pricing, which is derived from a theory and has been shown to be a good approximation, but not so good that practically important improvements cannot be made.
- Most often some simple restrictions are placed on the behaviour we expect to find, for example linear models, $AR(p)$ processes, factorial models with limited interactions. These are sometimes called *empirical models*. Note that these can be very broad classes if transformations of variables (on both sides) are allowed.
- We now have model classes that can approximate any reasonable model, for example neural networks [38]. Nowadays we may have the data and the computational resources to fit such models, if not necessarily the understanding to fit them well.

The main distinction I would draw is between *explanation* and *prediction*. Generally with the mechanistic models we are concerned with explaining how the world works, even though the philosophy of science teaches that we test models by their ability to predict. The third class of models is unambiguously designed to give good predictions.

For the second class, we might be doing either. When people first started to do agricultural experiments they were (it seems) both trying to find out which factors had an effect, and for those that did, how large the effect was. Nowadays many experiments are done with *microarrays* to find out which few (out of thousands) of genes are expressed differently in different experimental conditions. But regression and time-series models are most commonly used for their predictive abilities.

For *explanation*, Occam's razor applies and we want

> an explanation that is as simple as possible, but no simpler
>
> attrib Einstein

and we do have a concept of a 'true' model, or at least a model that is a good working approximation to the truth. Simplicity helps both with communi-

cating the concepts embodied in the model and in what psychologists call *generalization*, the ability to 'work' in scenarios very different from those in which the model was studied.

On the other hand, *prediction* is like doing engineering development. All that matters is that it works, and if the aim is prediction, model selection should be based on the quality of the predictions. Workers in pattern recognition have long recognised this, and used *validation sets* to choose between models, and *test sets* to assess the quality of the predictions from the chosen model. Because the model may be used in scenarios very different from those in which it was tested, generalization is still important, and *other things being equal* a mechanistic model or a simple empirical model has more chance of reflecting the data-generation mechanism and so of generalizing. But other things rarely *are* equal.

We should ask why we do want to *choose* a model. It does seem a widespread misconception that model choice is about 'choosing the best model'. For explanation we ought to be alert to be possibility of there being several (roughly) equally good explanatory models: when I was a young Lecturer at Imperial College I learnt this from David Cox, having already done a lot of informal model choice in applied problems in which I would have benefited from offering several alternative solutions.

For prediction I find a good analogy is that of choosing between expert opinions: if you have access to a large panel of experts, how would you use their opinions? (See Cooke [11].) People do tend to pick one expert and listen to him/her, but it would seem better to seek a consensus view, which translates to *model averaging* rather than model choice. Our analogy is with experts, which implies some prior selection of people with a track record: one related statistical idea is the *Occam's window* [27] which keeps only models with a reasonable record.

A major reason to choose a model appears still to be computational cost, a viewpoint of Geisser [21, §4.1]. This has become less relevant, and we discuss model averaging in a later section. Note, though, that taking a consensus view only helps sometimes with generalization. For example, Draper [14, p. 48] has a graph of predictions of oil prices for 1981–90 made in 1980. The analysts were all confident, differed considerably from each other, and were all way off! Almost all of the uncertainty is in the 'correct' model for oil price movements, and the analysts' models all seem to be incorrect as prices went down when all the analysts predicted them to rise.

Ein-Dor and Feldmesser [16] provide an example of the confusion between explanation and prediction that is one of my favourite teaching ex-

amples. The title says they give a *relative performance prediction model*, yet they select[a] a subset of transformed variables in seeking an explanation.

3. A historical perspective

Let us look back 25 years to when I started to learn about this area. The set of models one could consider was severely limited by computational constraints, although packages such as GLIM 3.77 were becoming available.

Stepwise selection was the main formal tool, using hypothesis tests between a pair of nested models, e.g. F tests for regressions. Almost no one did enough tests to worry much about issues of multiple comparison. Nowadays people do tens of thousands of significance tests (see Marchini and Ripley [29] for an example of mine).

Residual plots were used to help assess the fit of models, but they were crude plots and limited to small datasets.

There was very little attempt to deal with choosing between models that were genuinely different explanations: Cox [12, 13]'s 'tests of separate families of hypotheses' existed but were little known and less used.

Formal methods of model choice were becoming available. Schwarz [41] had proposed a criterion sometimes called SBC or BIC, although it seems to be due to Jeffreys in the 1930's. Papers by Allen [3, 4] and Akaike [1, 2] had introduced PRESS and AIC (Akaike's An Information Criterion) respectively. Cross-validation goes back at least as far as Mosteller and Wallace [32], and Stone [43] was read to the Royal Statistical Society, to a less than appreciative audience.

Perhaps the only formal criterion that was in common use was Mallows' C_p criterion for regression, which I am told was well known long before Mallows' first publication [28].

My impression is that these developments were held back by the lack of computational resources to try out large classes of models, and by the lack of large datasets to present challenging problems.

4. Cross-validation

Cross-validation (CV) is a much misunderstood topic in the neural networks and machine learning community.

The idea of *leave-one-out CV* is that, given a dataset of N points, we use our model-building procedure on each subset of size $N-1$, and predict

[a] the paper is a wonderful example of how **not** to do that, too.

the point we left out. Then the set of predictions can be summarized by some measure of prediction accuracy. Allen's PRESS (prediction sum-of-squares) used this to choose a set of variables in linear regression. Stone [43] and Geisser [20] pointed out we could apply this to many aspects of model choice, including parameter estimation. It is often confused with jackknifing *a la* Quenouille and Tukey.

Having to do model-building N times can be prohibitive unless there are computational shortcuts.

In V-fold cross-validation we divide the data into V sets, amalgamate $V-1$ of them, build a model and predict the result for the remaining set. Do this V times leaving a different set out each time. How big should V be? We want the model-building problem to be realistic, so want to leave out a small proportion. We do not want too much work, so usually V is 3–10. One early advocate of this was Breiman *et al.* [9].

Leave-one-out CV does not work well in general, as it makes too small changes to the fit. Ten-fold CV often works well, but sometimes the result is very sensitive to the partitioning used, and it is often better for comparisons than for absolute values of performance. How prediction accuracy is measured can be critical. We can now afford to average the results over several random partitions.

Stone [43, pp. 126–7] mentioned the idea of using cross-validation not to choose between models but to combine them. This has been developed by Wolpert [47] under the name of *stacked generalization.*

5. AIC , BIC and all that

Akaike [1, 2] introduced a criterion for model adequacy, first for time-series models and then more generally. He relates how his secretary suggested he call it 'An Information Criterion', AIC. Two books largely about this criterion are Sakamoto *et al.* [39] and Burnham and Anderson [10].

This has a very appealing simplicity:

$$AIC = -2\log(\text{maximized likelihood}) + 2p$$

where p is the number of estimated parameters. Choose the model with the smallest AIC (and perhaps retain all models within 2 of the minimum). (Despite the simplicity, quite a few people have managed to get it wrong, for example the `step` function in S-PLUS.) This is similar to Mallows' C_p criterion for regression,

$$C_p = \text{RSS}/\sigma^2 + 2p - N$$

and is the same if σ^2 is known. Both are of the form

$$\text{measure of fit} + \text{complexity penalty}$$

Schwarz's criterion, often called BIC, replaces 2 by $\log n$ for a suitable definition of n, the size of the dataset. In the original regression context this is just the number of cases.

Derivation of AIC

Suppose we have a dataset of size N, and we fit a model to it by maximum likelihood, and measure the fit by the *deviance* D (constant minus twice maximized log-likelihood). Suppose we have m (finite) nested models.

Hypothetically, suppose we have another dataset of the same size, and we compute the deviance D^* for that dataset *at the MLE for the first dataset*. We would expect that D^* would be bigger than D, on average. In between would be the value D' obtained if we had evaluated the deviance at the true parameter values. Some Taylor-series expansions (e.g. Ripley [38], pp. 31–4) show that

$$E\,D^* - E\,D' \approx p, \qquad E\,D' - E\,D \approx p$$

and hence $AIC = D + 2p$ is (to this order) an unbiased estimator of $E\,D^*$. The latter is a reasonable measure of performance, the Kullback-Leibler divergence between the true model and the plug-in model (at the MLE).

These expectations are over the dataset under the assumed model.

Crucial assumptions

The assumptions needed for this argument are much less well known than they should be, and AIC is often proposed (and used) to select between m very different models.

(1) The model is true! Suppose we use this to select the order of an $AR(p)$ model. If the data really came from an $AR(p_0)$ model, all models with $p \geq p_0$ are true, but those with $p < p_0$ are not even approximately true. This assumption can be relaxed. Takeuchi [45] did so, and his result was rediscovered by Stone [44] and many times since. However, p gets replaced by a much more complicated formula that is not simple to measure.

(2) The models are nested[b] – AIC is widely used when they are not.

[b] see the bottom of page 615 in the reprint of Akaike [1].

(3) Fitting is by maximum likelihood. Nowadays many models are fitted by penalized methods or Bayesian averaging .... That can be worked through too, in NIC [33, 34, 35] or p_{eff} [30, 31].

(4) The Taylor-series approximations are adequate. People have tried various refinements, notably AICC (or AIC_c) given by

$$AICC = D + 2p\Big(\frac{N}{N-p+1}\Big)$$

Also, the MLEs need to be in the interior of the parameter space, even when a simpler or alternative model is true. (This is not likely to be true for variance components for example.)

(5) AIC is a reasonably good estimator of $E\,D^*$, or at least that differences between models in AIC are reasonably good estimators of differences in $E\,D^*$. This seems to be the Achilles' heel of AIC. For N independent samples we expect $AIC = O_p(N)$ but the variability as an estimate is $O_p(\sqrt{N})$. This reduces to $O_p(1)$ for differences between models *provided they are nested.*

AIC has been criticised in asymptotic studies and simulation studies for tending to over-fit, that is choose a model at least as large as the true model. That is a virtue, not a deficiency: this is a prediction-based criterion, not an explanation-based one.

AIC is asymptotically equivalent to leave-one-out CV for independent identically distributed samples using deviance as the loss function [44], and in fact even when the model is not true NIC is equivalent [38].

6. Bayesian Approaches

Note the plural — I think Bayesians are rarely Bayesian in their model choices. Assume m (finite) models, exactly one of which is true.

In the Bayesian formulation [5, 14], models are compared via $P\{M \mid \mathcal{T}\}$, the posterior probability assigned to model M given the dataset $\mathcal{T}$.

$$P\{M \mid \mathcal{T}\} \propto p(\mathcal{T} \mid M)p_M,$$

$$p(\mathcal{T} \mid M) = \int p(\mathcal{T} \mid M, \theta)p(\theta)\,\mathrm{d}\theta$$

so the ratio in comparing models M_1 and M_2 is proportional to $p(\mathcal{T} \mid M_2)/p(\mathcal{T} \mid M_1)$, known as the *Bayes factor*.

However, a formal Bayesian approach then averages predictions from models, weighting by $P\{M \mid \mathcal{T}\}$, unless a very peculiar loss function is in use.

Suppose we just use the Bayes factor as a guide. The difficulty is in evaluating $p(\mathcal{T} \mid M)$. Asymptotics are not useful for Bayesian methods, as the prior on θ is often very important in providing smoothing, yet asymptotically negligible. One approximation is to take $\widehat{\theta}$ as the mode of the posterior density and V as the inverse of the Hessian of $-\log p(\widehat{\theta} \mid \mathcal{T})$ (since for a normal density this is the covariance matrix); we can hope to find $\widehat{\theta}$ and V from the maximization of

$$\log p(\theta \mid \mathcal{T}) = L(\theta; \mathcal{T}) + \log p(\theta) + \text{const}$$

Let $E(\theta) = -L(\theta; \mathcal{T}) - \log p(\theta)$, so this has its minimum at $\widehat{\theta}$ and Hessian there of V^{-1}.

$$\begin{aligned} p(\mathcal{T} \mid M) &= \int p(\mathcal{T} \mid \theta)\, p(\theta)\, \mathrm{d}\theta = \int \exp -E(\theta)\, \mathrm{d}\theta \\ &\approx \exp -E(\widehat{\theta}) \int \exp[-\tfrac{1}{2}(\theta - \widehat{\theta})^T V^{-1} (\theta - \widehat{\theta})]\, \mathrm{d}\theta \\ &= \exp -E(\widehat{\theta})\ (2\pi)^{p/2} |V|^{1/2} \end{aligned}$$

via a Laplace approximation to the integral. Thus

$$\log p(\mathcal{T} \mid M) \approx L(\widehat{\theta}; \mathcal{T}) + \log p(\widehat{\theta}) + \tfrac{p}{2} \log 2\pi + \tfrac{1}{2} \log |V|.$$

It may be feasible to use this directly for model selection.

If we suppose θ has a prior which we may approximate by $N(\theta_0, V_0)$, we have

$$\begin{aligned} \log p(\mathcal{T} \mid M) \approx\ & L(\widehat{\theta}; \mathcal{T}) - \tfrac{1}{2}(\widehat{\theta} - \theta_0)^T V_0^{-1} (\widehat{\theta} - \theta_0) \\ & - \tfrac{1}{2} \log |V_0| + \tfrac{1}{2} \log |V| \end{aligned}$$

and V^{-1} is the sum of V_0^{-1} and the Hessian H of the log-likelihood at $\widehat{\theta}$. Thus

$$\log p(\mathcal{T} \mid M) \approx L(\widehat{\theta}; \mathcal{T}) - \tfrac{1}{2}(\widehat{\theta} - \theta_0)^T V_0^{-1} (\widehat{\theta} - \theta_0) - \tfrac{1}{2} \log |H|.$$

If we assume that the prior is very diffuse we can neglect the second term, so the penalty on the log-likelihood is $-\frac{1}{2} \log |H|$.

For a random sample of size n from the assumed model, this might be roughly proportional to $-(\frac{1}{2} \log n)\, p$ provided the parameters are identifiable. This is the proposal of Schwarz [41].

Crucial assumptions

(1) The data were generated as an independent, identically distributed random sample, and originally for linear models only. It is not clear what n should be for, say, a random effects model.

(2) Choosing a single model is relevant in the Bayesian approach.
(3) The model is true.
(4) The prior can be neglected. We may not obtain much information about parameters which are rarely effective, even in very large samples.
(5) The simple asymptotics are adequate and that the rate of data collection on each parameter would be the same. We should be interested in comparing different models for the same n, and in many problems p will be comparable with n.

Note that as this is trying to choose an explanation, we would expect BIC to neither overfit nor underfit, and there is some theoretical support for that.

7. Deviance Information Criterion

Named by Spiegelhalter *et al.* [42] in a Bayesian setting where prior information is not negligible, and the model is assumed to be a good approximation but not necessarily true.

In GLMs (and elsewhere) the *deviance* is the difference in twice maximized log likelihood between the *saturated* model and the fitted model, or

$$D(\theta) = \text{deviance}(\theta) = \text{const}(\mathcal{T}) - 2L(\theta; \mathcal{T})$$

and in GLMs we use $D(\widehat{\theta})$ as the (scaled) (residual) deviance.

Define

$$p_D = \overline{D(\theta)} - D(\bar{\theta})$$

The first overline means averaging θ over $p(\theta \mid \mathcal{T})$, and the second means our estimate of the 'least false' parameter value, usually the posterior mean of θ (but perhaps the median or mode of the posterior distribution). Then define

$$DIC = D(\bar{\theta}) + 2p_D$$

Clearly DIC is AIC-like, but

- Like NIC it allows for non-ML fitting, in particular for the regularization effect of the prior that should reduce the effective number of parameters.
- It is not necessary (but is usual) that $p_D \geq 0$.
- DIC is explicitly meant to apply to non-nested non-IID problems.

- DIC is intended to be approximated via MCMC samples from the posterior density of θ given $\mathcal{T}$. On the other hand, DIC needs an explicit formula for the likelihood (up to a model-independent normalizing constant).

8. Model Averaging

For prediction purposes (and that applies to almost all Bayesians) we should average the predictions over models. What do we average?

The probability predictions made by the models.

For linear regression this amounts to averaging the coefficients over the models (being zero where a regressor is excluded), and this becomes a form of shrinkage. Other forms of shrinkage like ridge regression may be as good at very much lower computational cost.

Note that we may not want to average over all models. We may want to choose a subset for computational reasons, or for plausibility.

How do we choose the weights?

In the Bayesian theory this is clear, via the Bayes factors. In practice this is discredited. Even if we can compute them accurately (and via Markov Chain Monte Carlo we may have a chance), we assume that one and exactly one model is true. In practice Bayes factors can depend on aspects of model inadequacy which are of no interest. I first encountered this in Ripley [37], where we fitted formal probability models to images (and therefore had tens of thousands of observations). There was a common noise model but different priors for the different models. We were able to calculate Bayes factors approximately by MCMC in a week or so, and we were pleased to see that that the factors were very decisive. After some checking, we discovered that they were very decisively picking the wrong model. There was a 'true' model (the models represented different species of nematodes) but a lot of investigation showed that the 'noise' model was interacting with the texture of the nematodes.

Alternative approaches are via cross-validation (goes back to Stone [43]) and via bootstrapping [25]. This can also be viewed as an extended estimation problem, with the weights depending on the sample via a model (e.g. a multiple logistic); so-called *stacked generalization* [47] and *mixtures of experts* [23].

Bagging, boosting, random forests

Model averaging ideas have been much explored in the field of classification trees.

In *bagging* [6, 7] models are fitted to bootstrap resamples of the data, and weighted equally. Breiman [7] motivates this for *unstable* methods such as classification trees in which a small change in the training set can lead to a large change in the classifier. A variant on this idea which has been suggested many times is to add 'noise' to the training set, randomly perturbing either the feature vectors $\mathbf{x}$ or the classes c (or both). Further along this line, we could model the joint distribution of $(\mathbf{X}, C)$ and create new training sets from this distribution. Bagging can be seen as the rather extreme form of this procedure in which the model is the empirical distribution. (Krogh and Vedelsby [24], use cross-validation rather than re-sampling, and consider designing training sets weighted towards areas where the existing classifiers are prone to disagree.)

In *boosting* [40, 17, 15, 18, 19] each additional model is chosen to (attempt to) repair the inadequacies of the current averaged model by resampling biased towards the mistakes. The idea is to *design* a series of training sets and use a combination of classifiers trained on these sets. (Majority voting and linear combinations have both been used.) There have been many papers on this topic, as well as empirical tests which tend to show [e.g., 36] that boosting often does well but occasionally does disastrously.

In *random forests* [8] the tree-construction algorithm randomly restricts itself at the choice of each split, to create a 'forest' of trees from a single training set.

9. Practical model selection in 2004

The concept of a model is much larger than it was 25 years ago. Even a decade ago, people attempted to fit neural networks with half a million free parameters. We are no longer so tied to maximum likelihood estimation, and fit models to much larger datasets. The latter almost inevitably means that we fit more complex models, and 'smooth' terms are often used in place of linear[c] terms.

Large model classes often overlap very considerably. There are many ways of obtaining a smooth curve like Figure 1. The traditional approach would be to fit a polynomial, and one of the curves is a degree-six polyno-

[c] or low-order polynomial

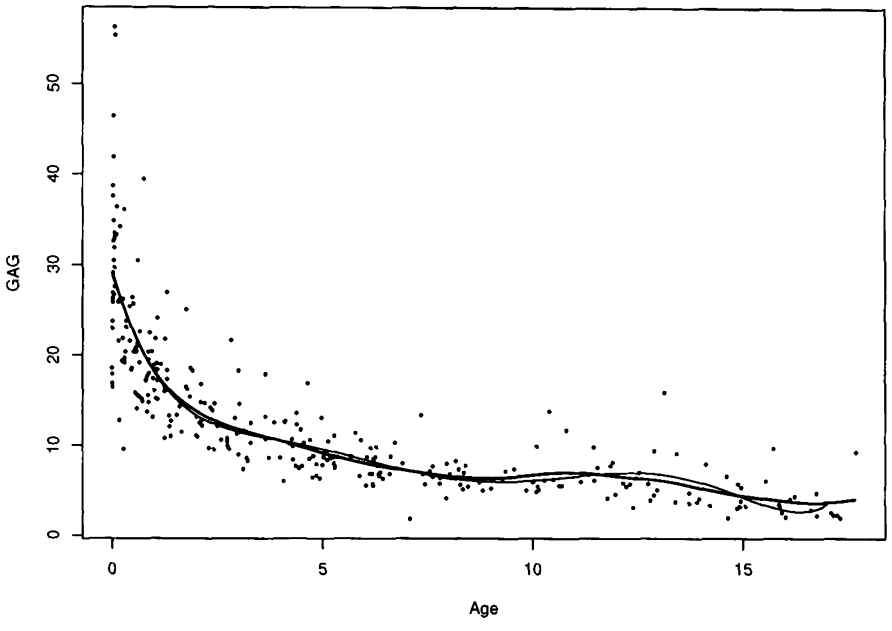


Fig. 1. Two smooth curves fitted to the concentration of the chemical GAG in the urine of 314 children aged 0–18 years.

mial chosen by forwards stepwise selection. The other is a smoothing spline, with the degree of smoothness chosen by GCV.[d] There are many alternative approaches, including neural networks and local polynomials [46, 26]. These can all fit very similar curves, and the issue of choosing between the model classes is rather a moot one.

Alternative explanations with roughly equal support are commonplace: model averaging seems a good solution. Selecting several models, studying their predictions and taking a consensus is also a good idea, *when time permits* and when *non-quantitative information is available.* As Figure 1 shows, we may need other information to choose between very different formulae with similar predictions, in so far as we can choose at all.

'Regression diagnostics' are often based on approximations to overfitting or case deletion. Now we can (and some of us do) fit extended models with smooth terms or use fitting algorithms that automatically downweight groups of points. (I rarely use least squares these days.) It is still all too easy to select a complex model just to account for a tiny proportion of aberrant observations.

[d] *generalized cross-validation*, which is not in fact cross-validation as defined here.

Although we do have more tools available than at the start of my career, it seems to me that model selection has actually got harder: as we explore more of the statistical model world we encounter more and more chasms awaiting the unwary. It worries me how casually AIC and its allies are used, and hope this paper will go some way to raising awareness of the limitations of formal methods of model selection.

References

1. Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (Eds B. N. Petrov and F. Cáski), pp. 267–281, Budapest. Akademiai Kaidó. Reprinted in *Breakthroughs in Statistics*, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. New York: Springer.
2. Akaike, H. (1974) A new look at statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.
3. Allen, D. M. (1971) Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 325–331.
4. Allen, D. M. (1974) The relationship between variable selection and data augmentation and a method of prediction. *Technometrics*, **16**, 467–475.
5. Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* Chichester: Wiley.
6. Breiman, L. (1996a) Bagging predictors. *Machine Learning*, **24**, 123–140.
7. Breiman, L. (1996b) The heuristics of instability in model selection. *Annals of Statistics*, **24**, 2350–2383.
8. Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.
9. Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees.* Monterey, CA: Wadsworth and Brooks/Cole.
10. Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference.* New York: Springer, second edition.
11. Cooke, R. M. (1991) *Experts in Uncertainty. Opinion and Subjective Probability in Science.* New York: Oxford University Press.
12. Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proc. 4th Berkeley Symposium* (Ed. J. Neyman), volume 1, pp. 105–123, University of California Press. University of California Press.

13. Cox, D. R. (1962) Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society series B*, **24**, 406–424.
14. Draper, D. (1995) Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society series B*, **57**, 45–97.
15. Drucker, H., Cortes, C., Jaeckel, L. D., LeCun, Y. and Vapnik, V. (1994) Boosting and other ensemble methods. *Neural Computation*, **6**, 1289–1301.
16. Ein-Dor, P. and Feldmesser, J. (1987) Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM*, **30**, 308–317.
17. Freund, Y. (1990) Boosting a weak learning algorithm by majority. In *Proceedings of the Third Workshop on Computational Learning Theory*, pp. 202–216. Morgan Kaufmann.
18. Freund, Y. (1995) Boosting a weak learning algorithm by majority. *Information and Computation*, **121**(2), 256–285.
19. Freund, Y. and Schapire, R. E. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the Second European Conference on Computational Learning Theory*, pp. 23–37. Springer.
20. Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association*, **70**, 320–328.
21. Geisser, S. (1993) *Predictive Inference: An Introduction.* New York: Chapman & Hall.
22. Harrell, Jr., F. E. (2001) *Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression and Survival Analysis.* New York: Springer-Verlag.
23. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. and Hinton, G. E. (1991) Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87.
24. Krogh, A. and Vedelsby, J. (1995) Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7. Proceedings of the 1994 Conference* (Eds G. Tesauro, D. S. Touretzky and T. K. Leen), pp. 231–238, Cambridge, MA. MIT Press.
25. LeBlanc, M. and Tibshirani, R. J. (1993) Combining estimates in regression and classification. Preprint, Depts of Preventive Medicine and Biostatistics and of Statistics, University of Toronto.
26. Loader, C. (1999) *Local Regression and Likelihood.* New York: Springer-Verlag.

27. Madigan, D. and Raftery, A. E. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the Royal Statistical Society*, **89**, 1535–1546.
28. Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–675.
29. Marchini, J. L. and Ripley, B. D. (2000) A new statistical approach to detecting significant activation in functional MRI. *NeuroImage*, **12**, 366–380.
30. Moody, J. E. (1991) Note on generalization, regularization and architecture selection in nonlinear learning systems. In *First IEEE-SP Workshop on Neural Networks in Signal Processing*, pp. 1–10, Los Alamitos, CA. IEEE Computer Society Press.
31. Moody, J. E. (1992) The *effective* number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In *Advances in Neural Information Processing Systems 4. Proceedings of the 1991 Conference* (Eds J. E. Moody, S. J. Hanson and R. P. Lippmann), pp. 847–854, San Mateo, CA. Morgan Kaufmann.
32. Mosteller, F. and Wallace, D. L. (1963) Inference in an authorship problem. *Journal of the American Statistical Association*, **58**, 275–309.
33. Murata, N., Yoshizawa, S. and Amari, S. (1991) A criterion for determining the number of parameters in an artificial neural network model. In *Artificial Neural Networks. Proceedings of ICANN-91* (Eds T. Kohonen, K. Mäkisara, O. Simula and J. Kangas), volume I, pp. 9–14, Amsterdam. North Holland.
34. Murata, N., Yoshizawa, S. and Amari, S. (1993) Learning curves, model selection and complexity of neural networks. In *Advances in Neural Information Processing Systems 5. Proceedings of the 1992 Conference* (Eds S. J. Hanson, J. D. Cowan and C. L. Giles), pp. 607–614, San Mateo, CA. Morgan Kaufmann.
35. Murata, N., Yoshizawa, S. and Amari, S. (1994) Network information criterion—determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, **5**, 865–872.
36. Quinlan, J. R. (1996) Bagging, boosting, and C4.5. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press.
37. Ripley, B. D. (1992) Classification and clustering in spatial and image data. In *Analyzing and Modeling Data and Knowledge* (Ed. M. Schader), pp. 93–105, Berlin. Springer.

38. Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.
39. Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986) *Akaike Information Theory Statistics.* Dordrecht: Reidel.
40. Schapire, R. E. (1990) The strength of weak learnability. *Machine Learning*, **5**(2), 197–227.
41. Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.
42. Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society series B*, **64**, 583–639.
43. Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society series B*, **36**, 111–147.
44. Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society series B*, **39**, 44–47.
45. Takeuchi, K. (1976) Distribution of informational statistics and a criterion of fitting. *Suri-Kagaku*, **153**, 12–18. [In Japanese].
46. Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing.* London: Chapman & Hall.
47. Wolpert, D. H. (1992) Stacked generalization. *Neural Networks*, **5**, 241–259.

PRINCIPLES OF DESIGNED EXPERIMENTS IN J. A. NELDER'S PAPERS

R. A. Bailey
School of Mathematical Sciences
Queen Mary, University of London
Mile End Road
London E1 4NS

1. Experimental protocol

When I am involved as a statistician in the design of a scientific experiment, I usually keep a written record under the following headings.

(1) How many experimental units are there, and how were they structured before the treatments were applied?
(2) How many treatments are there, and how were they structured before being allocated to experimental units?
(3) What is the systematic, or combinatorial, design used (for example, incomplete-block design with specified efficiency factors for a complete basis of treatment contrasts, split-plot design with certain factors on whole plots and others on subplots, fractional factorial design with specified aliasing, etc)? Why was it chosen?
(4) What method of randomization was used, and what *layout* did it produce?
(5) What will be the assumed expectation structure of the responses?
(6) What will be the assumed covariance structure of the responses?
(7) What is the proposed analysis of the data?

John Nelder has contributed to the thinking on all of these topics. In the sections below, I enlarge on his contributions and give some account of the work that followed from his. Section 2 emphasizes that structure

on the experimental units is different from structure on treatments. Sections 3 and 4 describe families of structures that may be suitable for either experimental units or treatments. Section 5 relates randomization to the structure on the experimental units, and Section 6 in turn relates both of these to the covariance structure. Families of expectation models defined by qualitative factors are covered in Section 7, while Section 9 touches briefly on quantitative factors. Section 8 discusses properties of the combinatorial design, in the sense of how treatments are allocated to plots. The proposed analysis, and its implications for design, are briefly discussed in Section 10.

2. Plot structure is different from treatment structure

Many books on the design of experiments, from [27] to [30, 35, 40], use notation such as y_{ij} to denote the response on the j-th experimental unit to which treatment i is applied. In other words, the experimental units are seen to have no intrinsic names before treatments are applied. This labelling ensures that the trial is not blind; it also encourages the scientist to collect data, or conduct mid-trial operations, in treatment order rather than according to any structure on the units.

In [43], John Nelder remarked that "experimental units ... have an internal structure regardless of whether any differential treatments are applied to them or not". He called this structure *block structure*; I call it *plot structure*, as *plot* is my shorthand for *experimental unit*. The plot structure should be specified, and the individual experimental units named or numbered, before treatments are applied. Information about treatment allocation can be added later, to show the full layout.

For example, a completely randomized trial for three pig feeds might be laid out as follows.

Pig	1	2	3	4	5	6	7	8	9	10	11	12
Feed	3	1	1	2	3	1	3	2	1	2	3	2

Here the experimental units are the pigs and the plot structure is *no structure*. The pigs should be weighed in the order 1–12, and the weight of the ninth pig recorded as y_9, not as y_{14}.

Alternatively, the pigs might comprise four pigs from each of three litters. Then the grouping into litters provides the plot structure. Many people would call the litters *blocks*. The layout might then be as follows.

Litter	1	1	1	1	2	2	2	2	3	3	3	3
Pig	1	2	3	4	5	6	7	8	9	10	11	12
Feed	3	1	1	2	3	1	3	2	1	2	3	2

Now the ninth pig can be called pig 9, giving weight y_9, or pig 1 in litter 3, giving weight y_{31}. In both cases it is the plot structure that provides the pig's label.

Combinatorial books, such as [15, 29, 58] typically have a different problem with this example. They see three treatments, and three blocks of size four, but they see no pigs at all. It is hard to convey the importance of plot structure to people steeped in this point of view.

Plot structure is usually determined by blocks, rows, columns etc. *Treatment structure* means the structure inherent in the set of treatments before they are applied to plots. Typical treatment structures are: factorial; new treatments plus control; treatments grouped into types. The distinction between the plot structure and the treatment structure must be made at the design stage and should be maintained when the data are analysed. It is common for otherwise admirable books for non-statisticians, such as [32], to ignore the distinction at the analysis stage.

Following [43], Genstat forces the user to specify plot structure and treatment structure separately using the directives `blockstructure` and `treatmentstructure`. See [50, Sections 4.1–4.2].

Some more complicated experiments, with several phases or with later treatments superimposed on previous ones, involve three or more structured sets rather than two. Brien [20] pointed out that proper design and analysis must be based on recognition of these distinct structures. Nelder's ideas for two structures are generalized to three or more by Brien and Bailey [21].

3. Crossing and nesting

John Nelder introduced the crossing and nesting operators ($*$ and $/$) in his important pair of papers [43, 44] on simple orthogonal block structure. They form the basis of the model structure formulas for both plots and treatments in Genstat analysis of variance [50, Chapter 4], and are also used in Genstat regression [50, Section 3.3]. As Heiberger [33] shows, equivalent operators are used in many statistical computing packages, although sometimes the notation is different. Nelder himself originally used $\rightarrow$ for nesting. However, the definitions in the statistical computing packages are subtly but importantly different from Nelder's original definitions. I shall call the original usage the *structural* one and the other the *modelling* one.

The differences between the two can be illustrated by the formula $R*C$. Here I use U to denote the trivial factor with a single level, which gives the model term for the grand mean, called μ by the Iowa school [36], and which

is typically omitted from analysis-of-variance tables (interestingly, it is not omitted by Nelder [47] nor always by Mead [38]); I use E to denote the other trivial factor, which is called ε by the Iowa school and `**units**` in Genstat. I shall explain later why I call some things *pre-factors* rather than factors.

The two usages of the formula $R*C$ differ not only in the interpretation of the formula but also in what information must be available when the formula is used. In the modelling view you must have already declared (i) a set of units and (ii) factors R and C on those units, i.e. the level of R and C on each unit. If $R*C$ is a treatment formula, it is an instruction to *fit* the following model terms, in order: U; R allowing for U; C allowing for R; $R.C$ allowing for R and C. If it is a plot formula, it is an instruction to decompose the data space into the following subspaces : U; R allowing for U; C allowing for R; $R.C$ allowing for R and C; E allowing for $R.C$. On the other hand, for the structural view you must have declared simply the sets of levels of a pre-factor R and a pre-factor C. Now $R*C$ means first *create* a set of units, being all ordered pairs of the form (level i of R, level j of C) and then *note* that the factors that you want on this set of units are

$$\begin{aligned} U &\text{ — no coordinates} \\ R &\text{ — first coordinate} \\ C &\text{ — second coordinate} \\ R.C &\text{ — both coordinates.} \end{aligned}$$

Thus the modelling view is that, *given* some factors on a set, the formula $R*C$ is a convenient shorthand for an ordered list of model terms. This formula not only saves us writing out $R+C+R.C$ but also helps us to think about the factors and their relationship to each other. It is extremely useful for breaking down treatment terms. For example (jumping ahead a little to a mixture of crossing and nesting), consider the treatment structure shown in Figure 1. This is often described wrongly, by people who fail to realise that time of application makes no difference to zero nitrogen, as having structural formula *timing* $*$ *quantity*. However, for a meaningful analysis, the modelling formula *control*/(*timing* $*$ *quantity*) should be used, where *control* is a 2-level factor distinguishing the treatment with zero nitrogen from the other six treatments.

To use the formula $R*C$ in the modelling way we do not need either or both of R and C to be uniform (also called *balanced* or *equi-replicate*), nor do we need R and C to be orthogonal to each other, nor do we need all aliasing between R and C to be in the grand mean, nor do we require that

neither be marginal to the other.

Before continuing with the structural view, let me change notation a little. I write $R \wedge C$ for the factor whose model term is $R.C$. Thus R and C are both marginal to $R \wedge C$, and $R \wedge C$ has the smallest number of levels subject to this restriction. One reason for using this notation is to make clear the connection with, and difference from, the factor which I write as $R \vee C$ and which is defined very similarly to $R \wedge C$ but in a dual sense. Thus $R \vee C$ is the unique factor which is marginal to both R and C and which has the maximum number of levels subject to this constraint. In particular, if R is marginal to C then $R \wedge C = C$ and $R \vee C = R$. Using this notation, the condition *all aliasing between R and C is in the grand mean* can be written in symbols as $R \vee C = U$.

Now consider the structural view. This is that, *given* some sets of levels of pre-factors, the formula $R * C$ denotes a *new* structured set with various specified factors. Strictly speaking, the pre-factor R that we start with is not quite the same as the factor R on the new set, which is why I use two distinct words; the distinction should be clearer when we consider nesting. Thus the formula $R * C$ tells us to *construct* a set and four factors on that set; moreover, it tells us how to do it. This is useful, because you do not need to declare the set explicitly, nor to generate factors R and C in standard order. The construction automatically ensures that R and C are both uniform, that they are orthogonal to each other, that $R \vee C = U$ and that neither of the factors R and C is marginal to the other.

Now compare the two usages for nesting. In the modelling view it is assumed that you have already declared (i) a set of units and (ii) factors B and P on those units. If B/P is a treatment formula, it is interpreted as an instruction to *fit* the following model terms, in order: U; B allowing

timing	amount of nitrogen			
	0	80	160	240
early				
late				

Fig. 1. An example where the modelling view is useful

for U; $B \wedge P$ allowing for B. If it is a plot formula, it is an instruction to decompose the data space as follows: U; B allowing for U; $B \wedge P$ allowing for B; E allowing for $B \wedge P$. On the other hand, the structural view is that you have declared simply the sets of levels of a pre-factor B and a pre-factor P. Now the formula B/P means first *create* a set of units, being all ordered pairs of the form (level i of B, level j of P) and then *note* that the factors that you want on this set of units are

$$\begin{aligned} U &\quad \text{— no coordinates} \\ B &\quad \text{— first coordinate} \\ B \wedge P &\quad \text{— both coordinates.} \end{aligned}$$

Thus P is *not a factor*. The modelling view allows unequal block sizes (if the levels of B are called blocks) and plots bigger than single units (if the levels of P are called plots), whereas the structural view does not. Likewise, the structural view can never give the factorial-treatments-plus control structure exemplified in Figure 1.

Very many of the designed experiments that we analyse have plot structures that can be constructed using $*$ and $/$ in the structural way. For example, the plot structure of a row-column design is $row * column$: all we need to know is *how many* rows and how many columns. Thus a formula such as

$$6\ rows * 10\ columns \tag{1}$$

gives complete information. Similarly, the plot structure for the second version of the pigs example is

$$3\ litters/4\ pigs. \tag{2}$$

4. Orthogonal structures

Structure on either the experimental units or the treatments can often be specified by iterated crossing and nesting in the structural sense. Thus the plot structure in a consumer experiment might be

$$3\ months/(4\ weeks * 4\ housewives)/2\ washloads \tag{3}$$

or the treatment structure in a horticultural experiment might be

$$3\ varieties * 3\ composts * 3\ watering\ regimes. \tag{4}$$

In [43], Nelder called these *simple orthogonal block structures*. He showed that such a structure gives a unique orthogonal decomposition of the data

space: there is a subspace for fitting each factor, allowing for each factor marginal to it.

Marginality relations are conveniently shown on a Hasse diagram. Each factor is represented by a black dot. If A is marginal to B then A is joined to B by a line or series of lines running generally downwards. The number of levels of each factor is written beside it. Its number of degrees of freedom, shown after this, is calculated by subtracting all the degrees of freedom above from the number of levels. The Hasse diagrams for structures (1)–(4) are in Figures 2–5 respectively (where convenient, names of pre-factors are abbreviated to their initial letters).

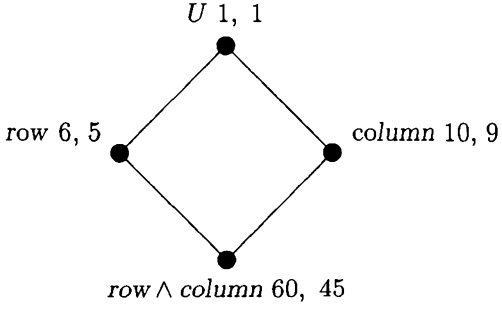


Fig. 2. Hasse diagram for the plot structure in (1)

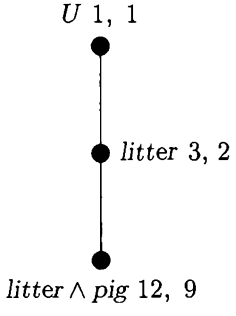


Fig. 3. Hasse diagram for the plot structure in (2)

In the structure $R*C$, different levels of C are important whether or not the units under consideration have the same level of R. On the other hand, in the structure B/P we do not care about levels of the pre-factor P unless we know that the levels of B are equal. Let us say that B *dominates* P. (Many people would say that B *nests* P, but this is subtly different from the two meanings of nesting as an operator.) Then dominance is a partial order, so the pre-factors form a partially ordered set (poset), which can also be shown on a Hasse diagram, this time using white dots for pre-factors. These for structures (1)–(4) are in Figures 6–9. Every simple orthogonal block structure gives a poset of pre-factors. There is a genuine factor in the structure for each collection of pre-factors which satisfies the condition that if B is in the collection and A dominates B then A is in the collection.

If we start with a poset of pre-factors then we obtain a structure which gives an orthogonal decomposition just as above. These new structures, which I now call *poset block structures*, were developed in this form in [11, 56, 57]. They are more general than simple orthogonal block structures.

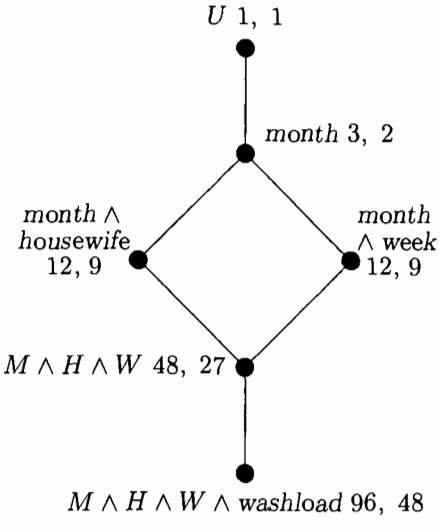


Fig. 4. Hasse diagram for the plot structure in (3)

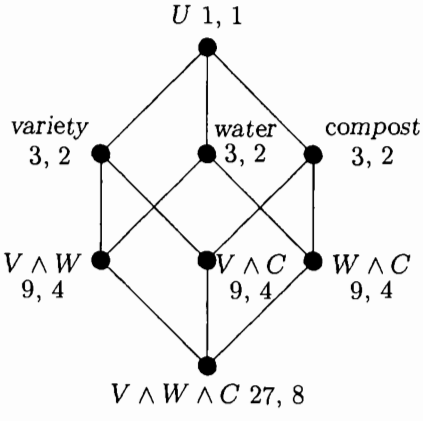


Fig. 5. Hasse diagram for the treatment structure in (4)

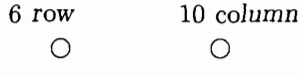


Fig. 6. Poset of pre-factors in the simple orthogonal block structure (1)

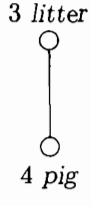


Fig. 7. Poset of pre-factors in the simple orthogonal block structure (2)

For example, a trial on chemical cleaners in milking parlours had the plot structure defined by the poset in Figure 10. This cannot be obtained by iterated crossing and nesting.

Nelder gave an iterative procedure for obtaining the orthogonal decomposition of a simple orthogonal block structure. This does not extend to poset block structures, but the Hasse-diagram procedure works for these, and hence for simple orthogonal block structures. In fact, the Iowa school [36], especially Zyskind [63], had effectively invented poset block structures before Nelder, but they confused the two sorts of Hasse diagram and their description did not lead to straightforward algorithms.

Speed and Bailey [56] gave a further generalization, which I now call an *orthogonal block structure* [5]. Each of these consists of a collection of

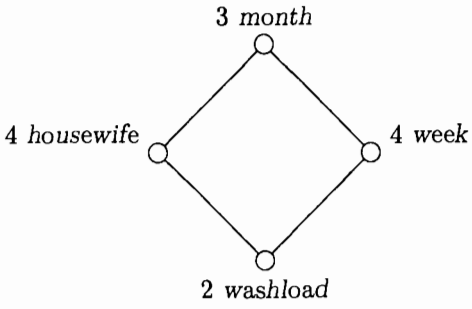


Fig. 8. Poset of pre-factors in the simple orthogonal block structure (3)

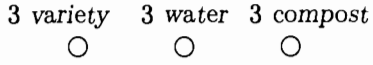


Fig. 9. Poset of pre-factors in the simple orthogonal block structure (4)

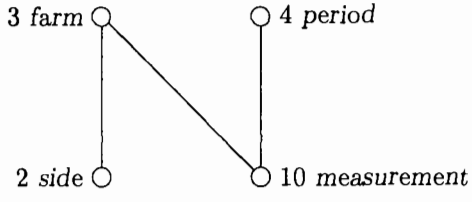


Fig. 10. A poset that cannot be obtained from iterated crossing and nesting

uniform factors, including the two trivial ones, such that if A and B are included then so are $A \wedge B$ and $A \vee B$, and A is orthogonal to B in the sense that their averaging matrices commute. These conditions ensure that the Hasse-diagram algorithm gives an orthogonal decomposition. The least complicated orthogonal block structure which is *not* a poset block structure is a Latin square.

Tjur [59] generalized further by dropping the condition of uniformity and the requirement that $A \wedge B$ is included if A and B are. The treatment structure in Figure 1 is a Tjur structure, whose Hasse diagram is in Figure 11. The Hasse-diagram algorithm still works.

Just as crossing and nesting can be applied to a pair of structures which are not both simple orthogonal block structures, so each poset of size n gives an n-ary operator for combining structures [9]. However, neither orthogonal block structures nor Tjur structures can, in general, be defined by recursive use of such operators.

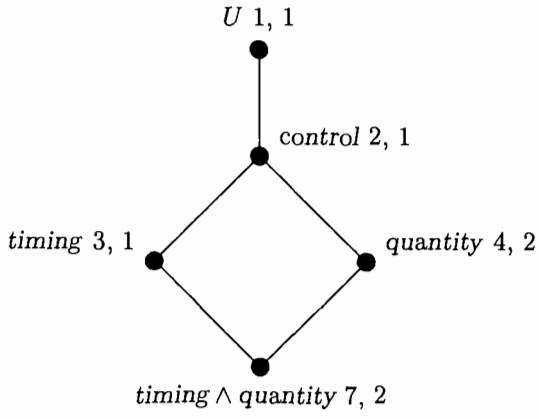


Fig. 11. The Tjur structure from Figure 1

5. Randomization

The experimenter typically associates randomization with treatments, often thinking of choosing a plan at random. Nelder [43] pointed out that "The close association of randomization with the allocation of treatments can obscure the essential independence of the block structure ... from the treatment structure." He argued that randomization should consist of choosing a random permutation of the experimental units from all those permutations that preserve the plot structure. This procedure justifies the assumption of a randomization distribution for the null responses, to which treatment effects are added. Bailey [2] spelt this out in more detail.

Because randomization should normally be defined by the plot structure, Genstat makes the link explicit. Once the `blockstructure` directive has been given, the procedure `arandomize` does the job, provided that the formula given for the plot structure is a simple orthogonal block structure. Alternatively, the `randomize` directive may be used; in this case the plot structure must be specified as an option. See [50, Section 4.10].

Bailey [2, 3] showed that randomization gives the covariance matrix $\mathbf{C}$ of the responses in the patterns-of-covariance form; that is,

$$\mathbf{C} = \sigma^2 \sum_i \rho_i \mathbf{A}_i, \tag{5}$$

where the $\mathbf{A}_i$ are symmetric $(0,1)$-matrices whose sum is $\mathbf{J}$. If $\mathbf{A}_i(\alpha,\beta) = 1$ then $\mathrm{Cov}(Y_\alpha, Y_\beta) = \sigma^2\rho_i$. Moreover, for each i, if $\mathbf{A}_i(\alpha,\beta) = 1$ then the set of pairs (γ,δ) for which $\mathbf{A}_i(\gamma,\delta) = 1$ is precisely the set of pairs (α^g, β^g) as g ranges over the permutations that preserve the plot structure.

6. Variance components

Nelder [43] showed that, for a simple orthogonal block structure, (5) could be rewritten in two equivalent forms: the components-of-variance form

$$\mathbf{C} = \sigma^2 \sum_i f_i \mathbf{B}_i, \tag{6}$$

where $\mathbf{B}_i$ is the totalling matrix for factor i, so that $\mathbf{B}_i(\alpha, \beta) = 1$ if α and β have the same level of factor i; and the spectral form

$$\mathbf{C} = \sum_i \xi_i \mathbf{P}_i, \tag{7}$$

where the $\mathbf{P}_i$ are the orthogonal projectors onto the eigenspaces of $\mathbf{C}$, irrespective of the values of the correlations ρ_i. He wrote that "these matrices define the *strata*", from which I concluded that the strata are the mutual eigenspaces of all matrices of type (5). The ξ_i are called the *stratum variances*.

Nelder [44] and Bailey [2] showed that if the data are projected onto any single stratum then the projected data effectively have a scalar covariance matrix, so one may use the standard results from the simple textbook model. Some treatment contrasts may be estimable in more than one stratum; Nelder gave a method for combining the different estimators in [46]. These methods (stratum projection followed by combination of information) work for any covariance matrix $\mathbf{C}$ whose eigenspaces are independent of the unknown (co)variance parameters. Nelder [43, page 153] called such a covariance matrix *orthogonal*; following this, Houtman and Speed [34] defined an orthogonal block structure to be any such covariance matrix. I prefer to reserve the term for particular combinatorial structures defined by one or more partitions into blocks, as in [5].

How is this connected with randomization? Nelder claimed in [43] that randomization of a simple orthogonal block structure gives Equation (5) for a specific form of the matrices $\mathbf{A}_i$: namely, $\mathbf{A}_i(\alpha, \beta) = 1$ if α and β have the same level of factor i but not of any factor to which i is marginal (although he did not express it in this form). However, he did not actually prove that the group of all permutations preserving a simple orthogonal block structure has this property. This was done later, by Bailey, Praeger, Rowley and Speed [11], who proved it for the larger class of poset block structures. By considering randomization of Latin-square plot structures, Preece, Bailey and Patterson [53] showed that the analogous result does not hold for all orthogonal block structures, although it does hold for some orthogonal block structures that are not poset block structures. Nevertheless, the equivalence

between (5), (6) and (7) does hold for all orthogonal block structures. In a Tjur structure with uniform factors, expressions (6) and (7) are equivalent, but it is not possible, in general, to express $\mathbf{C}$ in the form (5); that is, the covariance between Y_α and Y_β cannot be completely specified in terms of which factors take the same levels on α and β.

Bailey [3] explored the question of which structures have the property that their randomization model (5) can be expressed as (7) with known strata. The answer includes circles (such as the edge of a petri dish) and unordered pairs (such as people participating in experiments to compare telephone equipment), as well as structures obtained from these by crossing, nesting and poset operators. Such structures are now called *stratifiable* [14].

If any system of blocks in the plot structure has blocks of unequal size then the preceding theory breaks down. Clearly one cannot interchange a larger block with a smaller block in randomization. Nor are expressions (6) and (7) equivalent if blocks have different sizes. For a structure defined solely by one system of blocks of different sizes, if (6) holds then the within-blocks space is a stratum but the between-blocks space is not a stratum in the sense of eigenspace, even though writers such as Pearce [52, Section 4.8] call it one. Caliński and Kageyama [24, Section 3.2.2] demonstrate the difficulties that can occur in this case.

Because $\mathbf{C}$ must be non-negative definite, and (7) is its spectral form, the stratum variances ξ_i must be non-negative. The usual estimators of these from analysis of variance, possibly augmented by combination of information, are mean squares or positive linear combinations of mean squares, which are never negative. However, the components of variance f_i are linear combinations of the ξ_i, and their estimates may be negative. Nelder [42, 47] calls these *components of excess variance* and argues that negative estimates should be accepted at face value. Others, such as John and Williams [35, page 150] recommend that such estimates should be replaced by zero. Of course, this replacement procedure biases the estimators upwards, as [13] shows; moreover, Wolde-Tsadik and Afifi [62] conclude from an empirical study that such pooling of mean squares is inadvisable.

7. Model fitting, estimation and testing

There is some debate about whether the analysis of a designed experiment should focus on estimation or hypothesis testing. I think that we should do both. The first step is model fitting. We start with a full model and successively perform hypothesis tests to see if we can reduce it to a more

parsimonious model. Each of these tests should use a residual term that is orthogonal to the space for the full model. Once we have chosen a model, we parameterize it suitably and then estimate those parameters. We do not need to provide parameterizations for all the other models that we might have chosen.

For example, given the seven treatments in Figures 1 and 11, the models entertained for the effects of treatments are probably those in Figure 12. The inclusion relationships between these models, *as* well *as* their dimensions, are shown in the model diagram in Figure 13, with the usual convention that larger models are at the top, which is the opposite to the convention for analysis-of-variance tables. The diagram is not simply the inversion of Figure 11, because we include the model where *timing* and *quantity* are additive.

Model testing starts at the top of the model diagram. We use the difference between the sum of squares for *treatment*, SS(*treatment*), and the sum of squares for *timing* + *quantity*, SS(*timing* + *quantity*), to assess whether we can reduce the full model to the five-dimensional one. If we cannot, we accept the full treatment model and estimate its seven parameters. If we can, then we use a similar procedure to assess whether we can simplify the model further. However, there is now a choice about which branch to choose first: do we compare the five-dimensional model with *timing* or with *quantity*? If the six non-zero treatments are equally replicated then the factors *timing* and *quantity* are orthogonal to each other, which has the consequence that

$$\mathrm{SS}(\textit{timing} + \textit{quantity}) - \mathrm{SS}(\textit{timing}) = \mathrm{SS}(\textit{quantity}) - \mathrm{SS}(\textit{control})$$

so that we will choose the same model no matter which route we trace down the model diagram. As is well known, if two factors are not orthogonal to

zero	$E(Y_\alpha) = 0$
null	$E(Y_\alpha) =$ some constant κ for all α
control	$E(Y_\alpha) = \lambda_i$ if α has level i of *control*
timing	$E(Y_\alpha) = \tau_j$ if α has level j of *timing*
quantity	$E(Y_\alpha) = \gamma_k$ if α has level k of *quantity*
timing + *quantity*	$E(Y_\alpha) = \tau_j + \gamma_k$ if α has level j of *timing* and level k of *quantity*
treatment	$E(Y_\alpha) = \delta_l$ if α has level l of *treatment*

Fig. 12. List of models for the treatment structure in Figure 1

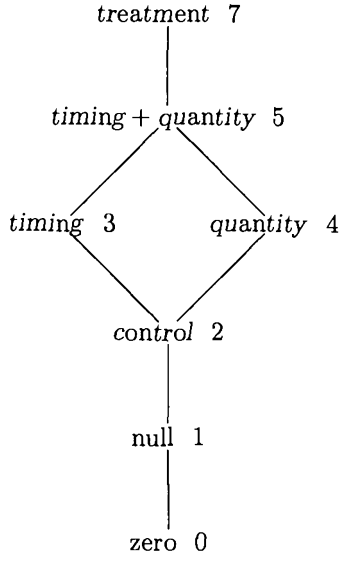


Fig. 13. Model diagram for the models listed in Figure 12

each other then there are some data which can be explained by either factor alone and other data which can be explained by both factors even though neither alone seems to have an effect. Thus orthogonality is important not only to make arithmetic easy but to make inference unambiguous.

Of the seven models that we might fit in this example, only one has any ambiguity about the parameters. For *timing* + *quantity* we can replace any set of estimates $\hat{\tau}_j$ and $\hat{\gamma}_k$ by $\hat{\tau}_j + c$ and $\hat{\gamma}_k - c$ for any constant c. In addition, $\hat{\tau}_0$ and $\hat{\gamma}_0$ can be replaced by $\hat{\tau}_0 + d$ and $\hat{\gamma}_0 - d$ for a different constant d. There is no need for this potential ambiguity to muddle the way that we think about model fitting. For this model, the results can be presented as a two-way table of fitted values, together with standard errors of differences.

I have laboured this example in some detail as an introduction to Nelder's famous campaign against " the neglect of marginality " [47, Section 2.1]. Most expositions of model fitting do not show a model diagram like Figure 13. Instead, they present a single composite equation for $E(\mathbf{Y})$ which can be used to specify the parameters no matter which model is chosen. Unfortunately, it is usually possible to use this equation to specify some extra, incoherent models, simply by setting some items in this equation to

be zero.

A slightly more complicated, but rather common, example makes this abundantly clear. The treatment structure in Equation (4) and Figure 5 gives the family of models in Figure 14. Model fitting proceeds from the top downwards, as described before. Any model in the diagram has a perfectly straightforward parameterization, and the parameters can be estimated once the model is chosen. Yet many books, including [30, 35], present a composite equation such as

$$E(Y_\omega) = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

when plot ω has level i of *variety*, level j of *water* and level k of *compost*. This encourages people to think that there is a sensible model with, say $\alpha_i = \beta_j = \gamma_k = (\alpha\beta)_{ij} = (\alpha\gamma)_{ik} = (\alpha\beta\gamma)_{ijk} = 0$ and $(\beta\gamma)_{jk}$ nonzero.

The problem is made worse by the different sorts of sums of squares (and hence mean squares) that are routinely produced by statistical software packages. These are based not on extra fits according to the model diagram, but on extra fits according to the composite parameters in the multipurpose equation for $E(\mathbf{Y})$. According to [30, 31, 40], a Type I sum of squares for a factor shows the magnitude of the extra fit when that factors' composite parameters are allowed to be non-zero, given the order in which the factors have been fitted. In Figure 13, it is appropriate for *timing* if *control* is fitted before it; it is not appropriate for *quantity* if the two composite parameters for *treatment* are fitted first. A Type II sum of squares shows the magnitude of the extra fit for a factor if all factors except those marginal to it have been fitted first. In Figure 13, this sum of squares for *timing* is appropriate for comparing the model *timing* + *quantity* with *quantity* but not necessarily for comparing *timing* with *control*. A Type III sum of squares gives the magnitude of the extra fit after all other composite parameters have been fitted. In Figure 13, this is appropriate for *treatment* but for no other term.

In my opinion, these sums of squares are all nonsense, because they are attached to model terms and hence to composite parameters. Instead, they should be attached to lines in the model diagram. If model M_1 is immediately above model M_2 then the test for reducing M_1 to M_2 uses the following mean square:

$$\frac{\mathrm{SS}(M_1) - \mathrm{SS}(M_2)}{\dim(M_1) - \dim(M_2)}.$$

This is the number that should be attached to the line joining M_1 to M_2. Of course, if all treatment factors are orthogonal to each other then the same number will be attached to several different lines, so there is redundancy

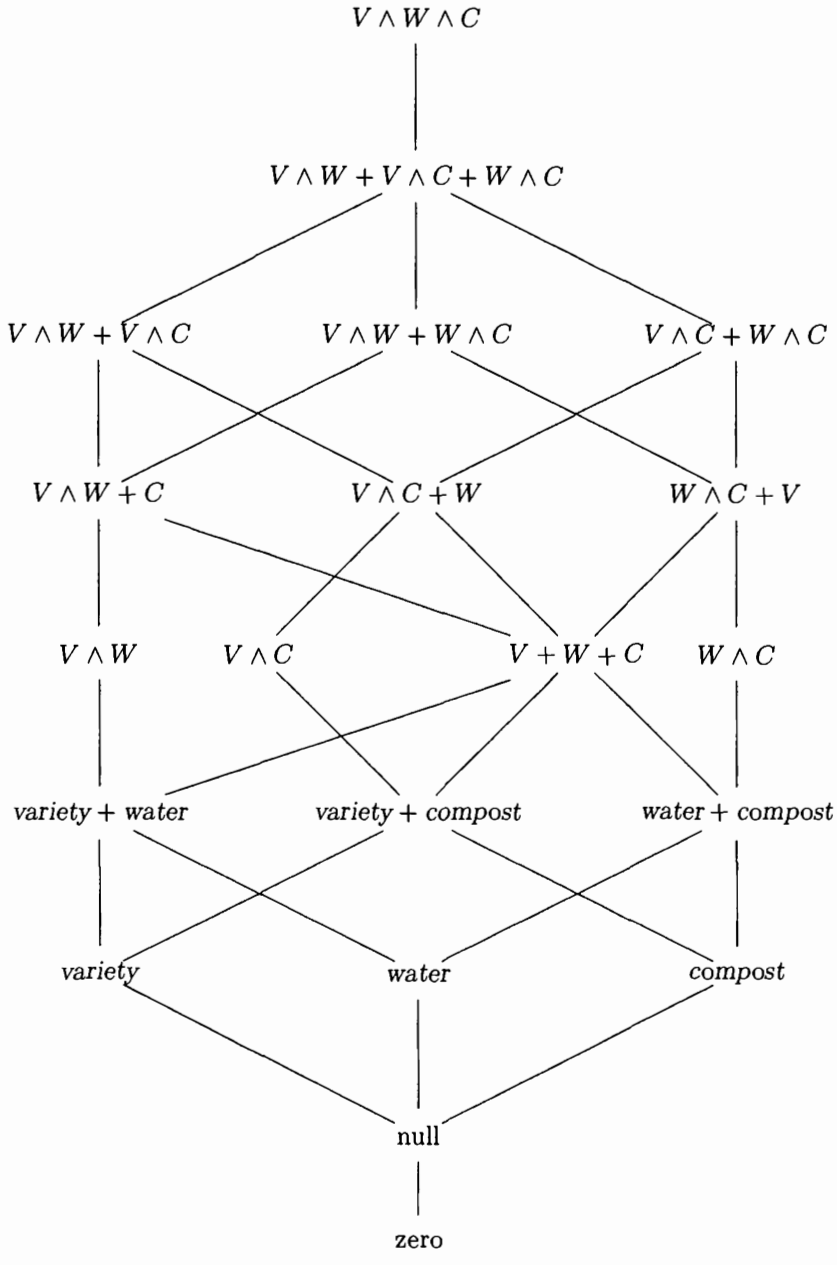


Fig. 14. Family of models for three treatment factors in Figure 9

in the diagram. Nevertheless, this approach should discourage the fitting of incoherent models, as even the relatively small examples in Figures 13 and 14 show.

Returning to his campaign in [48], Nelder described three false steps.

"1. The putting of constraints on parameters because constraints are put on their estimates.
"2. Neglecting marginality relations between terms in factorial models.
"3. Confusing non-centrality parameters in the expectation of sums of squares with the corresponding hypotheses that they might be used to test."

I believe that all of these could be avoided by use of the model diagram, by attaching each mean square to a line in the model diagram, and by refusal to parameterize until the model is chosen.

8. Combinatorial design

Given a stratifiable plot structure, we have a decomposition of the data space into strata, with orthogonal projectors $\mathbf{P}_i$. Given the treatment structure, we have a family of models and an orthogonal decomposition of the treatment space. If possible, the allocation of treatments to experimental units should be done in such a way that this orthogonality is maintained when the treatment space is regarded as a subspace of the data space. Equal replication is one way of ensuring this, but not the only way. For the seven treatments in Figure 1, the replication of the zero treatment has no effect on orthogonality. Let $\mathbf{T}_j$ be the orthogonal projectors onto the spaces decomposing the treatment space when it is embedded in the data space.

Given all the above, Nelder [44] defined what he meant for a design to be balanced. By [46] he had refined this definition, as follows. The design (in the sense of treatment allocation) is *generally balanced* if there are constants λ_{ij} such that

$$\mathbf{T}_j\mathbf{P}_i\mathbf{T}_j = \lambda_{ij}\mathbf{T}_j \tag{8}$$

and

$$\mathbf{T}_j\mathbf{P}_i\mathbf{T}_k = \mathbf{0} \qquad \text{if } j \neq k. \tag{9}$$

For such a design, $\mathbf{T}_j^2 = \mathbf{T}_j \sum_i \mathbf{P}_i\mathbf{T}_j = \sum_i \lambda_{ij}\mathbf{T}_j$, so $\sum_i \lambda_{ij} = 1$ for each j. The constant λ_{ij} is called the *proportion of information* on treatment term j in stratum i. If every λ_{ij} is equal to 0 or 1, the design is said to be *orthogonal* [46, p. 304].

How do these definitions help us to design good experiments? If we can make the design orthogonal then each treatment term is estimated in a single stratum. We use prior knowledge of the likely relative magnitudes of the stratum variances to ensure that treatment terms whose effects are most important (alternatively, those whose magnitude may be hard to detect) are assigned to strata whose variances are small.

If the design is generally balanced and the relative sizes of the ξ_i are known, then the best linear unbiased estimator of a normalized treatment contrast in $\mathrm{Im}(\mathbf{T}_j)$ is the weighted sum of the estimators from those strata i with $\lambda_{ij} \neq 0$, with weights proportional to λ_{ij}/ξ_i: see [34, 44, 46]. The variance of this combined estimator is $1/\sum_i(\lambda_{ij}/\xi_i)$. Once again, we can use prior knowledge of the ξ_i to compare one proposed design with another.

For example, Leeming [37] compares the two designs in Figure 15. The plot structure is 3 *blocks*/(2 *rows* $*$ 2 *columns*) in each case. Interest is in comparing each of three new treatments with the control treatment, 0. It is assumed that the blocks stratum variance can be very large, but that information from the other three strata can be combined. Which is the better design depends on whether or not $\xi_{B\wedge R\wedge C}/\xi_{B\wedge R} + \xi_{B\wedge R\wedge C}/\xi_{B\wedge C}$ is greater than 1. Morgan and Uddin [41] make a similar comparison for this plot structure. Bailey [6] uses similar techniques for the plot structure *large blocks/small blocks/plots* to compare designs over a wide range of values of the ratio $\xi_{L\wedge S\wedge P}/\xi_{L\wedge S}$. It is hard to see how designs can be compared analytically if they are not generally balanced.

1	0	2	0	3	0	1	0	2	0	3	0
0	2	0	3	0	1	0	1	0	2	0	3

Design 1 Design 2

Fig. 15. Two designs for comparing three new treatments with a control (0)

Nelder's definition of general balance explicitly depended on both the given $\mathbf{P}_i$ and the given $\mathbf{T}_j$. Houtman and Speed [34] relaxed this: given the stratum projectors $\mathbf{P}_i$, they defined a design to be generally balanced if *there exists* any treatment decomposition $\sum_j \mathbf{T}_j$ for which (8) and (9) hold. Since this is equivalent to the commutativity of the information matrices from the different strata [55], all block designs (with equal block sizes) are generally balanced by this weaker definition. Pearce [52] objected that such a definition is useless.

General balance has been further explored by the Polish school [16, 17, 23, 24, 39], as well as [4, 34, 51]. Caliński [23] argued that general balance (given the treatment decomposition) is important in *interpreting* the results of an experiment rather than in designing it. If the plot structure is more complicated than a single system of blocks, there are designs which are not generally balanced even according to the weaker definition. Houtman and Speed [34] gave an example for a row-column design, Bailey [6] for nested blocks. Indeed, known generally balanced designs for these more complicated plot structures all seem to be partially balanced in the sense of [18]: this includes cyclic designs [35] and designs generated by permutation groups [12] as special cases. An exposition of partially balanced designs with many strata is in [8, Chapter 7].

Bailey [3] and Houtman and Speed [34] suggested that these ideas could be extended to plot structures that are association schemes [19] other than orthogonal block structures. For example, the experimental units may consist of 6 positions around the circumference of several petri dishes, or of all unordered pairs from among 10 people. Further details are in [7].

Thus Nelder's general balance enables statisticians to compare proposed designs in terms of their efficiencies for various treatment contrasts. Since these are usually chosen to be the contrasts of interest, general balance also helps the statistician to interpret the results of the experiment. However, he gave no guidance on the combinatorial process of *finding* a good design. Theoretical results are available in [54], which can give an optimal block design in some circumstances. In other circumstances, the statistician needs to consult tables such as [26], or to use a program such as [60] or the advanced design features of Genstat [50, Section 4.8], or to collaborate with combinatorialists. A current project at Queen Mary, funded by EPSRC, is to make good designs available on the web: see [10, 25].

9. Quantitative treatments

Although most of Nelder's work on designed experiments concerns qualitative treatment factors, he did propose the family of inverse polynomials in [45] as suitable models when the explanatory variables are quantitative. The response Y satisfies an inverse polynomial in $x_1, \ldots, x_n$ if there is a polynomial $\mathcal{P}(x_1, \ldots, x_n)$ such that

$$\frac{x_1 \cdots x_n}{Y} = \mathcal{P}(x_1, \ldots, x_n). \tag{10}$$

Unlike ordinary polynomial models, these are not invariant to the addition of a constant to any of the x_i, so the zero points of the x_i have to be known

in advance or estimated along with the other parameters. Whether or not these are estimated, model (10) does not give a linear model for Y. Nelder did not address the question of choosing levels of the x_i for an experiment designed to fit (10). However, this clearly belongs to the part of the protocol for specifying and choosing treatments. There has been much work on this problem in the last three decades: see Atkinson and Donev [1].

10. Algorithms for analysis

Nelder's vision for Genstat included a single algorithm for the analysis of designed experiments. There is really no need to treat each combination of plot structure and treatment structure as a special case.

The anova directive in Genstat is built on previous work by Yates [28] and by Wilkinson [61]. The central simple idea is that the data space is first decomposed according to the plot structure, then further decomposed according to the treatment structure. Brien and Bailey [22] show how to extend this to three structured sets.

Although this algorithm has wide applicability, it does, of course, depend on (i) stratifiable plot structure, in particular, blocks of equal size; (ii) orthogonal treatment structure; (iii) general balance. Patterson and Thompson [49] introduced REML to analyse data where these conditions are not met. When they are met, the anova algorithm followed by combination of information should give the same results as REML, except possibly if there are any negative estimates of components of variance. Thus one might argue that analysis of variance and conditions (i)–(iii) should be ignored. I do not take this view. Stratifiability is needed for a randomization justification of the analysis; orthogonal treatment structure and general balance are both needed to help the statistician to design the experiment and to interpret the results. The dummy analysis of variance that can be done in Genstat [50, p. 250] in advance of putting in any data is a wonderful tool to help designers of experiments: it produces the numbers λ_{ij} and all relevant degrees of freedom. I usually run such a dummy analysis before advising the experimenter to go ahead. So, although I do not insist that (i)–(iii) are absolutely essential, I do think that you should have a good reason to design an experiment that does not satisfy them.

References

1. A. C. Atkinson and A. N. Donev, (1992). *Optimum Experimental Designs*, Oxford University Press, Oxford.

2. R. A. Bailey, (1981). A unified approach to design of experiments. *Journal of the Royal Statistical Society, Series A* **144** 214–223.
3. R. A. Bailey, (1991). Strata for randomized experiments (with discussion). *Journal of the Royal Statistical Society, Series B* **53** 27–78.
4. R. A. Bailey, (1994). General balance: artificial theory or practical relevance? In *Proceedings of the International Conference on Linear Statistical Inference LINSTAT '93* (T. Caliński and R. Kala, eds.), Kluwer, Amsterdam. 171–184.
5. R. A. Bailey, (1996). Orthogonal partitions in designed experiments. *Designs, Codes and Cryptography* **8** 45–77.
6. R. A. Bailey, (1999). Choosing designs for nested blocks. *Listy Biometryczne* **36** 85–126.
7. R. A. Bailey, (2003). Designs on association schemes. In *Science and Statistics: A Festschrift for Terry Speed* (Darlene R. Goldstein, ed.), Institute of Mathematical Statistics Lecture Notes—Monograph Series, 40, IMS, Beachwood, Ohio.
8. R. A. Bailey, (2004). *Association Schemes: Designed experiments, Algebra and Combinatorics*, Cambridge, Cambridge University Press.
9. R. A. Bailey, (2004). Generalized wreath products of association schemes. *European Journal of Combinatorics*, in press.
10. R. A. Bailey, P. J. Cameron, P. Dobcsányi, J. P. Morgan and L. H. Soicher, (2004). Designs on the web. Submitted to *Discrete Mathematics.*
11. R. A. Bailey, Cheryl E. Praeger, C. A. Rowley and T. P. Speed, (1983). Generalized wreath products of permutation groups. *Proceedings of the London Mathematical Society* **47** 69–82.
12. R. A. Bailey and C. A. Rowley, (1990). General balance and treatment permutations. *Linear Algebra and its Applications* **127** 183–225.
13. T. A. Bancroft and C.-P. Han, (1983). A note on pooling variances. *Journal of the American Statistical Association* **78** 981–983.
14. A. Bardin and J.-M. Azaïs, (1990). Une hypothèse minimale pour une théorie des plans d'expériences randomisés. *Revue de Statistique Appliquée* **38** 5–20.
15. T. Beth, D. Jungnickel and H. Lenz, (1999). *Design Theory* (2nd edition) Volumes 1 and 2, Cambridge University Press, Cambridge.
16. B. Bogacka, (1995). On information matrices for fixed and random parameters in generally balanced experimental block designs. In *MODA 4—Advances in Model-Oriented Data Analysis* (C. P. Kitsos and W. G. Müller, eds.), Physica-Verlag, Heidelberg. 141–149.

17. B. Bogacka and S. Mejza, (1994). Optimality of generally balanced experimental block designs. In *Proceedings of the International Conference on Linear Statistical Inference LINSTAT '93* (T. Caliński and R. Kala, eds.), Kluwer, Amsterdam. 185–194.
18. R. C. Bose and K. R. Nair, (1939). Partially balanced incomplete block designs. *Sankhyā* **4** 337–372.
19. R. C. Bose and T. Shimamoto, (1952). Classification and analysis of partially balanced incomplete block designs with two associate classes. *Journal of the American Statistical Association* **47** 151–184.
20. C. J. Brien, (1983). Analysis of variance tables based on experimental structure. *Biometrics* **39** 51–59.
21. C. J. Brien and R. A. Bailey, (2004a). Multitiered experiments: I. Design and randomization. Submitted for publication.
22. C. J. Brien and R. A. Bailey, (2004b). Multitiered experiments: II. Structure and analysis. Submitted for publication.
23. T. Caliński, (1993). The basic contrasts of a block experimental design with special reference to the notion of general balance. *Listy Biometryczne* **30** 13–38.
24. T. Caliński and S. Kageyama, (2000). *Block designs: A randomization approach. Volume I: Analysis*, Lecture Notes in Statistics, 150, Springer-Verlag, New York.
25. P. J. Cameron, P. Dobcsányi, J. P. Morgan and L. H. Soicher, (2003). *The External Representation of Block Designs*, version 1.1, `http://designtheory.org/library/extrep/`
26. W. H. Clatworthy, (1973). *Tables of Two-Associate-Class Partially Balanced Designs.* Applied Mathematics Series, 63. National Bureau of Standards, Washington, D.C..
27. W. G. Cochran and G. M. Cox, (1957). *Experimental Designs*, Wiley, New York.
28. W. G. Cochran, D. J. Finney and M. J. R. Healy, (1970). Foreword to *Experimental Design* by F. Yates. Griffin, London.
29. C. J. Colburn and J. H. Dinitz (editors), (1996). *The CRC Handbook of Combinatorial Designs*, CRC Press, Boca Raton, Florida.
30. A. Dean and D. Voss, (1999). *Design and Analysis of Experiments*, Springer-Verlag, New York.
31. W. P. Gardiner and G. Gettinby, (1998). *Experimental Design Techniques in Statistical Practice*, Horwood, Chichester.
32. A. Grafen and R. Hails, (2002). *Modern Statistics for the Life Sciences*, Oxford University Press, Oxford.

33. R. M. Heiberger, (1989). *Computation for the Analysis of Designed Experiments*, Wiley, New York.
34. A. M. Houtman and T. P. Speed, (1983). Balance in designed experiments with orthogonal block structure. *Annals of Statistics* **11** 1069–1085.
35. J. A. John and E. R. Williams, (1995). *Cyclic and Computer-Generated Designs*. Chapman and Hall, London.
36. O. Kempthorne, G. Zyskind, S. Addelman, T. N. Throckmorton and R. F. White, (1961). *Analysis of Variance Procedures*, Aeronautical Research Laboratory, Ohio. Report No. 149.
37. J. A. Leeming, (1997). Comparison of two nested row-column designs containing a control. *Listy Biometryczne* **34** 45–62.
38. R. Mead, (1988). *The Design of Experiments; Statistical Principles for Practical Applications*, Cambridge University Press, Cambridge.
39. S. Mejza, (1992). On some aspects of general balance in designed experiments. *Statistica* **2** 263–278.
40. D. C. Montgomery, (1997). *Design and Analysis of Experiments*, fourth edition, Wiley, New York.
41. J. P. Morgan and N. Uddin, (1993). Optimality and construction of nested row and column designs. *Journal of Statistical Planning and Inference* **37** 81–93.
42. J. A. Nelder, (1954). The interpretation of negative components of variance. *Biometrika* **41** 544–548.
43. J. A. Nelder, (1965a). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proceedings of the Royal Society of London, Series A* **283** 147–162.
44. J. A. Nelder, (1965b). The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proceedings of the Royal Society of London, Series A* **283** 163–178.
45. J. A. Nelder, (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics* **22** 128–141.
46. J. A. Nelder, (1968). The combination of information in generally balanced designs. *Journal of the Royal Statistical Society, Series B* **30** 303–311.
47. J. A. Nelder, (1977). A reformulation of linear models. *Journal of the Royal Statistical Society, Series A* **140** 48–77.
48. J. A. Nelder, (1994). The statistics of linear models: back to basics. *Statistics and Computing* **4** 221–234.

49. H. D. Patterson and R. Thompson, (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58** 545–554.
50. R. W. Payne, D. B. Baird, A. R. Gilmour, S. A. Harding, P. W. Lane, D. A. Murray, D. M. Soutar, R. Thompson, A. D. Todd, G. Tunnicliffe Wilson, R. Webster and S. J. Welham, (2000). *The Guide to Genstat, Part 2 Statistics*, VSN International Ltd, Oxford.
51. R. W. Payne and R. D. Tobias, (1992). General balance, combination of information and the analysis of covariance. *Scandinavian Journal of Statistics* **19** 3–23.
52. S. C. Pearce, (1983). *The Agricultural Field Experiment*, Wiley, Chichester.
53. D. A. Preece, R. A. Bailey and H. D. Patterson, (1978). A randomization problem in forming designs with superimposed treatments. *Australian Journal of Statistics* **20** 111–125.
54. K. R. Shah and B. K. Sinha, (1989). *Theory of Optimal Designs*, Springer, New York.
55. T. P. Speed, (1983). General balance. In *Encyclopedia of Statistical Sciences*, Volume 3 (S. Kotz, N. L. Johnson and C. B. Read, eds.), Wiley, New York. 320–326.
56. T. P. Speed and R. A. Bailey, (1982). On a class of association schemes derived from lattices of equivalence relations. In *Algebraic Structures and Applications* (P. Schultz, C. E. Praeger and R. P. Sullivan, eds.), Marcel Dekker, New York. 55–74.
57. T. P. Speed and R. A. Bailey, (1987). Factorial dispersion models. *International Statistical Review* **55** 261–277.
58. A. P. Street and D. J. Street, (1987). *Combinatorics of Experimental Design*, Oxford University Press, Oxford.
59. T. Tjur, (1984). Analysis of variance models in orthogonal designs. *International Statistical Review* **52** 33–81.
60. D. Whitaker, E. R. Williams and J. A. John, (2002). CycDesigN, CSIRO, `http://www.ffp.csiro.au/tigr/software/cycdesign/`
61. G. N. Wilkinson, (1970). A general recursive procedure for analysis of variance. *Biometrika* **57** 19–46.
62. G. Wolde-Tsadik and A. A. Afifi, (1980). A comparison of the "sometimes pool", "sometimes switch" and "never pool" procedures in the two-way ANOVA random effects model. *Technometrics* **22** 367–373.
63. G. Zyskind, (1962). On structure, relation, sigma, and expectation of mean squares. *Sankyhā, Series A* **24** 115–148.

LIKELIHOOD-BASED MODELS BEYOND GLMS

Youngjo Lee
Department of Statistics
Seoul National University
Seoul 151-747,
Korea
e-mail: youngjo@plaza.snu.ac.kr

Nelder and Wedderburn (1972) introduced generalized linear models (GLMs) for the analysis of non-Gaussian data such as proportions, counts, etc. In GLMs the likelihood plays a key role for their definition and likelihood methods are thus natural tools for routine analysis. The likelihood framework has advantages such as generality of application, plug-in method, algorithmic experience, consistency, asymptotic efficiency, etc. In GLMs a computationally efficient iterative weighted least squares (IWLS) procedure suffices for estimation. Wedderburn (1974) was the first to apply the IWLS procedure of GLMs to models that do not allow an exact GLM likelihood. In this paper I explain how Professor Nelder and I have been further extending the likelihood-based model classes to allow both fixed and random effects not only in the mean and but also in the dispersion. This class will, among other things, enable models of types widely used in the analysis of data from quality improvement experiments, longitudinal studies, financial models, temporal and spatial models, etc., to be explored, and should give rise to new extended classes of models within a likelihood framework. A single algorithm suffices to fit all members of the class. The Fisher likelihood needs to be extended to maintain a likelihood framework for such extended classes.

1. Introduction

GLMs were introduced by Nelder and Wedderburn (1972) for the analysis of non-Gaussian data. GLMs assume the likelihood to come from a one-parameter exponential family, for which the IWLS procedure suffices to obtain the maximum likelihood (ML) estimators and their standard errors. The IWLS procedure yields statistically and computationally efficient inferences. For inferences about the mean parameters, Wedderburn (1974) showed that the IWLS procedure can still be used for models that do not

have a GLM likelihood. This quasi-likelihood method for inferences about the mean parameters is widely accepted. However, it does not allow inferences about the dispersion parameters.

When I first collaborated with Professor Nelder in 1989 he was working on extending likelihood-based models to dispersion analysis. This led to joint GLM s (JGLMs) for mean and dispersion (Nelder and Lee, 1991). We had been further extending the model classes to allow random effects in the linear predictor for the mean. This led to HGLMs (Lee and Nelder, 1996). HGLMs have become increasingly popular since the initial synthesis of GLMs, random-effect models, and structured-dispersion models was found to be extendable to include models for temporal and spatial correlations (Lee and Nelder, 2001a, 2001b). Heteroscedasticity of means between clusters (so-called between-cluster variation) can be modelled by introducing random effects in the mean. To form these models we need to extend the Fisher likelihood. The standard likelihood consists of two classes of objects, observable random variables y and unknown fixed parameters. However, HGLMs have an additional object, the unobservable random effects (or parameters), so that for three classes of objects we need an extended likelihood, so-called hierarchical likelihood (or h-likelihood). Unlike fixed parameters, for which transformation of scale does not matter in likelihood inferences, the scale of random parameters does matter, due to a Jacobian, so that the new likelihood needs careful definition (Lee and Nelder, 1996).

In current HGLMs both fixed and random effects are allowed for the mean but only fixed effects for the dispersion. We introduced double HGLMs (DHGLMs, Lee and Nelder, 2003b) which allow both fixed and random effects not only for the mean but also for the dispersion. This means that heteroscedasticity of dispersion between clusters can be similarly modelled by introducing random effects in the dispersion. It also leads to a systematic way of generating heavy-tailed distributions for various types of data such as counts, proportions, etc. We use the word "volatility" to mean heterogeneity arising from the presence of random effects in the dispersion. We have found that rapid changes among repeated measures can be explained by introducing random effects in the dispersion.

Ever since Fisher introduced the concept of likelihood in 1921, it has played an important part in the development of both the theory and the practice of statistics. The likelihood framework has advantages such as generality of application, plug-in method, algorithmic wiseness (Efron, 2003), consistency, efficiency, etc., which can be summarized as statistical and computational efficiency. The h-likelihood is a proper extension of Fisher

likelihood and produces similar statistically and computationally efficient procedures. The h-likelihood plays a key role in the synthesis of the inferential tools needed for the extended classes of models; it gives a new definition of conjugate families, leading to data augmentation, and leads to the decomposition of an extended model into component GLMs (Lee and Nelder, 1996, 2001a,b). Thus, these extended models can be fitted as an interconnected set of component GLMs. A single algorithm can be used throughout all these extended classes of models and requires neither prior distributions of parameters nor multi-dimensional quadratures. This formulation means that a great variety of models can be fitted by a single algorithm and compared using extensions of standard GLM procedures. Thus we can change the link function, allow various types of term in the linear predictor and use model-selection methods for adding or deleting terms. Furthermore various model assumptions can be checked by applying GLM model-checking procedures to the component GLMs.

2. GLMs

A GLM is defined as having the following two components.
(i) The responses y follow a one-parameter exponential family, satisfying

$$E(y) = \mu \quad \text{and} \quad var(y) = \phi V(\mu).$$

The one-parameter exponential family has an exact (log-)likelihood whose kernel is

$$\sum \{y\theta - b(\theta)\}/\phi, \tag{1}$$

where $\theta = \theta(\mu)$ is the GLM canonical parameter.
(ii) The linear predictor for the mean for μ takes the form

$$\eta = g(\mu) = X\beta.$$

Here, ϕ is the dispersion parameter, $V()$ is the GLM variance function and $g()$ is the GLM link function.

In addition to the normal distribution the GLM family includes binomial, Poisson, gamma and inverse-Gaussian distributions. In GLMs the constant-variance assumption of Gaussian linear models is weakened by allowing a non-constant variance function. Additivity can be achieved by choosing an appropriate link function. Within GLMs the variance function suffices to characterize a family of distributions for the response variables. The GLM above is mainly for modelling the mean μ, and the dispersion ϕ

occurs as the reciprocal of the prior weight. In summary, a GLM is specified by a response variable y, a variance function $V()$, a link function $g()$, a linear predictor $X\beta$ and a prior weight as shown in Table 1.

The GLM likelihood (1) leads to the IWLS procedure, in which the ML estimators for β are obtained by solving

$$X^t\Sigma^{-1}X\hat{\beta} = X^t\Sigma^{-1}z,$$

where $z = \eta + (y - \mu)(\partial\eta/\partial\mu)$ is the GLM-adjusted dependent variable and $\Sigma = \phi W^{-1}$ with $W = (\partial\mu/\partial\eta)^2 V(\mu)$; their variance estimators are obtained from

$$var(\hat{\beta}) = (X^t\Sigma^{-1}X)^{-1} = \phi(X^tWX)^{-1}.$$

This IWLS procedure for GLMs is within the likelihood framework and so is statistically and computationally efficient.

Table 1. Attributes for GLMs.

Components	β (fixed)
Response	y
Mean	μ
Variance	$\phi V(\mu)$
Link	$\eta = g(\mu)$
Linear Pred.	$X\beta$
Dev. Comp.	d
Prior Weight	$1/\phi$

$d_i = 2\int_{\hat{\mu}_i}^{y} (y - s)/V(s)\, ds$

In GLMs the dispersion parameter ϕ is assumed to be constant or known a priori. For example, the binomial and Poisson distributions assume $\phi = 1$. However, in practice, extra-binomial or extra-Poisson variation ($\phi > 1$) is often necessary, for which an exact GLM likelihood (1) may no longer be available. Also, for general variance functions $V(\mu)$, an exact GLM likelihood is usually unavailable. For inferences about the mean parameters β in such models, Wedderburn (1974) showed that the IWLS procedure, characterized by the variance function, can be still used.

Since 1989 I have collaborated with Professor Nelder in extending the likelihood-based models beyond GLMs. For this purpose, the Fisher likelihood needs to be extended. The extended likelihood should give rise to inferential procedures that keep within the likelihood framework, if possible. The minimum requirement we imposed for such extensions was to keep

the IWLS procedure, maintaining computational efficiency. For extended classes of likelihood-based models the GLM will provide building blocks.

3. JGLMs

The Taguchi analysis of data from quality-improvement experiments often aims to find control factors that reduce the dispersion ϕ, while keeping the mean μ on target. This leads immediately to the joint modelling of mean and dispersion parameters. To allow structured dispersion, Nelder and Lee (1991) introduced joint GLMs (JGLMs) as having the following three components:
(i) the responses y has a GLM family characterized by

$$E(y) = \mu \quad \text{and} \quad var(y) = \phi V(\mu).$$

(ii) the model for the mean μ has the linear predictor

$$\eta = g(\mu) = X\beta.$$

(iii) the model for the dispersion ϕ has the linear predictor

$$\xi = h(\phi) = G\gamma,$$

where $h()$ is the GLM link function for the dispersion.

In a GLM the variance, $var(y) = \phi V(\mu)$, splits into two components; the variance function $V(\mu)$, which describes the functional dependence between the mean and variance, and the dispersion parameter ϕ (Lee and Nelder, 2003). By modelling ϕ instead of the whole variance, we achieve Box's (1988) separation criterion, namely the elimination of unnecessary complications in the model due to functional dependence between the mean and variance (sometimes called the elimination of *cross talk* between location and dispersion effects). Parameters μ and ϕ satisfy the orthogonality condition of Cox and Reid (1987), and this is crucial for extending restricted maximum likelihood (REML) estimation to non-normal errors.

3.1. *EQL and quasi-models*

JGLMs need to be extended to arbitrary variance functions, in which there is no longer an exact GLM likelihood. There are two possibilities, either to use the extended quasi-likelihood (EQL) of Nelder and Pregibon (1987),

$$q = -\sum[d_i/\phi_i + \log\{2\pi\phi_i V(y_i)\}]/2,$$

where $d_i = 2\int_{\hat{\mu}_i}^{y_i} (y_i - s)/V(s)\,ds$ is the GLM deviance component, or the pseudo-likelihood (PL),

$$-\sum[(y_i - \mu_i)^2/\{\phi_i V(\mu_i)\} + \log\{2\pi\phi_i V(\mu_i)\}]/2,$$

based on a normal likelihood. For normal errors, they are the same, but in general they are different and there is considerable disagreement about their relative advantages. The EQL yields the quasi-likelihood estimator for β, while the PL does not, so that if the PL is to be used it should be only for inferences about dispersions.

Maximum PL (MPL) estimators for dispersions satisfy sample-size asymptotics and are asymptotically consistent as the sample size increases (Davidian and Carroll, 1988), while maximum EQL (MEQL) estimators satisfy parameter asymptotics and therefore are consistent in the limit for a given sample size as parameter values tend to certain limits. Many theoretical statisticians appear to think that an estimator ought to be asymptotically consistent in order to be statistically useful. However, in finite samples the mean square error (MSE) criterion seems to us a more relevant measure of the efficiency of estimators. Nelder and Lee (1992) showed that the sample-size inconsistency of the MEQL estimator would often be offset by the small MSE in finite samples: see also Piegorsh (1990). We have used the EQL because it has better finite-sampling properties; this is important in the analysis of data from quality-improvement experiments, where the data are often obtained from highly fractional factorial designs.

The EQL is the saddle-point approximation to the GLM family of distributions with a given variance function if one exists. The approximation is exact for normal and inverse Gaussian distributions, and differs for the gamma distribution and the discrete GLM distributions by the replacement respectively of the gamma function and factorials by their Stirling approximations. The EQL gives identical inferences to those from Efron's (1986) double exponential family (Lee and Nelder, 2000a).

Some may prefer to use the EQL for its finite sampling property, whereas others may prefer the PL for consistency. This means that the EQL loses consistency while the PL loses statistical efficiency in the likelihood framework; an exact GLM likelihood can maintain both if it exists. Thus, even though we can maintain computational efficiency by keeping the IWLS procedure we may no longer maintain statistical efficiency. This difficulty in dispersion inference is caused by the fact that neither the EQL nor the PL is an exact likelihood since they do not correspond to the probability of an observed response y. This does not cause a problem in mean inferences

(Wedderburn, 1974), showing that mean and dispersion inferences are quite different. In dispersion inferences the EQL has smaller MSE than the ML estimator in finite samples for wide ranges of parameters, which may not happen in mean inferences (Lee and Nelder, 1992, Piegorsh, 1990).

3.2. *Fitting algorithm for JGLMs*

Let l be a likelihood with nuisance effects δ. Consider a function $p_\delta(l)$, defined by

$$p_\delta(l) = (l - [\log\det\{D(l,\delta)/2\pi\}]/2)|_{\delta=\tilde{\delta}} \tag{2}$$

where $D(l,\delta) = -\partial^2 l/\partial\delta^2$ and $\tilde{\delta}$ solves $\partial l/\partial\delta = 0$. $p_\delta(l)$ is an adjusted profile likelihood eliminating nuisance parameters δ. Here the use of $p_\beta(q)$ is equivalent to the first-order approximation to the restricted (or residual) likelihood obtained by conditioning on $\tilde{\beta}$ under parameter orthogonality (Cox and Reid, 1987).

Lee and Nelder (1998) showed that the JGLM can be fitted conveniently via two interconnected GLMs as follows:
(i) The mean GLM: given $\hat{\phi}$, inferences about β (based upon q) can be made by the IWLS algorithm for the GLM, characterized by a response y, variance function $V()$, a link $g()$, a linear predictor $X\beta$, and a prior weight $1/\hat{\phi}$.
(ii) The dispersion GLM: given $\hat{\beta}$, inferences about γ (based upon $p_\beta(q)$) can be made by the IWLS algorithm for the GLM, characterized by a response $d_i^* = d_i/(1-q_i)$, a gamma error, a link $h()$, a linear predictor G, and a prior weight $(1-q_i)/2$, where q_i is the GLM leverage.
Instead of d_i, use of the Pearson deviance

$$(y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i)$$

in the above algorithm gives an extension of the MPL procedure of Davidian and Carroll (1988).

Suppose that the responses y have a normal distribution, i.e. $V(\mu) = 1$. If β were known each $d_i^* = (y_i - x_i\beta)^2$ would have a prior weight $1/2$. This is because

$$E(d_i^*) = \phi_i \text{ and } var(d_i^*) = 2\phi_i^2$$

and 2 is the dispersion for the $\phi_i\chi_1^2$ distribution, a special case of the gamma distribution. With β unknown, the responses d_i^* have $1-q_i$ degrees of freedom instead of 1, because they have to be estimated. For normal models our method provides the ML estimators for β and the REML estimators

for ϕ. For the dispersion link function $h()$ we usually take the logarithm. The standardized deviance components d^* become the responses for the dispersion GLM. Then the reciprocals of the fitted values from the dispersion GLM provide prior weights of the next iteration for the mean GLM; these connections are marked in Table 2. The resulting see-saw algorithm is very fast.

Table 2. GLM attributes for JGLMs.

Components	β (fixed)	γ (fixed)
Response	y	d^*
Mean	μ	ϕ
Variance	$\phi V(\mu)$	$2\phi^2$
Link	$\eta = g(\mu)$	$\xi = h(\phi)$
Linear Pred.	$X\beta + Zv$	$G\gamma + Fb$
Dev. Comp.	d	gamma(d^*, ϕ)
Prior Weight	$1/\phi$	$(1-q)/2$

$d_i = 2\int_{\widehat{\mu}_i}^{y} (y-s)/V(s)\, ds$,
$d^* = d/(1-q)$,
$d_m^* = d_m/(1-q_m)$,
$d_d^* = d_d/(1-q_d)$,
gamma(d^*, ϕ)= $2\{-\log(d^*/\phi) + (d^* - \phi)/\phi\}$ and
q is the GLM leverage (Lee and Nelder, 1998).

4. HGLMs

GLMs are extended to generalized linear mixed models (GLMMs) to allow Gaussian random effects in the linear predictor η for GLMs (Pierce and Sands, 1975). Model classes of GLMMs can be further extended to HGLMs by allowing non-Gaussian random effects and structured dispersion. Lee and Nelder (1996, 2001a) introduced HGLMs as having the following three components:

(i) Conditional on random effects u, the responses y follow a GLM family of distributions, characterized by

$$E(y|u) = \mu \quad \text{and} \quad var(y|u) = \phi V(\mu).$$

(ii) A random-effects model for the mean μ : given random effects u, the linear predictor for the mean model for μ takes the GLM form

$$\eta = g(\mu) = X\beta + Zv, \tag{3}$$

where $v = g_m(u)$, for some monotone function $g_m()$, are the random effects and β are the fixed effects; parameters λ for the random effects u have the GLM form

$$\xi_m = h_m(\lambda) = G_m \gamma_m, \tag{4}$$

where $h_m()$ is the GLM link and γ_m are fixed effects.

(iii) A GLM for the dispersion ϕ : the linear predictor for the dispersion model for ϕ takes the GLM form

$$\xi = h(\phi) = G\gamma. \tag{5}$$

4.1. *H-likelihood*

The Fisher likelihood is composed of two classes of objects, observable random variables y and unobservable fixed parameters. HGLMs have an additional class of objects, namely unobservable random parameters u. However, many people have considered the marginal likelihood, obtained by integrating out the random effects, as the proper likelihood of such classes of models. Lee and Nelder (1996) found that the Fisher likelihood needs to be extended beyond its use in (fixed) parametric inference to inference from models of a more general nature that may include fixed parameters, random parameters, and unobserved variables. Under their framework the marginal likelihood is the adjusted profile likelihood for the fixed effects after eliminating the nuisance random parameters by integration. As a basis for the likelihood for HGLMs, we may consider the joint density of the responses y and the random effects u, written by

$$L(u, y|\beta, \phi, \lambda) = f_{\beta,\phi}(y|v(u)) f_\lambda(u). \tag{6}$$

In (1) $f_{\beta,\phi}(y|v(u))$ is a density with a distribution from a one-parameter exponential family for GLMs; the second term $f_\lambda(u)$ is the density function of the random effects u with parameter λ. Even though $f_{\beta,\phi}(y|v(u)) \equiv f_{\beta,\phi}(y|u)$, mathematically we write the conditional density as $f_{\beta,\phi}(y|v(u))$, to mean that the function $v(u)$ defines the scale on which the random effects combine additively with the fixed effects β in the linear predictor. For inference from HGLMs, Lee and Nelder (1996) proposed to use a subclass of joint likelihoods of the form (6)

$$L(v, y|\beta, \phi, \lambda) = f_{\beta,\phi}(y|v) f_\lambda(v),$$

defined by a particular scale $v = v(u)$. Let

$$h = \log f_{\beta,\phi}(y|v) f_\lambda(v)$$

be the (log-)h-likelihood.

There have been several alleged counter examples purporting to show that h-likelihood provides qualitatively different (i.e. non-invariant) inferences for trivial re-expressions of the underlying model; however, all of them use a wrong scale in defining the h-likelihood. Thus, in the extended likelihood the scales of the random parameters are important in its definition, while the scales of fixed parameters are not. Another criticism has been about the statistical efficiency of the h-likelihood method; however, it gives statistically efficient estimates if properly implemented (Noh and Lee, in press).

To summarize, we maintain that the h-likelihood is a natural extension of Fisher likelihood to models with random parameters because it maintains the likelihood framework in respect of having both computational and statistical efficiency.

4.2. *Fitting algorithm for HGLMs*

This form of the h-likelihood gives a nice representation for conjugate HGLMs, in which the random effects u follow the conjugate distribution of the GLM distribution (1) for the $y|u$ component. Lee and Nelder (2001a) showed that the kernel of $\log f_\lambda(v)$ in h for the conjugate model with $v = \theta(u)$ takes the GLM form

$$\sum\{\psi v - b(v)\}/\lambda,$$

so that ψ (known constants for each conjugate distribution) can be treated as quasi-data and u (and hence $v = \theta(u)$) as quasi-fixed parameters, satisfying the purely formal relationships

$$E(\psi) = u \quad \text{and} \quad var(\psi) = \lambda V(u).$$

Lee and Nelder (2001a) showed further that various combinations of GLM distributions and links for $y|v$ and any conjugate GLM distribution and link for v can be used to construct HGLMs. For example GLMMs, assuming normally distributed random effects, are HGLMs with $V(u) = 1$. With the use of h-likelihood, the random-effect model (5) can be viewed as an *augmented GLM* with the response variables $(y^t, \psi_m^t)^t$, assuming

$$\mu = E(y), \;\; u = E(\psi), \;\; var(y) = \phi V(\mu), \;\; var(\psi_m) = \lambda V_m(u)$$

and the augmented linear predictor

$$\eta_{ma} = (\eta^t, \eta_m^t)^t = T\omega,$$

where $\eta = g(\mu) = X\beta + Zv$, $\eta_m = g_m(u) = v$, $\omega = (\beta^t, v^t)^t$ are fixed unknown parameters and quasi-parameters, and the augmented model matrix is $T = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix}$. Thus, for estimating $\omega = (\beta, v)$ the maximization of h leads to IWLS equations from the augmented GLMs (Lee and Nelder, 2001a).

For estimation of (β, ϕ, λ) Lee and Nelder (1996, 2001a) proposed to use an extended restricted likelihood $p_{v,\beta}(h)$ (2), which gives the following fitting algorithm: given $\hat{\beta}$,
(i) we estimate γ by solving the IWLS equations for a GLM, characterized by a response $d_i^* = d_i/(1-q_i)$, a gamma error, a link $h()$, a linear predictor G, a prior weight $(1-q_i)/2$ and q_i is a generalization of the GLM leverage.
(ii) we estimate γ_m by solving the IWLS equations for a GLM, characterized by a response $d_{mi}^* = d_{mi}/(1-q_{mi})$, a gamma error, a link $h_m()$, a linear predictor G_m, and a prior weight $(1-q_{mi})/2$, where q_{mi} is a generalization of the GLM leverage.

The fitting algorithm is summarized in Table 3.

Table 3. GLM attributes for HGLMs.

Components	β (fixed)	γ (fixed)
Response	y	d^*
Mean	μ	ϕ
Variance	$\phi V(\mu)$	$2\phi^2$
Link	$\eta = g(\mu)$	$\xi = h(\phi)$
Linear Pred.	$X\beta + Zv$	$G\gamma$
Dev. Comp.	d	gamma(d^*, ϕ)
Prior Weight	$1/\phi$	$(1-q)/2$
Components	u (random)	λ (fixed)
Response	ψ_m	d_m^*
Mean	u	λ
Variance	$\lambda V_m(u)$	$2\lambda^2$
Link	$\eta_m = g_m(u)$	$\xi_m = h_m(\lambda)$
Linear Pred.	v	$G_m\gamma_m$
Deviance	d_m	gamma(d_m^*, λ)
Prior Weight	$1/\lambda$	$(1-q_m)/2$

$$d_i = 2\int_{\hat{\mu}_i}^{y} (y-s)/V(s)\, ds,$$
$$d_{mi} = 2\int_{\hat{u}_i}^{\psi} (\psi - s)/V_m(s)\, ds,$$
$$d^* = d/(1-q),$$
$$d_m^* = d_m/(1-q_m),$$
gamma$(d^*, \phi) = 2\{-\log(d^*/\phi) + (d^* - \phi)/\phi\}$ and

(q, q_m) extends the idea of GLM leverage to HGLMs (Lee and Nelder, 2001a).

5. DHGLMs

The model class of HGLMs was further extended by Lee and Nelder (2003b) to the DHGLM model class, having the following three components:
(i) Conditional on random effects (u, a), the responses y assume a GLM family of distributions, satisfying

$$E(y|u,a) = \mu \quad \text{and} \quad var(y|u,a) = \phi V(\mu).$$

(ii) Given random effects u, the linear predictor for the mean model for μ takes the GLM form

$$\eta = g(\mu) = X\beta + Zv.$$

Parameters λ for the random effects u have the GLM form with

$$\xi_m = h_m(\lambda) = G_m\gamma_m.$$

(iii) Given random effects a, the linear predictor for the dispersion model for ϕ takes the GLM form

$$\xi = h(\phi) = G\gamma + Fb,$$

where $b = g_d(a)$, for some monotone function $g_d()$, are the random effects and γ are the fixed effects. Parameters α for the random effects a have the GLM form with

$$\xi_d = h_d(\alpha) = G_d\gamma_d,$$

where $h_d()$ is the GLM link and γ_d are fixed effects.
For the likelihood Lee and Nelder (in press b) use

$$L(v, b, y|\beta, \phi, \lambda, \alpha) = f_{\beta,\phi}(y|v, b) f_\lambda(v) f_\alpha(b).$$

Random effects in the mean describe heterogeneity, such as extra-binomial or extra-Poisson variation or correlation between repeated measures for the same subject. Random effects in the dispersion describe volatility among repeated measures. The use of DHGLMs allows heterogeneity and volatility to be distinguished. There are two kinds of heterogeneity, one to be explained by extra-variation and the other by sudden changes among repeated measures, which could be described by heavy-tailed distributions. We may call the former heterogeneity and the latter volatility. Lee and Nelder (in press b) showed that DHGLM can be again fitted by

interlinked GLMs as summarized in Table 4. Note that DHGLMs could be further extended by allowing random effects for the dispersion components α and λ.

Table 4. GLM attributes for DHGLMs.

Components	β (fixed)		γ (fixed)	
Response	y		d^*	
Mean	μ		ϕ	
Variance	$\phi V(\mu)$		$2\phi^2$	
Link	$\eta = g(\mu)$		$\xi = h(\phi)$	
Linear Pred.	$X\beta + Zv$		$G\gamma + Fb$	
Dev. Comp.	d		gamma(d^*, ϕ)	
Prior Weight	$1/\phi$		$(1-q)/2$	
Components	u (random)	λ (fixed)	a (random)	α (fixed)
Response	ψ_m	d_m^*	ψ_d	d_d^*
Mean	u	λ	a	α
Variance	$\lambda V_m(u)$	$2\lambda^2$	$\alpha V_d(a)$	$2\alpha^2$
Link	$\eta_m = g_m(u)$	$\xi_m = h_m(\lambda)$	$\eta_d = g_d(a)$	$\xi_d = h_d(\alpha)$
Linear Pred.	v	$G_m\gamma_m$	b	$G_d\gamma_d$
Deviance	d_m	gamma(d_m^*, λ)	d_d	gamma(d_d^*, α)
Prior Weight	$1/\lambda$	$(1-q_m)/2$	$1/\alpha$	$(1-q_d)/2$

$d_i = 2\int_{\widehat{\mu}_i}^{y} (y-s)/V(s)\, ds,$

$d_{mi} = 2\int_{\widehat{u}_i}^{\psi} (\psi - s)/V_m(s)\, ds,$

$d_{di} = 2\int_{a_i}^{\psi_d} (\psi_d - s)/V_d(s)\, ds,$

$d^* = d/(1-q_0),$

$d_m^* = d_m/(1-q_m),$

$d_d^* = d_d/(1-q_d),$

gamma$(d^*, \phi) = 2\{-\log(d^*/\phi) + (d^* - \phi)/\phi\}$ and

(q, q_m, q_d) extends the idea of leverage to HGLMs (Lee and Nelder, 2001a).

6. Random effects for temporal and spatial correlations

When the data consist of a series of repeated measurements made on each of a set of subjects, it is usual to find that the observations on any one subject, or, with spatial data, those in neighbouring areas, are correlated, and these temporal and spatial correlations must be reflected in the model used for the analysis. For this purpose Lee and Nelder (2001b) allowed random effects to satisfy $v = L(\rho)r$, where the elements r are independent, with joint distribution given by

$$r = L(\rho)^{-1} v \sim MVN(0, \Lambda)$$

with Λ =diag(λ_i); MVN stands for the multivariate normal distribution. Here, ρ denotes parameters for temporal and spatial correlations. Previously developed random effects in the statistical literature can be classified into three categories.

6.1. *Random effects described by fixed L matrices*

An obvious example is the random walk,

$$r_t = v_t - v_{t-1}.$$

Various temporal models such, as the state-space models of Harvey (1989) and Durbin and Koopman (2000), and various spatial models of Besag and Higdon(1999), fall into this category.

6.2. *Random effects described by a covariance matrices*

The multivariate normal models of Laird and Ware (1982) and Diggle et al. (1994) are examples. Autoregressive models, their extended form by Diggle et al. (1994, 1998), compound symmetric, and Toeplitz models belong in this category.

6.3. *Random effects described by a precision matrices*

The antedependent structures for temporal correlation of Gabriel (1962) and Markov-random-field models for spatial correlation (Cressie, 1993) belong to this category.

6.4. *Financial models for dispersion*

These temporal and spatial models can also be applied to the dispersion. In financial models the responses are often mean-corrected to allow zero means to be assumed: see for example, Kim et al. (1998). Suppose that we have a process

$$y_t = e_t,$$

where $e_t = \sigma_t z_t$ and $z_t \sim N(0,1)$ are independent. The simplest autoregressive conditional heteroscedasticity of order 1 model (ARCH(1), Engel, 1995) takes the form

$$\phi_t = \gamma_0 + \gamma y_{t-1}^2,$$

where $\phi_t = \sigma_t^2$. The natural analogue of exponential ARCH takes the log-link

$$\xi_t = h(\phi_t) = \log \phi_t = \gamma_0 + \gamma y_{t-1}^2.$$

This is a JGLM under our framework. The most popular stochastic volatility (SV) model, originating from Harvey, Ruiz and Shephard (1994), is

$$\xi_t = h(\phi_t) = \log \phi_t = \gamma_0 + b_t, \tag{7}$$

where $b_t = \rho b_{t-1} + r_t \sim AR(1)$; this is a DHGLM in our framework. If we take positive-valued responses y^2 these SV models become HGLMs with shape parameter 1/2, satisfying

$$E(y^2|b) = \phi \quad \text{and} \quad var(y^2|b) = 2\phi^2,$$

which is equivalent to assuming that $y^2|b \sim \phi\chi_1^2$. Here, ϕ plays the part of a mean parameter rather than a dispersion parameter.

6.5. *Fitting correlated random effects*

Suppose that we have serially correlated AR(1) random effects

$$\eta_t = g(\mu_t) = x_t\beta + v_t, \tag{8}$$

where $v_t \sim AR(1)$, i.e. $v_t - \rho v_{t-1} = r_t \sim N(0, \lambda)$. Lee and Nelder (2001b) showed that such models can be fitted by using a HGLM with

$$\eta = g(\mu) = X\beta + Z^* r, \tag{9}$$

where $Z^* = ZL(\rho)$, and in which Z^* is updated iteratively. This fitting method applies to correlated random effects for the dispersion as well.

7. Quasi extended GLM classes of models

The h-likelihood, a particular form of the joint likelihood, provides a natural extension of the Fisher likelihood to models with random effects. It provides a simple unified framework for fitting random-effect models and allows a computationally and statistically efficient IWLS procedure (Lee and Nelder, 2001a, 2003b). Models can be further extended by assuming the quasi-GLM for each component, characterized by a variance function. We can use EQLs as likelihoods for component GLMs. For these extended quasi-models we can maintain computational efficiency by having the IWLS procedure, but lose statistical efficiency because the EQLs are no longer exact likelihoods, since they do not correspond to the probabilities of observables y and unobservables v. For a detailed description of the algorithm see Lee and Nelder (2001a, in press b).

8. Frailty models via HGLMs

Frailty models have been widely used for the analysis of correlated survival data in the form of recurrent or multiple-event times. Given the unobserved frailties v_i, the hazard function for the j-th recurrent event time of the i-th patient has the form

$$\lambda_{ij}(t|v_i) = \lambda_0(t)\exp(\eta_{ij}),$$

where $\lambda_0(\cdot)$ stands for the non-parametric baseline hazard function,

$$\eta_{ij} = x_{ij}^T\beta + v_i$$

is the linear predictor for hazards, β the regression parameters, and x_{ij} (not including a constant) the covariates. This model is an extension of Cox's (1972) proportional hazards model to allow random effects. Ma et al. (2003) showed that the frailty model above can be fitted by using the following Poisson HGLM. Let $y_{(k)}$ be the k-th smallest distinct event time among the observed event times or censored times $t_{ij}$and let $y_{ij,k}$ be 1 if the event time t_{ij} occurs at $y_{(k)}$ and 0 otherwise. Given frailties v_i, assume $y_{ij,k}$ to be conditionally independent with

$$y_{ij,k}|v_i \sim \text{Poisson}(\mu_{ij,k}) \text{ for } (i,j) \in R(y_{(k)}),$$

where

$$\log\mu_{ij,k} = \alpha_k + x_{ij}^T\beta + v_i = x_{ij,k}^T\gamma + v_i,$$

$R(y_{(k)}) = \{(i,j) : t_{ij} \geq y_{(k)}\}$ is the risk set at time $y_{(k)}$, $x_{ij,k} = (e_k^T, x_{ij}^T)^T$, e_k is a vector of components 0 and 1 such that $e_k^T\alpha = \alpha_k$, and $\gamma = (\alpha^T, \beta^T)^T$.

Ha et al. (2001) have extended the h-likelihood to frailty models and showed that the resulting h-likelihood procedure gives a statistically and numerically efficient fitting algorithm. Ha and Lee (in press) showed that it is the h-likelihoods that coincide for both frailty models and the Poisson HGLM above. Thus, the likelihood inferences from them are identical, so that h-likelihood inferences for Poisson HGLMs can be used for the analysis of frailty models. We do not assume that $y_{ij,k}|v_i$ follows the Poisson distribution; this is not possible because the Poisson distribution assigns probabilities to values greater than one. An alternative to frailty models for the analysis of survival data is mixed linear models allowing censoring, for which marginal likelihood inference is computationally very hard. Ha et al. (2002) showed that the h-likelihood again provides a computationally and statistically efficient procedure.

9. Application of DHGLMs

The use of extended likelihood-based models provides new solutions to various problems. These include:
(1) joint modelling of mean and dispersion (Lee and Nelder, 2001a);
(2) the analysis of temporally and spatially correlated data (Lee and Nelder, 2001b);
(3) the provision of model checking to see if the postulated pattern of random effects is supported by the data (Lee and Nelder, 2001a);
(4) meta analysis (Lee and Nelder, 2002);
(5) analysis of survival data (Ha, Lee and Song, 2001; Ha, Lee and Song, 2002);
(6) implicit implementation of an EM-type algorithm to yield good estimators for censored linear mixed models (Ha and Lee, in press);
(7) the prediction of future observations (Pawitan, 2001, Chapter 16);
(8) a simple alternative to kernel smoothing (Pawitan, 2001);
(9) the use of information from concordant pairs without making parameter assumptions regarding the random-effect distribution (Lee, 2001);
(10) new robust sandwich variance estimates for fixed effect estimators, which cannot be obtained from marginal likelihood (Lee, 2002);
(11) a likelihood-based alternative to the *ad-hoc* approach of generalized estimating equations (Zeger et al., 1988);
(12) analysis with missing data (Lee, Noh and Ryu, 2003);
(13) the provision of inferential tools for a much wider class of models, the so-called double HGLMs. Among other things, they extend stochastic volatility models in financial modelling, and provide natural extensions for heavy-tailed distributions for counts and proportions, etc. (Lee and Nelder, 2003b).

10. Concluding remarks

The h-likelihood is a natural extension of the Fisher likelihood, meets the need for general inferences from extended models, and provides a unified framework of inference with an implementable algorithm. With the h-likelihood apparatus extensive classes of new models can be brought together within a single framework. With extended likelihood-based models we can model heterogeneities in both the means and variances between clusters and analyze them using single algorithm. The decomposition of the complete model into several components provides insights into the development, extension, analysis and checking of the models. Statistical inferences

for a complicated phenomenon can then be made from an integrated model built upon GLMs. In the extended likelihood framework the marginal likelihood appears as a profile likelihood similar to the restricted likelihood, which has been recommended to reduce the bias.

It is perhaps unfortunate that Bayesians, from Lindley and Smith (1972) onwards, seem to have made a take-over bid for all hierarchical models, implying that one has to be Bayesian to deal with them. The availability of Markov-chain Monte Carlo, making all problems seem more easily solvable via Bayesian computations, has appeared to justify this. However, by using h-likelihood, we can deal with such models directly in a likelihood framework because there is an explicit analytic form for that kind of likelihood. Furthermore inferences for unobservables are possible without resorting to an empirical Bayesian framework. H-likelihood gives a powerful and practical tool for statistical inference; being a natural extension of Fisher likelihood to models with unobservables, it will become, we believe, widely used for inference from hierarchical models. This is what Professor Nelder has been establishing with me for the last 15 years.

Acknowledgments

This research was supported by a grant from Statistical Research Center for Complex Systems of Korean Science and Engineering Foundation.

References

Besag, J. E. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *J. Roy. Statist. Soc. Ser.* B, **61**, 691-746.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformation (with discussion). *Technometrics*, **30**, 1-40.

Cox, D. R. (1972). Regression models and life tables. *J. R. Statist. Soc.* B, **74**, 187-220.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, **32**, 1-18.

Cressie, N. (1993). *Statistics for spatial data.* Oxford: Clarendon Press.

Davidian, M. and Carroll, R. J. (1988). A note on extended quasi-likelihood. *J. R. Statist. Soc.* B, **50**, 74-82

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of longitudinal data.* Oxford: Oxford University Press.

Diggle, P. J., Tawn, J. A. and Moyeed R. A. (1998). Model-based geostatistics. *Appl. Statist.*, **47**, 299-350.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives, (with discussion). *J. R. Statist. Soc.* B, **62**, 3-56.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Ass.*, **81**, 709-21.

Efron, B. (2003). A conversation with good friends. *Statistical Science*, **18**, 268-281.

Engel, R. E. (1995). *ARCH.* Oxford: Oxford University Press.

Gabriel, K. R. (1962). Ante-dependence analysis of an ordered set of variables. *Annals of Mathematical Statistics*, **33**. 201-212.

Ha, I. D. and Lee, Y. (in press). Multilevel mixed linear models for survival data. to appear in *Lifetime Data Anal.*

Ha, I. D., Lee, Y. and Song, J. K. (2001). Hierarchical likelihood approach for frailty models, *Biometrika*, **88**, 233-243 (2001).

Ha, I. D., Lee, Y. and Song, J. K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Anal.*, **8**, 163-176.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter.* Cambridge: Cambridge University Press.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994). Multivariate stochastic variance models. *Review of Economic Studies*, **61**, 247-264.

Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **98**, 361-393.

Laird, N. M. and Ware, J. H. (1982). Random-effects Models for Longitudinal Data. *Biometrics*, **38**, 963-974.

Lee, Y. (2001). Can we recover information from concordant pairs in binary matched paired? *J. Appl. Statist.*, **28**, 239-246.

Lee, Y. (2002). Robust variance estimators for fixed-effect estimates with hierarchical-likelihood. *Statist. and Comput.*, **12**, 201-207.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.* B, **58**, 619-678.

Lee, Y. and Nelder, J. A. (1998). Generalized linear models for the analysis of quality-improvement experiments. *Can. J. Statist.*, **26**, 95-105.

Lee, Y. and Nelder, J. A. (2000a). The relationship between double exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ratio data. *Appl. Statist.*, **49**, 413-419.

Lee, Y. and Nelder, J. A. (2001a). Hierarchical generalised linear models: a

synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika*, **88**, 987-1006.

Lee, Y. and Nelder, J. A. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modelling*, **1**, 3-16.

Lee, Y. and Nelder, J. A. (2002). Analysis of the ulcer data using hierarchical generalized linear models. *Statistics in Medicine*, **21**, 191-202.

Lee, Y. and Nelder, J. A. (2003). Robust design via generalized linear models. *J. Qual. Tech.*, **35**, 2-12.

Lee, Y. and Nelder, J. A. (2003a). Likelihood for random effect models. manuscript prepared for publication.

Lee, Y. and Nelder, J. A. (2003b). Double hierarchical generalized linear models. manuscript prepared for publication.

Lee, Y., Noh, M. and Ruy, K. (in press). HGLM Modelling of Dropout Process using Frailty Model. manuscript prepared for publication.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1-44.

Ma R., Krewski D. and Burnett R. T. (2003). Random effects Cox models: a Poisson modelling approach, *Biometrika*, **90**, 157-169.

Nelder, J. A. and Lee, Y. (1991). Generalized linear models for the analysis of Taguchi-type experiments. *Applied Stochastic Models and Data Analysis*, **7**, 107-120.

Nelder, J. A. and Lee, Y. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons. *J. R. Statist. Soc.* B, **54**, 273-284.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221-231.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370-384.

Noh, M. and Lee, Y. (2003). Laplace approximation based methods for binary data in generalised linear mixed models. manuscript prepared for publication.

Pawitan Y. (2001). *In all likelihood: statistical modelling and inference using likelihood.* Oxford: Oxford University Press.

Piegorsch, W. G. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* **46**, 863-867.

Pierce, D. A. and Sands, B. R. (1975). Extra-Bernoulli variation in regression of binary data. Technical Report 46, Statistics Department, Oregon State, Cornwallis, OR.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

A STATISTICAL EXAMINATION OF THE HASTINGS RARITIES

J. A. Nelder

THIS PAPER[a] IS an attempt to establish the consistency or otherwise of the great flood of rarities from east Sussex and west Kent during the first two decades of this century, from the internal evidence presented by the statistical aspects of the records themselves. The consequences of various hypotheses which assume the validity of all the records will be tested against the numerical evidence.

METHODS

For the analysis, all records of rarities for the counties of Kent and Sussex for the years 1895-1954 inclusive have been extracted from Walpole-Bond (1938), Harrison (1953), the *South-Eastern Bird Reports* for 1936-47, the *Kent Bird Reports* for 1952-54 and the *Sussex Bird Reports* for 1948-54. Between them these publications cover the period and region required.

The region has been split up into three parts: area X is contained inside a circle with centre Hastings Pier and radius 20 miles, except that the whole of Romney Marsh (apart from Hythe) is included; area YS is the rest of Sussex not in area X and area YK is the rest of Kent similarly. The inclusion of the whole of Romney Marsh in X is necessary because a number of records in the sources do not specify exact places in the Marsh, and the allocation of these records to the correct area would be problematical if the 20-mile radius definition were strictly adhered to.

The 60-year span has been divided into two eras, 1895-1924 inclusive, called A, and 1925-1954 inclusive, called B. The records dealt with here thus fall into one of six categories, XA, XB, YSA, YSB, YKA or YKB. These six combinations of areas and eras will be termed area-eras for short.

[a]Republished with permission from *British Birds*, 1962, **55**, 283–298

For the purposes of this paper a rarity is defined as a species whose recorded occurrences have been completely enumerated in the books and reports mentioned, and which has not occurred on the average more than once per year in any of the six area-eras.

The reduction to tabular form of such a heterogeneous collection of data as these reports of occurrences of bird rarities over the 60 years is not easy. The records exhibit all gradations from virtual certainty about the identity of the bird to considerable vagueness, and some rules for their acceptance or rejection are essential. Within the limits necessarily imposed by the "strictness" or "leniency" of the sources, I have tended towards strictness and the reduction of acceptances to a minimum. No record has been accepted for the purpose of the following analysis unless all the following conditions are satisfied:

(i) The name of the observer in the field or identifier (if bird dead), must be given in the source.
(ii) The date must be given to within a year.
(iii) No doubt must be expressed by the author (or editor) about validity of the record; if a record occurs in more than one source no doubt must be expressed by any of the authors (or editors). All square-bracketed records have thus been rejected. (In a few cases it was not quite obvious whether the author was expressing doubt or not, but, in accordance with the general principle, such records were rejected.)
(iv) The bird must have been *seen*; records based on birds heard but not seen have been rejected.
(v) The bird must have been seen or taken from the land; no records of birds observed from ships (including lightships) have been admitted.

Since occurrences of rarities other than singly are important in these data, some formal definition of an occurrence is required. In this paper, two birds are said to have occurred together (and so to constitute one occurrence) if they were seen or taken within five mile and within seven days of each other. A set of more than two records of individual birds forms a single occurrence, if, when arranged in chronological order, every adjacent pair satisfies the condition for a single occurrence. A set of records also forms a single occurrence if the birds were specifically recorded as having come from a flock, even though successive records were not all within seven days of each other. Occurrences relating to one bird, two birds and more than two birds will be called *singular*, *dual* and *plural* respectively, while *multiple* will be used to cover dual and plural combined. Sometimes reports

are vague about numbers and in such cases the minimising rule is brought into play: thus "several" is taken to mean "three" (i.e., the smallest integer greater than two), "a small flock" is taken as "four", and if the author or editor expresses a belief that several records refer to the same individual, this is taken to be so and only a single occurrence is allowed.

TABLE 1 — HYPOTHETICAL EXAMPLES OF TWO-WAY TABLES (SEE TEXT BELOW)

	(a) Exact Proportionality		(b) With random errors added (not significantly different from proportionality)		(c) Significant departure from proportionality	
	Winter	Summer	Winter	Summer	Winter	Summer
Class II	100	300	98	310	100	200
Class I	50	150	54	145	50	150
	$\chi_1^2 = 0$		$\chi_1^2 = 0.21$ $(0.7 > P > 0.5)$		$\chi_1^2 = 7.94$ $(P < 0.01)$	

For the purpose of analysis an index of the rarity of a species is required. In this paper the number given for England in *The Handbook* is used as an index and, as before, when any doubt is expressed there about numbers the smaller one is taken. Again records from sea-based observers have been rejected. Certain objections to this index can be raised; in particular, it does not cover the whole period under investigation, and hence weights records in favour of the earlier period. Its main advantage is that it was compiled by one man, independently of the present investigation, and is, therefore, consistent and objective. In the main, species and subspecies will be divided into three classes: class I rarities, which have less than 20 accepted English examples in *The Handbook*; class II, with 20-99 examples (inclusive); and class III, consisting of those whose occurrences are not enumerated in *The Handbook*. It is possible that some class III species have actually less than 100 records over the period covered by *The Handbook*, and so overlap class II, because *The Handbook* does not appear to be entirely consistent in this matter. However, this overlap, if it exists, is small and unimportant.

The only statistical, in the sense of probabilistic, techniques used in the following analysis are the χ^2 (chi-squared) goodness-of-fit test and the Poisson distribution. The χ^2 test is applied here mostly to frequencies arranged

in two-way tables, for example the frequencies of the occurrence of rarities of different classes in different seasons. The simplest situation in such a table occurs when the relative frequencies in one set of categories (e.g. rarity classes) are the same for all categories in the other set (seasons). The first part (a) of Table I shows such an ideal situation. Each rarity group has three times as many summer records as winter, and each season has twice as many class II records as class I. The hypothesis that the relative frequencies are of this simple type is called in statistical parlance the *null hypothesis*. Of course in any particular sample the frequencies would almost never be exactly proportional, even if the null hypothesis were true, because of random errors in them. These random errors might give rise to something like Table 1(b). Here the hypothesis of proportionality is not disproved. However, these random errors can only distort the picture to a certain degree and χ^2 can be regarded as a measure of whether this distortion has reasonably been exceeded in any particular case.

Under certain conditions the relative frequencies of different values of χ^2 turning up can, if the null hypothesis true, be calculated. The average value of χ^2 equals a quantity called the number of degrees of freedom, which itself depends only on the form of the table, not on the numbers in it. A value of χ^2 much in excess of the average value means that a very unlikely event has taken place, if the null hypothesis is true, and hence that it should be discarded for some other hypothesis more in accordance with the facts. Thus in Table I(c), while class I has three times as many summer as winter records, class II has only twice as many. This gives a large χ^2 and tends to discredit the null hypothesis. Similarly, for a χ^2 with two degrees of freedom (written χ^2_2), a value of six would be exceeded in only 5% of cases if the null hypothesis were true. Thus values of χ^2_2 greater than six are said to be significant at the 5% level, or significant $P = 0.05$ and provide considerable evidence that the null hypothesis is false. It should be pointed out that the null hypothesis can fail to be true in two rather different ways. In one situation, the true frequencies may not be proportional, so that occurrences among class III rarities might have a relatively greater frequency in winter than occurrences in the other two classes; this is a systematic deviation from the null hypothesis. The other situation occurs when the random deviations are unusually large, but the true frequencies are still proportional; this may occur if the thing being measured comes from a heterogeneous population, made up of several sub-populations with unequal chances of being represented. Thus our class III rarities comprise a number of species of which some are relatively much commoner than others,

and this may produce a random deviation larger than average. In practice it is often possible to distinguish the two kinds of deviations, since one has a pattern while the other has not. In the analysis which follows we shall meet examples where the null hypothesis is well supported, and where there are deviations both random and systematic from it.

The Poisson distribution is a theoretical probability distribution, often useful in the description of the frequencies of rare events. It is completely specified by its mean value. For a general description of χ^2 and this distribution a standard statistical textbook should be consulted (e.g. Snedecor 1946).

The Results

The results to be discussed embrace 1,015 occurrences, involving 1,360 birds of 168 species and subspecies. Subspeciation is as given in *The Handbook*. This is generally satisfactory for our purposes, but the Yellow Wagtail (*Motacilla flava*) complex has presented difficulties. In particular the "Sykes" type (resembling *beema*) must be a class I rarity by our definition, though modern records make it much commoner and this bird actually makes up nearly 10 of the class I records for the rest of Kent in the years 1925-54 (YKB). However, since *The Handbook* is being used for the rarity index, no exceptions are made to its classification of subspecies and records.

Table 2 — Total occurrences in different rarity classes
An explanation of the area-eras will be found on page 215, and of the rarity classes on page 217

Area-era	Class I	Class II	Class III	Total
XA	243	108	165	516
XB	54	51	103	208
YSA	15	16	45	76
YSB	19	13	32	64
YKA	11	11	22	44
YKB	26	28	53	107
Total	368	227	420	1,015

The complete list of records used (which is not given in full here, but is being deposited at the Edward Grey Institute, Oxford) has been split up in various ways for the investigation, and the following aspects will be presented and discussed: the relative frequencies of singular, dual, plural and total occurrences in the three rarity classes for the six area-eras; also

the distribution of occurrences in the various seasons of the year and in different years throughout the periods concerned.

The distribution of the total number of occurrences

We consider first the total number of occurrences in each rarity class for each area-era, the relevant figures being shown in Table 2. The most obvious feature of these figures is that the distribution of the occurrences among the rarity classes in the Hastings Area for the period 1895-1924 (XA) is quite different from the distribution in the remaining area-eras. A χ^2 test carried out on these remaining area-eras gives $\chi^2_8 = 3.55$, showing no significant difference in the proportions of the three rarity classes. Considering the heterogeneous nature of the data, the agreement is remarkably good. Table 3, however, compares Hastings (XA) with the total of the remaining area-eras and it will be seen immediately that XA has nearly twice the proportion of class I rarities that the remainder has, balanced by a deficiency of class III rarities.

TABLE 3 — TOTAL OCCURRENCES FOR HASTINGS 1895-194 COMPARED WITH ALL OTHER AREA-ERAS COMBINED

Area-era	Class I	Class II	Class III	Total
Hastings (XA)	243	108	165	516
Remainder	125	119	255	499

In contrast to the homogeneity of the remainder of the area-eras, these discrepancies are highly significant, producing the enormous χ^2_2 value of 57.40.

So far we have considered only the distribution of the numbers in the different rarity classes, without looking at the total number of occurrences in the different area-eras. It is clear from inspection of the figures for the rest of Sussex and Kent (YS and YK) that the trend in the two regions over the period of time concerned is quite different. While the total number of records for YS has actually declined slightly for era B compared with era A, that for YK has markedly increased. (It should not be assumed from the YS figures that the amount of bird-watching has gone down in that area over the period considered, because if a species has too many records in the second period for it to be enumerated completely in the sources, or if there are more than 30 records in that period, its contribution is automatically eliminated from these figures by the rules previously laid down. This tends to minimize the number of records for the second era, but does not bias the other comparisons we are making.) In the absence of agreement between the trends for these two areas we cannot say, with any

conviction, what the figures for XA ought to be. Incidentally, even if YS and YK had agreed in their trends over the two eras, no significance test comparing them with X would have been valid, since we have deliberately chosen XA for investigation on account of its unusually large total of rarities (the fact of this choice does not invalidate significance tests on the other aspects we are considering). It is fair to note, however, that the trend for the Hastings Area does not agree with either of the other areas. It is nearer to YS, but to be comparable the XA figure should be about 247 instead of the 516 actually recorded.

The distribution of numbers at each occurrence

Considering class I rarities first, and dividing occurrences into singular and multiple (there being insufficient records in most area-eras to divide the multiple occurrences into dual and plural), we get Table 4. The proportion of multiple records in XA (25.1%) is much higher than in the other area-eras (average 12.0%). A χ^2 test excluding XA gives $\chi^2_4 = 2.00$, indicating homogeneity among the "remainder" group, while comparison of XA with the remainder gives $\chi^2_1 = 8.65$ ($P < 0.01$), showing that XA disagrees with the remainder. This is even more marked if we divide the multiples into duals and plurals as shown in Table 5. Here $\chi^2_2 = 12.76$, a more extreme value than the previous $\chi^2_1 = 8.65$. The remainder group has only one plural occurrence for class I rarities– the Paddock Wood Snow Finches of 1906 (*Handbook*, I: 155).

TABLE 4 — DISTRIBUTION OF SINGULAR AND MULTIPLE OCCURRENCES FOR CLASS I RARITIES

An explanation of the area-eras will be found on page 215 of singular and multiple occurrences on page 216 and of the rarity classes on page 217 (see Appendix on pages 233-234)

Area-era	Singular	Multiple	Total
XA	182	61	243
XB	46	8	54
YSA	14	1	15
YSB	18	1	19
YKA	10	1	11
YKB	22	4	26
Total	292	76	368

The situation with class II rarities is very much the same as with class I; the proportion of multiple records for XA is 26.9%, while for the remainder

group it is 12.6%, with YSA the highest at 18.8%. Again the remainder group gives a low $\chi^2_4 = 1.20$, indicating homogeneity, while comparison of XA with the remainder gives a significant $\chi^2_2 = 7.36$ ($P < 0.05$). With class III rarities the situation is less clear cut, for the XA proportion of multiple records, here 22.8%, is slightly less than that for YSA, which has 24.4%. This difference of YSA from the rest of the remainder group is almost entirely due to records for one species, the Glossy Ibis, which contributes four out of the multiple occurrences for YSA. The same species contributes three out of the 38 multiple occurrences for XA. The result of this is that XA and YSA do not differ significantly, though XA differs from the remaining four area-eras ($\chi^2_2 = 7.29$). One other aspect of the data deserves mention. In the remainder group, the percentages of plural occurrences for class III, II, and I rarities are 5.1, 2.5, and 0.8 respectively; that is, they fall steadily, being greatest in the least rare class. This is what one might expect *a priori*. However, in the XA group, the percentages (in the same order) are 9.1, 5.6 and 9.9, and show no such trend.

TABLE 5 — DISTRIBUTION OF SINGULAR, DUAL AND PLURAL OCCURRENCES FOR CLASS I RARITIES

Area-era	Singular	Dual	Plural	Total
Hastings (XA)	182	37	24	243
Remainder	110	14	1	125
Total	292	51	25	368

The distribution of occurrences by season

For nearly 97% of the occurrences, the month of occurrence is given in the source. Where it is not given, the occurrence is excluded from the analysis in this section. Where a single bird stayed for several months, the first month is taken. If a flock was present and members were shot from it or seen in more than one month, then the month of the first record is again used. The numbers for most of the months in most of the area-eras are too small to allow any accurate comparisons, so they have been grouped in four seasons of winter (December-February), spring (March-May), summer (June-August) and autumn (September-November).

TABLE 6 — DISTRIBUTION OF OCCURRENCES BY SEASONS
An explanation of the area-eras will be found on page 215, and of the rarity classes on page 217.

Area-Era and rarity class		Number of Occurrences: Spring (Mar/May)	Summer (Jun/Aug)	Autumn (Sep/Nov)	Winter (Dec/Feb)	Total
XA	I	101	42	63	36	242
	II	25	38	29	13	105
	III	42	35	56	27	160
	Total	168	115	148	76	507
XB	I	16	15	16	4	51
	II	12	11	27	1	51
	III	35	16	41	10	102
	Total	63	42	84	15	204
YSA	I	5	0	7	3	15
	II	6	1	7	2	16
	III	7	3	18	7	35
	Total	18	4	32	12	66
YSB	I	6	4	9	0	19
	II	3	2	6	2	13
	III	9	11	8	3	31
	Total	18	17	23	5	63
YKA	I	2	3	3	3	11
	II	4	2	3	0	9
	III	1	3	6	7	17
	Total	7	8	12	10	37
YKB	I	11	9	3	3	26
	II	12	8	7	1	28
	III	20	6	14	10	50
	Total	43	23	24	14	104

The frequency of occurrences for all area-eras, rarity classes and seasons is given in Table 6. The distribution by seasons is much more variable in the remainder group than the previous distributions considered. A remarkable feature is the growth of spring records in the rest of Kent from 18.9% in era A to 41.3% in era B, while the rest of Sussex shows no such change though summer records have increased there. Both these areas agree, however, in showing a decline in the proportion of autumn and winter records as we pass from era A to era B. YS shows a fall of 22.3% from 66.7% to 44.4%, and YK a fall of 23.0% from 59.5% to 36.5%. By contrast, the autumn and winter records for the Hastings Area rise slightly from 44.2% to 48.5%. It is also noticeable that, while the seasonal distributions of total occurrences for XB and YSB are very similar (giving $\chi^2_3 = 1.26$), those for XA and YSA are

quite unlike each other ($\chi^2_3 = 15.70$). The XA records have another property not shared by any of the other area-eras in that they have a considerably greater proportion of spring records for class I rarities than for classes II and III.

The distribution of records by years

The relevant data on distribution by years are given in Table 7 for all area-eras and rarity classes. Considering first the earlier era A, we find that for both YS and YK the distribution of the number of class I rarities is very close to a Poisson; the actual frequencies and the theoretical ones of the Poisson distributions with the same means are shown in Table 8.

These good fits to the theoretical distributions suggest strongly that there were no large differences in the numbers of class I rarities reaching these areas each year during this period, or in the intensity of observations made on them; for, if there had been any such large differences, the actual frequency distributions would have had longer "tails" and the Poisson model would no longer have fitted well. The distribution of class I rarities in XA is obviously quite unlike the last two considered. In the first place it shows strong time trends, there being a sharp increase in the early 1900s followed by an equally sharp decrease after 1916. In such circumstances it is unreasonable to expect a theoretical distribution to fit well and, in fact, the Poisson distribution is a very bad fit here. It is somewhat surprising that the peak years in X, namely 1905, 1914 and 1915, do not correspond with any peaks in the other two regions.

Differences between regions are much less remarkable for classes II and III. In YSA and YKA the Poisson fits less well, due doubtless to increasing heterogeneity in the population sampled, while the distributions for XA are less extreme than that for XA class I. In era B, we see a number of trends in time which make any agreement with a simple theoretical model out of the question. The major factor is the post-war increase in bird-watching, with its resulting effect on the number of rarities seen from 1946 onwards. Conversely, the war itself has depressed the number of records in X and YS below that of the pre-war years (although YK does not seem to show this), while during the period 1925-1939 there seems to be a trend towards an increasing number of records. In spite of these effects of the number of observers (for that is what they most likely are), the figures show two points of interest. One is that the post-war boom in bird-watching has increased total records in all regions by much the same proportion when compared with the 1925-1939 period.

TABLE 7 — DISTRIBUTION OF OCCURRENCES BY YEARS
An explanation of the area-eras will be found on page 215,
and of the rarity classes on page 217.

Era A

Area-era and rarity class		1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
XA	I	2	4	2	0	1	4	7	7	10	7	19	9	13	8	15	6
	II	3	2	0	2	1	1	7	3	4	8	7	4	3	3	3	2
	III	3	2	5	4	3	6	3	9	11	9	16	9	3	6	4	7
	Total	8	8	7	6	5	11	17	19	25	24	42	22	19	17	22	15
YSA	I	0	0	0	1	0	1	0	1	1	0	1	0	0	1	3	1
	II	0	0	1	1	0	0	1	0	0	3	3	1	2	0	0	0
	III	0	1	1	0	4	1	3	1	3	1	0	3	3	4	5	2
	Total	0	1	2	2	4	2	4	2	4	4	4	4	5	5	8	3
YKA	I	0	1	1	0	1	0	0	2	0	1	1	1	2	0	0	0
	II	0	0	0	1	2	0	0	0	0	1	0	0	3	1	0	0
	III	1	3	1	1	0	0	0	0	2	0	0	4	1	1	1	0
	Total	1	4	2	2	3	0	0	2	2	2	1	5	6	2	1	0

TABLE 7 — DISTRIBUTION OF OCCURRENCES BY YEARS
An explanation of the area-eras will be found on page 215,
and of the rarity classes on page 217.

Era A

Area-era and rarity class		1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
XA	I	2	4	2	0	1	4	7	7	10	7	19	9	13	8	15	6
	II	3	2	0	2	1	1	7	3	4	8	7	4	3	3	3	2
	III	3	2	5	4	3	6	3	9	11	9	16	9	3	6	4	7
	Total	8	8	7	6	5	11	17	19	25	24	42	22	19	17	22	15
YSA	I	0	0	0	1	0	1	0	1	1	0	1	0	0	1	3	1
	II	0	0	1	1	0	0	1	0	0	3	3	1	2	0	0	0
	III	0	1	1	0	4	1	3	1	3	1	0	3	3	4	5	2
	Total	0	1	2	2	4	2	4	2	4	4	4	4	5	5	8	3
YKA	I	0	1	1	0	1	0	0	2	0	1	1	1	2	0	0	0
	II	0	0	0	1	2	0	0	0	0	1	0	0	3	1	0	0
	III	1	3	1	1	0	0	0	0	2	0	0	4	1	1	1	0
	Total	1	4	2	2	3	0	0	2	2	2	1	5	6	2	1	0

Era A

Area-era and rarity class		1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	TOTAL
XA	I	14	14	8	27	22	15	6	4	3	5	0	3	6	2	243
	II	3	2	7	9	7	5	1	5	2	3	2	4	2	3	108
	III	10	5	5	5	5	4	4	4	4	4	1	5	2	7	165
	Total	27	21	20	41	34	24	11	13	9	12	3	12	10	12	516
YSA	I	1	0	2	0	0	1	0	0	0	0	0	1	0	0	15
	II	0	0	0	0	0	0	0	1	0	0	1	2	0	0	16
	III	1	3	0	1	0	0	1	1	1	1	0	2	0	2	45
	Total	2	3	2	1	0	1	1	2	1	1	1	5	0	2	76
YKA	I	0	1	0	0	0	0	0	0	0	0	0	0	0	0	11
	II	2	0	0	0	0	0	0	0	0	0	1	0	0	0	11
	III	0	1	1	0	1	1	0	0	1	2	0	0	0	0	22
	Total	2	2	1	0	1	1	0	0	1	2	1	0	0	0	44

Era A

Area-era and rarity class		1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	1924	TOTAL
XA	I	14	14	8	27	22	15	6	4	3	5	0	3	6	2	243
	II	3	2	7	9	7	5	1	5	2	3	2	4	2	3	108
	III	10	5	5	5	5	4	4	4	4	4	1	5	2	7	165
	Total	27	21	20	41	34	24	11	13	9	12	3	12	10	12	516
YSA	I	1	0	2	0	0	1	0	0	0	0	0	1	0	0	15
	II	0	0	0	0	0	0	0	1	0	0	1	2	0	0	16
	III	1	3	0	1	0	0	1	1	1	1	0	2	0	2	45
	Total	2	3	2	1	0	1	1	2	1	1	1	5	0	2	76
YKA	I	0	1	0	0	0	0	0	0	0	0	0	0	0	0	11
	II	2	0	0	0	0	0	0	0	0	0	1	0	0	0	11
	III	0	1	1	0	1	1	0	0	1	2	0	0	0	0	22
	Total	2	2	1	0	1	1	0	0	1	2	1	0	0	0	44

Era B

Area-era and rarity class		1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940
XB	I	0	1	0	0	0	0	2	1	2	3	3	6	3	1	0	0
	II	2	1	1	1	1	0	2	2	0	1	1	0	2	2	0	1
	III	5	1	0	3	1	0	1	0	2	3	3	6	5	4	2	1
	Total	7	3	1	4	2	0	5	3	4	7	7	12	10	7	2	2
YSB	I	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0
	II	0	0	0	0	0	0	1	0	0	2	0	0	0	1	0	0
	III	0	1	1	2	3	0	2	2	2	0	0	1	4	0	0	0
	Total	0	2	1	3	3	0	4	3	2	2	0	1	4	1	0	0
YKB	I	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	0
	II	0	1	1	0	0	0	0	1	0	2	1	1	1	0	0	0
	III	0	0	0	1	1	0	0	2	0	0	2	2	1	0	1	0
	Total	0	1	1	1	1	0	0	3	1	3	3	4	2	1	2	1

Era B

Area-era and rarity class		1925	1926	1927	1928	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939	1940
XB	I	0	1	0	0	0	0	2	1	2	3	3	6	3	1	0	0
	II	2	1	1	1	1	0	2	2	0	1	1	0	2	2	0	1
	III	5	1	0	3	1	0	1	0	2	3	3	6	5	4	2	1
	Total	7	3	1	4	2	0	5	3	4	7	7	12	10	7	2	2
YSB	I	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0
	II	0	0	0	0	0	0	1	0	0	2	0	0	0	1	0	0
	III	0	1	1	2	3	0	2	2	2	0	0	1	4	0	0	0
	Total	0	2	1	3	3	0	4	3	2	2	0	1	4	1	0	0
YKB	I	0	0	0	0	0	0	0	0	1	1	0	1	0	1	1	0
	II	0	1	1	0	0	0	0	1	0	2	1	1	1	0	0	0
	III	0	0	0	1	1	0	0	2	0	0	2	2	1	0	1	0
	Total	0	1	1	1	1	0	0	3	1	3	3	4	2	1	2	1

Era B

Area-era and rarity class		1941	1942	1943	1944	1945	1946	1947	1948	1949	1950	1951	1952	1953	1954	Total
XB	I	0	1	0	1	0	3	3	3	2	4	5	4	3	3	54
	II	2	0	2	0	0	0	2	3	2	5	7	4	2	5	51
	III	0	2	0	1	1	2	6	1	6	11	7	10	7	12	103
	Total	2	3	2	2	1	5	11	7	10	20	19	18	12	20	208
YSB	I	0	0	0	0	0	0	2	2	0	4	2	1	2	2	19
	II	0	1	0	0	0	1	0	3	1	2	0	1	0	0	13
	III	0	0	0	1	0	0	1	2	3	2	1	2	1	1	32
	Total	0	1	0	1	0	1	3	7	4	8	3	4	3	3	64
YKB	I	0	2	2	0	1	0	0	0	1	4	2	1	4	4	26
	II	0	0	1	0	0	2	0	3	0	2	1	1	4	5	28
	III	2	2	2	3	1	1	1	0	1	5	6	3	10	6	53
	Total	2	4	5	3	2	3	1	3	2	11	9	5	18	15	107

TABLE 8 — THE NUMBER OF CLASS I RARITIES PER YEAR IN THE REST OF SUSSEX AND KENT DURING 1895-1924
An explanation of the area-eras will be found on page 215, and of the rarity classes on page 217.

	Number per year	Years with this number	Poisson frequencies
Rest of Sussex (YSA)	0	18	18.2
	1	10	9.1
	2	1	2.3
	3	1	0.4
Rest of Kent (YKA)	0	21	20.8
	1	7	7.6
	2	2	1.4
	3	0	0.2

The figures are 3.3 to I for X, 3.0 to I for YS and 3.9 to I for YK. The other point is that, although the post-war records for class III rarities in the X area are now running at a level higher than the mean for era A, while class II rarities are about equal, the post-war bird watchers have not managed to average even half the number of class I rarities per year that XA shows, while their best effort, five in 1951, is less than a fifth of the peak year (1915) in XA.

DISCUSSION

We began this investigation by noting the remarkable number of rarities recorded from the Hastings Area during the earlier part of this century. No attempt has been made to assess the intrinsic probability of obtaining so many rarities from a relatively small area in such a short time, for the obvious reason that the information necessary to determine such a probability – such as numbers of observers, intensity of observation and actual totals of rare birds to be seen – is almost wholly lacking. Instead we have classified the records in various ways and compared the distributions for the Hastings Area so obtained with those for two neighbouring areas, and during two eras. A number of striking differences in these distributions has been obtained, and most of them have been in the direction of making XA, the Hastings Area for 1895-1924, the odd one out. We now consider what hypotheses would have to be adopted to explain these differences, assuming the validity of all the records.

For the total number of occurrences in the three rarity classes, we found

XA to be quite different from the remainder of the area-eras which did not differ significantly among themselves. This discrepancy in XA is unlikely to be due simply to more or more enthusiastic observers, since the effect of this in YK, as shown in the differences between YKA and YKB, has been to leave the proportions in the rarity classes almost unchanged. Nor can X be a specially good area for class I rarities, judging by its performance during era B and since (when, in spite of the establishment of an observatory at Dungeness, there has still been no exceptional proportion of class I rarities). We must thus postulate observers who failed to report many class II and III rarities while recording all class I rarities. Also the evidence from the distribution by years shows that, to obtain the number of class I rarities actually recorded for XA, something more than twice the activity of post-war observers would be required. Whether there is any direct evidence either of the suppression of lesser rarities or of this enormously increased activity in the XA area-era I must leave others better qualified to say, but the possibility seems inherently unlikely. The distribution of the numbers at each occurrence for class I and class II rarities shows XA to have an excessive number of multiple occurrences when compared with the rest of the area-eras. Here again a mere change in the number of observers cannot account for it, since the proportion of multiple occurrences has remained effectively unchanged for YS and YK in both eras, even though the type of observation has largely changed from shooting to watching and the number of observers has greatly increased. Again to judge by the performance of XB, X has not recently been a specially good area for multiple occurrences. Hence we must suppose XA to have had observers exceptionally skilled in detecting and collecting multiple occurrences. Now although our definition allows a certain separation in both space and time for the birds in a multiple occurrence, in fact the birds in most multiple occurrences were from the same place and date, or from what was stated to be the same flock at different dates. It is difficult to conceive of an observer who will produce markedly more multiple occurrences than average. For if a person is skilled enough to track down one rarity he surely will not omit to look around for the possible presence of others of the same species. Nevertheless, the presence of such unlikely types seems to be the only suitable explanation, assuming that we can discard the possibility, even among class I rarities, that some single occurrences were suppressed. The changes in the distribution of rarities by seasons, although not exactly the same for YS and YK, are in one respect similar : the percentage of spring and summer records has risen as we pass from era A to era B. This might perhaps be the expected consequence of a changeover

from shooting, which is primarily an autumn and winter activity, to bird-watching, which is much more an all-the-year-round activity. The greater rise of spring records in Kent than in Sussex is probably a reflection of a real difference in the numbers of spring migrants passing through the two counties, which seems likely for reasons of geography. From the position of the X area, one would expect it to behave more like the rest of Sussex than the rest of Kent. This is so in era B where, as we have shown above, the distribution by seasons of records in XB and YSB do not differ significantly. In era A, by contrast, the spring and summer percentages are both greater for X than for YS, and slightly greater for XA than for XB. Thus, once again, the XA records need a special hypothesis to account for them. The agreement between XB and YSB suggests, also once again, that it is the observers whose activities must be different. For their era they were more active in the spring and summer than observers in the rest of the two counties.

The distribution of records by years adds a further anomaly to the XA records, in that the frequencies for yearly numbers of class I rarities fit well to a simple theoretical distribution for YSA and YKA, but not to XA. The YSA and YKA records thus suggest a more or less static situation with regard to both numbers of rarities and numbers and activities of observers, while XA suggests violent fluctuations in one or the other or both. The era B records are interesting in showing that a trend like the post-war increase in bird-watching is reflected very similarly in all three areas, which we might expect *a priori*, in contrast to the situation in era A when area X is so different from the other two regions.

It will now be clear from the foregoing discussion that if we accept all the XA records as genuine we are led to postulate an extraordinary situation regarding the activities of observers operating in this area-era. While the apparent results of their activities cannot be proved to be impossible, they appear so inherently unlikely as to call very seriously in question the basic assumption that all the XA records are genuine. I conclude that the data themselves constitute a strong *prima facie* case for a thorough investigation into the circumstances in which the Hastings Rarities came into existence.

SUMMARY

(1) A statistical investigation has been made of certain aspects of the many rare birds recorded in east Sussex and west Kent in the era 1894-1924 (the "Hastings Rarities"), using other areas in Kent and Sussex and a later era (1925-1954) for comparison.

(2) The basic unit for the analysis is an occurrence, which may involve one, two or more birds. Species and subspecies are classified into three rarity classes based on the number of English occurrences given in *The Handbook*. The distribution of the total number of occurrences of birds in three different classes of rarity shows the Hastings records in the era 1895-1924 to be anomalous, the remaining area-eras being consistent with one another.

(3) The distribution of the numbers at each occurrence for species of the greatest rarity is also shown to be anomalous for Hastings 1895-1924 when compared with the remaining area-eras.

(4) The proportion of spring and summer records for the two areas excluding Hastings is shown to have increased from era 1895-1924 to era 1925-1954, but to have decreased for Hastings. Other anomalous results involving the Hastings 1895-1924 records are pointed out.

(5) The distribution of occurrences year-by-year over the period 1895-1924 is shown to fit a simple theoretical distribution for the two areas excluding Hastings, but not to fit any such distribution for Hastings. Certain trends common to all areas for the period 1925-1954 are pointed out and the results compared with those for the earlier period.

(6) Auxiliary hypotheses necessary to account for these anomalous results are considered, on the assumption that all the records are genuine.

(7) It is concluded that these hypotheses are exceedingly unlikely to be true and that the basic assumption of the validity of all the records must be questioned.

Acknowledgments

I am most grateful to my wife and Hilary Fry for much assistance in extracting and checking the records, to D. D. Harber and E. H. Gillham for helping to track down places, and to I. J. Ferguson-Lees and E. M. Nicholson for discussion and criticism.

References

Harrison, J.M. (1953) : *The Birds of Kent.* London. Vols. 1-2.

Snedecor, G. W. (1946): *Statistical Methods.* Iowa State College Press. 4th edition.

Walpole-Bond, J. (1938): *A History of Sussex Birds.* London. Vols. I-III.

Witherby, H. F., Jourdain, F. C. R., Ticehurst, N. F., and Tucker, B. W. (1938-41): *The Handhook of British Birds.* London. Vols. 1-5.

Appendix-Rarity classes of species and subspecies analysed

A full explanation of the rarity classes will be found on page 217.

Class I (rarities with 1-19 English examples accepted in *The Handbook*)

Wilson's Petrel
Madeiran Petrel
Madeiran Little Shearwater
Cape Verde Little Shearwater
Audubon's Shearwater
Mediterranean Shearwater
North Atlantic Shearwater
Upland Sandpiper
Slender-billed Curlew
Solitary Sandpiper
Spotted Sandpiper
Greater Yellowlegs
Lesser Yellowlegs
Marsh Sandpiper
Grey-rumped Sandpiper
Terek Sandpiper
Baird's Sandpiper
White-rumped Sandpiper
Semipalmated Sandpiper
Buff-breasted Sandpiper
Broad-billed Sandpiper
Black-winged Pratincole
Bulwer's Petrel
Little Egret
Great White Heron
American Bittern
Blue-winged Teal
King Eider
Kite
Lesser Kestrel
Red-rumped Swallow
Thick-billed Nutcracker
Wallcreeper
Dusky Thrush
Black-throated Thrush
Alpine Ring Ouzel
Rock Thrush
Desert Wheatear
Western Desert Wheatear
Western Black-eared Wheatear
Eastern Black-eared Wheatear
Isabelline Wheatear
Black Wheatear
Sociable Plover
Semipalmated Ringed Plover
Killdeer
Caspian Plover
American Golden Plover
Asiatic Golden Plover
Dowitcher
Olivaceous Warbler
Orphean Warbler
Rüppell's Warbler
Sardinian Warbler
Rufous Warbler
Brown-backed Warbler
Dusky Warbler
Brown Flycatcher
Collared Flycatcher
Masked Wagtail
Grey-headed Wagtail
Black-headed Wagtail
"Sykes's" Wagtail
South European Grey Shrike

Ivory Gull
Great Black-headed Gull
Mediterranean Black-headed Gull
Bonaparte's Gull
Sooty Tern
Bridled Tern
Yellow-billed Cuckoo
Black Lark
Calandra Lark
White-winged Lark
Short-toed Lark
Crested Lark
North African Black Wheatear
Siberian Stonechat
Thrush Nightingale
White-spotted Bluethroat
Cetti's Warbler
Savi's Warbler
Moustached Warbler
Great Reed Warbler
Eastern Great Reed Warbler
Melodious Warbler
Icterine Warbler
Lesser Grey Shrike
Corsican Woodchat Shrike
Masked Shrike
Pine Grosbeak
Black-headed Bunting
Rock Bunting
Rustic Bunting
Little Bunting
Western Large-billed Reed Bunting
Eastern Large-billed Reed Bunting
Snow Finch

CLASS II (rarities with 20-99 English examples accepted in *The Handbook*)

Balearic Shearwater
Purple Heron
Squacco Heron
Surf Scoter
Red-footed Falcon
Little Crake
Pectoral Sandpiper
Pratincole
Cream-coloured Courser
Whiskered Tern
Gull-billed Tern
Caspian Tern
Scops Owl
Snowy Owl
Tengmalm's Owl
Slender-billed Nutcracker
White's Thrush
Aquatic Warbler
Barred Warbler
Yellow-browed Warbler
Red-breasted Flycatcher
Alpine Accentor
Richard's Pipit
Tawny Pipit
Red-throated Pipit
Woodchat Shrike
Serin

CLASS III (rarities with occurrences not enumerated in *The Handbook*)

Great Shearwater
Night Heron
White Stork
Glossy Ibis
Red-crested Pochard
Ferruginous Duck
Ruddy Shelduck
Goshawk
White-tailed Eagle
Gyr Falcon
Baillon's Crake
Eastern Little Bustard
Great Snipe
Black-winged Stilt
Red-necked Phalarope
Pomarine Skua
Long-tailed Skua
Iceland Gull
Sabine's Gull
White-winged Black Tern
Roseate Tern
Black Guillemot
Pallas's Sandgrouse
Bee-eater
Roller
Golden Oriole
Chough
British Dipper
Red-spotted Bluethroat
Scandinavian Chiffchaff
Siberian Chiffchaff
Water Pipit
Rose-coloured Starling
Northern Bullfinch
Two-barred Crossbill
Ortolan Bunting

THE WORKS OF JOHN NELDER

Years refer to completion of work.

BOOKS

1971 *Computers in Biology.*
London and Winchester: Wykeham Publications (London) Ltd., pp. 149.

1983 (with P. McCullagh)
Generalized Linear Models. [A Monograph].
London: Chapman and Hall, pp. 261.

1989 (with P. McCullagh)
Generalized Linear Models, second edn.
London: Chapman and Hall, pp. 511.

PAPERS

1951 A note on the statistical independence of quadratic forms in the analysis of variance. *Biometrika*, **38**, 482–483.

1952 Some genotypic frequencies and variance components occurring in biometrical genetics. *Heredity*, **6**, 387–394.

1953 Statistical models in biometrical genetics. *Heredity*, **7**, 111– 119.

1954 A note on missing plot values. *Biometrics*, **10**, 400–401.

1954 The interpretation of negative components of variance. *Biometrika*, **41**, 544–548.

1955 (with J. M. Hammersley) Sampling from an isotropic Gaussian process. *P. Camb. Philol. S.*, **51**, 652–662.

1956 Notes on the planning and executing of field experiments. *NAAS Quart. Rev.*, **34**, 139–146.

1956 (with N. Moss) The spacing of lettuce in heated glasshouses. *J. Hortic. Sci.*, **31**, 177–187.

1956 The spacing of lettuce in Dutch light frames and cold structures. *Exp. Hortic.*, No. 4, 20–30.

1959 National Vegetable Research Station new laboratory building. *Nature*, **184**, 1368–1369.

1960 The estimation of variance components in certain types of experiment on quantitative genetics (ed., O. Kempthorne). *Biometrical Genetics*, London: Pergman Press, 139–158.

1960 (with R. B. Austin, J. K. A. Bleasdale & P. J. Salter) An approach to the study of yearly and other variation in crop yields. *J. Hortic. Sci.*, **35**, 73–82.

1960 (with J. K. A. Bleasdale) Plant population and crop yield. *Nature*, **188**, 342.

1961 The fitting of a generalization of the logistic curve. *Biometrics*, **17**, 89–110.

1961 A note on some growth patterns in a simple theoretical organism. *Biometrics*, **17**, 220–228.

1962 (with Patricia Cooke, R. Morley Jones, K. Mather and G. W. Bonsall) Estimating the components of continuous variation. I. Statistical. *Heredity*, **17**, 115–133.

1962 New kinds of systematic designs for spacing experiments. *Biometrics*, **18**, 283–307.

1962 An alternative form of generalized logistic equation. *Biometrics*, **18**, 614–616.

1962 Quantitative genetics and growth analysis. Statistical genetics and Plant Breeding, Washington National Academy of Sciences, 1963, 445–545.

1962 A statistical examination of the Hastings rarities. *Brit. Birds*, **55**, 283–297.

1963 Identification of contrasts in fractional replicates of 2^n experiments. *Appl. Stat.*, **12**, 38–43.

1963 Yield-density relations and Jarvis's lucerne data. *J. Agr. Sci.*, **61**, 427–429.

1963 The use of response surfaces in the interpretation of groups of experiments. Presented - 5th International Biometric Conference, Cambridge, 1963.

1964 (with R. B. Austin and G. Berry) The use of a mathematical model for the analysis of manurial and weather effects on the growth of carrots. *Ann. Bot.*, **28**, 153–162.

1964 (with D. J. Greenwood) The effect of drugs on the growth of Lemna minor. *C. Ann. Bot. N. S.*, **28**, 711–715.

1964 Appendix to 'Effect of shape on oxygen diffusion and aerobic respiration in soil aggregates'. *J. Sci. Food Agr.*, **11**, 781–790.

1965 The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proc. R. Soc. A*, **283**, 147–162.

1965 The analysis of randomized experiments with orthogonal block structure. II. Treatment structure and the general analysis of variance. *Proc. R. Soc. A*, **283**, 163–178.

1965 (with R. Mead) A simplex method for function minimization. *Comput. J.*, **7**, 303–333.

1966 Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128–141.

1966 Evolutionary experimentation in horticulture. *J. Sci. Food Agr.*, **17**, 7–9.

1966 (with G. Berry, T. J. Cleaver and P. J. Salter) Methods of recording data in the laboratory and field. *Exp. Agr.*, **2**, 69–80.

1967 (with B. E. Cooper (co-editors)) Statistical programming. *Appl. Stat.*, **16**, (2).

1967 The application of computers to agricultural and horticultural research. (Middleton Memorial Lecture). *Agric. Prog.*, **42**, 7–23.

1967 Epilogue to a meeting on 'Statistical Programming'. *Appl. Stat.*, Vol XVI, No. **2**, 133–151.

1968 Weighted regression, quantal response data, and inverse polynomials. *Biometrics*, **24**, 979–985.

1968 Construction of additive tables. *Appl. Stat.*, **17**, 279–283.

1968 Main effects from a multiway table. *Appl. Stat.*, **17**, 277– 279.

1968 The combination of information in generally balanced designs. *J. Roy. Stat. Soc. B*, **30**, 303–311.

1968 Regression, model building and invariance. *J. Roy. Stat. Soc. A*, **131**, 303–329.

1969 The efficient formation of a triangular array with restricted storage for data. *Appl. Stat.*, **118**, 202–206.

1969 The description of data structures for statistical computing. In: *Statistical Computation*, New York: Academic Press, 13– 36.

1971 Statistical computing and computer languages. *Appl. Stat.*, **20**, 25–32.

1971 (with B. E. Cooper) Input/output in statistical programming. *Appl. Stat.*, **20**, 56–73.

1971 (with J. C. Gower) Statistical systems and general-purpose languages. *Bull. Inst. Int. Stat.*, **44**, 296–301.

1971 Discussion 'On some desirable patterns in block designs' by T. Calinski. *Biometrics*, **27**, 275–292.

1972 Summary and assessment: a statistician's point of view. In: *Mathematical Models in Ecology.* London: Blackwell's Scientific Publications, 367–373.

1972 Mathematics for biologists. *Bull. Inst. Math. App.*, **8**, 217–219.

1972 (with R. W. M. Wedderburn) Generalized Linear Models. *J. Roy. Stat. Soc. A*, **135**, 370–384.

1974 Log linear models for contingency tables: a generalization of classical least squares. *Appl. Stat.*, **23**, 323–329.

1974 Statistical packages. Introductory remarks. *Bull. Inst. Math. App.*, **10**, 165–166.

1974 A user's guide to the evaluation of statistical packages and systems. *Int. Statist. Rev.*, **42**, 291–298.

1974 Genstat: a statistical system. In: COMPSTAT: Proceedings in Computational Statistics. Vienna: Physica-Verlag.

1975 An introduction to Genstat. *Mathematical Scientist*, Supplement to Vol. **1**, No. 1, 57–58.

1975 GLIM (Generalized Linear Interactive Modelling). *Appl. Stat.*, **24**, 259–261.

1975 (with R. W. Payne) Data structures in statistical computing. Proceeds of the 9th Int. Biometrics Conf., Vol. II, 1991–207.

1975 Obituary. Robert William MacLagan Wedderburn, 1947-1975. *J. Roy. Stat. Soc. A*, **138**, 587.

1975 My kind of statistics. *Bias* **4**, 77–81.

1976 A simple algorithm for scaling graphs. *Appl. Stat.*, **25**, 94–96.

1976 Intelligent programmes, the next stage in statistical computing. Proceedings of the European Congress of Statisticians, Grenoble, 1976.

1977 A reformulation of linear models. *J. Roy. Stat. Soc. A*, **140**, 48–77.

1977 Multi-dimensional contingency table with one factor as a response. *The Statistician*, **26**, 41–42.

1978 The future of statistical software. *Compstat 1978*, Proc. in Computational Statistics, Leiden: Physica-Verlag, 11–19.

1978 (with R. J. Baker) The GLIM System Manual, Release 3, Oxford: Numerical Algorithms Group.

1978 Comment on Review of user Guides to BMDP and SPSS. *J. Am. Stat.Assoc.*, **73**, 89–90.

1979 Experimental design and statistical evaluation. Proc. of I.F.I.P. Working Conference on Performance of Numerical Software, Baden, Austria: North Holland Publishing Co., 309–315.

1980 Iterative weighted least squares; an algorithm for many occasions. Proc. of the Second I.R.I.A. Symposium Data Analysis and Informatics, Versailles, 1979. North Holland Publishing Co., 75–81.

1980 Les qualites souhaitables dans des systemes statistiques en se servant de Genstat comme reference. Statistique et Analyse de Données, **5** 17–27.

1982 (with P. W. Lane) Analysis of Covariance and Standardization as Instances of Prediction. *Biometrics*, **38**, 613– 621.

1982 Generalized Linear Models; a useful synthesis in statistics. *I.H.S. Journal*, Vol. **5**, 191–201. Vienna: Physica-Verlag.

1982 Linear Models and Non-Orthogonal Data. *Utilitas Mathematica*, **21B**. Special issue for 80th birthday of Dr. F. Yates, 141–152.

1982 (with R. J. Baker) GLIM. *Encyclopedia of Statistical Sciences*, **3**, 439–442.

1982 (with R. J. Baker) Generalized Linear Models. *Encyclopedia of Statistical Sciences*, **3**, 343–348.

1982 Analysis of Generalized Linear Models using GLIM. *Metron*, **40**, 243–248.

1983 The role of models in official statistics. Eurostat Seminar on recent developments in the analysis of large scale data sets, 16–18 November 1982. Item No. 01, Secretariate: PB 1907-Luxembourg.

1984 Do statisticians need special interactive languages? *Data Analysis and Informatics, III* (eds. E. Diday et al). North Holland: Elsevier Science Publishers, 511–516.

1984 Present Position and Potential Developments: Some Personal Views. Statistical Computing. 150th Anniversary issue. *J. Roy. Stat. Soc. A*, **147**, 151–160.

1984 (with R. J. Baker) Statistical software; progress and prospects. Computer Science and Statistics: Proceedings of the 16th Symposium on the Interface, 33–37.

1984 An alternative interpretation of the singular-value decomposition in regression. *Am. Stat.*, **39**, 63–64.

1984 Models for rates with Poisson errors. *Biometrics*, **40**, 1159– 1162.

1985 Statistical Models for Qualitative data. Measuring the Unmeasurable, Martinus Nijhoff, Dordrecht, 31–38.

1986 Statistics, Science and Technology. *J. Roy. Stat. Soc. A*, **149**, 109–121.

1986 (with D. E. Wolstenholme) A front end for GLIM. Expert Systems in Statistics, Gustav Fischer, Stuttgart.

1986 (with D. E. Wolstenholme) A front end for GLIM. Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface (ed. J. Boardman), 155–177. American Statistical Association, Washington D.C.

1987 (with D. Pregibon) An extended Quasi-likelihood function. *Biometrika*, **74**, 211–232.

1987 Artificial Intelligence and Generalized Linear Modelling: An expert System for GLIM. *Interactions in Artificial Intelligence and Statistical Methods*, 36–44. London: Unicom Technical Press.

1988 Comment on Paper by Streitberg, B. Statistical Software Newsletter, in *Comput. Stat. Data Anal.*, **14**, 73.

1988 The Role of Expert Systems in Statistics. Fortschritte der Statistik-Software **1**, 175–182.

1988 (with D. E. Wolstenholme and C. M. O'Brien) GLIMPSE: a knowledge-based front end for statistical analysis. *Knowledge-Based Systems*, **1**, 173–178.

1988 How should the statistical expert system and its user see each other? *Compstat 88*, 107–116. Heildelberg: Physica-Verlag.

1989 Contribution to discussion of paper by Godambe and Thompson. *J. Stat. Plan. Infer.*, **22**, 153–172.

1990 Nearly parallel lines in residual plots. *Am. Stat.*, **44**, 221–222.

1991 (with Y. Lee) Generalized linear models for the analysis of Taguchi-type experiments. *Appl. Stoc. Model. Data Anal.*, **7**, 107–120.

1991 GLIMPSE, a knowledge-based front end for GLIM. In *Computing and Graphics in Statistics*, 125–131. New York: Springer- Verlag.

1991 (with R. W. Payne) GENSTAT as a computing environment. In *Computing and Graphics in Statistics*, 133–138. New York: Springer-Verlag.

1992 (with Y. Lee) Likelihood, quasi-likelihood and pseudo- likelihood: some comparisons. *J. Roy. Statist. Soc. B*, **54**, 273–284.

1992 Generalized linear models for enzyme-kinetic data. *Biometrics*, **47**, 1605–1609.

1992 The computer as statistical assistant. In *Programming Environments for High-Level Scientific Problem Solving*, 69–75. Amsterdam: North Holland.

1992 Contributions to 'Taguchi's parameter design: a panel discussion'. *Technometrics*, **34**,127–161.

1992 Pseudo-likelihood and quasi-likelihood. Letter to *Appl. Stat.*, **41**, 595–597.

1992 Modelling in statistics. *Math. Comput. Model.*, **16**, 131–136.

1992 Models for curves through the origin. *J. Appl. Stat.*, **19**, 551–552.

1992 Joint modelling of mean and dispersion. In *Statistical Modelling*, 263–272. Elsevier Science Publishers B.V.

1992 Statistical packages and unbalanced data. Statistical Software Newsletter, in *Comput. Stat. Data Anal.*, **14**, 403–406.

1993 Aspects of statistical computing, past, present and future. *Stat. Neerl.*, **47**, 3–8.

1993 The most important areas of statistical research in the next ten years. *Stat. Comput.*, **3**, 202–203.

1994 An alternative view of the splicing data. *Appl. Stat.*, **43**, 469–476.

1994 (with Y. Lee) Double generalized linear models. Technical Report, Department of Mathematics, Imperial College.

1994 A re-analysis of the pump-failure data. *Scand. J. Stat.*, **21**, 187–191.

1994 The statistics of linear models: back to basics. *Stat. Comput.*, **4**, 221–234.

1995 (with P. W. Lane) The computer analysis of factorial experiments: in memoriam Frank Yates. *Am. Stat.*, **49**, 382–385.

1995 Rejoinder to comments on 'The statistics of linear models: back to basics'. *Stat. Comput.*, **5**, 109–111.

1995 Generalized linear models: a powerful tool for the analysis of quality-improvement experiments. Proc. Intl. Conf. on Statistical Methods and Statistical Computing for Quality and Productivity Improvement, Seoul.

1996 Computing: in *Advances in Biometry.* (Ed. David & Armitage) John Wiley & Sons.

1996 (with Y. Lee) Hierarchical Generalized Linear Models. *J. Roy. Stat. Soc. B*, **58**, 619–656.

1996 (with H. Su, P. Wolbert & R. Spence) Application of generalized linear models to the design improvement of an engineering artifact. *Qual. Reliab. Eng. Int.*, **12**, 101–112.

1997 (with M. Hamada) Generalized linear models for quality- improvement experiments. *J. Qual. Technol.*, **29**, 292–304.

1997 Functional marginality is important. Letter to *Appl. Stat.* ***46***, 281–2.

1997 The great mixed-model muddle is alive and flourishing - alas! *Food Qual. Prefer.*, **9**, 157–159.

1997 (with R. J. Verrall) Credibility theory and generalized linear models. *Astin Bulletin*, **27**, 71–82.

1998 A large class of models derived from generalized linear models. *Stat. Med.*, **17**, 2747–2753.

1998 (with Y. Lee) Generalized linear models for the analysis of quality-improvement experiments. *Canad. J. Stat.*, **26**, 95–105.

1998 (with Y. Lee) Joint modeling of mean and dispersion. Letter to *Technometrics*, **40**, 168–171 .

1998 How strong is the weak-heredity principle? *Am. Stat.*, **52**, 315–318.

1998 From statistics to statistical science. *J. Roy. Stat. Soc. D*, **48**, 269.

1998 (with Y. Lee) Extended quasi-likelihood and estimating equations. Selected Proc. of the Symp. on Estimating Equations: IMS Lecture Notes, Vol.**32**, 139–148.

1998 (with N. T. Longford) Statistics vs. statistical science in the regulatory process. *Stat. Med.*, **18**, 2311–2320.

1998 (with Z. Malik & H. Su) Informative experimental design for electronic circuits. *Qual. Reliab. Eng. Int.*, **14**, 1–10.

1998 (with Z. Malik) Modified inverse polynomials as a reponse- surface model for magnetic flux density. *Electron. Lett.*, **34**, 2252–2253.

1998 (with Z. Malik, L. Tweedie, A. J. Smith & R. Spence) Modelling for problem holders: the modellers' guide. *Compstat98*: Proc. in Computational Statistics, Bristol, UK R. Payne and P. Lane (eds), pp:185–186.

1998 (with Z. Malik, D. Dyck, R. Spence & D .Lowther) Response-surface models using function values and gradient information, with application to the design of an electromagnetic device. Engineering Design Conference '98, 199–209.

1999 (with A. J. Smith, Z. Malik, & R. Spence) Visualisation tools for model making. Proc. 14th International Workshop on Statistical Modelling. Graz, Austria, July 19–23, 1999.

1999 Comments on 'Experimental design for product and process design and development'. *J. Roy. Stat. Soc. D*, **49**, 107–109.

1999 (with Y. Lee) The robustness of the quasi-likelihood estimator. *Canad. J. Stat.*, **27**, 321–327.

1999 (with A. J. Smith, A. Buja, Z. Malik, L. Tweedie & R. Spence) Sampling schemes for model visualisation. *J. Comput. Graph. Stat.*, **10**, 545–554.

1999 (with Y. Lee) Joint modelling of the mean and dispersion for the analysis of quality-improvement experiments. *Statistical Process Monitoring and Optimization* , edited by Park & Vining, Chapter 23, 387–394, Marcel Dekker, New York.

2000 (with Y. Lee) The relationship between double exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex-ratio data. *Appl. Stat.*, **49**, 413–419.

2000 (with Y. Lee) Two ways of modelling overdispersion in non-normal data. *Appl. Stat.*, **49**, 591–598

2000 How helpful are packages in teaching statistics? *Maths and Stats*, **11**, 15–17.

2000 (with Y. Lee) HGLMs for analysis of correlated non-normal data. *Compstat 2000* (ed. Bethlehem & van der Heijden), Physica-Verlag.

2000 Quasi-likelihood and pseudo-likelihood are not the same thing. *J. Appl. Stat.*, **27**, 1007–1011.

2000 There are no outliers in the stackloss data. *Student*, **3**, 211–216.

2000 The analysis of contingency tables with one factor as the reponse factor: round two. *The Statistician*, **49**, 383–388.

2000 Functional marginality and response-surface fitting. *J. Appl. Stat.*, **27**, 109–112.

2001 (with Y. Lee) Hierarchical generalized linear models: a synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

2001 (with Y. Lee) Modelling and analysing correlated non-normal data. *Statistical Modelling*, **1**, 3–16.

2001 (with A. J. Smith, Z. Malik & R. Spence) A visual interface for model-fitting. *Qual. Reliab. Eng. Int.*, **17**, 85–91.

2002 (with Y. Lee) Analysis of ulcer data using hierarchical generalized linear models. *Stat. Med.*, **21**, 191–202.

2003 (with Y. Lee) Extended REML estimators. *J. Appl. Stat.*, **30**, 845–856.

2003 (with Y. Lee) False parsimony and its detection in GLMs. *J. Appl. Stat.* **30**, 477–484.

2003 (with Y. Lee) Robust design via generalized linear models. *J. Qual. Technol.*, **35**, 2–12.

2004 (with Y. Lee) Conditional and marginal models: another view. (With discussion) to appear in *Stat. Sci.*

2004 (with Y. Lee) Fitting via alternative random-effect models. submitted to *Stat. Comput.*

2004 (with Y. Lee) Double hierarchical generalized linear models. submitted to *Scand. J. Stat.*

INDEX