ФАКТОРНЫЙ, ДИСКРИМИНАНТНЫЙ И КЛАСТЕРНЫЙ АНАЛИЗ

Перевод
с английского
А.М. ХОТИНСКОГО
С.Б. КОРОЛЕВА

Под редакцией
И.С. ЕНЮКОВА

МОСКВА
"ФИНАНСЫ И СТАТИСТИКА"
1989

ББК 16.2.9
Ф18

Авторы: Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка,
М. С. Олдендерфер, Р. К. Блэшфилд

Факторный, дискриминантный и кластерный анализ: Пер.
Ф18 с англ./Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.; Под ред. И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.: ил.
ISBN 5-279-00247-X

Книга представляет сборник работ американских ученых, в которых рассмотрен аппарат факторного, дискриминантного и кластерного анализа, широко применяемый в социально-экономических классификациях и анализе неявных закономерностей в экономических и социальных процессах.

Для научных работников, преподавателей, студентов, аспирантов экономических вузов

Ф $\frac{0702000000-068}{010(01)-89}$ 111—89

ББК 16.2.9

ISBN 5-279-00247-X

ОТ ИЗДАТЕЛЬСТВА

В настоящее время методы многомерной статистики интенсивно развиваются. Углубляется понимание смысла возникающих задач, разрабатываются новые, более эффективные методы. С появлением персональных компьютеров существенно расширился круг исследователей и практических работников, применяющих аппарат многомерного статистического анализа. Этот процесс не может быть отнесен только к математической или прикладной статистике. Использование методов многомерной статистики предполагает обращение к системному анализу рассматриваемого явления, основных его составляющих и их связей, принятие решения о характере установленных закономерностей. Кроме того, программно-алгоритмическое обеспечение такого анализа имеет отношение к методам искусственного интеллекта (обобщение данных с помощью факторного и кластерного анализа, распознавание с помощью дискриминантного анализа).

Именно этот многогранный характер современного развития многомерной статистики, выводящий на стык собственно статистики, системного анализа и принятия решений, информатики и искусственного интеллекта, обеспечивает постоянный и достаточно бурный рост числа специалистов, использующих эту методологию.

Данное издание призвано послужить практическим интересам этой обширной и быстро растущей аудитории специалистов и так называемых «пользователей», которые имеют в своем распоряжении программные средства многомерной статистики, но подчас не знают, как их применять.

В настоящую книгу вошли три выпуска серии «Quantitative Applications in the Social Sciences», издаваемой американской фирмой «Sage Publications, Inc.», представляющие некоторые важнейшие методы многомерного статистического анализа.

СОДЕРЖАНИЕ

Дж.-О. Ким, Ч. У. Мьюллер

ФАКТОРНЫЙ АНАЛИЗ: СТАТИСТИЧЕСКИЕ МЕТОДЫ И ПРАКТИЧЕСКИЕ ВОПРОСЫ

Jae-On Kim, Charles W. Mueller. **Factor Analysis:** Statistical Methods and Practical Issues (Eleventh Printing, 1986).

ПРЕДИСЛОВИЕ

Настоящая работа является продолжением книги Джэй-Он Кима и Чарльза У. Мьюллера «Введение в факторный анализ: что это такое и как им пользоваться», также опубликованной в серии «Quantitative Applications in the Social Sciences». Последняя является введением в метод факторного анализа; в ней даются ответы на вопросы читателя: «Для чего используется факторный анализ?» и «Какие предположения делаются при использовании этого метода?», но не затрагиваются вопросы применения факторного анализа к конкретным данным. В работе «Факторный анализ: статистические методы и практические вопросы» более подробно рассматриваются специфические примеры анализа данных, различные виды факторного анализа и ситуации, когда его применение наиболее полезно. Различие между конфирматорным и разведочным факторным анализом здесь обсуждается более детально, чем во «Введении в факторный анализ». Например, рассматриваются различные критерии для факторного вращения. Особенно полезным является обсуждение различных форм косоугольных вращений и интерпретации коэффициентов в факторном анализе. Дж.-О. Ким и Ч. У. Мьюллер также ставят вопрос о числе факторов, фигурирующих в разведочном факторном анализе, разбирают методы проверки гипотез в конфирматорном анализе и рассматривают проблему вычисления значений факторов. Предлагается словарь специальных терминов, а также ответы на вопросы, наиболее часто возникающие у пользователей факторного анализа, которые могут предостеречь их от многих ошибок. Математический аппарат достаточно скромный — приводятся только сведения из матричной алгебры.

ISBN 0-8039-1161-1

Факторный анализ использовался в экономических задачах, в которых наличие сильно коррелированных параметров приводило к неверным результатам в регрессионном анализе. Ученые, занимающиеся общественно-политическими проблемами, сопоставляли всевозможные признаки наций с разными политическими и социально-экономическими характеристиками, пытаясь определить, какие из них наиболее важны при классификации наций (например, благосостояние и численность); социологи определяли «дружественные группы», изучая группы людей, симпатизирующих именно друг другу (а не другим индивидуумам). Психологи использовали метод факторного анализа для определения того, как люди воспринимают всевозможные «стимулы» и классификации людей в группы, соответствующие различным реакциям, а издатели применяли факторный анализ для изучения способов связывать отдельные элементы языка*.

Как утверждают авторы, их работа не охватывает всех аспектов факторного анализа, так как он постоянно развивается. Тем не менее если читатель получит достаточно полное представление о том, как этот метод может быть использован, то можно считать, что авторы выполнили свою задачу.

Е. М. Асланер, редактор серии

---

* Более подробно одно из таких исследований описано в разд. «Кластерный анализ». — *Примеч. ред.*

I. ВВЕДЕНИЕ

Основная концепция факторного анализа проста и несложна для изучения. Тем не менее существует несколько причин, по которым овладение методом для практического использования может быть достаточно трудным. Во-первых, для понимания принципов статистического оценивания, как правило, требуется бóльшая искушенность в математике, чем это необходимо для понимания постановки задачи. Во-вторых, в литературе были описаны многочисленные методы получения факторных решений, и даже относительно простая компьютерная программа, вероятно, предусматривает различные варианты на каждой стадии анализа. Эти обстоятельства могут ошеломить начинающего и вызвать затруднения даже у специалиста. В-третьих, практическая задача почти всегда является более сложной, чем предполагается в факторной модели. Например, (1) организация измерений некоторых или всех переменных не соответствует требованиям, принятым в факторном анализе; (2) некоторые предположения модели, такие, как независимость ошибок измерений, могут не выполняться для определенных данных или (3) могут существовать второстепенные факторы, идентификация которых непосредственно не нужна, но присутствие которых влияет на идентификацию основных общих факторов. Трудность состоит в том, что исследователь должен в конце концов принять по собственному усмотрению некоторые «внестатистические» решения. К счастью, как будет показано, эти трудности преодолимы.

Исследователь для решения проблемы в большей или меньшей степени должен положиться на существующие компьютерные программы, которые часто предусматривают различные варианты вычислений, принятые по умолчанию. Последние устраивают пользователя по крайней мере до тех пор, пока задача не потребует некоторых изменений. Более того, по мере знакомства с разнообразными вариантами факторного анализа становится ясно, что различия между ними большей частью поверхностны. Фактически это разнообразие обусловлено расхождением в небольшом числе основных предположений.

Еще более существенно, что применение различных методов и критериев к одним и тем же данным приводит к эквивалентным, с практической точки зрения, результатам. Короче говоря,

читателю не обязательно изучать и использовать все варианты немедленно. Вместе с тем необходимо, чтобы пользователь знал наиболее распространенные алгоритмы факторного анализа и осознавал с самого начала тот факт, что большинство проблем не имеет единственного, окончательного (или наилучшего) решения.

Надеемся, что читатель имеет общее представление о концепции факторного анализа, а также знаком с различием между неоднозначностью вывода скрытой (латентной) факторной структуры из наблюдаемых ковариаций (логическая задача) и разбросом значений оценок параметров генеральной совокупности по выборке (статистическая задача). Хотя при получении решения задачи факторного анализа эти две проблемы в целом переплетаются, важно представлять концептуальные различия. Прежде чем мы изложим статистические методы и практические вопросы, нам кажется, что будет полезно обратиться к основам факторного анализа.

ОБЗОР ОСНОВ ФАКТОРНОГО АНАЛИЗА

В факторном анализе предполагается, что наблюдаемые переменные являются линейной комбинацией некоторых латентных (гипотетических или ненаблюдаемых) факторов. Некоторые из этих факторов допускаются общими для двух и более переменных, а другие — характерными для каждого параметра в отдельности. Характерные факторы — ортогональны друг другу (по крайней мере в разведочном факторном анализе). Следовательно, характерные факторы не вносят вклад в ковариацию между переменными. Другими словами, только общие факторы, число которых предполагается гораздо меньшим числа наблюдаемых переменных, вносят вклад в ковариацию между ними.

Принимаемая в факторном анализе линейная система такова, что структура ковариаций может быть идентифицирована без ошибок, если известна матрица нагрузок латентных факторов. Тем не менее однозначное восстановление латентной факторной структуры исходя из наблюдаемой ковариационной структуры всегда проблематично. Эта неопределенность не имеет никакого отношения к статистическому оцениванию и должна разрешаться с помощью «внестатистических» постулатов: принципа факторной причинности и принципа экономии.

При использовании этих постулатов и свойств линейной системы можно точно идентифицировать латентную факторную структуру путем исследования результирующей ковариационной матрицы, если структура не является слишком сложной и если она удовлетворяет требованиям простой факторной структуры. Модель с двумя общими факторами (рис. 1) может быть восстановлена из матрицы корреляций, представленной в нижнем треугольнике табл. 1. Любая компьютерная программа (какой бы алгоритм в ней ни был заложен) позволяет достаточно хорошо восстановить данную модель[1].

На практике тем не менее на исследуемую матрицу корреляций оказывают влияние различные случайные и неслучайные ошибки, и в результате она будет отлична от корреляционной матрицы, обусловленной факторной структурой генеральной совокупности. Над главной диагональю табл. 1 помещены элементы корреляционной матрицы, вычисленной для выборки объема 100 с использованием факторного отображения, приведенного на рис. 1 (т. е. с использованием матрицы корреляции под диагональю табл. 1). Обратите внимание на отличие между соответствующими наддиагональными и поддиагональными элементами таблицы и на тот факт, что каждая выборочная корреляционная матрица будет отличаться в некоторой степени от корреляционной матрицы для генеральной совокупности и от любой другой выборочной матрицы для других выборок из той же самой генеральной совокупности. Таким образом, на практике невозможно получить точную структуру факторной модели, можно только пытаться найти оценки параметров факторной структуры, с использованием определенных статистических и (или) практических критериев.

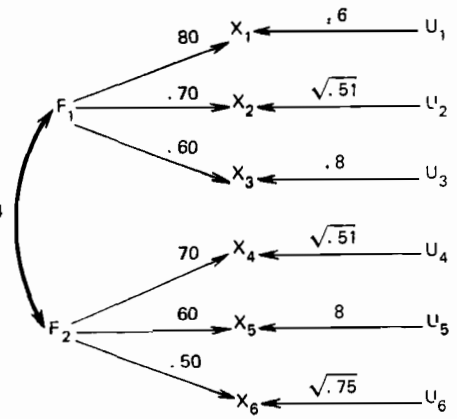

Рис. 1. Граф факторной структуры с шестью переменными и двумя косоугольными общими факторами, где наблюдаемые переменные означают:

X_1 — правительство должно тратить больше средств на школы;
X_2 — правительство должно тратить больше средств на сокращение процента безработных;
X_3 — правительство должно контролировать большой бизнес;
X_4 — правительство должно устранять сегрегацию через занятость населения;
X_5 — правительство должно обеспечивать национальным меньшинствам соответствующую квоту рабочих мест;
X_6 — правительство должно выполнять программу борьбы с кризисами

При решении задач разведочного факторного анализа исследователь обычно делает три шага: (1) подготовка соответствующей ковариационной матрицы; (2) выделение первоначальных (ортогональных) факторов и (3) вращение с целью получения окончательного решения. Подчеркнем, что исходную информацию для факторного анализа получить сравнительно просто.

ОСНОВНЫЕ АЛГОРИТМЫ И МЕТОДЫ

В зависимости от задач исследователя следует воспользоваться либо *разведочным,* либо конфирматорным факторным анализом. В обоих случаях существуют три основных этапа: подготовка соответствующей матрицы ковариаций, выделение перво-

Таблица 1

Коэффициенты корреляции для генеральной совокупности (поддиагональные элементы) и модельной выборки объема 100 (наддиагональные элементы), относящиеся к модели с двумя общими факторами, представленной на рис. 1

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	—	0,6008	0,4984	0,1920	0,1959	0,3466
X_2	0,560	—	0,4749	0,2196	0,1912	0,2979
X_3	0,480	0,420	—	0,2079	0,2010	0,2445
X_4	0,224	0,196	0,168	—	0,4334	0,3197
X_5	0,192	0,168	0,144	0,420	—	0,4207
X_6	0,160	0,140	0,120	0,350	0,300	—

начальных факторов и вращение с целью получения окончательного решения. Хотя на практике для получения окончательного решения не всегда требуются все эти шаги (особенно при проверке специальных гипотез), тем не менее удобно обсуждать разнообразие методов факторного анализа в связи с данными этапами. Таким образом, первая часть этой работы так или иначе касается этих трех этапов анализа.

Перед проведением факторного анализа необходимо решить: использовать ли как исходную матрицу ковариации (корреляции) между переменными или использовать корреляции между индивидуумами (объектами). В данной работе мы будем обсуждать только первый из этих подходов*.

На первом этапе может применяться модель *общих факторов,* а также *анализ главных компонент,* цель которого отлична от цели факторного анализа. В то же время оба метода широко используются и являются эффективными способами исследования «взаимосвязей» между переменными. Основное отличие между этими двумя методами заключается в том, что главные компоненты являются линейными функциями от наблюдаемых переменных, в то время как общие факторы не выражаются через комбинацию наблюдаемых переменных. Альтернативой анализу первоначальных факторов служит анализ образов-факторов, в котором предполагается, что наблюдаемые переменные выбраны из бесконечного множества переменных, причем вводятся «образы-факторы», являющиеся линейными комбинациями переменных. Сопоставление этих подходов будет рассмотрено ниже. Кроме того, существует несколько путей выделения первоначальных факторов. Из них в этой работе рассматриваются следующие: 1) решение, получаемое методом максимального правдоподобия (включая канонический факторный анализ); 2) решение по ме-

* Второй подход, так называемый *Q*-техника, кратко рассматривается в разд. «Кластерный анализ» — *Примеч. ред.*

тоду наименьших квадратов (включая метод минимальных остатков и метод главных факторов с итерациями по общностям) и 3) альфа-факторный анализ. Последний может рассматриваться либо как вариант метода с общими факторами, либо как альтернативная стратегия.

Шаг, связанный с вращением, включает два варианта: ортогональное и косоугольное вращение. Косоугольные вращения в свою очередь подразделяются на те, которые основаны на прямом упрощении матрицы коэффициентов факторного отображения, и на те, которые используют упрощение матрицы нагрузок на вторичные оси. Внутри этих вариантов существует множество подвариантов. О большинстве из них мы поговорим в следующих разделах. Вопрос о числе факторов рассматривается отдельно, что связано с необходимостью обсудить несколько эмпирических правил, которые многие практики находят полезными.

В разделе, посвященном конфирматорному факторному анализу, будет дано понятие эмпирического подтверждения факторных моделей, а затем мы проиллюстрируем его на двух простых, но важных практических примерах.

Далее мы обсудим вопрос вычисления значений факторов. Этот раздел помещен после обсуждения конфирматорного факторного анализа, поскольку используются некоторые его результаты.

В заключительном разделе рассматривается широкий спектр проблем в форме вопросов и ответов, причем многие из них в основном тексте вовсе не обсуждались. Здесь мы также даем некоторые практические советы для решений, по которым пока нет единого мнения.

Словарь, приложенный в конце работы, служит не для точного определения каждого термина, а лишь дает удобный способ представления контекста, в котором этот термин встречается.

И наконец, ссылки не предназначены ни для отражения исторического развития методов факторного анализа, ни для обзора последних достижений в этой области. Мы пользовались источниками, которые считали ценными, с точки зрения нашего собственного понимания предмета.

II. МЕТОДЫ ВЫДЕЛЕНИЯ ПЕРВОНАЧАЛЬНЫХ ФАКТОРОВ

Основная цель выделения первичных факторов в разведочном факторном анализе заключается в определении минимального числа общих факторов, которые удовлетворительно воспроизводят корреляции между наблюдаемыми переменными. При отсутствии ошибок измерений и случайности в выборке, а также при выполнении принципа факторной причинности, для заданной корреляционной матрицы существует точное соответствие между минимальным числом общих факторов и рангом редуцированной

корреляционной матрицы. (В редуцированной корреляционной матрице общности помещаются на главную диагональ.) Иными словами, в случае отсутствия ошибок в соответствии факторной модели данным число общих факторов и общности могут быть сколь угодно точно вычислены с помощью исследования ранга редуцированной корреляционной матрицы. Если же выборка является случайной, то проблема усложняется и возникает задача найти критерий, с помощью которого можно было бы оценить минимально необходимое число общих факторов. Но поскольку основной критерий определения минимального числа общих факторов заключается в хорошей воспроизводимости наблюдаемых корреляций с помощью отобранных факторов, то задачу можно переформулировать следующим образом: определить правило остановки при выделении общих факторов. Эта задача сводится к определению момента, когда расхождение между вычисленными и наблюдаемыми корреляциями может быть приписано случайности выборки.

Мы начнем с описания основной стратегии, которая является общей для ряда методов выделения. Она включает проверку гипотез о минимальном числе общих факторов, необходимых для воспроизведения наблюдаемых корреляций. При отсутствии априорных данных следует обратиться к однофакторной модели. Эта «гипотеза» (достаточности одного фактора) проверяется с помощью критерия, применяя который можно узнать, достигнуто ли удовлетворительное расхождение между предполагаемой моделью и данными. Если расхождение статистически значимо, то оценивается модель с еще одним дополнительным фактором и снова применяется критерий. Этот процесс продолжается до тех пор, пока расхождение не сможет быть приписано случайности выборки. Следует заметить, что реальные компьютерные программы могут явно не делать такую последовательную оценку, но принцип выделения первых k факторов, которые согласуются с наблюдаемыми ковариациями, остается в силе.

Хотя принцип этой основной стратегии прост, его применение — разнообразно, поскольку есть различные критерии наилучшего соответствия (или минимальной невязки). Существуют два главных метода решения, в которых фигурируют общие факторы: 1) метод максимального правдоподобия [Lawley, Maxwell, 1971; Jöreskog, 1967; Jöreskog, Lawley, 1968], варианты которого сводятся к каноническому факторному анализу [Rao, 1955] и к алгоритмам, основанным на минимизации детерминантов матрицы частных коэффициентов корреляции [Browne, 1968]; 2) метод наименьших квадратов, варианты которого включают метод главных осей с итерациями по общности [Thomson, 1934] и метод минимальных остатков [Harman, 1976]. Кроме того, существуют еще три основных метода выделения: 1) альфа-факторный анализ [Kaiser, Gaffrey, 1965]; 2) анализ образов [Guttman, 1953; Harris, 1962] и 3) анализ главных компонент [Hotelling, 1933].

ГЛАВНЫЕ КОМПОНЕНТЫ, СОБСТВЕННЫЕ ЗНАЧЕНИЯ И ВЕКТОРА

Мы начинаем обсуждение именно с анализа главных компонент по двум причинам: во-первых, он послужит в качестве базовой модели, с которой будут сравниваться и сопоставляться методы, где используются общие факторы. Во-вторых, он представляется наиболее простым для введения таких особых понятий, как корни характеристического уравнения (собственные числа) и собственные вектора, и дает возможность выявить их роли в алгоритмах факторного анализа. (Мы не отказывается от стремления применять наиболее простой математический аппарат, но знакомство с подобной терминологией необходимо для использования многих компьютерных программ. Мы настоятельно рекомендуем читателям ознакомиться с основными определениями.)

Анализ главных компонент — это метод преобразования данной последовательности наблюдаемых переменных в другую последовательность переменных. Наиболее простой способ пояснить внутреннюю логику метода сводится к его изучению в двумерном случае. Предположим, что есть две переменные X и Y с совместным нормальным распределением.

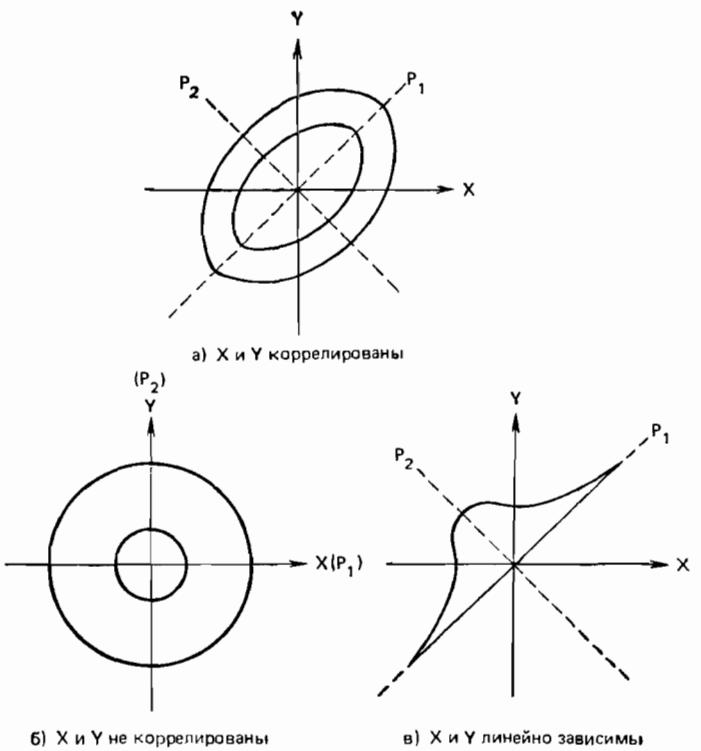

Рис. 2. Главные оси двумерных распределений

Совместное нормальное распределение величин, имеющих положительную корреляцию, представлено на рис. 2 с помощью кривых равных вероятностей. Эти кривые показывают, что благодаря положительной связи между X и Y данные представляют кластер, в котором бо́льшие величины X имеют тенденцию соответствовать бо́льшим величинам Y (и наоборот). Таким образом, в большинстве случаев точки попадают в первый и третий квадранты, и реже — во второй и четвертый. Кривые равных вероятностей имеют форму эллипсов, две оси которых изображены пунктирными линиями. Главная ось (P_1) проходит по линии, вдоль которой располагается основная часть данных; вторая ось (P_2) — по линии, вдоль которой расположена меньшая часть данных.

Теперь предположим, что нужно представить точки в терминах только одной размерности (оси). В этом случае естественно выбрать ось P_1, потому что в целом она ближе описывает данные наблюдений. Тогда первая главная компонента есть не что иное, как представление точек, расположенных вдоль выбранной главной оси. Например, точка с единичными значениями X и Y будет иметь координату, большую 1 по оси P_1 и меньшую 1 по оси P_2. Если мы описываем каждую точку в терминах P_1 и P_2 (в новой системе координат), потери информации не произойдет. Тем не менее можем сказать, что первая ось (и первая компонента) является более информативной в описании точек, так как связь между X и Y становится сильнее. В том случае, когда X и Y связаны линейной зависимостью, первая главная компонента будет содержать всю информацию, необходимую для описания каждой точки. Если X и Y независимы, то главная ось отсутствует и анализ главных компонент не способствует даже минимальному сокращению (сжатию) результатов наблюдений.

Понятие главных осей относится не только к нормальным распределениям. В общем случае главная ось задается линией, для которой сумма квадратов расстояний до всевозможных точек минимальна. Сравнение анализа главных компонент с принципом наименьших квадратов поможет объяснить это определение. При нахождении линии регрессии ($\hat{Y}=a+bX$) методом наименьших квадратов мы минимизируем сумму квадратов расстояний между Y и $\hat{Y}$, т. е. минимизируем* ($Y-\hat{Y}$), где расстояние измеряется по линии, параллельной оси Y и перпендикулярной оси X. При нахождении главной оси мы минимизируем расстояние** от точки до оси (т. е. расстояние по перпендикуляру к главной оси, а не к оси X). Это отличие показано на рис. 3. (В [Malinvand, 1970] описан метод наименьших квадратов с помощью ортогональной регрессии.)

---

* Более точно минимизируется среднее значение квадрата такой невязки. — *Примеч. ред.*

** Более точно минимизируется среднее значение квадрата этого расстояния. — *Примеч. ред.*

Поскольку первая компонента определена таким образом, что основная доля информации содержится именно в ней (дисперсия в направлении этой компоненты максимальна), вторая компонента определяется аналогичным образом при условии, что ее ось перпендикулярна первой. Следовательно, в двумерном случае после фиксирования первой компоненты вторая становится известна автоматически. Если Y не является линейной функцией от X, то главных компонент будет две (для полного описания совместного распределения необходимы две оси).

При определении главных компонент не обязательно предполагать существование гипотетических факторов. Новые оси являются математическими (линейными) функциями наблюдаемых переменных. Даже если с помощью анализа главных компонент достигается сжатие данных (выделение только нескольких первых компонент), задача состоит не в объяснении корреляции между переменными, а в объяснении максимальной доли дисперсии наблюдений. С другой стороны, для рассматриваемого двумерного случая в факторном анализе потребуется лишь один фактор, и главной задачей будет объяснение *корреляций* между переменными. Итак, первая задача относится к объяснению дисперсий, а вторая — к объяснению корреляций.

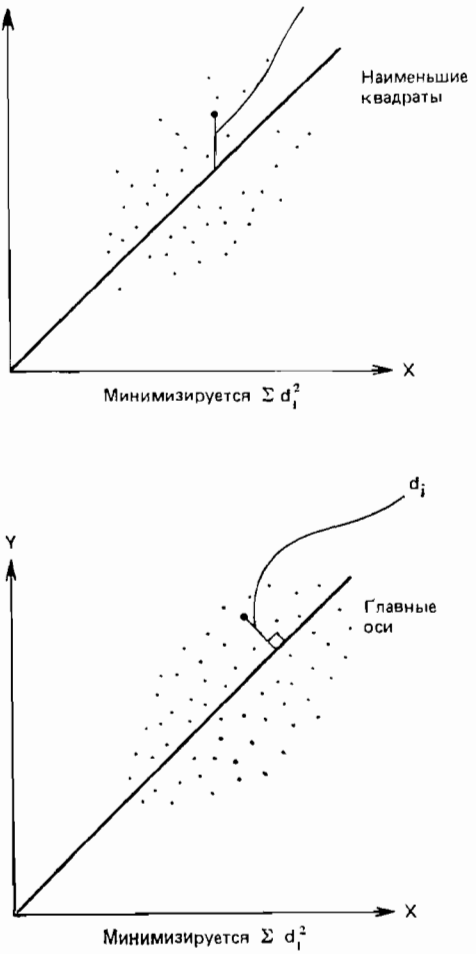

Рис. 3. Сравнение регрессий, полученных с помощью методов наименьших квадратов и главных осей

При наличии более двух переменных принцип определения главных компонент тот же. Например, для трехмерного нормального распределения поверхность равной вероятности будет ограничивать овальное тело (эллипсоид), где первая главная ось — его наибольший диаметр, вторая — пройдет по наибольшему диаметру в плоскости, перпендикулярной первой оси; третья ось будет самой короткой, перпендикулярной двум первым осям.

Основной математический метод получения направлений главных осей основан на нахождении собственных чисел и векторов корреляционной (ковариационной) матрицы. Для определения собственных чисел и векторов уравнение с использованием матричной записи имеет следующую форму:

$$RV=\lambda V, \tag{1}$$

где R — матрица, для которой ищется решение; V — искомый собственный вектор, а λ — собственное число. Решение базируется* на более простой форме в виде детерминанта матрицы:

$$\mathrm{Det}\,(R-I\lambda)=0, \tag{2}$$

что дает для квадратной матрицы уравнение

$$\mathrm{Det}\begin{pmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{pmatrix}=0, \tag{3}$$

которое по определению детерминанта может быть представлено в виде

$$(1-\lambda)(1-\lambda)-r_{12}(r_{12})=0. \tag{4}$$

Раскрывая скобки и группируя члены, получаем:

$$\lambda^2-2\lambda+(1-r_{12}^2)=0. \tag{5}$$

Собственные числа теперь могут быть получены при решении квадратного уравнения. Для двумерной корреляционной матрицы собственные числа имеют вид

$$\lambda_1=1+r_{12}, \tag{6}$$

$$\lambda_2=1-r_{12}. \tag{7}$$

Если между двумя переменными имеется линейная зависимость, то одно собственное число будет 2, а другое — 0. Для некоррелированных переменных оба собственных числа будут равны 1.

Заметим также, что сумма собственных чисел $\lambda_1+\lambda_2=(1+r_{12})+(1-r_{12})=2$ равна числу переменных, а произведение $\lambda_1\lambda_2=(1-r_{12}^2)$ равна детерминанту корреляционной матрицы. Эти свойства сохраняются для корреляционных матриц любой размерности, причем первое (большее) собственное число представляет величину дисперсии, соответствующую первой главной оси, а второе собственное число — величину дисперсии, соответствующую второй главной оси и так далее. Так как при использовании корреляционной матрицы сумма собственных чисел равна числу переменных, то, разделив первое собственное число на m (число переменных), можем получить долю дисперсии, соответствующую данному направлению или компоненте:

$$\begin{pmatrix}\text{Доля соответствующая}\\ \text{данной компоненте}\end{pmatrix} = \begin{pmatrix}\text{Соответствующее}\\ \text{собственное число}\end{pmatrix} \Big/ m. \tag{8}$$

* Использование детерминантного уравнения типа (2) эффективно только для матриц небольшого порядка (небольшого числа переменных). Гораздо результативнее различные итерационные схемы. — *Примеч. ред.*

Таблица 2

Две первые главные компоненты корреляционной матрицы, представленной поддиагональными элементами табл. 1

Переменная	Главные компоненты		h^2 [1]
	F_1	F_2	
X_1	0,749	—0,395	0,713
X_2	0,706	—0,405	0,666
X_3	0,651	—0,417	0,597
X_4	0,595	0,579	0,623
X_5	0,548	0,529	0,581
X_6	0,488	0,526	0,514
Собственные значения	2,372	1,323	Сумма = =3,695
Доля дисперсии, соответствующая компоненте F_i (%)	39,5	22,1	
Доля дисперсии, соответствующая компоненте (F_λ) и двум компонентам (F_1 и F_2) вместе (%)	39,5	61,6	

[1] Величины h^2, строго говоря, не являются оценками общностей, так как в анализе главных компонент не предполагается существование общих факторов.

При определении соответствующих собственных векторов есть дополнительное ограничение, состоящее в том, что их длина должна быть единичной. По этой причине коэффициенты нагрузок для главных компонент получаются делением коэффициентов собственных векторов на квадратный корень соответствующих собственных чисел, что правильно отражает относительную долю дисперсии наблюдений.

Для дальнейшего сравнения анализа главных компонент с другими методами произведем вычисления для корреляционной матрицы, представленной в табл. 1. Мы используем модельные данные с целью выявления характеристик без статистических флуктуаций. В табл. 2 сведены результаты анализа главных компонент. Следует выделить три момента: 1) имеется шесть компонент (последние четыре являются второстепенными и в таблице не представлены); 2) первые две компоненты объясняют бо́льшую долю дисперсии, чем первые два общих фактора (61,6 и 41% соответственно); 3) первые две компоненты в отличие от первых двух факторов не объясняют наблюдаемые корреляции. Например, $(b_{11}b_{21}) + (b_{12}b_{22}) = (0{,}747 \cdot 0{,}706) + (-0{,}395) \times (-0{,}409) = 0{,}6890$, что значительно больше, чем скрытая корреляция, равная 0,56.

Сходство анализа главных компонент и факторного анализа заключается в том, что в обоих методах происходит сокращение данных. Зная величину собственных чисел, исследователь

может принять, например, решение использовать только две первые компоненты. Но снова отметим, что эти компоненты не объясняют корреляции. Существует еще одно сходство двух методов — они применяются при исследовании взаимной зависимости переменных. Заметим, что в случае некоррелированных переменных главных компонент не существует, так как все они равноправны: каждой соответствует одинаковая доля дисперсии. Если же корреляция между переменными увеличивается, то доля, объясняемая несколькими первыми компонентами, возрастает.

Одним из отличий между двумя рассматриваемыми методами является следующее. Факторный анализ представляет ковариационную структуру в терминах гипотетической модели, в то время как анализ главных компонент сокращает данные посредством использования нескольких линейных комбинаций наблюдаемых переменных. Выбор метода определяется целью исследования. Объяснение корреляций в терминах небольшого числа факторов возможно лишь при введении гипотетической модели. Если же иметь дело с линейными комбинациями переменных, то обращаться к какой-либо модели нет необходимости, при этом латентная факторная структура остается «вещью в себе».

Таким образом, анализ главных компонент ориентирован на несколько другие задачи по сравнению с факторным. Тем не менее стоит повторить, почему мы уделили ему особое внимание. Во-первых, анализ главных компонент часто рассматривается как один из методов факторного анализа. Во-вторых, при описании метода главных факторов используются аналогичные понятия и вычислительные процедуры (нахождение собственных значений и векторов). Более того, знание анализа главных компонент помогает понять методы факторного анализа. В-третьих, и это самое важное, некоторая статистика, встречающаяся в анализе главных компонент, очень часто применяется на практике для определения числа факторов. (Речь идет о критерии «собственных чисел, бóльших единицы», на котором мы остановимся ниже.)

МЕТОДЫ ФАКТОРНОГО АНАЛИЗА

Наиболее ранним методом факторного анализа является *метод главных факторов,* в котором методика анализа главных компонент используется применительно к редуцированной корреляционной матрице. На главной диагонали последней располагают общности, для оценивания которых обычно пользуются квадратом множественного коэффициента корреляции между соответствующей переменной и совокупностью остальных переменных. Также может применяться наибольший по абсолютной величине коэффициент корреляции в соответствующей переменной строке корреляционной матрицы.

После размещения оценок общностей на главной диагонали корреляционной матрицы выделяются факторы таким же спосо-

бом, что и в анализе главных компонент. Другими словами, факторный анализ проводится исходя из характеристического уравнения, как и в анализе главных компонент (отсюда и название — метод главных факторов). Характеристическое уравнение в этом случае имеет вид

$$\det (R_1 - \lambda I) = 0, \tag{9}$$

где R_1 — редуцированная корреляционная матрица с оценками общностей на главной диагонали. Хотя настоящий подход еще широко распространен, он постепенно уступает место методу наименьших квадратов, к изложению которого мы и приступаем.

Метод наименьших квадратов

Метод наименьших квадратов в факторном анализе сводится к минимизации остаточной корреляции после выделения определенного числа факторов и к оцениванию степени соответствия вычисленных и наблюдаемых коэффициентов корреляции (берется сумма квадратов отклонений). Если взять количество факторов, равное числу переменных, то вычисленные и наблюдаемые коэффициенты корреляции совпадут. Кроме того, расхождение между ними уменьшается при увеличении числа предполагаемых факторов. Поэтому, используя метод наименьших квадратов, мы будем считать, что число факторов меньше числа переменных.

В общих чертах алгоритм состоит в следующем. На первом шаге предполагается, что число факторов есть некоторое k. (Можно начать с однофакторной гипотезы, а затем, увеличивая число факторов, получить приемлемое решение.) На втором шаге производится оценка общностей. (Применяется квадрат множественного коэффициента корреляции между данной переменной и остальными.) На третьем шаге выделяются k факторов, для которых вычисленные коэффициенты корреляции наилучшим образом приближают наблюдаемые корреляции (в смысле минимума суммы квадратов отклонений). На этом этапе решается уравнение, аналогичное (9). На четвертом шаге снова производится оценка общностей, причем используется матрица факторного отображения, полученная на предыдущем этапе. Процесс повторяется до тех пор, пока дальнейшее улучшение станет невозможным. Описанный алгоритм известен под названием: «Метод главных факторов с итерациями по общностям».

Метод минимальных остатков (Harman, 1976) также является итерационной процедурой, основанной на том же принципе, что и метод главных факторов, причем с вычислительной точки зрения данный подход более эффективен. Для метода минимальных остатков при большом объеме выборки применим критерий хи-квадрат. Харман утверждает, что этот приближенный критерий независим от метода выделения факторов и может использоваться не только в алгоритме минимальных остатков. Критерий хи-

Таблица 3

Метод главных факторов с итерациями по общностям (исследование политических взглядов)

Переменная	F_1	F_2	h^2
X_1	0,731	—0,320	0,637
X_2	0,642	—0,252	0,492
X_3	0,550	—0,241	0,360
X_4	0,513	0,473	0,487
X_5	0,441	0,409	0,362
X_6	0,367	0,340	0,251
Собственные значения	1,842	0,746	
Объясняемая доля дисперсии	30,7	12,4	

квадрат может быть применен для проверки окончания работы алгоритма (Harman, 1975; McDonald, 1975). Хотя этот критерий применяется для больших выборок, «ирония» заключается в том, что именно когда объем выборки велик, даже незначительная по величине сумма квадратов отклонений может быть статистически значима. Поэтому Харман предлагает рассматривать число факторов, получаемых с помощью критерия хи-квадрат, лишь как оценку сверху и выделять существенные, теоретически интерпретируемые факторы после анализа результатов вращения.

В табл. 3 представлены результаты вычислений по итерационному методу главных факторов для исходных данных, взятых из табл. 1.

Метод максимального правдоподобия

Метод максимального правдоподобия преследует ту же цель, что и метод наименьших квадратов — найти факторное решение, которое наилучшим образом объясняет наблюдаемые корреляции. Алгоритм можно представить следующим образом. Допустим, что наблюдаемые данные — это выборка из генеральной совокупности, которая точно соответствует k-факторной модели. Совместное распределение переменных (включая факторы) предполагается многомерным нормальным. Неизвестными являются значения нагрузок для каждой переменной. Задача сводится к оцениванию значений латентных переменных (нагрузок) генеральной совокупности, при которых в заданных предположениях функция правдоподобия для распределения элементов корреляционной матрицы максимальна. Несколько иной критерий заключается в нахождении факторных нагрузок, при которых общие факторы и наблюдаемые переменные находятся в канонической корреляции, т. е. коэффициент корреляции между ними максимален. Третий критерий, основанный на тех же принципах, сводится к определению факторных нагрузок, при которых детерминаит матрицы остаточных корреляций максимален. Все эти критерии достаточно сложны для практического применения, но существуют различные итерационные схемы для получения на их основе решений, существенно отличающихся друг от друга с точки зрения

вычислительной эффективности. В настоящее время метод, предложенный Йореско (Jöreskog, 1967), считается одним из лучших.

В принципе все варианты метода максимального правдоподобия сводятся к решению характеристического уравнения, которое может быть представлено в виде

$$\det\ (R_2-\lambda I)=0, \tag{10}$$

где R_2 определяется соотношением

$$R_2=U^{-1}(R-U^2)\,U^{-1}= \tag{11}$$

$$=U^{-1}R_1U^{-1} \tag{12}$$

причем U^2 — оценка дисперсии характерных параметров. Разница между уравнениями (4) и (10) в том, что в последнем используется редуцированная корреляционная матрица R_2 вместо корреляционной матрицы R. В отличие от метода наименьших квадратов в вычисляемую на каждом шаге оценку общностей с бо́льшим весом входят корреляции с переменными, имеющими меньшую характерность. Заметим, что выражение $(R—U^2)$ в (11) то же самое, что R_1 в (9), т. е. вся разница только в весовых множителях. В методе максимального правдоподобия характерность играет роль дисперсии «квази-ошибки»: больший вес имеют переменные с максимальной общностью (т. е. с минимальной

Таблица 4

Двухфакторное решение методом максимального правдоподобия для наддиагональных элементов табл. 1

Переменная	До вращения		Общность	После вращения критерия прямым методом облимин	
	F_1	F_2		F_1	F_2
X_1	0,747	—0,300	0,648	0,817	—0,027
X_2	0,701	—0,266	0,562	0,754	—0,009
X_3	0,599	—0,176	0,389	0,602	0,046
X_4	0,428	0,362	0,314	0,027	0,547
X_5	0,505	0,605	0,621	—0,113	0,833
X_6	0,534	0,248	0,367	0,202	0,468
Сумма квадратов[1]	2,132	0,749		1,652	1,215
Критерий χ^2 с четырьмя степенями свободы	0,825				

[1] В решенни, полученном до косоугольного вращения, суммы квадратов — это собственные числа, которые после деления на число переменных m дают долю дисперсии, объясняемую соответствующими факторами. В решении, получаемом после вращения, суммы квадратов можно представить как «непосредственный» вклад каждого фактора. Общий вклад (включая корреляции между факторами) в решении до вращения по-прежнему равен сумме собственных величин.

характерностью). Это соответствует основному принципу статистического оценивания, по которому менее точные наблюдения учитываются с меньшим весом.

Мы упомянули о том, что с помощью оптимальных алгоритмов можно точно оценить переменные генеральной совокупности для модельных данных в отсутствии ошибок. Хорошие программные реализации таких алгоритмов позволяют практически использовать эти потенциальные возможности.

В табл. 4 представлены результаты применения метода максимального правдоподобия к выборочным корреляциям, являющимся наддиагональными элементами табл. 1. Как мы и ожидали, гипотеза адекватности для полученного решения подтверждается.

Формула для вычисления статистики χ^2 показывает, что ее значение определяется объемом выборки, в то время как число степеней свободы от выборки не зависит:

$$V_k = N\{\ln|C| - \ln|R| + \mathrm{tr}(RC^{-1})\} - n, \qquad (13)$$

где ln — натуральный логарифм; tr — след матрицы; N — объем выборки; n — число переменных, R — матрица ковариаций; $C = FF' + U^2$; F — матрица факторных нагрузок; U^2 — характерности. (Это же соотношение используется для проверки адекватности решения методом наименьших квадратов, отличие только в оценках F и U.) Важно отметить, что при фиксированной корреляционной матрице, величина U_k пропорциональна объему выборки N. Соответствующее число степеней свободы равно:

$$df_k = 1/2[(n-k)^2 - (n+k)], \qquad (14)$$

где k — число гипотетических факторов, а n — число переменных. Как видно, df_k не зависит от объема выборки N.

Существенное преимущество метода максимального правдоподобия состоит в том, что для большой выборки он позволяет получить критерий значимости. Если критерий χ^2 показывает значимое отклонение наблюдений от k-факторной модели, то в рассмотрение вводится модель с $(k+1)$ факторами.

В разведочном анализе вычисления, как правило, начинают с одного фактора, а заканчивают, когда отклонение наблюдений от модели становится статистически незначимо. Хотя эти последовательные проверки гипотез находятся в зависимости друг от друга, на практике это несущественно (Lawley, Maxwell, 1971).

Если при оценивании числа факторов положиться на один только критерий значимости, то возникает опасность получить факторов больше, чем нужно. Там, где модель всего лишь приближена к реальности, неизбежные невязки обусловливают появление дополнительных значимых факторов. В разд. IV мы вернемся к вопросам, связанным с определением числа факторов.

Альфа-факторный анализ

Предполагается, что и в методе наименьших квадратов и в методе максимального правдоподобия существует генеральная

совокупность объектов*, на которую распространяются результаты статистического анализа выборки. В альфа-факторном анализе используемые переменные считаются выборкой из некоторой совокупности переменных, о которой можно судить на основании наблюдаемой *совокупности объектов*. Таким образом, в альфа-факторном анализе выводы носят психометрический, а не статистический характер.

Кайзер и Кэффри (Kaiser, Caffrew, 1965) утверждают, что этот метод основан на выделении таких факторов, которые имеют максимальные корреляции с соответствующими факторами генеральной совокупности переменных. С другой стороны, характерные факторы при данном подходе можно рассматривать как ошибки, обусловленные психометрической выборкой переменных. Следовательно, оценки общностей в этом контексте имеют смысл «надежностей». На первом шаге образуется «подправленная» корреляционная матрица вида

$$R_3 = H^{-1}(R - V^2)H^{-1} \quad , \tag{15}$$

где V^2 и H^2 — диагональные матрицы характерностей и общностей соответственно. (H^{-1} — диагональная матрица, элементами которой являются обратные величины к квадратным корням из общностей.) Тогда характеристическое уравнение, связанное с этой «подправленной» матрицей, представляется следующим образом:

$$\det(R_3 - \lambda I) = 0. \tag{16}$$

Сопоставим уравнения (16) и (10), а также уравнения (15) и (11). В методе максимального правдоподобия матрица нормируется с помощью характерностей, а в альфа-факторном анализе — дисперсий общностей. Другими словами, в первом случае больший вес имеют переменные с большей общностью, а во втором, — наоборот, с меньшей. Как правило, в обоих случаях решение осложняется тем, что значения общностей пересчитываются в процессе итераций.

В альфа-факторном анализе число выделяемых факторов определяется с помощью критерия, заключающегося в том, что соответствующие собственные величины должны быть больше 1. Этот критерий эквивалентен критерию выделения факторов с помощью коэффициента обобщенности α (квадрат коэффициента корреляции данного фактора с соответствующими факторами, взятыми из генеральной совокупности. — *Примеч. пер.*). Выделяются факторы, для которых коэффициент α положителен. Разумеется, при этом подходе не используются обычные критерии значимости, так как совокупность объектов предполагается известной.

---

* Альфа-факторный анализ был разработан для упорядочения данных в области психологии. В частности, объектом исследования в нем являются индивидуумы. *Примеч. ред.*

Результаты применения альфа-факторного анализа к матрице коэффициентов корреляции, представленных наддиагональными элементами табл. 1, сведены в табл. 5. Здесь же даются результаты анализа образов, к обсуждению которого мы приступаем.

АНАЛИЗ ОБРАЗОВ

В анализе образов определение общей и характерной части переменной отличается от принятого в классическом факторном анализе. Под общей частью переменной подразумевается та ее составляющая, которая выражается через линейную комбинацию других переменных. Эта доля переменной называется *«образ-переменной»*. Вторая составляющая переменной, независимая от остальных, называется *«антиобразом»*. Причем считается, что мы имеем дело с генеральными совокупностями переменных и объектов; все вопросы, связанные с выборкой, не рассматриваются.

В анализе образов предполагается, что потенциальное множество переменных бесконечно. Для сравнения обратимся к двухфакторной модели на рис. 1. Шесть переменных, рассматриваемых там, образуют некоторую совокупность. Но в анализе образов эти переменные считаются выбранными из бесконечного множества переменных, удовлетворяющих двухфакторной модели.

Если бы у нас была возможность наблюдать все переменные этого пространства, средний квадрат образа был бы равен *общности* переменной, определяемой в факторном анализе, а средний квадрат антиобраза — характерности. (Подразумевается, что мы имеем дело с нормированными переменными.) Другими словами, квадрат множественного коэффициента корреляции между одной переменной и остальными переменными совокупности равен общности данной переменной.

Образы и антиобразы, определяемые для некоторого набора наблюдаемых переменных, называются соответственно частными образами и частными антиобразами. Хотя частные образы являются только приближением к полным образам, они (частные образы) полностью задаются наблюдаемыми переменными. В этом смысле анализ образов в корне отличается от классического факторного анализа, в котором общая часть переменной является линейной комбинацией гипотетических факторов и не может быть явной функцией наблюдаемых переменных.

Методика анализа образов предполагает введение матрицы ковариаций частных образов:

$$R_4 = (R - S^2)R^{-1}(R - S^2), \qquad (17)$$

где R — корреляционная матрица, а S^2 — диагональная матрица, элементами которой являются доли дисперсии каждой переменной, не объясняемые другими параметрами (т. е. доли дисперсии антиобразов). Получение матрицы (17) сводится, во-первых, к замене диагональных элементов матрицы R на квадраты мно-

жественных коэффициентов корреляции каждой переменной с совокупностью всех остальных переменных, и, во-вторых, к преобразованию недиагональных элементов для получения матрицы Грама. Характеристическое уравнение для этой матрицы имеет вид

$$\det(R_4 - \lambda I) = 0. \qquad (18)$$

Число выделяемых факторов определяется количеством собственных чисел, больших 1, но не для матрицы R_4, а для матрицы $S^{-1}RS^{-1}$. Обычно число выделяемых таким методом факторов велико — приблизительно половина числа исходных параметров. Кайзер предлагает после соответствующих вращений отбрасывать незначимые и неинтерпретируемые факторы. В табл. 5 даны сравнительные результаты применения анализа образов и альфа-факторного анализа.

Таблица 5

Факторные нагрузки, вычисленные с помощью альфа-факторного анализа и анализа образов, для модельной корреляционной матрицы, приведенной в табл. 1[1]

Переменная	Матрица факторного отображения до вращения					
	Альфа-факторный анализ			Анализ образов		
	F_1	F_2	общность	F_1	F_2	общность
X_1	0,669	0,437	0,633	0,575	0,133	0,348
X_2	0,586	0,334	0,490	0,538	0,139	0,309
X_3	0,502	0,329	0,361	0,477	0,131	0,245
X_4	0,585	—0,352	0,489	0,372	—0,270	0,211
X_5	0,502	—0,329	0,360	0,335	—0,263	0,182
X_6	0,419	—0,274	0,251	0,287	—0,239	0,140

[1] Общности, полученные с помощью альфа-факторного анализа, весьма близки к истинным общностям; анализ образов дает менее точные оценки.

III. МЕТОДЫ ВРАЩЕНИЯ

Как уже отмечалось, на первом этапе анализа определяется минимальное *число* факторов, адекватно воспроизводящих наблюдаемые корреляции, а также значения общностей каждой переменной. Следующий шаг состоит в нахождении легко интерпретируемых факторов с помощью процедуры вращения. При этом число факторов и значения общностей переменных фиксируются.

Применение методов, рассмотренных в предыдущем разделе, приводит к набору ортогональных факторов, упорядоченных в порядке убывания их значимости. Эти два ограничения являются в некотором смысле, искусственными. Они принимаются, чтобы обеспечить единственность решения. В результате этих ограничений,

во-первых, факторная сложность переменных, скорее всего, будет больше единицы, независимо от вида истинной факторной структуры, т. е. переменные будут иметь нагрузки более чем на один фактор;

во-вторых, все факторы, за исключением первого, являются биполярными, другими словами, некоторые переменные должны иметь положительную нагрузку на этот фактор, а некоторые — отрицательную.

существуют три различных подхода к проблеме вращения. Первый подход — графический*. Вращение заключается в проведении новых осей, которые соответствуют некоторому критерию простой, легко интерпретируемой структуры. Если в пространстве факторов есть явные скопления (кластеры) точек (переменных), легко отделяемые друг от друга, простая структура получается в том случае, когда оси проведены через эти скопления. Но если такое разделение не очевидно или число факторов велико, графический метод неприменим.

Второй подход связан с аналитическими методами. В этом случае выбирается некоторый объективный критерий, которым надо руководствоваться при выполнении вращения. В рамках этого подхода различают два вида вращения — ортогональное и косоугольное. А они в свою очередь имеют многочисленные вариации. В этом разделе мы остановимся на наиболее известных из них.

Третий подход заключается в задании априорной целевой матрицы. Цель вращения — нахождение факторного отображения, наиболее близкого к некоторой заданной матрице. Так как при задании целевой матрицы делаются определенные предположения о факторной структуре, третий подход схож с конфирматорным факторным анализом, в котором проверяются гипотезы о матрице факторного отображения.

ГЕОМЕТРИЧЕСКИЙ МЕТОД ВРАЩЕНИЯ, ПРОСТАЯ СТРУКТУРА И ВТОРИЧНЫЕ ОСИ

Геометрический метод вращения практически неприменим, когда скопления точек трудно разделимы или когда число факторов больше двух. Мы рассматриваем этот подход только потому, что он дает возможность лучше понять аналитический способ вращения. Хорошим введением в геометрический метод служит работа Мьюлейка (Mulaik, 1972).

Целью всех вращений является получение наиболее простой факторной структуры. К сожалению, концепция простоты неоднозначна, и поэтому не существует единых формальных критериев. Наиболее полное определение простой структуры дано

---

* В этом случае исходные переменные рассматриваются как точки в факторном пространстве координаты которых равны нагрузкам на факторы, а размерность определяется числом факторов. — *Примеч. ред.*

Тэрстоуном (Thurstone, 1947), но в последнее время уже выяснилось, что не все его критерии формализуются в аналитическом виде. Поскольку Тэрстоун использует понятие гиперплоскости или подпространства, мы остановимся на более простом подходе Мьюлейка (Mulaik, 1972), предполагающем знание лишь элементов теории векторных пространств. (В определении Мьюлейка через r обозначено число общих факторов, а V — матрица вторичной структуры, образованная координатами (нагрузками) вторичных факторов, получаемых в результате вращения.)

1) В каждой строке матрицы вторичной структуры V должен быть хотя бы один нулевой элемент. Это предположение является основным в определении простой структуры.

2) Для каждого столбца k матрицы вторичной структуры V должно существовать подмножество из r линейно-независимых наблюдаемых переменных, корреляции которых с k-м вторичным фактором — нулевые. Данный критерий сводится к тому, что каждый столбец матрицы должен содержать не менее r нулей.

3) У одного из столбцов каждой пары столбцов матрицы V должно быть несколько нулевых коэффициентов (нагрузок) в тех позициях, где для другого столбца они ненулевые. Это предположение гарантирует различимость вторичных осей и соответствующих им подпространств размерности r—1 в пространстве общих факторов.

4) При числе общих факторов больше четырех в каждой паре столбцов должно быть некоторое количество нулевых нагрузок в одних и тех же строках. Данное предположение дает возможность разделить наблюдаемые переменные на отдельные скопления.

5) Для каждой пары столбцов матрицы V должно быть как можно меньше значительных по величине нагрузок, соответствующих одним и тем же строкам. Это требование обеспечивает минимизацию сложности переменных.

Сформулированные критерии основаны на двух соображениях: а) необходимо определить признаки простой структуры и б) необходимо выяснить условия, при которых простая структура выделяется однозначно и объективно. В специальных работах по факторному анализу при обсуждении этого понятия преобладает второе соображение. Мы же оставим этот вопрос специалистам, и сосредоточим наше внимание на первом.

Хотя трудно определить минимальные требования к простой структуре, но если взять число факторов r и число переменных n, то всегда можно сказать, какая структура наиболее простая. Факторная структура является наипростейшей, когда все переменные имеют факторную сложность, равную 1, т. е. когда каждая переменная имеет ненулевую нагрузку только на один общий фактор. Если число факторов два и больше, то это означает, что в наиболее простой матрице факторной структуры, во-первых, каждая строка будет содержать только один ненулевой элемент, во-вторых, каждый столбец будет иметь несколько нулей и, в-третьих, для каждой пары столбцов нулевые элементы не совпадают.

Для реальных данных такая простая структура недостижима. Следовательно, задача состоит в том, чтобы «определить» факторную структуру, которая является самой «близкой» к простой структуре. Здесь специалисты расходятся в определении «простоты» для таких «несовершенных» структур, а также в вычислительных методах решения задачи. Как уже отмечалось, критерий Тэрстоуна дает эмпирические условия, при которых простая структура определяется однозначно. Одно из них состоит в следующем: для каждого фактора должны существовать по крайней мере три переменные, имеющие на этот фактор значительную нагрузку. Но определение простой структуры никак не зависит от этого эмпирического ограничения, принимаемого при анализе реальных данных. В разведочном факторном анализе исследователь вынужден довольствоваться теми переменными, которыми он располагает, и прежде чем начать интерпретировать факторы, заранее определить, что он понимает под «простой» структурой.

Первоначально простую факторную структуру определяли в терминах *вторичных осей*. Хотя это понятие не является абсолютно необходимым (так как есть методы косоугольных вращений, где вторичные оси не вводятся), мы остановимся на нем, поскольку в некоторых компьютерных программах для косоугольных вращений введение этих осей предполагается.

Заметим, что первичные факторные нагрузки — это не что иное, как проекции переменных на две оси (в случае двухфакторной модели), т. е. нагрузки определяются при опускании перпендикуляров из данной точки на первичные ортогональные оси. Простая структура получается в том случае, когда все значения переменных лежат на этих осях. В ортогональном случае простая структура задается множеством точек, имеющим ненулевые нагрузки (нулевые проекции) только на один фактор (на одну ось). Проекция будет ненулевой, если угол между скоплениями точек отличен от прямого угла. При этом следует провести вторичные оси перпендикулярно гиперплоскостям, проходящим через эти скопления, которые сами могут рассматриваться как первичные факторы (для двухфакторной модели гиперплоскость есть прямая; рис. 4).

Таким образом, можно предположить, что скопления точек находятся на первичных осях, или же что проекции точек на вторичных осях — нулевые. В нашем примере переменные X_1, X_2, X_3 имеют нулевые проекции на вторичной оси R_2, а переменные X_4, X_5, X_6 — нулевые проекции на оси R_1. Однако не ясно, зачем проводить вторичные оси, вместо того чтобы провести первичные оси прямо через скопления точек. Следует отметить, что метод, основанный на идентификации вторичных осей, при котором они рассматриваются как ортогональные, позволяет более точно определить первичные оси, если число факторов больше двух, а скопления точек не столь явные, как в модельных данных. Из всего сказанного можно сделать вывод, что основная

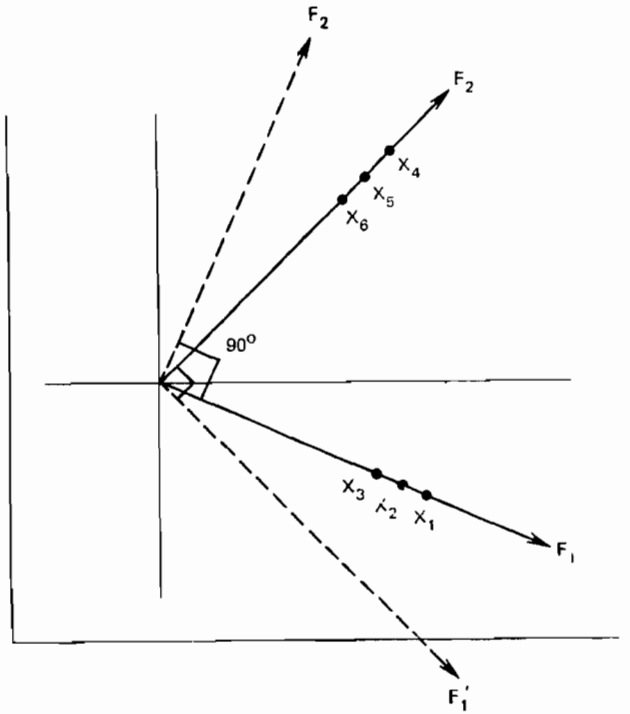

Рис. 4. F_1 и F_2 — первичные косоугольные факторы; F_1' и F_2' — соответствующие вторичные оси. Проекции X_1, X_2 и X_3 будут нулевыми на оси F_2', а проекции X_4, X_5 и X_6 — на оси F_1'

цель вращения заключается в нахождении матрицы факторного отображения, наиболее близкой к простейшей идеальной структуре, описанной выше.

МЕТОДЫ ОРТОГОНАЛЬНОГО ВРАЩЕНИЯ: КВАРТИМАКС, ВАРИМАКС И ЭКВИМАКС

Мы остановимся только на основных принципах каждого метода, так как предполагается, что читатель будет использовать какую-то готовую компьютерную программу. В предыдущем разделе описана простейшая структура при заданном числе общих факторов k и числе переменных n. Полезно еще раз повторить некоторые свойства такой матрицы.

Поскольку каждая переменная имеет нагрузку только на один фактор, интерпретация *переменных* не представляет труда. Но для численного использования эта характеристика степени сложности неудобна. Одной из возможных мер сложности модели является вариация квадрата факторной нагрузки для каждой стро-

ки (для каждой переменной). Мы рассматриваем квадрат нагрузок только для того, чтобы избежать осложнений, связанных с учетом знака. Известно, что дисперсия определяется как математическое ожидание квадрата отклонений от среднего, поэтому при фиксированном числе факторов и заданных общностях дисперсия максимальна, если одно из значений квадратов нагрузок равно общности, а все остальные элементы в строке нулевые. Иначе говоря, дисперсия квадратов факторных нагрузок переменной есть мера факторной сложности этой переменной:

$$\text{Факторная сложность переменной} = \frac{1}{r} \sum_{j=1}^{r} (b_{ij}^2 - \bar{b}_{ij}^2)^2, \tag{19}$$

где r — число столбцов факторной матрицы; b_{ij} — факторная нагрузка j-го фактора на i-ю переменную; $\bar{b}_{ij}$ — среднее значение квадратов факторных нагрузок в i-й строке. Соотношение (19) может быть представлено в следующем виде:

$$q_i = \frac{\sum_{j=1}^{r} (b_{ij}^4) - (\sum_{j=1}^{r} b_{ij}^2)^2}{r^2}. \tag{20}$$

Число факторов r и общности каждой переменной считаются известными в результате решения задачи выделения первоначальных факторов. Поэтому слагаемое, входящее в (20) с отрицательным знаком, является константой, ибо

$$\sum_{j=1}^{r} b_{ij}^2 = h_i^2$$

в случае ортогонального решения. Общей мерой сложности может служить сумма q_i всех переменных

$$q = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} \frac{\sum_{j=1}^{r} (b_{ij})^4 - (\sum_{j=1}^{r} b_{ij}^2)^2}{r^2}. \tag{21}$$

Использование критерия *квартимакс* основано на вращении осей таким образом, чтобы результирующие факторные нагрузки максимизировали q. При этом максимизация q эквивалентна максимизации следующего выражения:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{r} b_{ij}, \tag{22}$$

так как слагаемое со знаком минус в (21) является константой. Отсюда и название — квартимакс.

На практике, применяя этот критерий, можно достичь простоту интерпретации переменных за счет простоты интерпретации

факторов. В частности, описание переменной упрощается при уменьшении числа общих факторов, связанных с ней. В то же время описание фактора становится проще, если относительно небольшое число переменных имеют существенные нагрузки на этот фактор, а остальные переменные — нулевые нагрузки. В общем, метод квартимакс имеет тенденцию к выделению генерального фактора.

Метод *варимакс* использует несколько другой критерий, в котором добиваются упрощения описания столбцов факторной матрицы. Вместо дисперсии квадратов нагрузок переменной рассматривается дисперсия квадратов нагрузок фактора. Индекс сложности v_j фактора j равен:

$$v_j = \frac{n \sum_{i=1}^{n} b_{ij}^4 - \left(\sum_{i=1}^{n} b_{ij}^2 \right)^2}{n^2}. \qquad (23)$$

Заметим при этом, что выражение

$$\sum_{i=1}^{n} b_{ij}^2.$$

где суммирование происходит по номеру параметра i, не является константой. Общая мера простоты задается критерием

$$V = \sum_{j=1}^{r} v_j = \frac{\sum_{j=1}^{r} n \sum_{i=1}^{n} b_{ij}^4 - \sum_{j=1}^{r} \left(\sum_{i=1}^{n} b_{ij}^2 \right)^2}{n^2}, \qquad (24)$$

известным под названием критерия варимакс. Обычно нормированные факторные нагрузки применяют, чтобы избавиться от нежелательного влияния на результат вращения переменных с большой общностью, т. е. в выражении (24) квадраты нагрузок b_{ij}^2 заменяются на b_{ij}^2/h_i^2, а четвертые степени b_{ij}^4 — на b_{ij}^4/h_i^4.

В табл. 6 представлены результаты применения методов квартимакс и варимакс (с нормированием) к одним и тем же данным. Отметим, что, хотя, алгоритмически метод квартимакс проще, чем варимакс, последний дает лучшее разделение факторов. Эксперименты, проведенные Кайзером (Kaiser, 1958), пока-

Таблица 6

Результаты вращений по методам варимакс и квартимакс, применяемых к факториой матрице в табл. 4[1]

Переменная	Метод вращения варимакс		Метод вращения квартимакс	
	F_1	F_2	F_1	F_2
X_1	0,787	0,167	0,793	0,133
X_2	0,730	0,170	0,736	0,143
X_3	0,595	0,187	0,602	0,166
X_4	0,154	0,539	0,173	0,533
X_5	0,083	0,783	0,111	0,780
X_6	0,306	0,503	0,324	0,492

[1] В этом примере тенденция выделения генерального фактора методом квартимакс выражена слабо.

зывают, что факторная матрица, получаемая с помощью метода вращения варимакс, в большей степени инвариантна по отношению к выбору различных множеств переменных.

Учитывая, что критерий квартимакс основан на упрощении описания строк, а критерий варимакс — на упрощении описания столбцов, можно предложить некоторый совместный критерий, введя соответствующие веса. Обобщенный критерий имеет вид

$$\alpha Q+\beta V=\text{Maximum}, \qquad (25)$$

где Q — задается соотношением (22), а V — соотношением (24), умноженным на n для удобства представления и с учетом того, что умножение на константу не влияет на процесс нахождения максимума; α и β — веса. Полученный критерий запишем в форме:

$$\sum_{j=1}^{r}\sum_{i=1}^{n} b_{ij}^{4}-\gamma \sum_{j=1}^{r}\Big(\sum_{i=1}^{n} b_{ij}^{2}\Big)^2/n=\text{Maximum}, \qquad (26)$$

где $\gamma=\beta/(\alpha+\beta)$.

Если $\gamma=0$, то образуется критерий квартимакс, а если $\gamma=1$, то — варимакс. При $\gamma=r/2$ и $\gamma=0{,}5$ получаем особые критерии, названные *эквимакс* и *биквартимакс* соответственно.

МЕТОДЫ КОСОУГОЛЬНОГО ВРАЩЕНИЯ

Косоугольное вращение является более общим, чем ортогональное, так как здесь нет ограничений, связанных с некоррелированностью факторов. Преимущество косоугольного вращения состоит в следующем: когда в результате его выполнения получаются ортогональные факторы, можно быть уверенным, что эта ортогональность действительно им свойственна, а не привнесена методом вращения. Поскольку косоугольные вращения производятся с учетом корреляций между факторами, существуют многочисленные методы интерпретации результатов факторного анализа. Так, для объяснения корреляции между факторами в ряде случаев вводят факторы второго и более высокого порядков. Кроме того, существуют два подхода к косоугольному вращению — использование вторичных осей и первичной матрицы факторного отображения. Основные принципы получения простой структуры уже обсуждались, поэтому описание методов будет кратким.

Методы, основанные на введении вторичных осей

Обсуждаемые здесь методы основаны на том, что если существуют разделимые скопления точек, определяемые первичными факторами, то они будут иметь почти нулевые проекции на все вторичные оси, за исключением одной. Таким образом, можно определить критерий, называемый *квартимин,* который аналогичен квартимаксу:

$$N = \sum_{i=1}^{n} \sum_{j<k=1}^{r} a_{ij} a_{ik}, \qquad (27)$$

где a_{ij} и a_{ik} — проекции i-го параметра на j-ю и k-ю вторичные оси. Величина N будет нулевой, если все параметры имеют нагрузку только на один фактор. Цель вращения — нахождение таких факторных нагрузок, которые минимизируют N. Для ортогональных вращений этот критерий эквивалентен квартимаксу.

По аналогии с ортогональным критерием варимакс вводится критерий *ковариминн.* В этом случае минимизируется ковариация квадратов проекций на вторичные оси

$$C = \sum_{j<k=1}^{r} \left(n \sum_{i=1}^{n} a_{ij}^2 a_{ik}^2 - \sum_{i=1}^{n} a_{ij}^2 \sum_{i=1}^{n} a_{ik}^2 \right). \qquad (28)$$

Модификация этого критерия основана на нормировании — замене a_{ij}^2 на a_{ij}^2/h_i^2. Применительно к одним и тем же данным критерий ковариминн, как правило, дает меньше косоугольных факторов, чем квартимин. Объединение этих двух критериев приводит к обобщенному критерию

$$B = \alpha N + \beta C/n = \text{minimum}, \qquad (29)$$

где α и β — веса, назначаемые для N и для C соответственно. После умножения соотношения (29) на n и группировки членов, получаем общий критерий *облимин*

$$B = \sum_{j<k=1}^{r} \left(n \sum_{i=1}^{n} a_{ij}^2 a_{ik}^2 - \gamma \sum_{i=1}^{n} a_{ij}^2 \sum_{i=1}^{n} a_{ik}^2 \right), \qquad (30)$$

где $\gamma = \beta/(\alpha+\beta)$.

Этот общий критерий при $\gamma = 0$ переходит в квартимин (наибольшая косоугольность), при $\gamma = 0{,}5$ — в биквартимин, а при $\gamma = 1$ — в ковариминн (наименьшая косоугольность). Еще раз отметим, что, как правило, применяется критерий облимин в нормированной форме, т. е. когда a_{ij}^2 заменяется на a_{ij}^2/h_i^2.

Другой критерий, тесно связанный с принципами облимина, но используемый в совершенно другом вычислительном алгоритме, называется критерием бинормамин. В нем заложена идея объективного выбора значения γ в соотношении (30). По сравнению с критерием биквартимин, в котором $\gamma = 1/2$, бинормамин дает лучшие результаты для особо простых или особо сложных данных.

Прямой метод облимин

Дженрих и Сэмпсон (Jenrich, Sampson, 1966) предложили критерий, основанный на упрощении матрицы нагрузок первичных факторов (без использования вторичных осей). Этот критерий допускает эффективную программную реализацию. Минимизируемая функция имеет вид, аналогичный (30). Отличие только

в том, что используются нагрузки первичных факторов, а не нагрузки вторичной структуры. Критерий имеет вид

$$D=\sum_{j<k=1}^{r}\left[\sum_{i=1}^{n} b_{ij}^2 b_{ik}^2 - d\left(\sum_{i=1}^{n} b_{ij}^2 \sum_{i=1}^{n} b_{ik}^2\right)/n\right], \qquad (31)$$

где b_{ij} — элементы матрицы нагрузок первичных факторов. Заметим, что в соотношении (31) по сравнению с (30), член с отрицательным знаком дан с сомножителем $1/n$. Как и в традиционном критерии облимин, выбор параметра α регулирует «степень» косоугольности получаемого решения.

Бóльшие значения α соответствуют «наиболее» косоугольным решениям, а мéньшие отрицательные значения — «наиболее» ортогональным решениям. В наиболее простом случае однофакторной модели следует положить $\alpha=0$. Необходимо сделать предостережение о том, что выбор α в прямом критерии облимин отличается от выбора γ в (30). Подробно этот аспект рассматривался Харманом (Harman, 1975).

Другие методы косоугольного вращения

Существует много других методов косоугольного вращения. Мы упомянем некоторые наиболее известные.

Критерий *облимакс* (Saunders, 1953) основан на упрощении факторной структуры по принципу увеличения числа значительных и пренебрежимо малых нагрузок за счет остальных коэффициентов структуры. Этот критерий эквивалентен критерию квартимакс в случае ортогонального вращения, но приводит к решению, отличному от метода квартимин при использовании его без ограничения, связанного с ортогональностью.

Следует отметить еще два метода вращения. Это — метод *ортоблик* (orthoblique), предложенный Гаррисом и Кайзером (Harris, Kaiser, 1964), и метод *максплейн* (maxplane), рассмотренный Каттеллем и Мерлем и позднее Эбеном (Cattell, Muerle, 1960; Eben, 1966). Последний подход принципиально отличается ото всех упомянутых ранее.

ВРАЩЕНИЕ С ИСПОЛЬЗОВАНИЕМ ЦЕЛЕВОЙ МАТРИЦЫ

Еще один подход к вращению основывается на априорной информации о факторной структуре.

Во-первых, можно задать значения нагрузок для каждой переменной, а затем проводить вращения с целью обеспечения минимального отличия полученной матрицы факторной структуры от заданной матрицы (в смысле критерия наименьших квадратов). При этом можно налагать дополнительные ограничения типа ортогональности. Этот вид вращения обычно применяется для анализа соответствия двух факторных структур.

Во-первых, в качестве целевой матрицы можно использовать

некоторые функции от ортогонального решения. Этот подход, известный под названием *промакс*-метода косоугольных вращений (Hendrickson, White, 1964), основан на том, что ортогональные вращения, как правило, близки к косоугольным. Сводя некоторые меньшие нагрузки к почти нулевым, можно получить пригодную для дальнейшего анализа целевую матрицу. Затем находятся косоугольные факторы, для которых расхождение вычислительной матрицы факторной структуры с целевой — минимально. В рамках данного метода существуют различные алгоритмы, основанные на целевой матрице факторной структуры, но мы не будем их описывать.

Таблица 7

Целевая матрица, состоящая из нулей и единиц

Переменная	Фактор	
	1	2
X_1	1	0
X_2	1	0
X_3	1	0
X_4	0	1
X_5	0	1
X_6	0	1

В-третьих, можно задать целевую матрицу, состоящую из нулей и единиц. Этот подход часто соответствует действительной степени информированности исследователя, когда ему известно только то, что некоторые нагрузки должны быть велики, а другие — малы. В табл. 7 представлен пример такой целевой матрицы.

Можно воспользоваться более общим видом целевой матрицы: некоторые ее элементы полагаются нулевыми, некоторые — равными другим фиксированным величинам, а остальные элементы полагаются произвольными. Более подробно это будет обсуждаться в разделе, посвященном конфирматорному факторному анализу.

IV. ЕЩЕ О ПРОБЛЕМЕ ОПРЕДЕЛЕНИЯ ЧИСЛА ФАКТОРОВ

Хотя мы уже ранее рассмотрели ряд методов нахождения минимального числа факторов, обеспечивающих согласие с наблюдениями, однако существуют причины, чтобы вернуться к этому вопросу. Во-первых, при обсуждении метода выделения первоначальных факторов отмечалось, что число факторов можно оценивать достаточно приблизительно, поэтому мы не будем вдаваться в подробности, относящиеся к данной задаче. Во-вторых, некоторые первоначальные решения не дают достоверной информации о числе факторов, так как требуют последующего проведения вращений. В-третьих, мы можем столкнуться с затруднениями, связаниыми с неполным соответствием между факторной моделью и данными наблюдений. В-четвертых, надо быть готовыми к тому, что в большинстве компьютерных программ требуется предварительная оценка числа факторов.

Существует несколько часто употребляемых критериев определения числа факторов. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать вместе, чтобы один дополнял другой. Наиболее часто применяются: 1) критерии значимости, связанные с методами максимального правдоподобия и наименьших квадратов; 2) различные правила, формулируемые в терминах собственных чисел; 3) критерий, основанный на величине долей дисперсий факторов; 4) критерий отсеивания и 5) критерий интерпретируемости и инвариантности.

КРИТЕРИИ ЗНАЧИМОСТИ

При условии выполнения предложений, необходимых для метода максимального правдоподобия, с чисто статистической точки зрения предпочтительнее пользоваться критерием χ^2. Применение этого критерия показало, что для большой выборки при значительном количестве переменных число выделяемых факторов намного больше числа факторов, которое ожидает получить исследователь. Хотя это обстоятельство не является недостатком метода, в некоторых случаях оно заставляет исследователей после применения критерия статистической значимости использовать дополнительно критерий, основанный на величине доли воспроизводимой дисперсии.

Анализ с помощью метода Монте-Карло критерия максимального правдоподобия показывает, что последний особенно эффективен, когда модель генеральной совокупности известна и отсутствуют второстепенные факторы. Другими словами, данный метод хорошо приспособлен к отклонениям, связанным с выборкой, и гораздо хуже — к изменениям в модели. При достаточно большом объеме выборки любые отклонения в модели будут трактоваться как значимые факторы. Таким образом, после соответствующих вращений второстепенные факторы необходимо удалить с учетом величины долей их дисперсий.

Ранее было описано пошаговое использование критерия χ^2: начиная с однофакторной модели, постепенно увеличивают число факторов, если имеют место статистически значимые отклонения модели от наблюдений. Однако при большом числе параметров данная процедура может быть чрезмерно трудоемкой. Поэтому можно сочетать один из быстрых методов определения числа общих факторов, описанных ниже, с критерием максимального правдоподобия. После того как будет получено начальное число факторов, количество их следует увеличивать, если наблюдения значимо отличаются от модели, либо уменьшать, если эти отличия — незначимы. Со статистической точки зрения решение с помощью метода наименьших квадратов не столь эффективно, как решение с помощью метода максимального правдоподобия, но все сделанные замечания относятся и к нему.

КРИТЕРИИ, ОСНОВАННЫЕ НА СОБСТВЕННЫХ ЧИСЛАХ

При определении числа факторов часто применяют правило, которое позволяет оставлять факторы с собственными числами, большими 1. При этом используется корреляционная (нередуцированная) матрица. Этот простой критерий хорошо себя зарекомендовал, так как обычно дает результаты, совпадающие с теми, что ожидает получить исследователь. Кроме того, этот метод был тщательно проверен на модельных искусственных данных.

Для корреляционной матрицы, относящейся к генеральной совокупности, рассматриваемый критерий всегда дает нижнюю оценку числа общих факторов. Иначе говоря, число общих факторов, соответствующих данной корреляционной матрице, будет больше или равно числу факторов, выделяемых согласно этому критерию. Однако полученное неравенство не обязательно справедливо для выборочной корреляционной матрицы. Хотя Кайзер приводит несколько причин в пользу критерия собственных чисел, бóльших 1, тем не менее он носит эвристический характер. После исследования других, более «утонченных» методов, Кайзер все же отдает предпочтение именно этому критерию (Kaiser, 1974).

Другой метод, основанный на собственных числах, относится к редуцированной корреляционной матрице. Согласно этому критерию сохраняются факторы с собственными числами, большими нуля. Преимущество этого метода в том, что для корреляционной матрицы генеральной совокупности он дает более точные нижние оценки числа общих факторов. Но для выборочной корреляционной матрицы критерий обычно дает значительно большее число факторов.

Данный критерий может применяться, когда общности оцениваются и помещаются на главную диагональ. Как правило, некоторые собственные числа будут отрицательными. При этом не имеет смысла выделять все факторы с собственными числами, большими нуля. Хотя сумма отрицательных и положительных собственных чисел равна сумме всех общностей, (т. е. дисперсии, объясняемой общими факторами), отрицательные величины нельзя интерпретировать как дисперсии. Поэтому их присутствие является причиной «инфляции» суммы положительных собственных чисел в том смысле, что она становится больше суммы общностей. Харман (Harman, 1975) предлагает прекратить выделение общих факторов, когда сумма собственных чисел превысит сумму оценок общностей.

КРИТЕРИЙ, ОСНОВАННЫЙ НА ВЕЛИЧИНЕ ДОЛИ ВОСПРОИЗВОДИМОЙ ДИСПЕРСИИ

Критерии значимости «оперируют» с выборочной изменчивостью данных. Критерии, основанные на собственных числах, формулируются в терминах абстрактных характеристик матрицы.

Возможен третий подход — для каждого фактора оценивается доля дисперсии, воспроизводимая этим фактором. Данный критерий становится особенно наглядным, когда выделение первоначальных факторов производится с помощью нередуцированной корреляционной матрицы. Тогда в качестве статистики этого критерия выступает доля дисперсии, воспроизводимой последним выделяемым фактором по отношению к полной дисперсии, равной числу параметров. Следует напомнить, что рассмотренные выше методы выделения предполагают упорядочение факторов по убыванию их долей дисперсии. Обсуждаемый критерий определяется уровнем (порогом) для минимальной доли воспроизводимой дисперсии. Например, это может быть один, пять или десять процентов. Заметим, что критерий «собственных чисел, бо́льших единицы», эквивалентен данному критерию для $100/n\%$-го уровня.

Во всех упомянутых выше методах, кроме анализа главных компонент, используется редуцированная корреляционная матрица. При этом доля воспроизводимой дисперсии равна отношению собственного числа к сумме всех собственных чисел (сумме элементов на главной диагонали матрицы). Основной недостаток критерия, основанного на величине доли воспроизводимой дисперсии, состоит в определенной его субъективности. Однако он основан на легко поддающейся интерпретации статистике и в этом преимущество данного метода.

КРИТЕРИЙ ОТСЕИВАНИЯ

Настоящий метод предложен Каттеллом (Cattell, 1965). Рассматривается графическое изображение собственных чисел корреляционной матрицы, которые наносятся на график в порядке их убывания. Выделение заканчивается на том факторе, после которого исследуемая зависимость близка к почти горизонтальной прямой линии. Эту прямую Каттелл и предлагает использо-

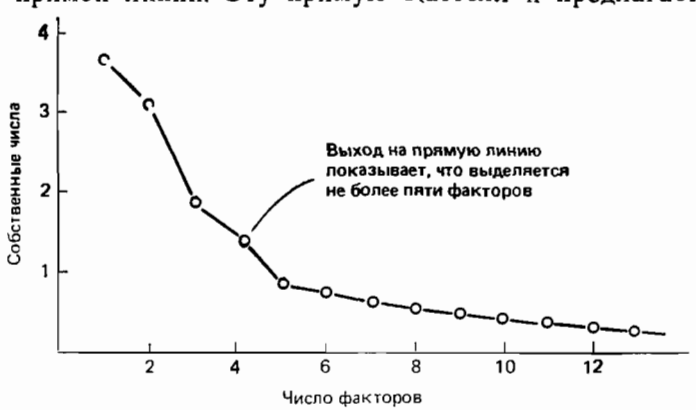

Рис. 5. Критерий отсеивания

вать для выделения факторов. Пример применения критерия отсеивания представлен на рис. 5. Как видно, выделяется не более пяти факторов. Моделирование по методу Монте-Карло показывает, что при наличии второстепенных факторов, данный критерий предпочтительнее по сравнению с другими (Tucker, Koopman, Linn, 1969; Linn, 1968). Кайзер скептически относится к критерию отсеивания (Kaiser, 1970), так как на графике можно получить более чем один излом, и тогда выделение какой-либо прямой становится субъективным.

КРИТЕРИЙ ИНТЕРПРЕТИРУЕМОСТИ И ИНВАРИАНТНОСТИ

Для исключения сомнительных результатов можно попытаться применить к одним и тем же данным комбинацию различных независимых критериев и принимать только те результаты, которые подходят ко всем критериям (Harris, 1967). Окончательное решение должно базироваться на его приемлемости с точки зрения научных представлений в данной области. Этот подход является «обходным маневром», но, к сожалению, а может быть и к счастью, мы вынуждены принять его, если хотим, чтобы нашими результатами могли воспользоваться другие исследователи.

V. ВВЕДЕНИЕ В КОНФИРМАТОРНЫЙ ФАКТОРНЫЙ АНАЛИЗ

Рассматривая разведочный факторный анализ, мы выделяли те предположения, которые необходимы для его применения. Наиболее важные из этих предположений — принципы факторной причинности и экономии. Методика факторного анализа состоит в том, что априори принимается определенная модель взаимосвязи между наблюдаемыми переменными, а затем находится решение, наиболее полно согласующееся с наблюдениями. Возникает законный вопрос: существует ли возможность какого-либо подтверждения факторной модели? Как уже упоминалось, нет способа доказать существование определенной причинной структуры исходя из наблюдаемой ковариационной структуры. Тем не менее можно оценить, до какой степени правдоподобие факторной модели эмпирически подтверждено.

СТЕПЕНЬ ЭМПИРИЧЕСКОГО ПОДТВЕРЖДЕНИЯ ГИПОТЕЗ О ФАКТОРНОЙ МОДЕЛИ

По сравнению с разведочным анализом в конфирматорном факторном анализе рассматриваются более специфические гипотезы о факторной структуре. Следовательно, имеется вероятность, что если на самом деле данные не полностью соответствуют модели, то определенные гипотезы будут отвергнуты. В этом смыс-

ле модели в конфирматорном анализе являются самопроверяющимися. Если данная гипотеза подтверждается результатами наблюдений, появляется бо́льшая уверенность в том, что рассматриваемая факторная модель соответствует действительности. Перед тем как обсуждать конфирматорный факторный анализ, важно получить представление об эмпирическом подтверждении модели в целом, а также решить, можно ли использовать факторный анализ для наших данных или нет.

Пример

Применение факторной модели к двумерной корреляционной матрице не дает никакой новой информации, так как модель с одним общим фактором всегда совместима с ней. Таким образом, в этой ситуации факторный анализ неприменим вовсе не потому, что факторная модель несовместима с данными. Причина в другом — степень эмпирического подтверждения модели (или, короче, его информативность) нулевая, и, кроме того, нет единственного решения.

Рассмотрим зависимость между первыми двумя переменными в модели, представленной на рис. 1. Если задан произвольный коэффициент корреляции, можно выбрать первую факторную нагрузку на интервале от -1 до $+1$ (за исключением 0). При этом существует другая факторная нагрузка, обеспечивающая совместимость с наблюдаемой корреляцией. Короче говоря, всегда есть факторное решение, совместное с даннымн.

Ситуация несколько меняется, когда факторный анализ применяется к корреляционной матрице с тремя переменными. Если оказывается, что однофакторная модель совместима с данными, степень эмпирического подтверждения уже ненулевая, так как некоторые случайно выбранные корреляционные матрицы несовместимы с однофакторной моделью. В частности, для того, чтобы корреляционная матрица с тремя параметрами была совместима с однофакторной моделью, три коэффициента корреляции должны удовлетворять следующим условиям: 1) либо все коэффициенты корреляции положительные, либо четное число из них является отрицательным; 2) абсолютная величина любого коэффициента должна быть больше или равна абсолютной величине произведения остальных двух коэффициентов:

$$|r_{ij}| \geqslant |r_{ik} r_{jk}|. \qquad (32)$$

Интересно проанализировать условие (32). Рассмотрим верхнюю часть рис. 1 и введем такие обозначения:

$$r_{12} = b_1 b_2; \quad h_1 = b^2_1;$$
$$r_{13} = b_1 b_3; \quad h_2 = b^2_2;$$
$$r_{23} = b_2 b_3; \quad h_3 = b^2_3;$$

где $b_1 = 0{,}8$, $b_2 = 0{,}7$, $b_3 = 0{,}6$ — факторные нагрузки, h_1, h_2, h_3 — общности.

Перемножим два коэффициента корреляции:

$$r_{12}r_{13}=b_1b_2b_1b_3=b^2{}_1\,b_2b_3=h^2{}_1r_{23}\cdot \qquad (33)$$

Поскольку общности не превышают 1, то из (33) вытекает условие (32):

$$h^2{}_1=|r_{12}r_{13}|\,/\,|r_{23}|\leqslant 1.$$

Аналогичные рассуждения можно провести и для других пар коэффициентов корреляции. Но не все случайно выбранные корреляционные матрицы для трех переменных удовлетворяют сформулированным условиям. Поэтому тот факт, что экспериментальные данные согласуются с однофакторной моделью является информативным, однако не слишком информативным, так как условию (32) удовлетворяют достаточно много случайно выбранных корреляционных матриц для трех переменных.

Корреляционная матрица для четырех переменных, основанная на однофакторной модели, удовлетворяет трем дополнительным условиям:

$$\begin{aligned} r_{13}r_{24}&=r_{14}r_{23}\\ r_{12}r_{34}&=r_{14}r_{23}\\ r_{13}r_{24}&=r_{12}r_{34}\cdot \end{aligned} \qquad (34)$$

Эти условия легко получить:

$$r_{13}r_{24}=b_1b_3b_2b_4=(b_1b_4)\,(b_2b_3)=r_{14}r_{23}\cdot$$

Вообще, чем больше число переменных, тем больше число условий, которым должна удовлетворять корреляционная матрица для данной факторной модели. Таким образом, совместимость однофакторной модели с корреляционной матрицей для четырех переменных дает исследователю эмпирическое подтверждение, что факторные предположения не совсем произвольны. Следовательно, некоторое заключение о факторной структуре информативно только тогда, когда корреляционная матрица удовлетворяет некоторым ограничениям. Лишь в этом случае можно судить, соответствует ли данная факторная модель экспериментальным данным. Более того, чем больше отношение числа переменных к числу гипотетических факторов, тем весомее эмпирическое подтверждение факторной модели, поскольку увеличивается число структурных ограничений, накладываемых на корреляционную матрицу с целью согласования с данной моделью.

Вспомним теперь, что применение факторного анализа предполагает наложение различных допущений на экспериментальные данные. Поэтому можно отвергнуть факторную модель только на основе того, что эти предположения являются либо произвольными, либо неподходящими. Тем не менее такое суждение смягчается, когда степень эмпирического подтверждения высока, поскольку следует считаться со структурными ограничениями в данных. С одной стороны, можно сказать, что информативность факторного анализа зависит от условий его применения. С другой стороны, в факторном решении содержится информация о его

пригодности: чем больше число эмпирических ограничений, которым должно удовлетворять решение, тем больше степень уверенности в том, что факторная модель соответствует данным. С этой точки зрения даже разведочный факторный анализ дает информацию о пригодности и экономичности модели.

ЧИСЛО ЭМПИРИЧЕСКИХ ОГРАНИЧЕНИЙ ДЛЯ ФАКТОРНОЙ МОДЕЛИ

С учетом вышесказанного важной характеристикой информативности гипотезы является число ограничений, накладываемых данной факторной моделью (т. е. число условий, которым должны удовлетворять элементы корреляционной матрицы для возможного их восстановления с помощью факторной модели). Оказывается, это число равно количеству степеней свободы для критерия значимости решения максимального правдоподобия. Ясное понимание зависимости между факторной гипотезой и соответствующим ей числом степеней свободы является решающим моментом для понимания конфирматорного факторного анализа.

Существует несколько различных подходов к определению числа ограничений для элементов корреляционной матрицы. Один подход сводится к использованию теоремы о ранге. В этой теореме утверждается, что если на диагональ корреляционной матрицы поместить общности, соответствующие r-факторной модели, то ранг (число линейно-независимых строк или столбцов) редуцированной корреляционной матрицы будет равен r. При этом все миноры, содержащие больше, чем r строк и столбцов, будут иметь нулевой детерминант. Отсюда можно определить число условий, которым должна удовлетворять корреляционная матрица при заданном числе факторов и параметров (Harman, 1976). Другой подход связан с изучением степеней свободы для критерия значимости. По-видимому, второй подход является более общим.

Для примера предположим, что мы имеем дело с эмпирической корреляционной матрицей. Количество аппроксимируемых параметров, содержащихся в ней, равно $1/2n\ (n-1)$ — числу элементов над главной диагональю. Факторный анализ позволяет получить первоначальное решение с помощью варьирования $n \times r$ факторных нагрузок (r — число общих факторов) с тем, чтобы обеспечить наилучшее воспроизведение наблюдаемой корреляционной матрицы. Но для первоначального факторного решения требуется ортогональность полученных факторов. Это условие влечет за собой $1/2r\ (r-1)$ дополнительных связей. Поэтому число свободных параметров составит

$$nr - (1/2)r(r-1). \tag{35}$$

Итак, число условий, которым должны удовлетворять элементы корреляционной матрицы, задается соотношением

$$1/2n(n-1) - [nr - 1/2r(r-1)] = 1/2\,[(n-r)^2 - (n+r)]. \tag{36}$$

Таблица 8

Число степеней свободы для n переменных и r факторов[1]

Число переменных, n	Число факторов				Максимальное число факторов для положительного числа степеней свободы	Число независимых коэффициентов, $1/2n\ (n-1)$
	1	2	3	4		
3	0	−2	−3	—	нет	3
4	2	−1	−3	−4	1	6
5	5	1	−2	−4	2	10
6	9	4	0	−3	2	15
7	14	8	3	−1	3	21
8	20	13	7	2	4	28
9	27	19	12	6	5	36
10	35	26	18	11	5	45
11	44	34	25	17	6	55
12	54	43	33	24	7	66
20	170	151	133	116	14	190
40	740	701	663	626	31	780

[1] Общая формула для числа ограничений:

$$D=\frac{(n-r)^2-(n+r)}{2}.$$

Выражение (36) и определяет упомянутое выше число степеней свободы. Когда вместо корреляционной матрицы используется ковариационная матрица, число независимых элементов равно $1/2n\ (n+1)$, а не $1/2n\ (n-1)$. Однако число степеней свободы не меняется, поскольку возникают дополнительные условия, связанные с применимостью факторной модели к ковариационной матрице.

В табл. 8 представлены значения числа ограничений при различных комбинациях количества факторов и переменных. Следует выделить несколько аспектов. Во-первых, как правило, число эмпирических ограничений увеличивается при возрастании отношения числа переменных к числу факторов. Во-вторых, когда число ограничений отрицательно, эмпирическое подтверждение факторной модели невозможно. Таким образом, имеет смысл рассматривать только модели, которые накладывают на данные некоторые ограничения. Например, применение двухфакторной модели при четырех переменных и трехфакторной модели при шести переменных — неинформативно. В-третьих, число ограничений для фиксированного количества факторов быстро растет при увеличении числа переменных, т. е. добавление переменной заметно повышает информативность полученного факторного решения. В-четвертых, в эмпирическом подтверждении решения более существенное значение имеет разность мжду числом переменных и числом факторов, а не их отношение. Заметим, что количество ограничений практически одинаково для следующих комбинаций: 1 фактор при 7 переменных (14); 2 — при 8 (13);

3 — при 9 переменных (12) и так далее. Однако нет оснований считать разность между числом переменных и числом факторов непосредственной мерой степени эмпирического подтверждения. Альтернативой служит отношение количества ограничений к количеству независимых коэффициентов наблюдаемой матрицы. Хотя в таблице эти отношения не представлены (знаменатели их приведены в последнем столбце), следует отметить их достаточно высокую информативность.

При оценивании степени эмпирического подтверждения факторного решения следует принимать во внимание два осложняющих дело обстоятельства: 1) определенные свойства, присущие генеральной совокупности не обязательно могут проявиться в выборке; 2) даже при использовании генеральной совокупности факторная модель может не совсем точно соответствовать экспериментальным данным. Другими словами, свойства генеральной совокупности должны оцениваться с учетом этих расхождений. Более того, на практике не представляется возможным отделить действие одного из этих упомянутых обстоятельств от другого. Таким бразом, само по себе выражение (36) не может служить мерой степени эмпирического подтверждения. Решение, на которое накладывается большее число ограничений, обеспечивает более значительную степень подтверждения при заданной степени расхождения между факторным решением и наблюдениями. Поэтому необходимо научиться оценивать вышеупомянутое расхождение.

Степень эмпирического подтверждения или надежность

С помощью критерия значимости, применяемого для какого-либо первоначального факторного решения, оценивается возможность приписать расхождение между гипотетической моделью и наблюдениями статистической флуктуации в выборке. Критерий значимости непосредственно зависит от объема выборки; при достаточно большой выборке любые расхождения между моделью и экспериментальными данными могут стать значимыми. Это следует из того факта, что если модель точно соответствует наблюдениям, то чем больше объем выборки, тем меньше расхождения между выборочными параметрами и параметрами генеральной совокупности. Для очень большой выборки такие расхождения весьма малы.

Применение этого статистического принципа бывает затруднительным, когда исследователь подозревает наличие второстепенных факторов и не имеет возможности определить их природу. Тогда критерий значимости может не подтвердить адекватность модели. Даже если рассматриваемая факторная модель воспроизводит бóльшую долю наблюдаемых ковариаций и привносит определенный порядок в структуру наблюдений, критерий значимости может показать, что модель статистически неадекватна экспериментальным данным. Поэтому необходима мера адекват-

ности, которая концептуально независима от статистической значимости.

Итак, необходимо определить меру расхождения между наблюдаемой корреляционной матрицей и воспроизведенной матрицей. Один из возможных подходов описан Харманом. Он предлагает использовать среднее значение квадрата отклонения, при котором квадраты отклонений корреляций, полученных для окончательного факторного решения, от наблюдаемых корреляций суммируются и делятся на число этих коэффициентов:

$$\sum_{i \neq j}\sum (r_{ij} - \hat{r}_{ij})^2 / [n(n-1)],$$

где суммирование распространяется на все недиагональные элементы (Harman, 1976). Однако для этой величины не ясен выбор порогового значения.

Другая альтернатива, предложенная Такером и Левисом (Tucker, Lewis, 1973), рассматривает коэффициент надежности для факторного решения методом максимального правдоподобия. Этот подход основан на использовании частных коэффициентов корреляции, причем вводится нормировка на число степеней свободы с тем, чтобы учесть возможные расхождения между факторными решениями. Кроме того, в коэффициенте надежности происходит сопоставление соответствующих статистик со случаем отсутствия влияния факторов. Формула для коэффициента надежности

$$rho = \frac{M_0 - M_k}{M_0 - 1}, \qquad (37)$$

где M_0 — математическое ожидание статистики χ^2 в отсутствии влияния факторов, деленное на $^1/_2\ n\ (n-1)$, а M_k — математическое ожидание χ^2 для окончательного факторного решения, деленное на $(^1/_2)\ [(n-r)^2-(n+r)]$ (Sörbom, Jöreskog, 1976). Коэффициент *rho* принимает значения от 0 до 1, причем 0 означает наихудшее согласие модели и данных, а 1 — наилучшее. На практике чаще применяется приближенное значение *rho*, асимптотически эквивалентное (37) при возрастании объема выборки:

$$rho = 1 - \frac{E_1 - 1}{E_2 - 1}.$$

где

$$E_1 = \sum_{i \neq j}\sum (r_{ij \cdot F})^2 / df_k\ ;$$

$$E_2 = \sum_{i \neq j}\sum (r_{ij})^2 / [1/2n(n-1)];$$

r_{ijF} — частные коэффициенты корреляции без влияния факторов/ d_{fk} — число степеней свободы, равное $^1/_2\ [(n-r)^2-(n+r)]$, в разведочном факторного анализе. В конфирматорном анализе число степеней свободы несколько больше. Отметим, что частные коэф-

фициенты корреляции есть не что иное, как расхождения между воспроизведенными и наблюдаемыми корреляциями, представленные в стандартной форме.

ДРУГАЯ КОНЦЕПЦИЯ ЭМПИРИЧЕСКОГО ПОДТВЕРЖДЕНИЯ: ВЫБОРОЧНАЯ АДЕКВАТНОСТЬ

При использовании традиционных статистических критериев предполагается, что есть выборка объектов. Однако на практике в определенной мере имеет место и психометрическая выборка — анализируемые переменные почти всегда являются выбранными из некоторой совокупности. Возникает вопрос об адекватности рассматриваемой факторной модели по отношению к данному набору переменных. Напомним, что психометрический подход используется в анализе образов и альфа-векторном анализе, но предмет обсуждения относится и к любому другому методу факторного анализа.

При прочих равных условиях степень эмпирического подтверждения увеличивается, если: 1) возрастает число переменных; 2) уменьшается число общих факторов; 3) уменьшаются частные коэффициенты корреляции либо 4) увеличивается коэффициент детерминации. Первые два условия гарантируют увеличение эмпирических ограничений, накладываемых факторной моделью на экспериментальные данные. Третье — относится к измерению близости наблюдаемых и воспроизводимых коэффициентов корреляции. Четвертое условие состоит в том, чтобы увеличивалась доля общности в дисперсии каждой наблюдаемой переменной. Последнее непосредственно относится к выборочной адекватности, так как коэффициент детерминации возрастает при увеличении числа переменных и при уменьшении среднего значения коэффициентов корреляции.

Эмпирический критерий выборочной адекватности был предложен Кайзером (Kaiser, 1970, 1974). Он назвал этот критерий «мерой выборочной адекватности» (MBA):

$$MBA = \frac{\sum\sum_{j \neq k} r_{jk}}{\sum\sum_{j \neq k} r_{jk}^2 + \sum\sum_{j \neq k} g_{ik}^2} \qquad (38)$$

где r_{ij} — наблюдаемые коэффициенты корреляции, а g_{ij} — элементы корреляционной матрицы антиобразов, которая задается выражением

$$Q = SR^{-1}S,$$

причем R^{-1}—обратная корреляционная матрица, а $S = (\text{diag}\, R^{-1})^{1/2}$. MBA изменяется от 0 до 1. Данный критерий принимает значение 1 тогда и только тогда, когда все недиагональные элементы матрицы, обратной в корреляционной матрице Q, — нулевые. Это в свою очередь означает, что каждая переменная может быть выражена

без ошибок через остальные переменные. Пороговые значения для МВА по Кайзеру (Kaiser, 1974) следующие:

свыше 0,9 — отлично
» 0,8 — хорошо
» 0,7 — средне
» 0,6 — посредственно
» 0,5 — плохо
ниже 0,5 — неприемлемо

Кайзер, экспериментируя с модельными данными, показал, что величина МВА увеличивается при: 1) возрастании числа переменных; 2) уменьшении числа общих факторов; 3) увеличении объема статистики и 4) увеличении среднего значения коэффициентов корреляций (Kaiser, 1970).

Еще раз отметим, что степень эмпирического подтверждения факторной модели с помощью экспериментальных данных варьируется в зависимости от обстоятельств. Исследователь должен знать условия, при которых информативность факторного анализа повышается. Начинающий пользователь факторных методов при исследовании адекватности может положиться на такую эмпирическую меру, как МВА. Разумеется, окончательное решение должно приниматься на основе теоретически обоснованных выводов.

КОНФИРМАТОРНЫЙ ФАКТОРНЫЙ АНАЛИЗ

При использовании конфирматорного факторного анализа всегда должна выдвигаться гипотеза о числе общих факторов. Она должна быть основана на понимании природы рассматриваемых переменных и на информации о том, какой фактор имеет нагрузку и от каких переменных, если мы не хотим гадать на кофейной гуще. Разнообразие по форме этих факторных гипотез неограничено.

Конфирматорный анализ можно разделить на два вида: одногрупповой, который имеет дело с одной генеральной совокупностью, и многогрупповой, работающий с двумя и более генеральными совокупностями. Начнем обсуждение с первого случая.

Одна группа или генеральная совокупность

Применяя конфирматорный факторный анализ для заданной ковариационной матрицы, необходимо иметь гипотезу о соответствующей факторной структуре. Затем проводится оценивание, насколько «значимо» наблюдаемые данные отличаются от гипотетической структуры. В некоторых случаях гипотеза может включать следующую информацию: а) число общих факторов; б) природу зависимости между факторами (ортогональные или косоугольные) и в) величину факторных нагрузок для каждой переменной. В других случаях гипотеза касается только числа общих факторов. Разумеется, существует множество гипотез, занимающих среднее положение между этими двумя крайностями.

Поскольку простейшая форма конфирматорного факторного анализа (когда фиксируется только число общих факторов) мало отличается от разведочного анализа, потребуются лишь небольшие комментарии. Для этого вида гипотез неважно, будет ли использоваться ортогональная или косоугольная факторная модель, и годится ли любой критерий значимости или какой-либо другой критерий типа коэффициента надежности для оценивания адекватности первоначального факторного решения. Единственное отличие, о котором можно упомянуть, заключается в том, что в конфирматорном анализе число факторов выбирается исходя из априорных соображений (в разведочном берется произвольное число факторов) и если первый выбор оказался неудачным, оно изменяется. Следует сказать, что неразумно целиком полагаться на критерии значимости, если мы не хотим вводить в рассмотрение второстепенные, но статистически значимые факторы. Желательно провести вращение решения и определить, имеет ли полученная структура «физический» смысл.

Другой крайний случай гипотезы также не представляет сложностей для обсуждения. Если есть гипотеза о числе факторов, зависимости между факторами и о значении коэффициентов нагрузок, то можно проверить, близки ли элементы воспроизведенной корреляционной матрицы к наблюдаемым коэффициентам корреляции, либо использовать эту гипотезу как целевую матрицу. В последнем случае следует определить решение, которое аппроксимирует целевую матрицу и наиболее точно воспроизводит наблюдаемые корреляции. В первом из рассмотренных случаев проверка адекватности гипотезы опирается на некоторый критерий для оценивания близости ковариационных матриц. Во втором случае требуется критерий для оценивания близости двух факторных решений. Более подробно об этом можно прочесть в работе Левина (Levine, 1977). На практике вряд ли в нашем распоряжении окажется такая полная информация. Однако данная гипотеза может понадобиться при сравнении факторной структуры для одного набора экспериментальных данных со структурой, основанной на другом наборе.

Сёрбом и Йореско разработали программное обеспечение для конфирматорного факторного анализа (Sörbom, Jöreskog, 1976). Мы опишем основные переменные этой весьма гибкой программы. Существует несколько способов задания каждой переменной. Переменные, применяемые в факторном анализе, включают факторные нагрузки ($n \times r$ коэффициентов) и коэффициенты корреляции между факторами ($^1/_2 r(r-1)$ чисел). Каждая из этих переменных может быть *фиксирована* или оставлена для варьирования. Наиболее часто при фиксировании используется обнуление отдельных нагрузок. Например, если задать все корреляции между факторами нулевыми, полученное решение будет ортогональным. Другим способом определения переменных является *задание ограничений*, сводящееся к тому, что одна переменная должна быть равна другой.

Таблица 9

Три примера задания переменных в конфирматорном анализе[1]

Переменная	Пример 1			Пример 2			Пример 3		
	F_1	F_2	F_3	F_1	F_2	F_3	F_1	F_2	F_3
X_1	x	0	0	x	x	0	x	x	x
X_2	x	0	0	x	x	0	x	x	x
X_3	x	0	0	x	x	0	x	x	x
X_4	0	x	0	x	x	0	x	x	0
X_5	0	x	0	x	0	x	x	x	0
X_6	0	x	0	x	0	x	x	0	0
X_7	0	0	x	x	0	x	x	0	0
X_8	0	0	x	x	0	x	x	0	0

[1] x — свободная переменная, 0 — переменная, равная нулю.

В табл. 9 представлены три случая задания *свободных* и *фиксированных* переменных. Кроме нулевых значений, можно использовать, например, значения 1,0; 0,5 и т. д. Однако представляется более реальным, что исследователь располагает лишь информацией о том, велики или малы те или иные нагрузки. Первая гипотеза задает однофакторную структуру — наиболее простой вид для заданного набора переменных. Вторая выделяет генеральный фактор и два групповых фактора. Третья гипотеза задает некоторую иерархическую структуру. Разумеется, можно задавать многие модификации этих структур.

Необходимо также задавать зависимости между факторами. Обычно используются следующие формы зависимости: 1) задание всех факторных корреляций нулевыми — ортогональная гипотеза; 2) варьируемые корреляции — косоугольная гипотеза и 3) смешанная структура, когда некоторые факторы предполагаются ортогональными, а остальные произвольными.

В табл. 10 представлен пример задания гипотезы для конфирматорного факторного анализа, использующий выборочные данные из табл. 1. Предположим, что мы хотим задать следующую гипотезу: 1) существуют два общих фактора; 2) два фактора могут быть коррелированы и 3) один фактор имеет нулевые нагрузки на переменные X_4, X_5, X_6, а другой — на X_1, X_2, X_3.

Заметим, что в отличие от разведочного анализа в конфирматорном факторном анализе 6 факторных нагрузок из 12 (nr) фиксированы, и один коэффициент в факторной ковариационной матрице полагается свободным. Соответственно мы налагаем 5 дополнительных ограничений.

Не все из этих ограничений отражены при вычислении количества степеней свободы. В разведочном анализе подразумеваются $^1/_2r(r-1)$ ограничений для обеспечения единственности решения. Таким образом, число ограничений равно: $5-{}^1/_2r(r-1)=4$. В общем случае невязка между моделью с фиксированными вели-

Таблица 10

Фиксированные и свободные величины, задаваемые для получения косоугольного факторного решения[1]

Переменные	Фактор	
	F_1	F_2
X_1	×	0
X_2	×	0
X_3	×	0
X_4	0	×
X_5	0	×
X_6	0	×

Корреляции между факторами

	F_1	F_2
F_1	1	
F_2		1

[1] 0 — соответствует фиксированным величинам, а × — свободным. Единица в факторной корреляционной матрице — фиксированное число для входной матрицы. Однако при вычислении степеней свободы эти единицы уже не считаются фиксированными.

чинами и экспериментальными данными будет больше, чем невязка для модели со свободными величинами. Но увеличение невязки будет компенсировано увеличением числа степеней свободы, если гипотетическая модель соответствует действительности.

Отметим, что вряд ли целесообразно применять трехфакторную модель к матрице с шестью переменными. Однако такую модель вполне можно использовать, если гипотетическая факторная структура имеет достаточное число ограничений для обеспечения нескольких степеней свободы. Например, могут быть заданы следующие ограничения: переменные X_1 и X_2 имеют нагрузку только на первый фактор; X_3 и X_4 — на второй, а X_5 и X_6 — на третий фактор.

Принципы, изложенные в данном разделе, могут быть использованы не только в факторном анализе. Можно сочетать особенности факторного анализа с особенностями регрессионного и путевого (path) анализа. Предположим, рассматривается набор наблюдаемых переменных, которые связаны с латентной переменной (F_1), влияющей в свою очередь на другую летентную переменную (F_2). Последняя также имеет набор наблюдаемых (индикаторных) переменных. Такую систему зависимостей можно проанализировать с помощью средств конфирматорного анализа. В данном случае модель может быть представлена в рамках конфирматорного факторного анализа с двумя коррелированными факторами (рис. 6). Отметим, что эта модель совпадает со структурой, представленной в табл. 9 (пример 1), когда не накладываются ограничения на корреляции между факторами. Мы упомянули только о наиболее простом обобщении конфирматорного факторного анализа; заинтересованный читатель может обратиться за более подробной информацией к другим работам (Jöreskog, 1970; Sörbom, Jöreskog, 1976).

Сравнение факторных структур

Другое применение конфирматорного факторного анализа состоит в сравнении факторных структур для нескольких групп наблюдений. Например, можно выдвинуть гипотезу о том, что фак-

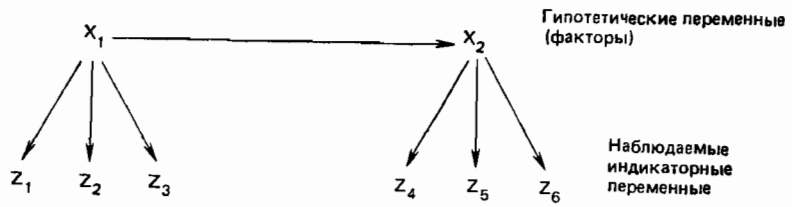

Рис. 6. Модель, включающая два гипотетических фактора и 6 наблюдаемых переменных

торная структура политических отношений чернокожего населения совпадает с аналогичной факторной структурой для белого населения, или о том, что структура одного общества эквивалентна структуре другого. Также возможно определить, что некоторые аспекты факторных структур совпадают, а другие для разных групп различны.

Существует компьютерная программа COFAMM (Confirmatory Factor Analysis with Model Modification), разработанная Йореско и Сёрбомом. Эта программа позволяет обращаться с весьма общими гипотезами. Например, она допускает всевозможные вариации при проверке факторной гипотезы для отдельной группы — некоторые величины могут фиксироваться или варьироваться, либо задаваться ограничениями типа совпадения одних величин с другими. В дополнение к использованию «ограничений» на величины любая часть величин, относящихся к структуре одной группы, может быть задана совпадающей с соответствующими величинами для другой группы.

В качестве примера рассмотрим сравнение структур политических отношений белого и чернокожего населения. Специальные гипотезы могут иметь следующий вид: 1) существуют два косоугольных фактора для белого и чернокожего населения; 2) переменные X_1 (финансирование образования), X_2 (выделение средств на уменьшение безработицы) и X_3 (контролирование большого бизнеса) имеют одинаковые нагрузки от одних и тех же факторов для белого и чернокожего населения; X_4 (программы занятости) и X_5 (квоты на профессии) имеют нагрузки от разных факторов для обеих рас; 3) однако переменная X_6 (программа борьбы с кризисами) имеет различные нагрузки для этих двух групп. В этом случае можно задать входные величины для белого населения так же, как в одногрупповом анализе, а величины для чернокожего населения (за исключением переменной, X_6) определить в виде ограничения — равенства соответствующим величинам для белой расы. Многочисленные примеры применения конфирматорного факторного анализа можно найти в работе Йореско (Jöreskog, 1976).

VI. ФАКТОРНОЕ ШКАЛИРОВАНИЕ*

После изучения результатов факторного анализа можно приступить к оценке факторных шкал. Для этого есть следующие основания. Во-первых, после определения скрытой факторной структуры измеряемых данных для объектов исследователю может понадобиться представить каждый из этих объектов в терминах значений факторов, а не измеряемых переменных. Во-вторых, может появиться необходимость использования одного или более факторов в качестве переменных для дальнейшего анализа. Действительно, за исключением психометрической литературы, факторный анализ применялся чаще в качестве средства создания новых факторных переменных (шкал) для других исследований, чем для изучения самой скрытой структуры. В этом разделе мы рассмотрим следующие методы оценки значений факторов: 1) регрессионные оценки; 2) оценки, основанные на искусственных переменных или критерии наименьших квадратов; 3) метод Бартлетта минимизации дисперсии ошибок и 4) оценки с ортогональными ограничениями. Дополнительно мы обсудим: 5) простой метод суммирования переменных с большими факторными нагрузками и 6) шкалирование с помощью главных компонент. Эти методы будут обсуждаться в связи с некоторыми важными аспектами факторного шкалирования.

НЕОПРЕДЕЛЕННОСТЬ ФАКТОРНОГО ШКАЛИРОВАНИЯ

Для начала рассмотрим модельные данные. Предположим, что мы их получили, воспользовавшись однофакторной моделью. Главной целью факторного шкалирования является определение значений общего фактора (F) через наблюдаемые переменные $(X_1, \ldots)$. Как уже говорилось, невозможно точно выразить общий фактор посредством наблюдаемых переменных, поскольку каждая из них содержит также и характерную компоненту, которую нельзя отделить от всей переменной. Можно получить лишь оценку значений общих факторов через наблюдаемые переменные. Поэтому шкалирование факторов всегда связано с некоторой неопределенностью.

Возьмем однофакторную модель с тремя переменными. Допустим, что все факторные нагрузки одинаковы (или что все коэффициенты корреляции равны). Этот пример показан на рис. 7 слева. Для нашей модели вычислить наблюдаемые коэффициенты корреляции между переменными можно с помощью перемножения факторных нагрузок, причем, поскольку все нагрузки одинаковы, коэффициент корреляции будет равен квадрату факторной нагрузки:

---

* Под факторным шкалированием понимается процедура, позволяющая присваивать каждому объекту некоторые числовые оценки значений выделенных факторов, используя значения наблюдаемых переменных для этого объекта. — *Примеч. ред.*

$$r_{ij}=b_ib_j=b^2_i=b^2_j=h^2. \qquad (39)$$

Выражение (39) показывает, что наблюдаемые корреляции совпадают в данном случае с общностью любой из переменных (все три общности здесь равны).

В качестве оценки значения фактора берется линейная комбинация параметров X_1, X_2, X_3. Так как каждая из этих переменных имеет одинаковую нагрузку от общего фактора, то естественно сложить их, беря соответствующие значения с одинаковым весом. Окончательное выражение будет иметь вид

$$\hat{F}=X_1+X_2+X_3,$$

а соответствующая диаграмма представлена в правой части рис. 7. Отметим, что оценка $\hat{F}$ фактически зависит от четырех переменных — общего фактора F и трех характерных факторов U_1, U_2 и U_3. Следовательно, из-за наличия характерных факторов, корреляция между F и $\hat{F}$ не равна 1. Ниже мы рассмотрим связь между скрытым общим фактором и его оценкой, т. е. получим надежность оценки.

Надежность факторного шкалирования

Дисперсию оценки $\hat{F}$ легко вычислить, используя свойства математических ожиданий:

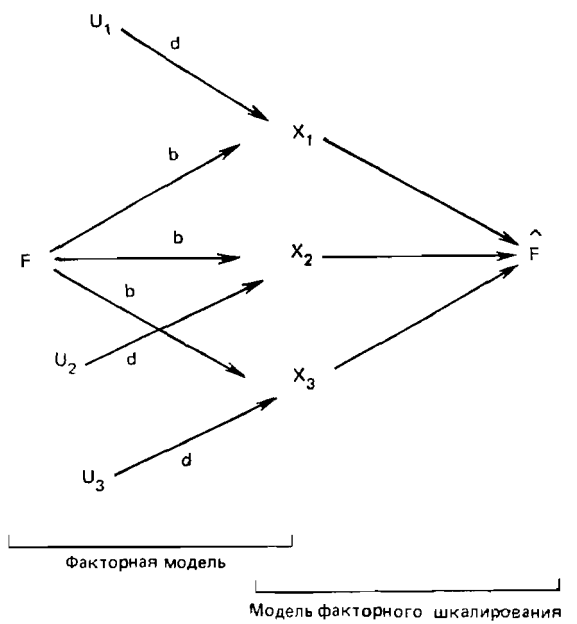

Рис. 7. Графическая модель, иллюстрирующая зависимость между фактором и его оценкой

$$\text{var}(\hat{F}) = \text{var}(X_1) + \text{var}(X_2) + \text{var}(X_3) + \\ + 2[\text{Cov}(X_1, X_2) + \text{Cov}(X \quad X_3) + \text{Cov}(X_2, X_3)]. \qquad (40)$$

Поскольку в этом примере взяты единичные веса, выражение упрощается. Дальнейшее упрощение достигается в том случае, если дисперсии каждой переменной будут единичными, а коэффициенты корреляции будут попарно равны друг другу:

$$\text{Var}(\hat{F}) = n + 2[r_{12} + r_{13} + r_{23}] = n + n(n-1)r = \\ = n[1 + (n-1)r] = n[1 + (n-1)h^2] \qquad (41)$$

(из формулы (39) следует, что $r_{12} = r_{13} = r_{23} = r = h_i^2$). Некоторая доля дисперсии F связана с характерными факторами. Их вклад равен: $\Sigma d_i^2 = \Sigma(1 - h_i^2) = n(1 - h^2)$, так как все общности в нашем примере равны. Таким образом, доля дисперсии $\hat{F}$, связанная с общим фактором F, получается из соотношения

$$r^2_{(F,\hat{F})} = \frac{\text{Var}(\hat{F}) - n(1-h^2)}{\text{var}(\hat{F})} = \frac{n[1+(n-1)h^2] - n(1-h^2)}{n[1+(n-1)h^2]} = \\ = \frac{nh^2}{1+(n-1)h^2} = \frac{nr}{1+(n-1)r} \qquad (42)$$

что соответствует формуле Спирмена — Брауна для надежности и специальному случаю альфа-параметра Кронбаха (Cronbach, 1951: Lord, Novick, 1968). Следует напомнить, что в данном случае h^2 можно заменить на r.

Для того чтобы показать степень неопределенности, или степень ожидаемой «надежности» факторного шкалирования, в табл. 11 представлены значения коэффициентов «надежности» для некоторых типичных значений общностей при различном числе переменных. Отметим, что при возрастании числа переменных для фиксированного значения общности (факторных нагрузок или корреляций) надежность возрастает. Кроме того, даже при весьма высокой факторной нагрузке (скажем, 0,8) надежность все же относительно низкая, если число переменных мало.

Следует иметь в виду, что при факторном шкалировании часто используют оценку $\hat{F}$ в стандартном виде — с нулевым математическим ожиданием и единичной дисперсией. Разумеется, принципиального значения это обстоятельство не имеет.

Неодинаковые факторные нагрузки

До сих пор мы ограничивались не только одинаковыми факторными нагрузками в однофакторной модели, но и брали лишь данные без ошибок. Теперь попробуем усложнить задачу.

Рассмотрим ситуацию, когда факторные нагрузки в однофакторной модели неодинаковы. Получаем корреляционную матрицу более общего вида. Если оценка фактора найдена в результате

Таблица 11

Коэффициент надежности (корреляции между фактором, и его оценкой) **для различных значений равных между собой факторных нагрузок и различного числа переменных**[1]

Факторные нагрузки	0,4	0,5	0,6	0,7	0,8	0,9
Общность (h^2) или корреляции между переменными	0,16	0,25	0,36	0,49	0,64	0,81
Число переменных						
2	0,276	0,400	0,529	0,658	0,780	0,895
3	0,364	0,500	0,628	0,742	0,842	0,927
4	0,433	0,571	0,692	0,794	0,877	0,945
6	0,533	0,667	0,771	0,852	0,914	0,962
8	0,604	0,727	0,818	0,885	0,934	0,972
12	0,696	0,800	0,871	0,920	0,955	0,981
20	0,792	0,870	0,918	0,951	0,973	0,988

[1] Формула для коэффициента надежности $(\alpha) = \frac{n(r)}{1+(n-1)r} = \frac{n(h^2)}{1+(n-1)h^2}$.

суммирования наблюдаемых параметров, надежность такой оценки будет равна:

$$\alpha = \frac{\text{сумма элементов редуцированной корреляционной матрицы}}{\text{сумма элементов корреляционной матрицы}} = \frac{\hat{\mathrm{Var}}(F) - \Sigma d^2}{\hat{\mathrm{Var}}(F)} = \frac{\hat{\mathrm{Var}}(F) - \Sigma(1-h^2_i)}{\hat{\mathrm{Var}}(F)}. \quad (43)$$

Если все общности одинаковые, то из соотношения (43) вытекает (42). При заданной средней общности (или среднем коэффициенте корреляции) коэффициент надежности будет больше, когда нагрузки одинаковые. Таким образом, в табл. 11 даны оценки сверху для коэффициентов надежности при различных нагрузках.

Более серьезным является вопрос, следует ли при шкалировании фактора суммировать переменные с одинаковыми весами, если известно, что коэффициенты нагрузки не равны друг другу. Рассмотрим крайний случай. Пусть одна общность равна 1, т. е. наблюдаемая переменная полностью определяется скрытым фактором. Тогда этот фактор можно оценить одной переменной, не учитывая остальные; добавление других параметров с общностями, отличными от 1, только ухудшит оценку. Поэтому и в общем случае при факторном шкалировании нельзя просто суммировать значения переменных. Если однофакторная модель точно описывает наблюдения, оптимальная оценка относительно проста; веса, назначаемые каждой переменной, получаются из соотношения

$$B'(R^{-1}), \quad (44)$$

где B — вектор факторных нагрузок, а R — корреляционная матрица измеряемых переменных. Соотношение (44), которое выводится из регрессии фактора на переменные, обеспечивает максимальную корреляцию между F и $\hat{F}$.

$$\text{Обобщенный коэффициент надежности} = \frac{\text{var}(\hat{F}) - \Sigma(1-h_i^2)w_i^2}{\text{var}(\hat{F})}, \qquad (45)$$

где w_i — регрессионные веса, задаваемые соотношением (44). При этом дисперсия оценки $\hat{F}$ равна

$$\text{var}(\hat{F}) = \sum_i \sum_j w_i w_j r_{ij}, \qquad (46)$$

что эквивалентно суммированию всех элементов редуцированной корреляционной матрицы, причем каждый элемент r_{ij} умножается на произведение соответствующих весов w_i и w_j. На диагонали редуцированной матрицы будут стоять квадраты весов переменных. Поскольку эта величина равна R^2, она не превосходит максимальной общности Следовательно, если некоторая переменная является точным повторением скрытого фактора, ее вес будет единичным, а веса остальных — нулевыми.

Важно также отметить, что при использовании различных весов для получения оценки значения фактора переменная с большой нагрузкой часто более существенна, чем остальные переменные с малыми нагрузками. Следует помнить, что коэффициент надежности оценки не превосходит квадрата наибольшей факторной нагрузки.

ВЫБОРОЧНЫЙ РАЗБРОС И РАЗЛИЧНЫЕ КРИТЕРИИ КАЧЕСТВА ОЦЕНОК

До сих пор мы рассматривали идеализированную ситуацию, когда однофакторная модель точно соответствует данным без разброса, вызванного выборкой. В этой ситуации скрытая модель идентифицируется абсолютно точно. Если же в наблюдениях появляется разброс, связанный с выборкой, зависимости, проявляющиеся в выборке, уже не будут точно соответствовать генеральной совокупности. Даже если однофакторная модель безошибочна для генеральной совокупности, она не будет абсолютно точно воспроизводить корреляции в выборочных данных. Поэтому мы вынуждены ввести критерии близости оценок и истинных значений факторов. Существуют три таких критерия.

Регрессионный анализ

Первый критерий сводится к нахождению оценки ($\hat{F}$) значения фактора (F), доставляющей максимум коэффициента корре-

ляции между F и $\hat{F}$. В другом представлении этот критерий сводится к минимизации суммы квадратов отклонений $\Sigma(F-\hat{F})^2$. Использование этого критерия обусловливает применение регрессионного анализа. Такой подход возможен, ибо факторный анализ дает значения факторных нагрузок, которые представляют собой корреляции между факторами (подлежащими оцениванию) и наблюдаемыми переменными (выступающими здесь в роли предикторов). При этом корреляции между предикторами являются не чем иным, как наблюдаемыми корреляциями. Эти две последовательности коэффициентов корреляции и представляют исходные данные для решения системы нормальных уравнений. Оценки значений факторов задаются тогда соотношением

$$\hat{F} = X(B'R^{-1}), \tag{17}$$

где B — матрица факторных нагрузок; X — вектор наблюдаемых переменных, а R — корреляционная матрица наблюдаемых переменных. Заметим, что весовые коэффициенты определяются из заранее введенного соотношения (44). Единственное отличие заключается в том, что в выражении (47) используются наблюдаемые значения корреляционной матрицы R, а для модельных данных без ошибок наблюдаемые значения корреляций совпадают с самими корреляциями. В общем случае воспроизводимые моделью корреляции не совпадают с наблюдаемыми. Ожидаемую надежность оценки факторов получаем с помощью выражения (45).

Критерий наименьших квадратов

В однофакторной модели каждая переменная считается взвешенной суммой общих и характерных факторов:

$$X_j = b_j F + d_j u_j.$$

Предположим, что вместо F взята его оценка $\hat{F}$. Поскольку критерий наименьших квадратов определяется оценкой F, минимизирующей сумму квадратов:

$$\sum_i \sum_j (X_{ij} - b_j \hat{F})^2 \tag{48}$$

то получаем следующую оценку:

$$\hat{F} = X(BB')^{-1}B. \tag{49}$$

Отличие (49) и (47) состоит в том, что в (49) входят воспроизведенные в модели корреляции BB' вместо R. Таким образом, регрессионный анализ и критерий наименьших квадратов приводят к одним и тем же оценкам, когда выборочные корреляции совпадают с корреляциями для генеральной совокупности. В противном случае эти оценки дают отличающиеся друг от друга результаты.

Критерий Бартлетта

Для данного подхода включается в рассмотрение выборочная изменчивость. Если характерную долю дисперсии отнести на счет условных ошибок наблюдений, то лучше уменьшать вес тех переменных, которые имеют большие дисперсии ошибок. Введем следующий критерий:

$$\sum_i \sum_j (X_{ij} - \hat{b}_j F)^2 / d_j^2. \qquad (50)$$

В результате параметры с меньшими общностями получают и меньший вес. Поэтому для неодинаковых коэффициентов факторных нагрузок оценка шкалы, полученная с помощью критерия Бартлетта, отличается от двух предыдущих:

$$\hat{F} = XU^{-2}B(B'U^{-2}B)^{-1} \qquad (51)$$

где U^{-2} — диагональная матрица характерностей. Наличие U^{-2} может рассматриваться как результат взвешивания.

НЕСКОЛЬКО ОБЩИХ ФАКТОРОВ И ДОПОЛНИТЕЛЬНЫЕ СЛОЖНОСТИ

Усложним ситуацию, предположив, что имеются два и более общих фактора. Три рассмотренных критерия можно обобщить для многофакторного случая как для ортогонального, так и для косоугольного решений. Все результаты, полученные для одного фактора, справедливы и для нескольких факторов. Тем не менее тот факт, что корреляция значения фактора с его оценкой не равна 1, порождает в многомерном случае следующие вопросы: 1) будут ли факторные шкалы ортогональны друг другу, если сами скрытые факторы являются ортогональными; 2) будет ли каждая шкала коррелировать только с соответствующим ей фактором (факторная шкала называется *монохроматической*, если ее частные коэффициенты корреляции с другими факторами нулевые)? В общем случае всем этим требованиям не удовлетворяет ни одна из оценок. Факторные шкалы будут коррелировать друг с другом, даже если скрытые факторы предполагаются ортогональными; кроме того, корреляции между факторными шкалами не совпадают точно с корреляциями между косоугольными факторами. Поэтому шкала некоторого фактора будет коррелировать с другими факторами.

Однако в частном случае перечисленные требования выполняются, во-первых, когда факторная модель точно соответствует экспериментальным данным и отсутствуют выборочная изменчивость и ошибки измерений и, во-вторых, каждая переменная имеет нагрузку только на один фактор. Если выполняются эти два условия, каждый фактор или размерность можно рассматривать отдельно, причем задача сводится к однофакторной модели для данных без ошибок. Кроме того, как уже было отмечено, в этих

условиях нет неопределенности при выборе критерия для оценки шкал — все они будут эквивалентны. К сожалению, такая идеализированная ситуация практически не осуществима.

Тем не менее есть еще и другие условия, когда для некоторых факторных шкал выполняются требования ортогональности и монохроматичности. Если первоначальные факторы (до вращения) были выделены с использованием критерия максимального правдоподобия, регрессионная оценка и оценка Бартлетта для факторных шкал будут ортогональны и монохроматичны. Правда, ортогональность в скрытой факторной модели проявляется далеко не всегда. К тому же после проведения ортогонального вращения для регрессионной оценки факторных шкал уже не выполняется ни одно из этих свойств, а для оценки Бартлетта остается справедливым только условие монохроматичности, т. е. ни тот, ни другой набор шкал после вращения не будет ортогональным.

Эти обстоятельства послужили мотивом появления четвертого критерия для шкалирования, введенного Андерсеном и Рубином (Anderson, Rubin, 1956). Критерий Андерсена — Рубина является модификацией подхода Бартлетта. Минимизируется взвешенная сумма квадратов, используемая в критерии Бартлетта, при условии, что получаемые шкалы ортогональны друг другу. Соответственно, независимо от того, вращаются факторы или нет, критерий дает некоррелированные шкалы. Тем не менее последние при вращении факторов не являются монохроматическими, даже если для выделения первоначальных факторов применяется метод максимального правдоподобия.

Выбор метода шкалирования

При выборе метода необходимо проанализировать свойства получаемых шкал. Если рассматривать корреляции между скрытыми факторами и их шкалами, то регрессионный метод предпочтительнее метода Бартлетта, а метод Бартлетта в свою очередь предпочтительнее метода наименьших квадратов.

С точки зрения требования монохроматичности оценка шкал по критерию Бартлетта является наилучшей, а если брать свойство ортогональности, то предпочтительнее критерий Андерсона — Рубина. Однако, так как чаще всего заранее неизвестно, ортогональны ли скрытые факторы, выбирать следует либо регрессионный анализ, либо метод Бартлетта.

Надо еще упомянуть некоторые обстоятельства, на которые нужно обратить внимание при выборе критерия. Во-первых, как правило, все введенные шкалы сильно коррелированы, поэтому на практике обоснование предпочтительности того или иного метода имеет лишь академический интерес. Для оценивания шкал хорош любой способ (Horn, 1965; Alwin, 1973). Во-вторых, выбор метода шкалирования зависит еще и от специфики решаемой задачи. Такер (Tucker, 1971) отмечает, что, если факторные шка-

лы используются совместно с какими-то новыми, внешними* переменными, некоторые методы являются более предпочтительными. Так, он показывает, что шкалирование с помощью регрессионного анализа не позволяет правильно оценивать корреляции между скрытыми факторами и внешними переменными, в то время как остальные методы это допускают.

С другой стороны, если задача состоит только в применении факторных шкал как предикторов для значений внешних переменных, регрессионный критерий является наилучшим.

И наконец, надо иметь в виду, что все приведенные выводы относились к случаю, когда модель точно соответствует генеральной совокупности, и расхождения между моделью и экспериментальными данными вызвано лишь случайностью выборки. Что же произойдет, если такое соответствие нарушится или если факторный анализ будет использоваться лишь в качестве эвристического метода выделения кластеров в экспериментальных данных? Тогда все сказанное выше о сравнении методов может иметь второстепенное значение, а основную роль будут играть какие-то другие, не относящиеся к факторному анализу, соображения.

НЕПОЛНЫЕ ФАКТОРНЫЕ ШКАЛЫ

Есть причины, по которым имеет смысл рассмотреть *шкалы, использующие только часть информации,* получаемой из факторного анализа. Можно предположить, что факторная модель точно соответствует данным для генеральной совокупности, а заданные конкретные значения, получаемые в факторном решении, можно считать обусловленными случайностью выборки. В этом случае, пренебрегая оценками значений факторных нагрузок, разумно учитывать лишь следующее: имеет ли переменная нагрузку на данный фактор или нет. Соответственно оценка значения фактора получается суммированием только тех переменных, которые имеют значительные коэффициенты нагрузки. Остальные переменные с небольшими коэффициентами нагрузки отбрасываются. Такую шкалу будем называть неполной факторной шкалой. При использовании таких шкал следует иметь в виду два обстоятельства: 1) даже если в генеральной совокупности для некоторых переменных факторные нагрузки нулевые в факторном решении, основанном на выборке, они будут отличны от нуля; 2) даже если факторные нагрузки принимают одинаковые значения в генеральной совокупности, их оценки по выборке могут не быть таковыми. На практике часто следуют эмпирическому правилу, по которому факторные нагрузки меньше 0,3 считаются несущественными.

Правомерность применения неполных факторных шкал определяется тем, насколько хорошо выполняются отмеченные пред-

* Внешние переменные — переменные, которые не использовались при проведении факторного анализа. — *Примеч. ред.*

положения. В идеале следует проверять их с помощью конфирматорного факторного анализа. Более того, если после проверки такая простая структура матрицы нагрузок подтверждается, то использование неполной факторной шкалы становится совершенно законным. Если все же обнаружены статистически значимые отклонения, необходимо выяснить степень этих отклонений, и в любом случае малые отклонения от простой структуры можно не учитывать.

Существует еще один подход для определения правомерности применения неполных факторных шкал (эти шкалы являются самыми простыми для вычислений, но основная причина их использования не в этом). Часто факторная модель неточно описывает экспериментальные данные: 1) неслучайные ошибки измерений переменных и 2) второстепенные факторы, не представляющие интерес для целей исследования и потому, не учитываемые в модели, могут внести вклад в наблюдаемые корреляции. А это в свою очередь влияет на получаемые значения. Следовательно, есть смысл не считать окончательными конкретные величины, получаемые в результате факторного решения. Консервативная точка зрения состоит в том, чтобы рассматривать найденные с помощью факторного анализа структуры лишь как гипотезы, отражающие в экспериментальных данных некоторые тенденции к скоплению переменных в кластеры (не более того). На наш взгляд, следует считать, что полученные значения содержат определенную ошибку, и разумно игнорировать некоторый уровень отклонений.

Возможны возражения против применения неполных факторных шкал. Так, можно заметить, что такая комбинация наблюдаемых переменных не является оптимальной, т. е. другое взвешивание переменных может увеличить корреляцию между шкалами и наблюдаемыми переменными. Однако здесь можно воспользоваться тем же доводом, который приводился в пользу простого суммирования: множественный коэффициент корреляции между шкалой (линейной комбинацией наблюдаемых переменных) и всем набором наблюдаемых переменных мало изменяется при небольших отклонениях в весах (Wang, Stanley, 1970; Wainer, 1976). Здесь следует сделать одно предостережение, о котором уже говорилось в разд. III. Если известно, что факторная модель точно соответствует данным, нельзя отбрасывать высокие факторные нагрузки (например, порядка 0,9) и приписывать им такое же значение, как и небольшим нагрузкам.

Итак, по-видимому, как обычные, так и неполные факторные шкалы имеют право на существование и могут использоваться на практике.

ЗНАЧЕНИЯ ГЛАВНЫХ КОМПОНЕНТ

Сделаем несколько замечаний по поводу шкал, соответствующих главным компонентам. Как уже отмечалось, принцип глав-

Таблица 12

Результат применения конфирматорного факторного анализа к корреляционной матрице, представленной наддиагональными элементами табл. 1 при использовании модели в табл. 7[1]

Переменные	Фактор		Общность
	F_1	F_2	h_2
X_1	0,792	0	0,624
X_2	0,756	0	0,571
X_3	0,663	0	0,501
X_4	0	0,577	0,333
X_5	0	0,669	0,448
X_6	0	0,635	0,404
$r_{F_1F_2}=0{,}501$			
$\chi^2=4{,}6534$ $df=8$ Вероятность=0,7939			

[1] Эти результаты получены с помощью программы LISREL III, а не COFAMM.

ных компонент отличается от принципа введения факторной модели. Поэтому ни один из этих подходов не может подменять другой. На практике применяются оба подхода. В некоторых задачах значения главных компонент могут быть предпочтительнее, чем факторные шкалы, в особенности если необходимо только сжать информацию, содержащуюся в данных, и факторная структура для этого не нужна. Именно поэтому стоит уделить этому вопросу немного внимания.

Как мы уже знаем, главные компоненты являются математическими функциями измеряемых переменных. Таким образом, компоненты можно непосредственно представлять в виде линейной комбинации переменных и говорить о значении компонент, а не об их оценках. Значения компонент получаются при суммировании значений переменных с весами, пропорциональными компонентным нагрузкам:

$$\text{Значение компоненты}=\sum_j [(b_{ij}/\lambda_i)X_j], \qquad (52)$$

где b_{ij} — нагрузка на j-ю переменную от i-й компоненты; λ_i — соответствующее собственное значение. Деление на собственное значение приводит к тому, что значение компоненты будет иметь единичную дисперсию.

ПРИМЕЧАНИЕ

1. Если скрытая структура является сложной, как, например, для бокс-проблемы Терстоуна (см. Иберла, 1980), трудно точно восстановить скрытую структуру из ковариационной матрицы только на основании какого-то аналитического критерия. Для этого могут потребоваться аппроксимация гиперплоскостями и применение визуальных вращений.

VII. КРАТКИЕ ОТВЕТЫ НА ЧАСТО ВОЗНИКАЮЩИЕ ВОПРОСЫ

ПРИРОДА ПЕРЕМЕННЫХ И ИХ ИЗМЕРЕНИЕ

а) *Какой способ измерений необходим в факторном анализе?*

В факторном анализе требуется, чтобы переменные измерялись по крайней мере на уровне шкалы интервалов (Stevens, 1946). Это требование обусловлено тем, что входной информацией для факторного анализа являются элементы ковариационной матрицы. Кроме того, представление переменных в виде линейной комбинации скрытых факторов и использование оценок факторов через линейные комбинации наблюдаемых переменных для порядковых переменных невозможны.

б) *Возможно ли использование тау-статистики Кендалла или гамма-статистики Гудмана и Крускала вместо обычных корреляций?*

Нет, невозможно. Как уже отмечалось, операции сложения для порядковых переменных не определены, поэтому не существует факторных моделей с порядковыми статистиками. Допускается лишь эвристическое использование таких моделей без статистической интерпретации результатов. (Существуют некоторые неметрические методы шкалирования, специально разработанные для оперирования с нечисловыми переменными.)

в) *Должен ли исследователь, учитывая данные выше ответы, всегда избегать использования факторного анализа в случаях, когда метризуемость пространства переменных не вполне ясна?*

Не обязательно. Многие переменные, такие, как меры отношений и мнений в социологии, различные переменные при обработке результатов тестирования, не имеют точно определенной метрической основы. Тем не менее часто предполагается, что порядковым переменным можно давать числовые значения, не нарушая их внутренних свойств. Окончательный ответ на этот вопрос основан на двух моментах: 1) насколько хорошо вспомогательные числовые значения отражают скрытые истинные расстояния и 2) велико ли искажение, вносимое в корреляции между параметрами (являющимися входными данными в факторном анализе) при введении шкалирования. К счастью, коэффициенты корреляции обладают свойством робастности по отношению к порядковым искажениям в измеряемых данных (Labovitz, 1967, 1970; Kim, 1975). Поэтому, если искажения корреляций, вносимые при шкалировании порядковых переменных, не слишком велики, вполне законно использовать эти переменные в качестве числовых. Тем не менее следует быть готовыми к появлению пусть даже незначительных, систематических ошибок в факторном решении.

г) *Расскажите о дихотомических переменных. Существует мнение, что факторный анализ вполне применим для таких переменных, во-первых, поскольку при использовании дихотоми-*

ческих переменных не требуется предположение об измерениях и, во-вторых, поскольку φ (фи), равное коэффициенту корреляции Пирсона, является адекватной мерой зависимости для факторного анализа. Поэтому, возможно ли применение факторного анализа к матрице значений φ?

Нет. Дихотомические переменные нельзя представить в рамках факторной модели. Действительно, вспомним о предположении, что каждая переменная является взвешенной суммой по крайней мере двух скрытых факторов (одного общего и одного характерного). Даже если эти факторы принимают лишь 2 значения (что вряд ли встретится на практике), наблюдаемая переменная будет принимать уже 4 возможных значения. Следовательно, никакие соображения, кроме чисто эвристических, не могут обосновать применение факторного анализа к дихотомическим переменным.

д) *Ответ на предыдущий вопрос озадачивает. Поскольку мы обычно предполагаем факторную модель непрерывной, следует ожидать и непрерывности измеряемых переменных. Однако переменные, с которыми мы имеем дело на практике, часто принимают лишь весьма ограниченный набор значений — да или нет; согласие или несогласие; в лучшем случае — целиком согласен, согласен, безразличен, полностью не согласен и т. д. Означает ли это, что мы применяем факторный анализ к данным, которые с ним не согласуются?*

В некотором смысле — да. Переменные, принимающие ограниченный набор значений, строго говоря, несовместимы с факторной моделью. Если предположить, что наблюдаемые переменные представляют собой результаты неточных измерений или результаты, полученные при объединении в одну группу близких значений, вопрос будет состоять не в том, применима ли факторная модель к данным, а в том, насколько неслучайные ошибки измерений искажают результаты факторного анализа.

Группирование близких значений, безусловно, сказывается на корреляциях, но степень этого влияния зависит от законов распределений, шага дескритизации и т. д. Тем не менее имеются некоторые обнадеживающие соображения по поводу использования факторного анализа как эвристического метода при наличии больших ошибок измерений (см. следующий вопрос).

е) *В каких случаях возможно применение факторного анализа к данным, содержащим дихотомические переменные или переменные с конечным множеством значений?*

В общем случае, чем шире множество значений, тем точнее результаты. В случае дихотомических переменных использование коэффициента φ может быть оправдано, если решается задача нахождения кластеров переменных и если корреляции между исходными переменными невелики*, скажем, не превосходят 0,6

* Здесь предполагается, что существуют некоторые скрытые переменные, порождающие наблюдаемые дихотомические переменные. Последние получаются делением интервалов значений этих скрытых переменных на 2 части. — *Примеч. ред.*

или 0,7. При переходе от непрерывных переменных к дихотомическим переменным корреляции уменьшаются. При этом на величину уменьшения влияет выбор точек деления. Если корреляции не очень велики, эффект, связанный с выбором точек деления, не столь значителен. Таким образом, группирование (дихотомизация) переменных в целом уменьшает корреляции между ними, но не влияет на кластерную структуру данных, поскольку факторный анализ основан на относительной величине корреляций. Если цель исследования состоит в нахождении кластерной структуры, использование факторного анализа оправдано (Kim, Nie, Verba, 1977).

ж) *Если отклонения, возникающие в решении из-за введения точек деления более значительны, чем отклонения, связанные с уменьшением корреляций при группировании, то почему бы не использовать относительные величины* φ/φ_{max} *или* R/R_{max} *вместо* φ *и* R*?*

Такой подход целесообразен только в том случае, когда распределение имеет какую-то особую (негауссову) форму (Carrol, 1961) или когда непрерывные переменные связаны функциональной зависимостью. В последнем случае не нужно применять факторный анализ. Поэтому данный подход нерационален (Kim, Nie, Verba, 1977).

з) *Существуют ли какие-либо более прямые методы решения этих задач?*

В литературе предложены два подхода. В каждом из них предполагается, что переменные, принимающие два либо несколько значений, являются индикаторными переменными для скрытых непрерывных переменных, к которым, безусловно, применима факторная модель. Соответственно для нахождения факторной структуры необходимо определить корреляции между скрытыми переменными. Первый путь связан с использованием тетрахорических корреляций вместо φ. Этот подход является эвристическим, поскольку вычисление таких корреляций не всегда возможно, и корреляционная матрица может не быть матрицей Грама (Bock, Lieberman, 1970). Другой подход непосредственно применяет скрытое многомерное распределение вместо вычисления тетрахорических корреляций исходя из двумерных таблиц. Данный метод является многообещающим, однако требует чрезмерно большого объема вычислений даже для современных компьютеров (Christoffersson, 1975)*.

ИСПОЛЬЗОВАНИЕ КОРРЕЛЯЦИОННЫХ ЛИБО КОВАРИАЦИОННЫХ МАТРИЦ

а) *Имеет ли значение, какую матрицу использовать — ковариационную или корреляционную?*

* Подход на основе *множественного анализа соответствий* описан в кн. М. Жамбю «Иерархический кластер-анализ». — М.: Финансы и статистика, 1988. — *Примеч. ред.*

Это зависит: 1) от того, имеются ли сравнимые метрики в пространстве переменных; 2) от применяемого метода выделения и 3) от того, есть ли необходимость в сравнении одной факторной структуры с другой. Если рассматривается только одна выборка (группа) и используется независимый от масштаба метод выделения, например, такой, как метод максимального правдоподобия, альфа-факторный анализ или анализ образов, то не имеет значения, какой матрицей воспользоваться, при условии, что необходимо идентифицировать соответствующие скрытые размерности (факторы). Если применяется ковариационная матрица и единицы измерения в значительной степени неоднородны, факторные шкалы будет сложно интерпретировать. Поэтому в случаях, когда дисперсии переменных существенно отличаются одна от другой и имеются разнородные единицы измерения, разумно использовать корреляционную матрицу. (Например, один параметр может измеряться в долларах, другой — в количестве лет, а третий — по шкале Ликерта.) Применение корреляционных матриц рекомендуется с практической точки зрения — некоторые компьютерные программы не допускают задания ковариационных матриц, и, кроме того, большинство примеров, приведенных в литературе, основано на матрицах корреляций.

б) *Когда использование ковариационных матриц предпочтительнее?*

Ковариационные матрицы предпочтительнее, когда производится сравненне факторных структур для различных выборок. Дело в том, что корреляционная матрица получается при масштабировании переменных с применением выборочных средних и дисперсий. По этой прнчине даже теоретически инвариантные параметры могут меняться от выборки к выборке. Обсуждение всевозможных осложнений, связанных с введением переменных в стандартной форме, приводится в работах (Kim, Mueller, 1979), а также (Sörbom and Jöreskog, 1976).

в) *Что делать, если задача состоит в сравнении факторных структур для различных выборок, и переменные измеряются в неодинаковых единицах?*

Один из методов заключается в нормировке переменных, т. е. в приведении их к стандартной форме, используя средние и дисперсии, вычисленные по совокупности выборок. Затем может быть вычислена ковариационная матрица для каждой выборки в отдельности. Этот подход отличен от получения корреляционной матрицы по одной выборке, когда переменные в каждой группе преобразуются с использованием частных выборочных средних и дисперсий.

КРИТЕРИИ ЗНАЧИМОСТИ И УСТОЙЧИВОСТЬ ФАКТОРНЫХ РЕШЕНИЙ

а) *В каких случаях используется метод максимального правдоподобия и связанные с ним критерии значимости, и каков минимальный объем выборки?*

Чем больше объем выборки, тем точнее χ^2-аппроксимация. Лоули и Максвелл (Lawley and Maxwell, 1971) считают, что этот критерий применим, когда выборка содержит по крайней мере на 51 наблюдение больше, чем рассматриваемое число переменных. Другими словами, это условие имеет вид $N-n-1 \geqslant 50$, где N — объем выборки, а n — число переменных. Разумеется, это — всего лишь эмпирическое правило.

б) *Сколько переменных должно приходиться на один гипотетический фактор?*

Тэрстоун считает, что на один фактор должно приходиться по крайней мере три переменные. Для конфирматорного факторного анализа эта пропорция, очевидно, меньше. Исследователи в целом сходятся на том, что переменных должно быть по меньшей мере вдвое больше, чем факторов. Минимальное число переменных для использования критерия значимости приводится в табл. 11 в разд. VI.

в) *Всегда ли необходимо предположение о многомерной нормальности закона распределения параметров?*

Сама по себе факторная модель не требует такого предположения. Например, возможно построить факторную модель, в которой факторы принимают значения 0 и 1. Однако в методе максимального правдоподобия и связанном с ним критерии значимости предположение о нормальности существенно. В общем случае, последствия нарушения этого допущения не вполне ясны.

ДРУГИЕ СТАТИСТИЧЕСКИЕ ВОПРОСЫ

а) *Что означает знак факторных нагрузок?*

Сам по себе знак не имеет внутреннего содержания и не несет информации о зависимости между переменной и фактором. Однако следует сопоставлять между собой знаки для различных переменных при данном факторе. Разумно перед применением факторного анализа так задать переменные, чтобы знаки коэффициентов нагрузок на данный фактор были одинаковы.

б) *Что означают собственные значения, связанные с факторами, полученными после вращения? Какова роль доли дисперсии, воспроизводимой каким-либо из этих факторов?*

Собственные значения, связанные с факторами до вращения, не совпадают с соответствующими величинами для вращаемых факторов; неизменна только сумма собственных значений. В первоначальном факторном решении величина собственного значения несет информацию об относительной важности каждого фактора. Для факторного решения после вращения это свойство не сохраняется. Поскольку в результате вращения определяются совсем другие факторы, не важно, какую долю дисперсии воспроизводит каждый из них.

в) *Возможно ли в факторном анализе, используя зависимости между факторными шкалами, получить факторное решение более «высокого» порядка?*

Нет. Корреляции между факторными шкалами не совпадают с корреляциями между скрытыми факторами. Для получения факторных решений более высокого порядка следует применять корреляционную матрицу, полученную в результате косоугольного вращения.

г) *Можно ли утверждать, что скрытая факторная структура является ортогональной, если экспериментальные данные не противоречат такому решению?*

Нет. Ортогональность вносится исследователем. Однако если ортогональность проявляется после косоугольных вращений или если графическое представление показывает, что скопления переменных составляют прямые углы, то свойство ортогональности, по-видимому, присуще скрытой структуре.

д) *Можно ли включать в анализ переменные, некоторые из которых являются причинными для других? Иначе говоря, необходимо ли, чтобы все переменные были на одном уровне причинности?*

В общем случае, переменные не должны быть причинными для других. В факторной модели предполагается, что все наблюдаемые переменные являются функциями скрытых факторов. Однако при достаточном опыте можно применять факторный анализ к причинным системам с более сложной структурой (Stinchcombe, 1971).

СПЕЦИАЛЬНАЯ ЛИТЕРАТУРА И КОМПЬЮТЕРНЫЕ ПРОГРАММЫ

а) *Есть ли книги или статьи по факторному анализу, доступные начинающим?*

По-видимому, нет. Большинство публикаций требует определенной технической подготовки. Вот наиболее простые работы: (Rummel, 1967; Schuessler, 1971; Cattell, 1952; Comrey, 1973; Fruchter, 1954).

б) *Какие книги предназначены для последующего, более глубокого изучения?*

(Harman, 1976; Mulaik, 1972; Lawley and Maxwell, 1971).

в) *В каких журналах регулярно публикуются работы по факторному анализу?*

Psychometrika; British Journal of Mathematical and Statistical Psychology; Educational and Psychological Measurement.

г) *Какие существуют пакеты прикладных программ, содержащие программы по факторному анализу?*

SPSS; OSIRIS; SAS; BMD.

д) *Есть ли какие-либо специализированные программы для задач факторного анализа?*

Little Jiffy, Mark IV (Kaiser, 1974); Cofamm (Sörbom, Jöreskog, 1976).

е) *Где можно прочесть об основных результатах по моделированию?*

Tucker, Koopman; Linn (1969); Browne (1968); Linn (1968); Hakstian (1971); Hakstian and Abell (1974).

ЛИТЕРАТУРА

ALWIN, D.F. (1973) "The use of factor analysis in the construction of linear composites in social research." Sociological Methods and Research 2:191-214.

ANDERSON, T.W. and H. RUBIN (1956) "Statistical inference in factor analysis." Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 5:111-150.

ASHER, H. (1976) Causal Modeling. Sage University Papers on Quantitative Applications in the Social Sciences, 07-003. Beverly Hills and London: Sage Pub.

BMDP-77: Biomedical Computer Programs (P-Series). W.J. Dixon, Series Editor, M.B. Brown, Editor 1977 edition. Los Angeles: Univ. of California Press, 1977.

BARGMANN, R.E. (1957) A Study of Independence and Dependence in Multivariate Normal Analysis. Mimeo Series No. 186. Chapel Hill, N.C.: Institute of Statistics.

BARTLETT, M.S. (1937) "The statistical conception of method factors." British Journal of Psychology 28:97-104.

BOCK, R.D. and R.E. BARGMANN (1966) "Analysis of covariance structure." Psychometrika 31:507-534.

BOCK, R.D. and M. LIEBERMAN (1970) "Fitting a response model for N dichotomously scored items." Psychometrika 26:347-372.

BOCK, R.D. and A.C. PETERSON (1975) "A multivariate correction for attenuation." Biometrika 62:673-678.

BROWNE, M.W. (1968) "A comparison of factor analytic techniques." Psychometrika 33:267-334.

COFAMM: Confirmatory Factory Analysis with Model Modification User's Guide. Sörbom, D. and Jöreskog, K.G. Chicago: National Educational Resources, Inc., 1976.

CARROLL, J.B. (1953) "Approximating simple structure in factor analysis." Psychometrika 18:23-38.

CARROLL, J.B. (1961) "The nature of data, or how to choose a correlation coefficient." Psychometrika 26:347-372.

CATTELL, R.B. (1952) Factor Analysis. New York: Harper and Bros.

CATTELL, R.B. (1965) "Factor analysis: an introduction to essentials. (I) the purpose and underlying models, (II) the role of factor analysis in research." Biometrics 21:190-215, 405-435.

CATTELL, R.B. (1966) Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally.

CATTELL, R.B. and J.L. MUERLE (1960) "The 'maxplane' program for factor rotation to oblique simple structure." Educational and Psychological Measurement 20:269-290.

CHRISTOFFERSSON, A. (1975) "Factor analysis of dichotomized variables." Psychometrika 40:5-32.

COMREY, A.L. (1973) A First Course in Factor Analysis. New York: Academic Press.

CRONBACH, L.J. (1951) "Coefficient alpha and the internal structure of tests." Psychometrika 16:297-334.

DUNCAN, O.D. (1966) "Path analysis: sociological examples." American Journal of Sociology 72:1-16.

EBER, H.W. (1966) "Toward oblique simple structure maxplane." Multivariate Behavioral Research 1:112-125.

FRUCHTER, B. (1954) Introduction to Factor Analysis. New York: Van Nostrand.

GREEN, B.F., Jr. (1976) "On the factor score controversy." Psychometrika 41:263-266.

GUILFORD, J.P. (1977) "The invariance problem in factor analysis." Educational and Psychological Measurement 37:11-19.
GUTTMAN, L. (1953) "Image theory for the structure of quantitative variates." Psychometrika 18:227-296.
GUTTMAN, L. (1954) "Some necessary conditions for common factor analysis." Psychometrika 19:149-161.
HAKSTIAN, A.R. (1971) "A comparative evaluation of several prominent methods of oblique factor transformation." Psychometrika 36:175-193.
HAKSTIAN, A.R. and R.A. ABELL (1974) "A further comparison of oblique factor transformation methods." Psychometrika 39:429-444.
HARMAN, H.H. (1976) Modern Factor Analysis. Chicago: University of Chicago Press.
HARMAN, H.H. (in press) "Minres method of factor analysis," in K. Enstein, A. Ralston, and H.S. Wilf (eds.) Statistical Methods for Digital Computers. New York: John Wiley.
HARMAN, H.H. and W.H. JONES (1966) "Factor analysis by minimizing residuals (Minres)." Psychometrika 31:351-368.
HARMAN, H.H. and Y. FUKUDA (1966) "Resolution of the Heywood case in the Minres solution." Psychometrika 31:563-571.
HARRIS, C.W. (1962) "Some Rao-Guttman relationships." Psychometrika 27:247-263.
HARRIS, C.W. (1967) "On factors and factor scores." Psychometrika 32:363-379.
HARRIS, C.W. and H.F. KAISER (1964) "Oblique factor analytic solutions by orthogonal transformations." Psychometrika 29:347-362.
HENDRICKSON, A.E. and P.O. WHITE (1964) "Promax: A quick method for rotation to oblique simple structure." British Journal of Mathematical and Statistical Psychology 17:65-70.
HORN, J.L. (1965) "An empirical comparison of various methods for estimating common factor scores." Educational and Psychological Measurement 25:313-322.
HORST, P. (1965) Factor Analysis of Data Matrices. New York: Holt Rinehart and Winston.
HOTELLING, H. (1933) "Analysis of a complex of statistical variables into principal components." Journal of Education Psychology 24:417-441, 498-520.
HOWE, W.G. (1955) Some Contributions to Factor Analysis. Report No. ORNL-1919. Oak Ridge, Tenn.: Oak Ridge National Laboratory. Ph.D. dissertation, University of North Carolina.
JENNRICH, R.I. (1970) "Orthogonal Rotation Algorithms." Psychometrika 35:229-235.
JENNRICH, R.I. (1974) "Simplified formulae in standard errors in maximum likelihood factor analysis." British Journal of Mathematical and Statistical Psychology 27:122-131.
JENNRICH, R.I. and P.F. SAMPSON (1966) "Rotation for simple loadings." Psychometrika 31:313-323.
JÖRESKOG, K.G. (1963) Statistical Estimation in Factor Analysis: A New Technique and Its Foundation. Stockholm: Almquist and Wiksell.
JÖRESKOG, K.G. (1966) "Testing a simple structure hypothesis in factor analysis." Psychometrika 31:165-178.
JÖRESKOG, K.G. (1967) "Some contributions to maximum likelihood factor analysis." Psychometrika 32:443-482.
JÖRESKOG, K.G. (1969) "A general approach to confirmatory maximum likelihood factor analysis." Psychometrika 34:183-202.

JÖRESKOG, K.G. (1970) "A general method for analysis of covariance structure." Biometrika 57:239-251.
JÖRESKOG, K.G. (1976) Analyzing Psychological Data by Structural Analysis of Covariance Matrices. Research Report 76-9. University of Uppsala, Statistics Department.
JÖRESKOG, K.G. and D.N. LAWLEY (1968) "New methods in maximum likelihood factor analysis." British Journal of Mathematical and Statistical Psychology 21:85-96.
KAISER, H.F. (1958) "The varimax criterion for analytic rotation in factor analysis." Psychometrika 23:187-200.
KAISER, H.F. (1963) "Image analysis," pp. 156-166 in C.W. Harris (ed.) Problems in Measuring Change. Madison: University of Wisconsin Press.
KAISER, H.F. (1970) "A second-generation Little Jiffy." Psychometrika 35:401-415.
KAISER, H.F. (1974) "Little Jiffy, Mark IV." Educational and Psychological Measurement 34:111-117.
KAISER, H.F. (1974) "An index of factorial simplicity." Psychometrika 39:31-36.
KAISER, H.F. and J. CAFFREY (1965) "Alpha factor analysis." Psychometrika 30: 1-14.
KIM, J.O. (1975) "Multivariate analysis of ordinal variables." American Journal of Sociology 81:261-298.
KIM, J.O. and C.W. MUELLER (1976) "Standardized and unstandardized coefficients in causal analysis: An expository note." Sociological Methods and Research 4:423-438.
KIM, J.O., N. NIE, and S. VERBA (1977) "A note on factor analyzing dichotomous variables: the case of political participation." Political Methodology 4:39-62.
KIRK, D.B. (1973) "On the numerical approximation of the bivariate normal (tetrachoric) correlation coefficient." Psychometrika 38:259-268.
LISREL III: Estimation of Linear Structural Equation Systems by Maximum Likelihood Methods. (User's Guide). Jöreskog, K.G. and Sörbom, D. Chicago: National Educational Resources, Inc., 1976.
LITTLE JIFFY, MARK IV. (See Kaiser, 1974).
LABOVITZ, S. (1967) "Some observations on measurement and statistics." Social Forces 46:151-160.
LABOVITZ, S (1970) "The assignment of numbers to rank order categories." American Sociological Review 35:515-524.
LAND, K.O. (1969) "Principles of path analysis," pp. 3-37 in E.F. Borgatta (ed.) Sociological Methodology. San Francisco: Jossey-Bass.
LAWLEY, D.N. (1940) "The estimation of factor loading by the method of maximum likelihood." Proceedings of the Royal Society of Edinburgh 60:64-82.
LAWLEY, D.N. and MAXWELL, A.E. (1971) Factor Analysis as a Statistical Method. London: Butterworth and Co.
LEVINE, M.S. (1977) Canonical Analysis and Factor Comparison. Sage University Papers on Quantitative Applications in the Social Sciences, 07-006. Beverly Hills and London: Sage Pub.
LI, C.C. (1975) Path Analysis–A Primer. Pacific Grove, Calif.: Boxwood Press.
LINN, R.L. (1968) "A Monte Carlo approach to the number of factors problems." Psychometrika 33:37-71.
LORD, F.M. and W.R. NOVICK (1968) Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

MALINVAND, E. (1970) Statistical Methods of Econometrics. New York: Elsevier.
MAXWELL, A.E. (1972) "Thomson's sampling theory recalled." British Journal of Mathematical and Statistical Psychology 25:1-21.
McDONALD, R.P. (1970) "The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis." British Journal of Mathematical and Statistical Psychology 23:1-21.
McDONALD, R.P. (1974) "The measurement of factor indeterminacy." Psychometrika 39:203-221.
McDONALD, R.P. (1975) "Descriptive axioms for common factor theory, image theory and component theory." Psychometrika 40:137-152.
McDONALD, R.P. (1975) "A note on Rippe's test of significance in common factor analysis." Psychometrika 40:117-119.
McDONALD, R.P. and E.J. BURR (1967) "A comparison of four methods of constructing factor scores." Psychometrika 32:380-401.
MULAIK, S.A. (1972) The Foundations of Factor Analysis. New York: McGraw-Hill.
NEUHAUS, J.O. and C. WRIGLEY (1954) "The method: an analytic approach to orthogonal simple structure." British Journal of Mathematical and Statistical Psychology 7:81-91.
OSIRIS Manual. Ann Arbor, Mich.: Inter-University Consortium for Political Research, 1973.
RAO, C.R. (1955) "Estimation and test of significance in factor analysis." Psychometrika 20:93-111.
RUMMEL, R.J. (1967) "Understanding factor analysis." Conflict Resolution 11:444-480.
RUMMEL, R.J. (1970) Applied Factor Analysis. Evanston: Northwestern University Press.
SAS: A User's Guide to SAS 76. Anthony J. Barr, James H. Goodnight, John P. Sall, and Jane T. Helwig. Raleigh, N.C.: SAS Institute, Inc., 1976.
SPSS: Statistical Package for the Social Sciences. Norman H. Nie, C. Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner, and Dale Bent. New York: McGraw-Hill, 1975.
SAUNDERS, D.R. (1953) An Analytic Method for Rotation to Orthogonal Simple Structure. Research Bulletin 53-10. Princeton, N.J.: Educational Testing Service.
SAUNDERS, D.R. (1960) "A computer program to find the best-fitting orthogonal factors for a given hypothesis." Psychometrika 25:199-205.
SCHUESSLER, K. (1971) Analyzing Social Data. Boston: Houghton Mifflin.
SÖRBOM, D. and K.G. JÖRESKOG (1976) COFAMM: Confirmatory Factor Analysis with Model Modification User's Guide. Chicago: National Educational Resources, Inc.
STEPHENSON, W. (1953) The Study of Behavior. Chicago: The University of Chicago Press.
STEVENS, S.S. (1946) "On the theory of scales of measurement." Science 103:677-680.
STINCHCOMBE, A.L. (1971) "A heuristic procedure for interpreting factor analysis." American Sociological Review 36:1080-1084.
THOMPSON, G.H. (1934) "Hotelling's method modified to give Spearman's g." Journal of Educational Psychology 25:366-374.
THURSTONE, L.L. (1947) Multiple Factor Analysis. Chicago: University of Chicago Press.

TRYON, C.R. and BAILEY, D.E. (1970) Cluster Analysis. New York: McGraw-Hill.
TUCKER, L.R. (1966) "Some mathematical notes on three mode factor analysis." Psychometrika 31:279-311.
TUCKER, L.R. (1971) "Relations of factor score estimates to their use." Psychometrika 36:427-436.
TUCKER, L.R., R.F. KOOPMAN, and R.L. LINN (1969) "Evaluation of factor analytic research procedures by means of simulated correlation matrices." Psychometrika 34:421-459.
TUCKER, L.R. and C. LEWIS (1973) "A reliability coefficient for maximum likelihood factor analysis." Psychometrika 38:1-8.
VELICER, W.F. (1975) "The relation between factor scores, image scores, and principal component scores." Educational and Psychological Measurement 36:149-159.
WAINER, H. (1976) "Estimating coefficients in linear models: it don't make no nevermind." Psychological Bulletin 83:213-217.
WANG, M.W. and J.C. STANLEY (1970) "Differential weighing: a review of methods and empirical studies." Review of Educational Research 40:663-705.

ГЛОССАРИЙ

Альфа-факторный анализ (alpha factoring): метод получения первоначального факторного решения, в котором переменные считаются выборкой из генеральной совокупности переменных; описан в работе (Kaiser and Caffrey, 1965).

Биквартимин (biquartimin criterion): критерий, применяемый при получении косоугольного решения.

Варимакс (varimax): метод получения ортогоиального решения, который сводится к упрощению факторной структуры с использованием критерия минимизации дисперсии столбца матрицы факторного отображения.

Вторичные оси (reference axes): оси, ортогональные первичным факторам; вводятся для упрощения косоугольного вращения.

Выделение факторов (extraction of factors): первоначальный этап факторного анализа; ковариационная матрица воспронзводится посредством небольшого числа скрытых факторов или компонент.

Главные компоненты (principal components): линейная комбинация наблюдаемых переменных, обладающая свойством ортогональности; первая главная компонента воспроизводит наибольшую долю дисперсии экспериментальных данных; вторая — следующую по величине долю и т. д.; главные компоненты часто считаются общими факторами, но более корректно предположение, что они противоположны им, поскольку общие факторы являются гипотетическими.

Главиых осей метод (principal axis factoring): метод получения первоначального факторного решеиия, при использовании которого редуцированная корреляционная матрица подвергается последовательной декомпозиции; метод главных осей с итерацнями по общности эквивалентен методу наимеиьших квадратов.

Грама матрнца (Gramian): квадратная симметрическая матрица, все собственные числа которой неотрицательны; корреляционная (нередуцированная) и ковариациоииая матрицы являются матрицами Грама.

Детерминант (determinant): характеристика квадратной матрицы; используется при определении ранга (числа независимых строк или столбцов) редуцированной корреляционной матрицы.

Дисперсия (variance): мера разброса параметра; определяется как сумма квадратов отклонений от среднего, деленная на число значений.

Значение фактора (factor score): оценка скрытого фактора в терминах наблюдаемых переменных; в факторном анализе имеет второстепенное значение.

Кайзера критерий (Kaiser criterion): критерий определения числа выделяемых факторов; предложен Гуттманом и Кайзером; также известен, как критерий «собственных чисел, больших 1».

Квартимакс (quartimax): критерий получения ортогонального решения; сводится к упрощению описания строк матрицы факторного отображения.

Квартимин (quartimin): критерий получения косоугольного решения; минимизируется тот же функционал, что и в критерии квартимакс без наложения ограничения ортогональности; требует введения вторичных осей.

Ковариаций анализ (covariance-structure analysis): метод анализа, в котором: 1) наблюдаемые коэффициенты ковариации описываются в рамках общей модели, включающей гипотетические факторы и наблюдаемые переменные; 2) исследователь затем определяет соответствующие значения, оценивая адекватность этого определения по отношению к структуре выборочных ковариаций.

Ковариации коэффициент (covariance): мера зависимости между двумя переменными; равен ковариации, деленной на число наблюдаемых значений; среднее значение сумм попарных произведений отклонений значений переменных от их среднего; для переменных в стандартной форме равен коэффициенту корреляции.

Ковариация (covariation): мера зависимости двух переменных; измеряется как сумма попарных произведений отклонений переменных от их среднего; используется как общий термин для описания зависимости между переменными.

Коваримин (covarimin): критерий, применяемый для получения косоугольного решения.

Конфирматорный факторный анализ (confirmatory factor analysis): факторный анализ, в котором проверяются гипотезы о числе факторов и их нагрузках.

Корреляция (correlation): мера зависимости между двумя переменными; обычно используется коэффициент корреляции Пирсона r, который равен ковариации двух параметров в стандартной форме; используется как общий термин для любого вида линейной зависимости между переменными; отметим, что переменная в стандартной форме имеет нулевое математическое ожидание и единичную дисперсию.

Косоугольное вращение (oblique rotation): преобразование, с помощью которого получается простая структура; факторы вращаются без наложения условия ортогональности, и результирующие факторы, вообще говоря, коррелируют друг с другом.

Косоугольные факторы (oblique factors): факторы, которые коррелируют друг с другом; получаются в результате косоугольного вращения.

Линейная комбинация (linear combination): сумма, в которую переменные входят с постоянными весами.

Линейная система (linear system): линейная зависимость между переменными; в факторном анализе — модель, в которой измеряемые величины линейно связаны со скрытыми факторами.

Максимального правдоподобия метод (maximum likelihood): метод статистического оценивания, в котором определяется значение переменных генеральной совокупности с использованием выборочного распределения; в факторном анализе — метод получения первоначального факторного решения, его варианты включают канонический факторный анализ и метод минимизации определителя матрицы остаточных коэффициентов корреляции.

Математическое ожидание (expectation): среднее значение случайной величины, определяемое как для дискретных, так и для непрерывных законов ее распределения; математическое ожидание является характеристикой данной величины.

Модельные данные (error-free data): данные, для которых скрытая факторная структура предполагается известной и достигается точное соответствие данных и модели.

Монте-Карло метод (Monte Carlo experiment): методика статистического моделирования выборочных характеристик.

Наименьших квадратов метод (least-squares solution): решение, для которого минимизируется сумма квадратов отклонений между наблюдаемыми и предполагаемыми значениями; в факторном анализе — метод получения первоначального факторного решения, варианты которого включают метод главных осей с итерациями по общностям и метод минимальных остатков.

Облимакс (oblimax): критерий получения косоугольного решения; эквивалентен критерию квартимакс при ортогональном вращении.

Облимин (oblimin): общий критерий получения косоугольного решения, для которого матрица отображения упрощается с использованием вторичных осей; его варианты включают критерии биквартимин, коваримин и квартимин.

Образов анализ (image factoring): метод получения первоначального факторного решения; наблюдаемая переменная представляется в виде образа и антиобраза вместо общей и характерной частей.

Общая часть (common part): часть наблюдаемой переменной, связанной с общими факторами.

Общий фактор (common factor): неизмеряемая (гипотетическая) скрытая величина, которая учитывает корреляцию по крайней мере между двумя наблюдаемыми переменными.

Общность (communality): доля дисперсии наблюдаемых переменных, обусловленная общими факторами; в модели с ортогональными факторами она равна сумме квадратов факторных нагрузок.

Ортогональное вращение (orthogonal rotation): преобразование, с помощью которого получается простая структура при выполнении ограничения ортого-

нальности (некоррелированности) факторов; факторы, выделяемые с помощью этого вращения по определению, некоррелированы.

Ортогональные факторы (orthogonal factors): факторы, которые не коррелируют друг с другом; получаются при ортогональном вращении.

Отсеивания критерий (scree-test): эвристический критерий определения числа факторов; основан на графическом изображении всех собственных значений корреляционной матрицы; применим при влиянии второстепенных (незначимых) факторов.

Ошибки дисперсия (error component): часть дисперсии наблюдаемой переменной, связанной с несовершенством измерений; входит в характерность.

Простая структура (simple structure): специальный термин, относящийся к факторной структуре, которая обладает определенными свойствами простоты: некоторые из этих свойств сводятся к тому, что переменные должны иметь нагрузку на минимальное число общих факторов, каждый общий фактор должен нагружать некоторые переменные и не нагружать остальные.

Прямой облимин (direct oblimin): метод получения косоугольного решения, в котором вращение выполняется без использования вторичных осей.

Разведочный факторный анализ (exploratory factor analysis): факторный анализ, который используется при исследовании скрытой факторной структуры без предположения о числе факторов и их нагрузок.

Ранг матрицы (rank of a matrix): максимальное число линейно-независимых строк или столбцов матрицы; является порядком наибольшего ненулевого детерминанта матрицы.

Редуцированная корреляционная матрица (adjusted correlation matrix): корреляционная матрица, в которой элементы главной диагонали соответствуют общностям: корреляционные или ковариационные матрицы, которыми пользуются перед выделением факторов.

Собственное число (eigenvalue): характеристика матрицы; используется при декомпозиции ковариационной матрицы одновременно как критерий определения числа выделяемых факторов и как мера дисперсии, соответствующей данному фактору.

Собственный вектор (eigenvector): вектор, связанный с соответствующим собственным числом; получается в процессе выделения первоначальных факторов; эти векторы, представленные в нормированной форме, являются факторными нагрузками.

Специфичность (specific component): доля дисперсии наблюдаемой переменной, соответствующая специфичному фактору; применяется для обозначения части характерности, получаемой при исключении дисперсии ошибки

Сумма квадратов отклонений (variation): мера разброса переменной; сумма квадратов отклонений от среднего.

Факторы (factors): гипотетические, непосредственно неизмеряемые, скрытые переменные, в терминах которых описываются измеряемые переменные; часто подразделяются на характерные и общие.

Факторной детерминации коэффициент (factorial determination): доля общности в дисперсии наблюдаемой переменной.

Факториая нагрузка (factor loading): общий термин, означающий коэффициенты матрицы факториого отображения или структуры.

Факторного отображения матрица (factor pattern matrix): матрица коэффициентов, в которой столбцы соответствуют общим факторам, а строки — наблюдаемым переменным; элементы матрицы факторного отображения представляют собой коэффициенты регрессии для общих факторов при условии, что наблюдаемые переменные являются линейной комбинацией факторов; для ортогонального решения матрица отображения содержит коэффициенты корреляции между переменными и факторами.

Факторная сложность переменной (factorial complexity): характеристика наблюдаемой переменной представляет собой число общих факторов с ненулевыми нагрузками для данной переменной.

Факториой структуры матрица (factor structure matrix): матрица коэффициентов корреляции между переменными и факторами; в случае некоррелированных (ортогональных) факторов совпадает с матрицей факторного отображения.

Факторной причинности принцип (postulate of factorial causation): предположение о том, что наблюдаемые переменные являются линейной комбинацией скрытых факторов и что ковариации между наблюдаемыми переменными воспроизводятся с помощью общих факторов.

Характерность (unique component): доля дисперсии наблюдаемой переменной, не связанная с общими факторами и свойственная именно данной переменной; она часто разделяется на специфичность и дисперсию ошибки.

Характерный фактор (unique factor): фактор, влияющий только на данную переменную; часто относится ко всем независимым факторам (включая ошибку измерений), характерным только для данной переменной.

Целевая матрица (target matrix): матрица коэффициентов, используемая при вращении в качестве целевой; первоначальное факторное решение вращается таким образом, чтобы результирующие факторные нагрузки в наибольшей степени приближали целевую матрицу.

Эквимакс (equimax): критерий, применяемый для получения косоугольного решения; сочетает свойства критериев варимакс и квартимакс.

Экономии принцип (postulate of parsimony): состоит в том, что из двух конкурирующих моделей выбирается наиболее простая; в факторном анализе принимаются модели, включающие минимальное число общих факторов.

У. Р. Клекка

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

William R. Klecka. **Discriminant Analysis** (Seventh Printing, 1986).

ПРЕДИСЛОВИЕ

Работа Уильяма Клекка «Дискриминантный анализ» представляет собой простое введение в круг вопросов, связанный со статистическими процедурами, объединенными общей идеей. Она носит практический характер и содержит многочисленные примеры приложений дискриминантного анализа. В первом разделе У. Клекка приводит основные предположения, в том числе с математическими выкладками. Рассмотрение дискриминантного анализа начинается с нескольких примеров из области социальных наук. Дискриминантный анализ представляет статистический аппарат для изучения различий между двумя и более группами объектов по отношению к нескольким переменным одновременно. Этот метод успешно использовался:

• психологами для разработки тестов; дискриминантный анализ особенно полезен для предсказывания (или объяснения) того, какие студенты будут иметь хорошую успеваемость, причем этот анализ проводится еще до того, как студенты прослушают какой-либо специальный курс;

• учеными, изучающими поведение горожан и законодателей во время выборов; выделяются переменные, позволяющие идентифицировать горожан, голосующих, например, за демократов и против республиканцев, или законодателей, голосующих обычно за либералов и против консерваторов;

• социологами и психологами, занимающимися изучением половых стереотипов в поведении детей;

• учеными, исследующими всевозможные закономерности, проявляющиеся в судебном производстве.

ISBN 0-8039-1491-1

Можно назвать много других примеров из области социальных наук. Дискриминантный анализ оказался весьма полезен в широком спектре исследовательских задач и в прогнозировании. Работа профессора Клекка является очень хорошим введением в предмет. Она доступна неспециалистам, и вместе с тем в ней дается фундамент для последующего, более глубокого изучения материала.

Он рассматривает канонические дискриминантные функции, алгоритмы и функции классификации, а также различные критерии выбора для включения переменных. Здесь приводится геометрическая интерпретация коэффициентов канонической дискриминантной функции, представлено использование коэффициентов в стандартной и нестандартной форме, а также алгоритмы определения числа значимых дискриминантных функций. Профессор Клекка начинает рассмотрение дискриминантного анализа с самых простых вещей, а затем постепенно переходит к более сложному материалу. В конце работы дается обсуждение возможных нарушений предположений, лежащих в основе дискриминантного анализа, которое послужит отправной точкой для тех, кто собирается впервые применять дискриминантный анализ в исследовательских задачах. Работа У. Клекка, безусловно, заслуживает самой высокой оценки.

Джон Л. Сулливан, редактор серии

I. ВВЕДЕНИЕ

Дискриминантный анализ является статистическим методом, который позволяет изучать различия между двумя и более группами объектов по нескольким переменным одновременно. Этот метод часто бывает полезен в социальных науках. Рассмотрим, например, такую ситуацию. Группа экспертов исследует возможность переговоров с террористами, захватившими заложников. Их интересуют те особенности ситуации, при которых было бы возможно безопасное освобождение заложников, даже если требования террористов не выполнены. В качестве альтернативы, что заложникам будет причинен вред, существует несколько переменных, предсказывающих их благополучное освобождение. Например, число террористов, наличие поддержки их местным населением, являются ли они независимой группой или принадлежат к большой военной организации, характер их устных заявлений, тип и количество оружия, отношение числа террористов к числу заложников и т. д. Изучая предыдущие инциденты, в которых власти отказались выполнить требования террористов, эксперты должны найти ответ на следующие вопросы: 1) какие из этих переменных могут быть полезными для предсказания судьбы заложников; 2) как эти переменные могут быть связаны в математическую функцию для предсказания наиболее вероятного исхода; 3) какова точность предсказания. Дискриминантный анализ может обеспечить получение необходимых данных. Если некоторые переменные, взятые из примера для случаев успешного освобождения заложников, отличаются от соответствующих переменных для случаев, когда заложники пострадали, то с помощью дискриминантного анализа можно помочь властям и в данной ситуации.

Есть и другие области применения дискриминантного анализа: тестирование при приеме на работу, анализ переписи населения, психологические тесты для детей, изучение эффекта от какого-либо метода лечения, исследование экономических различий между географическими районами, предсказание итогов голосования и др. Основным предположением дискриминантного анализа является то, что существуют две или более группы, которые по некоторым переменным отличаются от других групп, причем такие переменные могут быть измерены по интервальной шкале либо по шкале отношений[1]. Дискриминантный анализ помогает выявлять различия между группами и дает возможность классифицировать объекты по принципу максимального сходства.

КОГДА ИСПОЛЬЗУЕТСЯ ДИСКРИМИНАНТНЫЙ АНАЛИЗ — ОСНОВНЫЕ ПРЕДПОЛОЖЕНИЯ

Во-первых, *объекты (наблюдения)* должны принадлежать одному из двух (или более) *классов (групп)*. Объекты являются основными единицами анализа. Объектами изучения могут быть люди, животные, страны, экономика в различные моменты времени и вообще все, что угодно. В примере с террористами каждый предыдущий террористический акт есть объект. Класс должен быть определен таким образом, чтобы каждое наблюдение принадлежало одному и только одному классу. Последствия террористических актов могут быть отнесены к одному из двух классов: случаи успешного освобождения заложников и случаи, когда пострадали некоторые или все заложники.

В практических задачах допускаются объекты, которые нельзя отнести ни к какой группе. Например, иногда определенное число наблюдений не удается идентифицировать либо по какой-то причине откладывается анализ этих наблюдений. Такие объекты будут классифицироваться *позже,* на основе математических функций, полученных из анализа наблюдений с «известной» принадлежностью. В случае с террористами главная задача состоит в точном предсказании результатов будущих инцидентов. Поэтому будущие инциденты могут рассматриваться как «несгруппированные» и «нерасклассифицированные».

«Дискриминантный анализ» — это общий термин, относящийся к нескольким тесно связанным статистическим процедурам. В конкретных ситуациях не обязательно использовать все эти процедуры. Их можно разделить на методы интерпретации межгрупповых различий и методы классификации наблюдений по группам. Речь идет об *интерпретации,* когда рассматриваются различия между классами. Другими словами, при интерпретации необходимо ответить на вопросы: возможно ли, используя данный набор характеристик (переменных), отличить один класс от другого; насколько хорошо эти характеристики позволяют провести различение и какие из них наиболее информативны. Метод, относящийся к *классификации,* связан с получением одной или нескольких функций, обеспечивающих возможность отнести данный объект к одной из групп. Эти *функции, называемые дискриминантными,* зависят от значений характеристик таким образом, что появляется возможность отнести каждый объект к одной из групп. Например, если значения характеристик нового террористического акта близки к соответствующим значениям прошлых инцидентов, в которых все заложники были освобождены, дискриминантная функция покажет, что для рассматриваемого события более вероятен благоприятный исход. (После того как инцидент будет исчерпан, станет известно, оправдался ли прогноз, однако для многих других приложений подтвердить точность классификации не представляется возможным.) Разумеется, ди-

скриминантный анализ необходим и для интерпретации, и для классификации.

Характеристики, применяемые для того, чтобы отличать один класс от другого, называются *дискриминантными переменными.* Эти переменные должны измеряться либо по интервальной шкале, либо по шкале отношений. Таким образом, становится возможным вычисление математических ожиданий, дисперсий и правомерно использование дискриминантных переменных в математических уравнениях. В примере с террористами были упомянуты семь дискриминантных переменных (число террористов, степень поддержки, количество оружия и т. д.). В общем случае, число дискриминантных переменных неограничено, но в сумме число объектов должно всегда превышать число переменных по крайней мере на два.

Однако существуют определенные ограничения, касающиеся статистических свойств дискриминантных переменных. Во-первых, ни одна переменная не может быть линейной комбинацией других переменных. Линейная комбинация — это сумма одной или более переменных с постоянными весами. Таким образом, нельзя пользоваться суммой переменных или их средним арифметическим совместно с самими переменными. Соответственно недопустимы переменные, коэффициент корреляции которых равен 1. Переменная, являющаяся линейной комбинацией других, не несет какой-либо новой информации помимо той, которая содержится в компонентах суммы, поэтому она является лишней.

Другое предположение, принимаемое во многих случаях, заключается в том, что ковариационные матрицы для генеральных совокупностей (генеральные ковариационные матрицы) равны между собой для различных классов[2]. Часто используемой форме дискриминантного анализа присущи линейные дискриминантные функции, соответствующие просто линейной комбинации дискриминантных переменных. Этот метод наиболее элементарен, поскольку предположение об одинаковых ковариационных матрицах в классах упрощает формулы вычисления дискриминантных функций, а также облегчает проверку гипотез о статистической значимости.

Следующее допущение касается того, что закон распределения для каждого класса является многомерным нормальным, т. е. каждая переменная имеет нормальное распределение при фиксированных остальных переменных (Blalock, 1979; 452). Данное предположение позволяет получить точные значения вероятности принадлежности к данному классу и критерия значимости. При нарушении допущения о нормальности распределения значения вероятности вычислить точно уже нельзя, но соответствующие оценки могут быть полезны, если, конечно, соблюдать известную осторожность (Lachenbruch, 1975; 41—46).

Упомянутые выше допущения для дискриминантного анализа фундаментальны. Если экспериментальные данные для некоторой конкретной задачи не вполне удовлетворяют этим предполо-

жениям, то статистические выводы не будут точным отражением реальности. Нарушение основных предположений будет обсуждаться в разд. VI.

Из всего сказанного, должно быть ясно, что дискриминантный анализ используется для изучения различий между несколькими группами по определенному набору дискриминантных переменных (рис. 1). Рассматривая классы как значения некоторой классифицирующей переменной, измеренной по шкале наименований (когда каждому классу присваивается свое обозначение), мы представляем дискриминантный анализ в качестве метода сопоставления нескольких интервальных переменных одной номинальной переменной.

Заметим, что мы не сказали о причинности дискриминантной модели, и соответственно на рис. 1 связи приведены без указания их направления. Кроме того, не делается предположений о зависимости или независимости классифицирующей переменной и дискриминантных переменных. Если в конкретной ситуации классифицирующие переменные можно считать зависимыми от дискриминантных переменных, то задача аналогична задаче множественной регрессии. Основное отличие состоит в том, что в дискриминантном анализе зависимая переменная измеряется по шкале наименований (классов). Пример с террористами именно такого рода. Но когда предполагается, что значения дискриминантных переменных зависят от классов, дискриминантный анализ является обобщением многомерного дисперсионного анализа. Это типично для задач, в которых принадлежность переменных к некоторому классу вызывает изменения одновременно в нескольких переменных.

Теперь просуммируем математические допущения, которые принимаются в дискриминантном анализе. Сначала введем следующие обозначения:

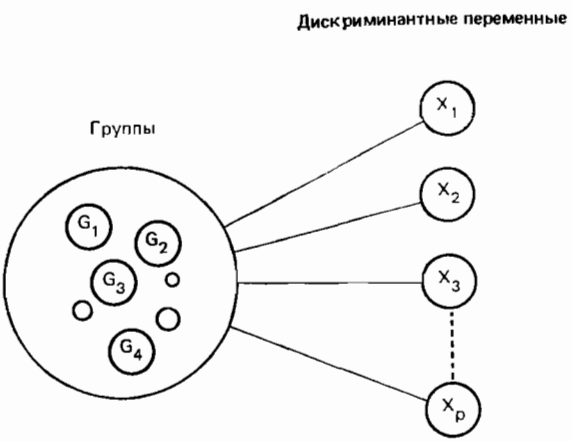

Рис. 1. Зависимость между группами и дискриминантными переменными

g — число классов;
p — число дискриминантных переменных;
n_i — число объектов (наблюдений) класса i;
$n.$ — общее число объектов всех классов.

В модели дискриминантного анализа должно быть:

1) два или более классов: $g \geqslant 2$;

2) по крайней мере два объекта в каждом классе: $n_i \geqslant 2$;

3) любое число дискриминантных переменных при условии, что оно не превосходит общее число объектов за вычетом двух: $0 < p < (n. - 2)$;

4) измерение дискриминантных переменных по интервальной шкале;

5) линейная независимость дискриминантных переменных;

6) приблизительное равенство между ковариационными матрицами для каждого класса (если не используются специальные формулы);

7) многомерная нормальность закона распределения дискриминантных переменных для каждого класса.

НЕСКОЛЬКО ПРИМЕРОВ ИЗ ОБЛАСТИ СОЦИАЛЬНЫХ НАУК

Применения дискриминантного анализа многочисленны. Впервые им воспользовался Фишер (Fisher, 1936), занимающийся проблемами антропологии и биологии. В социальных науках одно из первых приложений относится к психологическим и общеобразовательным тестам (Tatsuoka and Tiedeman, 1954). Ученые, проводящие исследования в области политики, применяли дискриминантный анализ при изучении поведения жителей городов во время выборов (Klecka, 1973), законодательных фракций (Kornberg and Frasure, 1971; Heyck and Klecka, 1973) и предрасположений судов к тем или иным истцам и ответчикам (Eisenstein and Jacob, 1977). Психологи широко используют дискриминантный анализ в области персональных тестов и тестов по специальным дисциплинам. Особенно полезна данная техника при анализе экспериментальных данных, когда предположение и принадлежность к определенной «испытуемой» группе влекут за собой изменение нескольких исследуемых переменных. Примером такого рода является изучение половых стереотипов в поведении детей (Klecka, 1974).

К сожалению, мы не можем остановиться на всех упомянутых приложениях. В данной работе постоянно будем обращаться к примеру, взятому из диссертации и статьи Бардес (Bardes, 1975; 1976). Речь идет об анализе голосований сенатских фракций по вопросу помощи иностранным государствам за период с 1953 по 1972 г. Бардес занималась исследованием фракций сената Соединенных Штатов, ее интересовала, насколько устойчивы были цели, которые отстаивались каждый год при голосовании, и как сказывались на это другие предметы обсуждения. Бардес было известно, что сенаторы не просто делились на группы «за» или «против» помощи иностранным государствам, и что несогласия порой пере-

ходили за пределы чисто партийной принадлежности. Часто дебаты возникали по поводу объема помощи, ее формы (наличные деньги, товары либо займы) и кто — президент или сенат — должен непосредственно заниматься данной проблемой. Изучая все ежеквартальные отчеты Congressional Quarterly, а также другую информацию о дебатах, Бардес выявила несколько фракций и познакомилась со многими сенаторами, которые придерживались той или иной фракции. Задачу осложняло то, что неизвестно было число фракций, существующих в данный момент, а также тот факт, что большинство сенаторов не проявляли явно свои склонности.

Бардес провела трехшаговое статистическое исследование по каждой из 10 рассмотренных сессий. Во-первых, она выбрала результаты голосования, относящиеся к внешнеполитическому законодательству, и используя кластерный анализ, свела их к ограниченному числу шкал. Это помогло выявить те вопросы, по которым наблюдались наибольшие разногласия. На втором шаге была проведена классификация всех сенаторов, проявивших свое отношение к данной проблеме. Число таких групп определялось с помощью имеющейся информации о раскладе мнений в сенате по рассматриваемой проблеме. На этой стадии сенаторы, не имеющие явно выдержанной позиции, объявлялись «нерасклассифицированными». И наконец, на третьем шаге Бардес применяла дискриминантный анализ, чтобы определить, возможно ли объединение групп при незначительных различиях в типе их поведения при голосовании. Дискриминантные функции также использовались для отнесения еще «нерасклассифицированных» сенаторов к одной из наиболее близких групп. Кроме того, удалось выявить моменты, являющиеся самыми существенными при классификации на группы. Рассматривая зависимость результатов анализа от времени, Бардес обнаружила рост численности одних фракций и уменьшение других, а также значительные изменения во взглядах сенаторов, связанные с приведением к присяге нового президента и прекращением вьетнамской войны.

На основании данных 1955—1956 гг.[3] Бардес выявила четыре фракции, существовавшие в этот период, и 19 сенаторов, явно примыкавших к этим фракциям. Они представляют собой «известные» или «расклассифицированные» объекты. Вот эти фракции (группы):

Группа	Число объектов	Описание
1	9	В целом за помощь иностранным государствам
2	2	В целом против помощи иностранным государствам
3	5	Против помощи государствам, испытывающим финансовые затруднения
4	3	Антикоммунисты

Для получения дискриминантных переменных, Бардес произвела разделение результатов голосования по следующим шкалам:

Шкала (переменная)	Описание
CUTAID	Сокращение фондов помощи
RESTRICT	Добавление ограничений в программу помощи
CUTASIAN	Сокращение фондов помощи азиатским государствам
MIXED	Смешанные взгляды: помощь некоторым государствам и никакой помощи коммунистам
ANTIYUGO	Неоказание помощи Югославии
ANTINEUT	Неоказание помощи нейтральным странам

Эти шкалы были определены как средние значения результатов голосования по даиным вопросам. Переменная CUTAID, например, была вычислена по результатам 10 голосований. Для каждого отдельного голосования сенаторам, явно выражающим свое мнение по данному вопросу, приписывалось значение 1. Значение 2 присваивалось воздерживающимся и отсутствующим сенаторам и значение 3 — тем, кто принимал положительное решение. В табл. 1 представлены средние значения для каждой из шести переменных во всех четырех группах. Как и следовало ожидать, группа 1 (за помощь) в целом возражала против мер, связанных с сокращением фондов помощи (среднее значение переменной CUTAID = 1,422), а группа 2 (против помощи) поддерживала эти меры (среднее значение CUTAID = 3,0), в то время как остальные груп-

Таблица 1

Зиачеиия перемеииых для «известных» сенаторов

Переменная	Группа				Среднее по группам
	1	2	3	4	
CUTAID	1,422	3,000	2,200	2,100	1,900
RESTRICT	1,944	1,000	2,000	2,333	1,921
CUTASIAN	1,000	3,000	2,000	1,333	1,526
MIXED	2,667	2,000	1,800	1,667	2,211
ANTIYUGO	1,556	2,500	2,600	3,000	2,158
ANTINEUT	1,259	1,667	2,133	2,444	1,719

пы занимали средние позиции. В общем, группы имеют тенденцию к различным значениям по каждой шкале[4], поэтому шкалы обладают свойствами дискриминантных переменных. Однако по приведенным одномерным статистикам трудно судить о возможностях многомерной классификации. В дальнейшем рассмотренный пример будет использоваться в качестве иллюстрации того, как с помощью дискриминантного анализа можно отличать одну группу от другой и «расклассифицировать» оставшихся 81 сенатора по четырем фракциям.

БИБЛИОГРАФИЧЕСКИЕ ЗАМЕЧАНИЯ

В работах (Tatsuoka, Tiedeman, 1954; Kendall, 1968) дается интересный материал по истории развития дискриминантного анализа. Первая из этих работ содержит многочисленные более ранние приложения в психологии, образовательных тестах и биометрике. Работы Моррисона (Morrison, 1969; 1974) предназначены для первоначального введения в предмет. В последней его работе дается обзор примеров использования дискриминантного анализа при исследовании торговли.

В нескольких книгах дискриминантный анализ обсуждается с точки зрения его применения в социальных науках (Lachenbruch, 1975; Cooley and Lohnes, 1971; Overall and Klett, 1972; Tatsuoka, 1971; Van de Geer, 1971). При их изучении требуется знание матричной алгебры. Однако они не столь сложные по сравнению с такими классическими работами, как (Anderson, 1958; Rao; 1952; 1965).

Каждому пользователю дискриминантного анализа полезно знакомство с компьютерными программами, разработанными в данной области. Как минимум надо ориентироваться в основных характеристиках и ограничениях этих программ. В некоторых руководствах дается обзор методов, приводятся основные формулы и библиографические источники. Полезные сведения содержатся в руководстве по пакету программ SPSS (Klecka, 1975), хотя там мало внимания уделяется формулам (в этом смысле более полна работа (Norusis, 1979), где обсуждаются алгоритмы, используемые в SPSS). В описаниях пакетов программ BMDP (Dixon, 1973) и SAS (Barr et al., 1976) даются только краткие сведения о самих программах без объяснения того, как интерпретировать результаты.

Вельдман (Veldman, 1967), Кули и Лохнес (Cooley and Lohnes, 1971) приводят тексты алгоритмов на языке Фортран для тех, кто собирается разрабатывать свои собственные программы. Следует иметь в виду, что программы постоянно совершенствуются, поэтому нужно ориентироваться на более поздние работы. Однако вполне можно использовать модельные данные и примеры для отладки собственных программ.

II. ПОЛУЧЕНИЕ КАНОНИЧЕСКИХ ДИСКРИМИНАНТНЫХ ФУНКЦИЙ

Прежде чем приступить к обсуждению вопроса классификации (его мы рассмотрим в разд. IV), проанализируем природу различий между классами. В данном разделе обсуждаются принципы, лежащие в основе вычисления канонических дискриминантных функций, и методы определения их числа.

Каноническая дискриминантная функция является линейной комбинацией дискриминантных переменных и удовлетворяет определенным условиям. Она имеет следующее математическое представление:

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \cdots + u_p X_{pkm}, \qquad (1)$$

где f_{km} — значение канонической дискриминантной функции для m-го объекта в группе k; X_{ikm} — значение дискриминантной переменной X_i для m-го объекта в группе k; u_i — коэффициенты, обеспечивающие выполнение требуемых условий.

Коэффициенты u_i для первой функции выбираются таким образом, чтобы ее средние значения для различных классов как можно больше отличались друг от друга. (Точное определение «максимального отличия между классами» будет дано несколько позднее.) Коэффициенты второй функции выбираются так же, т. е. соответствующие средние значения должны максимально отличаться по классам, при этом налагается дополнительное условие, чтобы значения второй функции были некоррелированы со значениями первой. Аналогично третья функция должна быть некоррелирована с первыми двумя и т. д. Максимальное число дискриминантных функций, которое можно получить описанным способом, равно числу классов без единицы или числу дискриминантных переменных, в зависимости от того, какая из этих величин ме́ньшая. В примере с голосованием в сенате число переменных равно шести, а классов — только четырем, поэтому максимальное число функций составит три.

ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ

Пусть дискриминантные переменные — оси p-мерного евклидова пространства. Каждый объект (наблюдение) является точкой этого пространства с координатами, представляющими собой наблюдаемые значения каждой переменной. Если классы отличаются друг от друга по наблюдаемым переменным, их можно представить как скопления точек в некоторых областях рассматриваемого пространства. Поскольку классы могут частично перекрываться, соответствующие им «территории» не совпадают. Для определения положения класса можно вычислить его «центроид». Центроид класса является воображаемой точкой, координаты которой есть средние значения переменных в данном классе. В примере с голосованием, наблюдения принадлежат 6-мерному прост-

ранству (имеются шесть переменных), а столбцы табл. 1 характеризуют координаты центроида для каждого из четырех классов. Центроид можно использовать для изучения различий между классами, так как он занимает положение типичных наблюдений соответствующего класса. Рассмотрение отдельных переменных не позволяет проводить многомерный анализ — число переменных может быть велико, и совокупную информацию поэтому трудно систематизировать. Оказывается, для того чтобы различать относительное положение центроидов, не нужна слишком большая размерность. Как правило, достаточно ограничиться размерностью, на единицу меньшей числа классов.

ЧИСЛО КАНОНИЧЕСКИХ ДИСКРИМИНАНТНЫХ ФУНКЦИЙ

Роль числа классов становится понятной, если обратиться к геометрическим аналогам. Для любых пространств, где применимы аксиомы евклидовой геометрии, две точки определяют положение прямой линии, три точки — плоскость, четыре — трехмерную поверхность и т. д. Принцип сводится к тому, что точки определяют пространство (линию, плоскость и так далее), имеющее размерность, на единицу меньшую, чем число точек.

Поскольку центроиды задают пространство, то соответственно имеется неограниченное число точек, где мы можем поместить систему координат. Наиболее удобна точка, в которой каждая ось имеет нулевое значение, — это «главный центроид». Главный центроид занимает положение, определяемое средними значениями совокупности объектов по каждой из осей. Относительно этого центра существует бесконечное множество ориентаций осей при условии, что они принадлежат пространству, «натянутому на центроиды». Теперь если мы направим одну из этих осей под углом, для которого средние значения классов разделяются в большей степени, чем для любого другого направления, то получим ось, которая, как нам кажется, должна быть особенно важной. Предполагая, что есть два и более класса, можно ориентировать вторую ось таким образом, чтобы было обеспечено максимальное разделение классов, но при дополнительном ограничении — вторая ось ортогональна первой (и принадлежит рассматриваемому пространству).

Аналогично проводятся другие оси. Расположение осей по такому принципу приводит нас к критерию для канонических дискриминантных функций. Соотношение (1) задает математическое преобразование p-мерного пространства дискриминантных переменных в q-мерное пространство канонических дискриминантных функций (где q — максимальное число функций). Каждой оси соответствует свое соотношение вида (1). Для данного наблюдения значение f_{km} интерпретируется как координата объекта в пространстве канонических дискриминантных функций.

Исключения из приведенного правила составляют случаи, когда один или несколько центроидов не определяют новое направ-

ление. Примером являются три точки, попадающие на одну прямую, либо четыре точки, лежащие в одной плоскости, т. е. может статься, что данная точка принадлежит пространству, которое задается другими точками. Можно пойти дальше и допустить ситуацию, когда четыре точки лежат на одной прямой. В дискриминантном анализе это случается. Как мы вскоре увидим, в примере с фракциями в сенате существуют две, а может быть даже всего одна дискриминантная функция, описывающая эти данные. В исследовательских задачах возможно появление лишних размерностей из-за ошибок выборки и измерений. Тем не менее каждую размерность можно проверить на статистическую значимость. Если она незначима, ее можно отбросить, так как маловероятно, что она имеет какое-то теоретическое или практическое значение. Такая проверка описана ниже.

В случае, когда число дискриминантных переменных p меньше числа классов, максимальное число функций q равно p. При этом уже не происходит преобразование из пространства с большей размерностью в пространство с меньшей размерностью. Мы только делаем замену координат, удовлетворяющую некоторому критерию.

ПОЛУЧЕНИЕ КОЭФФИЦИЕНТОВ КАНОНИЧЕСКОЙ ДИСКРИМИНАНТНОЙ ФУНКЦИИ

Рассмотрим основные принципы получения коэффициентов u_i канонической дискриминантной функции. Полное представление математических аспектов этой проблемы не входит в нашу задачу. Оно приводится в нескольких монографиях по многомерной статистике, например в (Cooley and Lohnes, 1971). Начнем с того, что необходим некий статистический метод для измерения степени различий между объектами (наблюдениями). Таблица групповых средних и стандартных отклонений недостаточна, так как не учитывает зависимости между переменными. Однако можно воспользоваться матрицей сумм квадратов и попарных произведений T, являющейся квадратной симметричной матрицей[5]. Для пояснения происхождения матрицы T введем следующие обозначения:

g — число классов;

n_k — число наблюдений в k-м классе;

$n.$ — общее число наблюдений по всем классам;

X_{ikm} — величина переменной i для m-го наблюдения в k-м классе;

$X_{ik.}$ — средняя величина переменной i в k-м классе;

$X_{i..}$ — среднее значение переменной i по всем классам (общее среднее).

Тогда элементы матрицы T задаются соотношением

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{i..})(X_{jkm} - X_{j..}). \qquad (2)$$

Выражения в скобках являются отклонениями значений переменных от общего среднего. Если $i=j$, то сомножители равны, и получается средне-квадратичное отклонение. Таким образом, диагональные элементы представляют собой сумму квадратов отклонений от общего среднего. Они показывают, как ведут себя наблюдения по отдельной переменной. При $i \neq j$ получаем сумму произведений отклонения по одной переменной на отклонение по другой. В этом состоит один из способов измерения корреляций (ковариаций) между двумя переменными, так как он показывает, насколько хорошо большое отклонение по одной переменной согласуется с большим отклонением по другой. Рассматривая целиком всю матрицу, мы имеем полную информацию о распределении точек по пространству, определяемому переменными.

Если разделить каждый элемент T на $(n.-1)$, получим ковариационную матрицу. В дискриминантном анализе чаще используется непосредственно матрица T, тем не менее в статистической литературе более распространена ковариационная матрица. Основываясь на наблюдениях, принадлежащих одному классу, можно вычислить ковариационные матрицы для него.

Степень зависимости двух переменных можно выяснить, исследуя их корреляцию. Для этого воспользуемся коэффициентом корреляции, поскольку он нормирован и принимает значения от -1 до $+1$. Можно легко преобразовать матрицу T в матрицу коэффициентов корреляции, деля каждый элемент на квадратный корень произведения двух соответствующих диагональных элементов. (Те же результаты могут быть получены из ковариационной матрицы; см. работу (Cooley and Lohnes, 1971.) В табл. 2 представлены коэффициенты корреляции по данным Бардес.

Таблица 2

Общая корреляционная матрица

	CUTAID	RESTRICT	CUTASIAN	MIXED	ANTIYUGO	ANTINEUT
CUTAID	1					
RESTRICT	0,43	1				
CUTASIAN	0,787	0,054	1			
MIXED	—0,732	—0,435	—0,677	1		
ANTIYUGO	0,534	0,470	0,493	0,638	1	
ANTINEUT	0,526	0,626	0,562	—0,829	0,776	1

Как видим, несколько переменных сильно коррелированы. Другими словами, значение наблюдения по одной переменной может быть предсказано по значению, соответствующему другой переменной.

Если расположения классов действительно различаются (т. е. их центроиды не совпадают), то степень разброса наблюдений внутри классов будет меньше общего разброса. Для измерения

разброса внутри классов служит матрица W, которая отличается от T только тем, что ее элементы определяются средними значениями переменных для отдельных классов, а не общими средними:

$$W_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.})(X_{jkm} - X_{jk.}). \qquad (3)$$

Если элементы матрицы W разделить на $(n. - g)$, получится внутригрупповая ковариационная матрица, она является взвешенным средним ковариационных матриц отдельных классов.

Матрицу W или внутригрупповую ковариационную матрицу легко преобразовать во внутригрупповую корреляционную матрицу, как это уже сказано по отношению общей корреляционной матрице. Каждый коэффициент корреляции является оценкой степени зависимости между соответствующей парой переменных внутри групп. Он обычно не совпадает с общей корреляцией, на величину которой сказываются межгрупповые различия. Если предположить, что наблюдения относятся к одной генеральной совокупности или к разным генеральным совокупностям, имеющим одинаковые статистические свойства, то в качестве оценок зависимостей между переменными предпочтительнее внутригрупповые корреляции, а не общие корреляции. В табл. 3 представлена матрица внутригрупповых корреляций для экспериментальных данных Бардес. Видно, что многие коэффициенты отличаются от значений, приведенных в табл. 2. Это обусловлено разбросом центроидов разных классов.

Таблица 3

Внутригрупповая корреляционная матрица

	CUTAID	RESTRICT	CUTASIAN	MIXED	ANTIYUGO	ANTINEUT
CUTAID	1					
RESTRICT	0,234	1				
CUTASIAN	0,692	0,562	1			
MIXED	—0,706	—0,547	—0,834	1		
ANTIYUGO	0,364	0,647	0,386	—0,411	1	
ANTINEUT	0,469	0,744	0,785	—0,748	0,645	1

Когда центроиды различных классов совпадают, элементы матриц W и T также будут равны (поскольку, тогда $X_{ik.} = X_{i..}$). Если же центроиды у классов разные, элементы W будут меньше соответствующих элементов матрицы T. Эта разница обозначается как матрица B ($B = T - W$, т. е. $b_{ij} = t_{ij} - W_{ij}$).

Матрица B называется межгрупповой суммой квадратов отклонений и попарных произведений. Величины элементов B по отношению к величинам элементов W дают меру различия между группами, как это будет выяснено позже.

Матрицы W и B содержат всю основную информацию о зависимости внутри групп и между группами. С помощью некоторых вычислений можно получить функцию, удовлетворяющую требуемым свойствам. Во-первых, необходимо решить систему уравнений:

$$\begin{aligned} \Sigma b_{1i}v_i &= \lambda \Sigma w_{1i}v_i \\ \Sigma b_{2i}v_i &= \lambda \Sigma w_{2i}v_i \\ &\vdots \quad\quad \vdots \\ \Sigma b_{pi}v_i &= \lambda \Sigma w_{pi}v_i \end{aligned} \tag{4}$$

где λ — собственное число, а v_i — последовательность p коэффициентов.

Как уже говорилось, b_{ji} и w_{ji} — элементы матриц B и W соответственно, которые получаются при обработке экспериментальных данных. Построение дискриминантной функции сводится к решению уравнений (4) относительно λ и v_i. Для получения единственно правильного решения дополнительно наложим условие, что сумма квадратов v_i должна быть равна 1. Максимально существует q нетривиальных решений этих уравнений. Каждое решение, которое имеет свое собственное значение λ и свою последовательность v_i, соответствует одной канонической дискриминантной функции. Коэффициенты v_i могут использоваться как коэффициенты требуемой дискриминантной функции:

$$u_i = v_i \sqrt{n_{\cdot} - g}, \quad u_0 = -\sum_{i=1}^{p} u_i X_{i\cdot\cdot} \tag{5}$$

Эти коэффициенты u_i и требовалось определить в соотношении (1). Применение u_i из (5) приводит величины f_{km} (значения дискриминантной функции) к стандартной форме. Это означает, что соответствующие дискриминантные значения по совокупности наблюдений (объектов) будут иметь нулевое среднее и единичное внутригрупповое стандартное отклонение[6]. Значение дискриминантной функции для данного объекта представляет положение этого наблюдения на оси, определяемой данной функцией.

КОЭФФИЦИЕНТЫ V_i

Решение системы уравнений (4) дает последовательность коэффициентов v_i для каждой функции. Эти коэффициенты могли бы быть непосредственно использованы при классификации. Однако их трудно интерпретировать, соответствующие им значения дискриминантной функции не имеют определенного смысла. Причина заключается в том, что данное решение не имеет ограничения по метрике дискриминантного пространства. Хотя это пространство вводится для обеспечения максимального разделения классов, последние могут располагаться в любой его области. Приведенная ситуация аналогична ситуации, когда игроки в бей-

сбол могут находиться в любой точке поля, лишь бы их взаимное расположение не противоречило правилам игры.

В некоторых компьютерных программах коэффициенты v_i распечатываются и могут использоваться при классификации (см. разд. IV). Однако более целесообразна их нормировка, задаваемая соотношением (5).

НЕСТАНДАРТИЗОВАННЫЕ КОЭФФИЦИЕНТЫ

Нормировка коэффициентов не меняет ни результат классификации, ни относительное расположение классов. Однако существенно то, что оси занимают более естественное положение, так как начало координат (точка, где проекции всех дискриминантных функций нулевые) совпадает с главным центроидом. Главный центроид, как мы уже говорили, является точкой пространства, в которой все дискриминантные переменные принимают средние (по всем наблюдениям) значения. Другими словами, это — центральное положение всех точек, представляющих наблюдения. Расположение начала координат в главном центроиде полезно, так как в данном случае рассматриваемые классы и объекты соотносятся с центром системы.

Нормировка коэффициентов влечет за собой и другие изменения. Они касаются единиц измерения расстояний. Нормированные коэффициенты приводят к дискриминантным значениям, измеряемым в единицах стандартного квадратичного отклонения, т. е. каждая ось растягивается или сжимается таким образом, что соответствующее дискриминантное значение для данного объекта представляет число стандартных отклонений точки от главного центроида[7]. Анализируя это значение, можно сразу отличить относительное расстояние от абсолютного и определить, насколько относительное расстояние велико по сравнению с размерами системы. Так, значение — 2,5 означает, что наблюдение располагается на расстоянии двух с половиной стандартных отклонений в отрицательном направлении от центра осей. Поскольку очень небольшое число точек может находиться вне окрестности радиуса, равного двум стандартным отклонениям, становится ясно, что данное наблюдение достаточно далеко отстоит от центра.

Способ приведения переменных к стандартной форме зависит от того, нормируются ли исходные значения наблюдений. Если исходные данные неприведены к стандартной форме, соответствующие им коээффициенты будем называть «нестандартизованными». Обозначение u как раз и относится к этим коэффициентам, а соотношение (5) показывает как значение v переходит в значение u. (Стандартизованные коэффициенты будут рассмотрены в следующем разделе.) Обычно нестандартизованные коэффициенты используются для вычисления дискриминантных значений.

В настоящем разделе мы рассмотрели получение канонических дискриминантных функций, постарались дать точное определение некоторых понятий, используемых в работе, и предложили

сведущим в математике читателям некоторые основные моменты статистического аппарата. Специалистам по приложениям, в общем-то, и необязательно досконально разбираться в этих вопросах. Им в первую очередь необходимо научиться применять и интерпретировать канонические дискриминантные функции. Это и является задачей следующего раздела.

III. ИНТЕРПРЕТАЦИЯ КАНОНИЧЕСКИХ ДИСКРИМИНАНТНЫХ ФУНКЦИЙ

Канонические дискриминантные функции определены, и теперь можно приступить к их интерпретации. Задача сводится, во-первых, к изучению относительных расстояний между объектами и центроидами классов и, во-вторых, к рассмотрению соотношений между отдельными переменными и функциями. Если существует более одной функции, мы также задаемся вопросом, все ли из них необходимы. Для большей конкретности начнем с изучения экспериментальных данных Бардес.

ВЫЧИСЛЕНИЕ ДИСКРИМИНАНТНЫХ ЗНАЧЕНИЙ

В табл. 4 представлены нестандартизованные дискриминантные коэффициенты для трех функций, полученных по данным Бардес. Эти функции определяют трехмерное пространство, в котором располагаются наблюдения, соответствующие отдельным сенаторам. Функция 1 определяет одну из осей. Если представить себе обычное трехмерное пространство, функцию 1 естественно считать горизонтальной осью. Способ получения функции 2 приводит к требованию ее перпендикулярности к функции 1, так что она должна представлять совершенно отличную информацию (две функции должны быть некоррелированы). Это будет вертикальная ось. Третья функция должна быть перпендикулярна первым двум[8].

Коэффициенты представляют положение наблюдений в дискриминантном пространстве. Формула для первой функции следующая:

$$f_{km} = 5{,}4243 + 0{,}8087\,X_{1km} + 0{,}7940\,X_{2km} - 4{,}6004\,X_{3km} - \\ - 0{,}6957\,X_{4km} - 1{,}1114\,X_{5km} + 1{,}4387\,X_{6km}, \qquad (6)$$

где f_{km} обозначает дискриминантные значения для наблюдения по функции 1; X_{jkm} — значение j-го дискриминантного параметра для m-го наблюдения из k-го класса. Формулы для двух других функций аналогичны.

Эти формулы сводятся к тому, что значение дискриминантной функции для каждого объекта получается путем умножения значений дискриминантных переменных на соответствующие коэффициенты, а затем сложения полученных произведений с некоторой постоянной. (Эта постоянная выбирается так, чтобы среднее зна-

Таблица 4

Нестандартизованные дискриминантные коэффициенты

Переменная	Нестандартизованные коэффициенты		
	Функция 1	Функция 2	Функция 3
Константа (u_0)	5,4243	3,5685	—4,3773
CUTAID	0,0878	—0,5225	1,6209
RESTRICT	0,7940	—1,1177	—0,3339
CUTASIAN	—4,6004	—1,1228	—1,1431
MIXED	—0,6957	—1,3160	1,1418
ANTIYUGO	—1,1114	1,1132	0,3781
ANTINEUT	1,4387	1,0422	0,2000

чение дискриминантной функции по всем наблюдениям было нулевым.)

Теперь вычислим значения дискриминантных функций непосредственно для одного из сенаторов в рассматриваемом примере. В табл. 5 приводятся результаты исследования позиции сенатора Айкена. Для каждой функции в таблице представлены нестандартизованные коэффициенты и соответствующие значения наблюдаемых переменных[9]. Произведение этих двух чисел вносит вклад в значение дискриминантной функции, характерной для позиции сенатора Айкена. Сумма вкладов и есть значение дискриминантной функции.

Последние определяют точку в пространстве дискриминантных функций. Ее координаты по позиции Айкена таковы: 2,25; —3,22; —0,90.

Можно также сделать заключение о том, насколько типично мнение Айкена среди других сенаторов. Помогают в этом дискриминантные значения, поскольку они выражены в единицах стандартного отклонения. По первой функции позиция Айкена — положительная (это означает, что он выступает за бо́льшие расходы на помощь иностранным государствам). По второй функции — резко отрицательная (он — за введение менее жестких ограничений). По третьей функции его позиция в некоторой степени отрицательна (он выступает против помощи государствам, испытывающим финансовые затруднения).

В качестве второго примера рассмотрим позицию сенатора Бриджеса, для которой значения наблюдаемых переменных следующие: 1,0; 2,5; 1,4; 2,0; 3,0; 3,0 соответственно. В пространстве дискриминантных функций точка, означающая позицию Бриджеса, занимает положение: 1,37; 2,51; —1,17.

Очевидно, мнения Бриджеса и Айкена очень далеки друг от друга в дискриминантном пространстве. По функции 1 они отличаются ненамного, по функции 2 занимают противоположные позиции, а по функции 3 позиция Бриджеса несколько более отрицательна, чем у Айкена.

Таблица 5

Вычисление дискриминантных значений для сенатора Айкена

Переменная	Функция 1			Функция 2			Функция 3		
	коэффициент × значение = вклад			коэффициент × значение=вклад			коэффициент × значение=вклад		
Константа			5,4243			3,5685			—4,3773
CUTAID	0,8078	1,0	0,8078	—0,5225	1,0	—0,5225	1,6209	1,0	1,6209
RESTRICT	0,7940	3,0	2,3820	—1,1177	3,0	—3,3531	—0,3339	3,0	—1,0017
CUTASIAN	—4,6004	1,0	—4,6004	—1,1228	1,0	—1,1228	—1,1431	1,0	—1,1431
MIXED	—0,6957	3,0	—2,0871	—1,3160	3,0	—3,9480	1,1418	3,0	3,4254
ANTIYUGO	—1,1114	1,0	—1,1114	1,1132	4,0	1,1132	0,3781	1,0	0,3781
ANTINEUT	1,4387	1,0	1,4387	1,4387	1,0	1,4387	0,2000	1,0	0,2000
Дискриминантное значение			2,2539			—3,2225			—0,8977

Нестандартизованные коэффициенты представляют собой изменение положения точки в дискриминантном пространстве при единичном приращении соответствующей переменной. Если представить себе, что некоторый сенатор меняет свое положение по переменной CUTAID от 1,0 до 2,0 (при всех прочих неизменных), его положение по функции 1 продвинется на 0,8078 единицы в положительном направлении. Разумеется, сенаторы не могут изменить свои позиции в прошлом, но нестандартизованные коэффициенты могут использоваться, чтобы различать одного сенатора от другого.

Позиции, занимаемые Айкеном и Бриджесом, имеют одно и то же значение переменной, а значение переменной ANTIYUGO Айкена составляет 1,0, а Бриджеса — 3,0. Это отличие в две единицы означает, что за счет переменной мнение Бриджеса будет отстоять от мнения Айкена на 2,2228 единицы в отрицательном направлении по функции 1 ($2 \times 1{,}1114 = 2{,}2228$). Поскольку позиции этих сенаторов отличаются также по другим переменным, необходимо рассмотреть все отличия, прежде чем мы узнаем их окончательное положение в дискриминантном пространстве. Однако часто представляет интерес изучение вклада данной переменной при фиксированных остальных.

В общем случае неэффективно рассматривать каждый объект отдельно, разве что число объектов очень мало. Чаще нас интересует положение центроида класса, т. е. «наиболее типичное» положение для каждой группы. Оно может быть вычислено с помощью групповых средних в формулах. По данным Бардес центроиды четырех классов имеют следующие координаты: (1,74; —0,94; 0,02), (—6,93; —0,60; 0,28), (—1,48, 0,69; —0,30) и (1,86; 2,06; 0,25). Хотя видно, что эти точки далеки одна от другой, нагляднее представить их геометрически.

ДВУХКООРДИНАТНЫЕ ГРАФИКИ

Для случая с двумя дискриминантными функциями легко изобразить графически положение центроидов и отдельных объектов. В нашем примере есть три функции, но двухкоординатный график все равно будет информативен, особенно если мы считаем первые две функции более важными. На рис. 2 показан такой график. Звездочками обозначены четыре групповых центроида, а числа соответствуют занимаемым позициям сенаторов, принадлежащих группе с данным номером. Позиция сенатора Айкена относится к группе 1 (за помощь иностранным государствам) и представлена единицей, находящейся в нижнем правом углу. Мнение сенатора Бриджеса обозначено цифрой 4, расположенной на графике около звездочки.

Изучение этого графика показывает, что группы вполне различимы. Цетроиды хорошо отделимы друг от друга, и нет явных перекрытий отдельных объектов, несмотря даже на то, что мнения двух сенаторов из группы 1 близки к группе 4. (В следующем разделе мы подробнее ими займемся.) Группы 1 и 4 занимают почти одно и то же положение по первой функции. Обе соответствуют позиции «за расширение помощи иностранным государст-

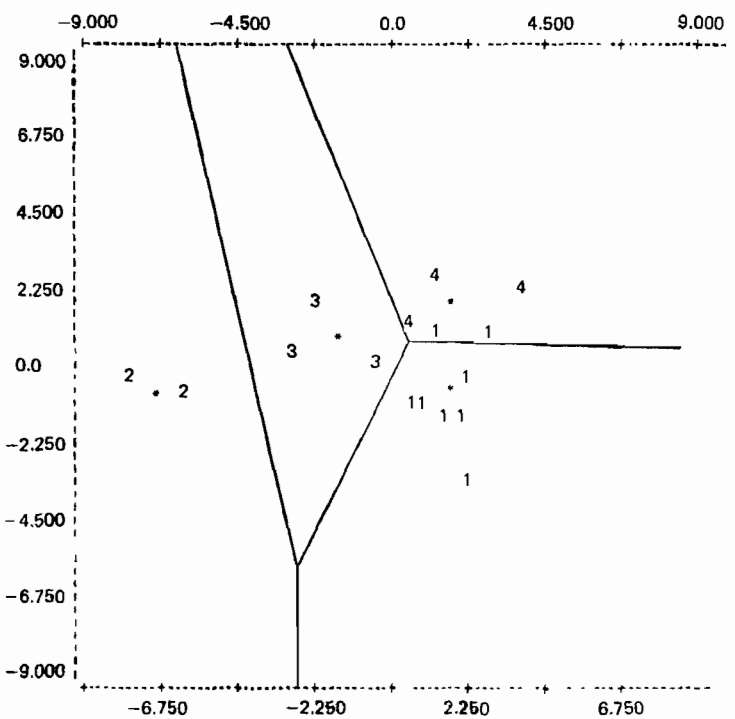

Рис. 2. Двухкоординатный график групповых центроидов и наблюдений. Ось абсцисс — функция 1; ось ординат — функция 2

вам». Однако они весьма различаются по второй дискриминантной фуикции (позиция «за» либо «против добавлений ограничений в программу помощи). Прямые линии, изображенные на графике, ограничивают «территории» соответствующих групп (см. следующий раздел).

Графики, аналогичные графику на рис. 2, могут быть полезны, когда пересечений между группами мало. Если группы становятся менее различимы, особеино когда число наблюдений велико, изображения точек сливаются в классы. В этом случае более полезно изучение положенний только центроидов классов либо построение графиков для каждой группы в отдельности.

При возрастаиии числа дискриминантных фуикций становится сложнее графически представить положения центроидов. Трехмерную модель еще можно наглядно изобразить, а четырехмерную — вряд ли. Поскольку две первые функции являются наиболее иинформативными для классификации, то можио ограничиться построением соответствующего двухмерного графика.

ОДНОКООРДИНАТНЫЕ ГРАФИКИ

В случае одной дискриминантной функции точки, соответствующие объектам, располагаются вдоль некоторой прямой. Положение точки характеризует долю функции, которая относится к даиному наблюдению; одиако при большом числе объектов теряется информация о плотности точек.

Альтернатива состоит в построении гистограммы для каждой группы. Во-первых, определяются интервалы, соответствующие

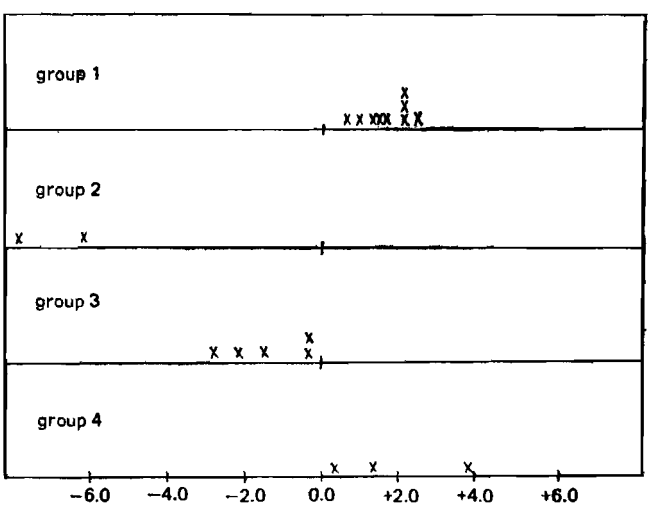

Рис. 3. Групповая гистограмма для данных Бардес. Знаки «X» обозначают позиции сенаторов. Ось абсцисс является первой каионической дискримннантной фуикцией, измеренной в еднницах стандартного отклонення

стандартному отклонению, равному, например, 0,1 Во-вторых, над интервалом помещается символ «X» (или какой-либо другой), если некоторое наблюдение попадает в данный интервал. Для последующих наблюдений, лежащих в рассматриваемом интервале, соответствующие символы «X» помещаются один над другим, поэтому высота получающихся в результате столбцов определяет число наблюдений в этом интервале.

Гистограмма наглядна для представления плотности и распределения группы. Расположив групповые гистограммы одну над другой, можно сравнивать относительное положение групп.

Данные Бардес, рассматриваемые относительно первой дискриминантной функции, представлены на рис. 3. В этом случае мы получим небольшое количество информации о гистограмме, так как в каждом интервале — недостаточное количество наблюдений.

Однако видно, что по этой выделенной функции отличие между группами 1 и 4 фактически отсутствуют. Действительно, их центроиды занимают одно и то же положение.

Гораздо лучший пример однокоординатных графиков дан в работе (Heyck, Klecka, 1973). Этот же пример приведен в «Руководстве по пакету программ SPSS» (Nie, 1975).

СТАНДАРТИЗОВАННЫЕ КОЭФФИЦИЕНТЫ

Переходя от изучения отдельных наблюдений и групповых центроидов к рассмотрению дискриминантных переменных становится важным вопрос о представлении дискриминантной функции коэффициентов в стандартной или нестандартной форме.

Поскольку коэффициент в нестандартной форме дает информацию об абсолютном вкладе данной переменной в значение дискриминантной функции, то при различных единицах измерений переменных (т. е. когда стандартные отклонения переменных различны) можно получить верную классификацию Если нас интересует относительный вклад переменной, то коэффициенты следует представлять в стандартной форме.

Стандартизованные коэффициенты получаются из соотношения (5), если наблюдения имеют единичные стандартные отклонения, что достигается их нормированием[10]. Вместо того чтобы приводить к стандартной форме наблюдения, а затем пересчитывать коэффициенты, можно их вычислить исходя из значений коэффициентов в нестандартной форме.

$$c_i = u_i \sqrt{\frac{w_{ii}}{n - g}} \qquad (7)$$

где w_{ii} — сумма квадратов i-й переменной, определяемая соотношением (3); n — общее число наблюдений; g — число групп. Стандартизованные коэффициенты полезно применять при выявлении тех переменных, которые вносят наибольший вклад в значение дискриминантной функции. Абсолютная величина коэф-

Таблица 6

Стандартизованные дискриминантные коэффициенты

Переменные	Стандартизованные коэффициенты		
	Функция 1	Функция 2	Функция 3
CUTAID	0,6094	—0,3942	1,2227
RESTRICT	0,7068	—0,9950	—0,2973
CUTASIAN	—2,1859	—0,5335	—0,5432
MIXED	—0,4760	—0,9004	0,7812
ANTIYUGO	—0,8077	0,8090	0,2748
ANTINEUT	1,0168	0,7365	0,1414

фициента анализируется в стандартной форме: чем она больше, тем больше вклад этой переменной.

В табл. 6 представлены стандартизованные коэффициенты для данных Бардес, относящихся к голосованию в сенате. Для функции 1 вклад переменной CUTASIAN максимален. Все остальные переменные по сравнению с CUTASIAN второстепенны. Переменные ANTINEUT и ANTIYUGO после CUTASIAN занимают следующее по значимости место. Можно отметить, что они приблизительно вдвое более значимы, чем переменная MIXED.

Для функции 2 четыре из шести переменных (RESTRICT, MIXED, ANTIYUGO и ANTINEUT) имеют относительно большие стандартизованные коэффициенты, поэтому они вносят приблизительно одинаковый вклад в значение дискриминантной функции. Переменная CUTAID, а вслед за ней MIXED являются доминантными переменными для функции 3[11].

При желании дискриминантное значение можно получить исходя из стандартизованных коэффициентов, однако тогда мы должны умножать эти коэффициенты на значение наблюдений в стандартной форме[12].

СТРУКТУРНЫЕ КОЭФФИЦИЕНТЫ

Для определения взаимной зависимости отдельной переменной и дискриминантной функции мы рассмотрим их корреляцию. Эти коэффициенты корреляции называются «полными структурными коэффициентамн». Как корреляции они являются косинусами углов между переменными и функцией. Поэтому, зная эти коэффициенты, мы имеем информацию о геометрической структуре пространства наблюдений[13].

Структурный коэффициент показывает, насколько тесно связаны переменные и дискриминантные функции. Когда абсолютная величина такого коэффициента велика, вся информация о дискриминантной функции заключена в этой переменной. Если же коэффициент близок к нулю — их зависимость мала. Можно дать «имя» дискриминантной функции, исходя из тех коэффи-

циентов, которые максимальны по абсолютной величине. Если соответствующие переменные измеряют похожие характеристики, то функции можно называть по этим характеристикам.

Например, рассмотрим полные структурные коэффициенты, относящиеся к даниым Бардес (табл. 7). Для функции 1 переменные CUTAID и CUTASIAN являются доминантными. Их отрицательные знаки говорят о том, что мы должны дать функции 1 название «за помощь». Для функции 2 полные структурные коэффициенты последних трех переменных имеют наибольшие абсолютные значения. Они относятся к ограничению помощи тем странам, которые ие связаиы альянсом с США, поэтому мы назовем эту функцию «против помощи государствам, не связанным альянсом». У функции 3 нет преобладающих больших коэффициентов, и она иаиболее трудна для интерпретации. Переменные RESTRICT и CUTAID имеют большие коэффициенты, но с противоположными знаками. Рассматривая расположения групповых центроидов по этой функции, можно получить некоторую дополнительную информацию.

Группы 2 и 4 («против помощи» и «антикоммунисты») расположены в положительном направлении; группа 3 («против помощи государствам, испытывающим финансовые затруднения») — в отрицательном направлении; группа 1 («за помощь иностранным государствам») — между ними. Поэтому функция 3 служит для различия всевозможных типов оппозиции к функции 1 («за помощь иностранным государствам»).

Таблица 7

Структурные коэффициенты

Переменная	Структурные коэффициенты		
	Функция 1	Функция 2	Функция 3
CUTAID	—0,565	0,355	0,326
RESTRICT	0,345	0,260	—0,429
CUTASIAN	—0,858	0,241	—0,160
MIXED	0,269	—0,671	0,254
ANTIYUGO	—0,293	0,785	0,076
ANTINEUT	—0,140	0,751	—0,269

Структурные коэффициенты, рассматриваемые здесь, основаиы иа понятии корреляции. Их полезно использовать при классификации групп. Однако иногда интересно узнать, как дискриминантные функции связаны с переменными в пределах отдельной группы. Таким образом, мы приходим к так называемым «внутригрупповым структурным коэффициентам», которые вычисляются следующим образом:

$$s'_{ij} = \sum_{k=1}^{p} r_{ik} c'_{kj} = \sum_{k=1}^{p} \frac{w_{ik} c_{kj}}{\sqrt{w_{ii} \cdot w_{kk}}}; \qquad (8)$$

где S'_{ij} — внутригрупповые структурные коэффициенты для переменной i и функции j; r'_{ik} — внутригрупповые структурные коэффициенты корреляции между переменными i и k; c_{kj} — стандартизованные коэффициенты канонической функции для переменной k и функции j.

В табл. 8 представлены внутригрупповые структурные коэффициенты, относящиеся к данным Бардес. Заметим, что эти коэффициенты меньше, чем описанные выше структурные коэффициенты, однако их расположения от бóльших к мéньшим схожи (хотя и не идентичны).

Такое положение является типичным, но не обязательным. Некоторые из внутригрупповых структурных коэффициентов могут быть больше или меньше, или иметь противоположный знак по сравнению со структурными коэффициентами. Кроме того, они могут иметь различный порядок расположения. Эти две последовательности коэффициентов относятся к разным сторонам структуры данных. Поэтому и не следует их интерпретировать одинаково, за исключением тех случаев, когда групповые центроиды совпадают:

Таблица 8

Внутригрупповые структурные коэффициенты

Переменная	Внутригрупповые структурные коэффициенты		
	Функция 1	Функция 2	Функция 3
CUTAID	—0,218	0,279	0,392
RESTRICT	0,115	0,176	—0,461
CUTASIAN	—0,483	0,276	—0,299
MIXED	0,102	—0,516	0,315
ANTIYUGO	—0,121	0,662	0,087
ANTINEUT	—0,054	0,588	—0,340

Структурные коэффициенты дают информацию, несколько отличную от той, которая относится к стандартизованным коэффициентам. Стандартизованные коэффициенты показывают вклад переменных в значение дискриминантной функции. Это является одним из подходов к определению значимости переменной. Однако этот подход имеет серьезное ограничение. Если две переменных сильно коррелированы, то их вклад в дискриминантное значение должен разделяться, даже при значительном совместном вкладе. Соответственно их стандартизованные коэффициенты могут быть меньше по сравнению с теми случаями, когда используется одна из этих переменных. Или, другими словами, вклад одного коэффициента частично погашается отрицательным вкладом другого[14]. Это происходит из-за того, что в стандартизованных коэффициентах одновременно принимается во внимание влияние всех переменных.

С другой стороны, структурные коэффициенты являются просто двуместными корреляциями, поэтому на них не влияют взаимные зависимости прочих переменных. Заметим, что CUTAID имеет небольшой стандартизованный коэффициент по функции 1, но относительно большой полный структурный коэффициент.

Возможно, это объясняется сильной корреляцией между переменными CUTAID и CUTASIAN (0,787). Поэтому переменная CUTASIAN дает большой отрицательный вклад в дискриминантные значения, а переменная CUTAID — небольшой положительный вклад. Структурные коэффициенты также помогают распознать вклад переменных ANTIYUGO и ANTINEUT в функцию 1. Эти две переменные имеют довольно малые структурные коэффициенты, т. е. они мало коррелируют с этой функцией. Функция 1 отличается от той, что мы получаем из стандартизованных коэффициентов, которые довольно велики и имеют противоположные знаки. Переменные ANTIYUGO и ANTINEUT сильно коррелированы (0,767), поэтому они дают большие вклады в противоположные направления и погашают друг друга. Анализ подобных ситуаций приводит нас к выводу, что структурные коэффициенты позволяют лучше интерпретировать канонические и дискриминантные функции, чем стандартизованные коэффициенты.

В работе (Overall and Klett, 1972; 292—295) описывается, как структурные коэффициенты могут использоваться для графического представления различия между групповыми центроидами в случае двух канонических дискриминантных функций. На графике с осями, которые относятся к этим двум функциям, представлены групповые центроиды и главный центроид, изображены векторы, исходящие из главного центроида и направленные в каждую дискриминантную переменную. Направляющие углы этих векторов вычисляются, исходя из структурных коэффициентов. Длина вектора определяется межгрупповыми и внутригрупповыми вариациями соответствующей переменной.

Полученная диаграмма дает наглядное представление о различиях групп с помощью дискриминантных переменных, а также о потенциальных возможностях этих переменных.

СКОЛЬКО ФУНКЦИЙ НАДО УЧИТЫВАТЬ

В разд. II было показано, что решению уравнения (4) соответствует собственное значение (лямбда) и множество коэффициентов для каждой канонической дискриминантной функции. Число возможных решений общей задачи в действительности равно числу дискриминантных переменных p. Однако некоторые из них будут математически тривиальными решениями, а другие — статистически малозначимыми. Все собственные значения лямбды будут положительными или равными нулю, причем чем больше значение лямбды, тем больше групп будет разделять соответствующая функция. Таким образом, функция с самым боль-

шим собственным значением является и самым мощным дискриминатором, а функция с наименьшим собственным значением — самым слабым дискриминатором.

Число функций

Предположив, что значение лямбда равно нулю, получим решение уравнения (4), которое не представляет интереса. Такое решение оказывается бесполезным, потому что оно допускает отсутствие различий между группами по этой функции. Однако, когда p меньше $(g-1)$, мы получаем $(p-g+1)$ решений, которые имеют нулевые собственные значения. По этой причине максимальное число канонических дискриминантных функций q меньше любого из чисел p и $(g-1)$. Возвращаясь к примеру о голосовании в сенате, имеем $p=6$, $(g-1)=3$, так что $q=3$. Среди этих q возможных решений мы все еще можем найти собственные значения, равные нулю. Это бывает в тех вырожденных случаях, когда один или несколько центроидов совпадают в пространстве, определенном другими центроидами. Более типичен случай не полного совпадения из-за ошибок выборки или ошибок измерения. Скорее всего, такое собственное значение функции будет малой величиной. Вопрос в следующем: как мала должна быть величина собственного значения лямбда, чтобы мы рассматривали ее как результат ошибки выборки или измерения, а не результат измерения величины, действительно отличной от нуля? Это вопрос о статистической значимости. Но даже если функция статистически значима, мы можем решить, что она не имеет самостоятельного значения, поскольку с ее помощью недостаточно хорошо различаются группы.

Прежде чем научиться проверять значимость, рассмотрим собственные значения функции, воспользовавшись примером о голосовании в сенате. Эти результаты приведены в табл. 9. Как и ожидалось, имеются три собственных значения, не равных нулю. Они даются в порядке убывания их величин. Так обычно делают потому, что величина собственного значения связана с дискриминирующими возможностями этой функции: чем больше собственное значение, тем лучше различение. Располагая их в порядке убывания, мы знаем, что первая функция обладает наибольшими возможностями: вторая функция обеспечивает максимальное различение *после* первой функции; третья дает наилучшее дополнительное различение *после* первой и второй и т. д. Все функции не обязательно дают идеальное различение, но мы, по крайней мере, знаем их порядок значимости.

Таблица 9

Собственные значения, соответствующие функции, и меры значимости

Конечная дискриминантая фуикция	Собственное значение	Относительное процентное содержание	Каноническая корреляция
1	9,65976	85,54	0,952
2	1,57922	13,93	0,782
3	0,05357	0,47	0,225

Относительное процентное содержание

Фактические числа, представляющие собственные значения, ни о чем нам не говорят. Их нельзя интерпретировать непосредственно. Если имеется более одной функции, желательно уметь сравнивать их дискриминантные возможности. Так, например, число 9,65976 для собственного значения, соответствующего первой функции, больше собственного значения, соответствующего второй, более чем в шесть раз. В случае когда первое собственное значение в 180 раз превосходит третье, то это доказывает, что третья функция обладает очень незначительными возможностями.

Чтобы облегчить такое сравнение, мы припишем собственным значениям относительное процентное содержание. Для этого сначала суммируем все собственные значения, чтобы установить размер общих возможностей различения. Затем разделим каждое собственное значение на общую сумму. Так, в приведенной системе уравнений первая функция содержит 85,54% общих дискриминантных возможностей.

Третья функция в этом примере иллюстрирует тот случай, когда она оказывается настолько мало значимой, что, по-видимому, ею можно пренебречь. К сожалению, нет правила, которое помогло бы определить, как велико должно быть относительное процентное содержание, чтобы функция представляла для исследователя интерес. Поэтому при дальнейшем рассмотрении может оказаться, что и функция 2 не удовлетворяет нас. Даже функция 1 иногда не имеет реальной значимости (согласно критерию, который рассматривается ниже), хотя она наиболее мощная[15]. Относительное процентное содержание только показывает что функция настолько слабее *по сравнению с другими,* что вряд ли она добавит что-либо к определению различий между группами.

Каноническая корреляция

Другой способ оценки реальной полезности дискриминантной функции можно получить, рассматривая коэффициент каноинческой корреляции, который является мерой связи (степени зависимости между группами и дискриминантной функцией). Нулевое значение говорит об отсутствии связи, а большие числа (всегда положительные) означают большую степень зависимости (максимальное значение равно 1,0). Каноническая корреляция (обозначаем ее r^*) связана с собственным значением следующей формулой:

$$r_i^* = \sqrt{\frac{\lambda_i}{1+\lambda_i}}, \qquad (9)$$

где i — номер соответствующей дискриминантной функции.

Понятие канонической корреляции взято из так называемого каноинческого корреляционного анализа (см. Levine, 1977). Ка-

нонической корреляция используется при изучении связей между двумя различными множествами переменных, измеренных по интервальной шкале. Анализ заключается в формировании q пар линейных комбинаций, где q — число переменных в меньшем множестве. Линейные комбинации в каждой паре (по одной из каждого множества) подбираются так, чтобы получить максимальную корреляцию между ними. Первая пара имеет самую высокую степень зависимости; вторая пара — следующую по величине степень зависимости при условии, что ее составляющие не коррелируют с первой парой и т. д. Канонический коэффициент корреляции, конечно, является мерой зависимости и идентичен смешанному моменту корреляции Пирсона между двумя линейными комбинациями в паре.

С помощью простого математического «фокуса» мы можем превратить дискриминантный анализ (по крайней мере, обсуждаемую часть его) в канонический корреляционный анализ. Очевидно, дискриминантные переменные образуют одно из «множеств». Тогда, если мы представим классы с помощью $(g-1)$ дихотомических переменных (известных так же, как «бинарные переменные» или «фиктивные переменные»), то получим другое «множество». Из них мы образуем q пар линейных комбинаций. В этом случае канонические коэффициенты корреляции можно интерпретировать в соответствии с приведенным выше определением как меру зависимости двух множеств переменных, найденную с помощью линейных комбинаций. Такой подход дает повод некоторым статистикам называть каноническую дискриминантную функцию «канонической переменной»[16].

Другая интерпретация канонического коэффициента корреляции заимствована из дисперсионного анализа (Iversen и Norpoth, 1976, 30—32), где он известен под именами «эта» (η) и «корреляционное отношение». Здесь классы рассматриваются как независимые переменные, которые влияют на величину дискриминантной функции, являющейся зависимой переменной. Коэффициент η измеряет степень различия средних значений дискриминантной функции для разных групп. Можно облегчить интуитивное понимание коэффициента η, если возвести его в квадрат. Коэффициент η^2 (т. е. каноническая корреляция в квадрате) является долей дисперсии дискриминантной функции, которая объясняется разбиением на классы.

Независимо от того, какой подход выбран, каноническая корреляция помогает получить представление о реальной полезности дискриминантной функции. Большая величина коэффициента, как например, у функции 1 в табл. 9, указывает на сильную зависимость между классами и первой дискриминантной функцией. С другой стороны, коэффициент для функции 3 имеет довольно малую величину, которая говорит о слабой связи, что и предсказывалось относительным процентным содержанием этой функции[17].

Анализируя данные табл. 9, не следует делать поспешного заключения о том, что первая дискриминантная функция будет всегда иметь большую каноническую корреляцию. Даже если функция 1 всегда «наиболее» значимая по сравнению с другими (судя по величине ее относительного процентного содержания), у нее может быть лишь слабая связь с классами (измеренная величиной канонической корреляции). По этой причине каноническая корреляция для нас более полезна, потому что она показывает насколько удачно выбрана дискриминантная функция. Если классы не очень хорошо различаются по исследуемым переменным, то все корреляции будут иметь малые значения, поскольку нельзя найти различия там, где их нет. Оценивая и относительное процентное содержание, и канонические корреляции, можно довольно точно узнать, как много дискриминантных функций имеют реальный смысл, и какую пользу они принесут при определении различий между группами.

Измерение остаточной дискриминации с помощью Λ-статистики Уилкса

До сих пор нас интересовало, сколько дискриминантных функций надо брать с точки зрения математических ограничений и их действительной значимости. В наших рассуждениях не учитывались выборочные свойства данных. Они равно справедливы как для генеральных данных (данных о генеральной совокупности)[18], так и для различных видов отбора (выборок). Когда мы анализируем генеральные данные, то ответы на вопросы о числе функций и их значимости даются с помощью относительного процентного содержания и канонической корреляции. В пределах ошибок измерения эти статистики полностью описывают различия между группами и дискриминантными функциями.

Когда же данные берутся из выборки (в противоположность данным, представляющим всю генеральную совокупность), то возникают дополнительные вопросы. Какова вероятность того, что данные о выборке покажут значительную степень различия, тогда как в генеральной совокупности различий между группами нет? Это вопрос статистической значимости, возникающей только в том случае, когда мы имеем дело с выборками[19]. Действительно, ответить на вопрос о статистической значимости можно, если выборочный процесс имеет вероятностную основу. Для многих статистик тесты значимости применимы лишь к простым случайным выборкам ввиду сложности получения тестов для других видов выборок. Таким образом, мы будем рассматривать лишь простые случайные выборки. При использовании каких-либо других процедур отбора, лучше всего к интерпретации тестов подходить консервативно и уделять больше внимания реальной значимости результатов.

Чаще всего статистическая значимость дискриминантных функций проверяется косвенным путем. Вместо проверки самой функции рассматривается остаточная дискриминантная способность

системы *до* определения этой функции. Под «остаточной дискриминантной способностью» мы понимаем способность переменных различать классы, если исключить информацию, полученную с помощью ранее вычисленных функций. Если остаточная дискриминация очень мала, то нет смысла продолжать вычисление очередных функций, даже если математически это возможно. Чтобы лучше усвоить это понятие, рассмотрим «Λ-статистику Уилкса», используемую для измерения дискриминации (так называемую *U*-статистику). Λ-статистика Уилкса — это мера различий между классами по нескольким переменным (дискриминантным переменным). Хотя существует несколько способов ее вычисления, мы воспользуемся следующей формулой:

$$\Lambda = \prod_{i=k+1}^{g} \frac{1}{1+\lambda_i}, \qquad (10)$$

где k — число уже вычисленных функций, а символ П означает, что для получения окончательного результата необходимо перемножить все члены.

Проиллюстрируем применение символа П. Сначала вычислим величину Λ-статистики Уилкса, для данных о голосовании в сенате до вычисления всех дискриминантных функций. Предположим, что $k=0$. Из табл. 9 мы получаем:

$$\Lambda = \left(\frac{1}{1+9{,}65976}\right) \cdot \left(\frac{1}{1+1{,}57922}\right) \cdot \left(\frac{1}{1+0{,}05357}\right) =$$

$$= (0{,}09381) \cdot (0{,}38771) \cdot (0{,}94915) = 0{,}03452.$$

Поскольку Λ является «обратной» мерой, этот результат означает, что шесть используемых переменных чрезвычайно эффективно участвуют в различении классов. Величины Λ, близкие к нулю, говорят о высоком различении (т. е. центроиды классов хорошо разделены и сильно отличаются друг от друга по отношению к степени разброса внутри классов). Увеличение Λ до ее максимального значения, равного 1, приводит к постепенному ухудшению различения, так как центроиды групп совпадают (нет групповых различий).

Очевидно, что позиции четырех групп сенаторов сильно различаются по выбранным переменным, так что имеет смысл найти дискриминантную функцию. После получения первой (и самой значимой) функции становится доступным большое количество информации, необходимой для различения групп. Теперь попытаемся ответить на вопрос: достаточен ли уровень остаточной дискриминантной способности для определения второй функции? Из табл. 10 видно, что Λ-статистика Уилкса равна 0,3680 (для $k=1$), т. е. все еще мала. Вычисление второй функции уменьшает количество оставшейся информации, и величина Λ становится равной 0,9492 (для $k=2$). Это значение (довольно высокое) говорит о том, что оставшуюся информацию о различиях классов уже не стоит искать. Мы пришли к такому же выводу, когда рассматри-

Таблица 10

Остаточная дискриминантная способность и проверка значимости

Номер функции, k	Λ-статистика Уилкса	Статистика хи-квадрат	Степени свободы	Уровень значимости
0	0,0345	43,760	18	0,001
1	0,3680	12,996	10	0,224
2	0,9492	0,678	4	0,954

вали относительное процентное содержание и канонические корреляции. Итак, остающиеся дискриминантные функции (в нашем случае только одна) либо не являются значимыми, либо они статистически недостоверны.

Проверка значимости с помощью Λ-статистики Уилкса

Мы рассматривали Λ-статистику Уилкса как еще одну меру зависимости, но то, что она принимает значения, обратные привычным, и оценивает остаточную дискриминантную способность, делает ее менее полезной, чем относительное процентное содержание и каноническая корреляция. Однако Λ-статистика может быть превращена в тест значимости. Таким образом, мы будем использовать ее скорее как вспомогательную статистику, а не как искомый конечный продукт.

На основе Λ-статистики Уилкса можно получить тест значимости, аппроксимируя распределение некоторой функции от нее либо распределением хи-квадрат (χ^2), либо F-распределением[20]. В дальнейшем можно пользоваться стандартными таблицами для этих распределений, чтобы определить уровень значимости, а некоторые компьютерные программы позволяют распечатать его точные значения. Если воспользоваться формулой

$$\chi^2 = -\left[n_{\cdot} - \left(\frac{p+g}{2}\right) - 1\right] \ln \Lambda_k \,, \qquad (11)$$

то полученное распределение и будет хи-квадрат распределением с $(p-k)(g-k-1)$ степенями свободы.

В табл. 10 приведены значения статистики хи-квадрат для данных примера о голосовании[21]. Как мы и предвидели, между позициями групп есть значимые различия еще до вычисления какой-либо из дискриминантных функций ($k=0$). Уровень значимости 0,001 показывает, что если в действительности между центроидами нет различий, то такое или большее значение статистики хи-квадрат мы получим только в одной из тысячи выборок (имеются в виду независимые, простые случайные выборки). Отбрасывая это невероятное событие, мы можем уверенно считать, что результаты получены из генеральной совокупности с различиями между

группами. Кроме того, это доказывает, что наша первая функция статистически значима.

После определения первой функции, снова проверим значимость оставшихся различий. Как и следовало ожидать, значение статистики хи-квадрат стало меньше, а уровень значимости стал равным 0,224 ($k=1$). Большинство исследователей будут считать этот результат незначимым, поэтому определять вторую и третью функции не следует, полагая таким образом, что вся значимая информация о различиях групп уже извлечена. Другими словами, одного-единственного измерения достаточно для представления всех замеченных различий между группами. Второе измерение (которое вместе с первым образует плоскость) не добавит никаких существенных различий.

Но если бы вместо этого была установлена значимость остаточной дискриминантной способности, то мы приступили бы к определению второй функции. Затем проверка значимости для новой остаточной дискриминантной способности была бы повторена ($k=2$). В нашем примере уровень значимости так велик (0,954), что никто не посчитал бы оставшиеся различия значимыми. Следовательно, нет абсолютно никакой необходимости вычислять третью функцию, так как она вряд ли что-либо добавит к объяснению различий между группами. Найденный результат помогает понять, почему у нас было так много трудностей при интерпретации структурных коэффициентов функции 3 и почему не было обнаружено больших различий между центроидами групп по этой функции.

В рассматриваемом примере число статистически значимых функций меньше того, которое допускается математикой. Однако так бывает не всегда. Во многих ситуациях остаточная дискриминантная способность для $k=g-1$* оказывается значимой. В таком случае нужно вычислить все возможные функции (вплоть до $k=g-1$), если, конечно, нет других причин не делать этого (таких, например, как низкая каноническая корреляция). Примем разумное решение — продолжить определение функций до тех пор, пока остаточная дискриминантная способность перестанет быть значимой. Таким образом, мы можем быть уверены в том, что полученные функции являются статистически значимыми в целом *как система*. Это не доказывает значимость какой-либо одной функции (если, конечно, она не была получена специально), а скорее дает значимость всех полученных функций. А поскольку мы используем функции как систему и наша цель — привести информацию, необходимую для разделения, к наименьшему числу размерностей, то этого вполне достаточно. Единственная реальная проблема, которая может быстро уничтожить любой исследовательский проект, возникает, если общее количество информации является незначимым, т. е. при $k=0$ (если только не нужно показать, что между классами нет различий).

---

* Более точно: $k=\min(g-1, p)$. — *Примеч. ред.*

Здесь мы рассмотрели все то, что обычно делает исследователь, но для лучшего усвоения — в обратном порядке. Логически исследователь должен начать с вопроса: «Какая из моих функций является статистически и реально значимой?» Нет необходимости продолжать анализ любой функции, исключенной из рассмотрения. Для выбранных функций исследователь должен сочетать рассмотрение структурных коэффициентов с определением положений центроидов классов, чтобы выявить значение каждой функции. Структурные коэффициенты дают, кроме того, информацию о том, как каждая из переменных участвует в различении классов в этой системе координат.

В некоторых исследованиях работа аналитика заканчивается вместе с окончанием интерпретации канонических дискриминантных функций. Более вероятно, исследователь продолжит классификацию объектов — либо для практических, либо для аналитических целей, что и является темой следующего раздела.

IV. ПРОЦЕДУРЫ КЛАССИФИКАЦИИ

Как уже было сказано, целью дискриминантного анализа является решение двух задач: интерпретации и классификации. До сих пор внимание фокусировалось в основном на задаче интерпретации, которая связана с определением числа и значимости канонических дискриминантных функций и с выяснением их значений для объяснения различий между классами. Классификация — это особый вид деятельности исследователя, в котором либо дискриминантные переменные, либо канонические дискриминантные функции используются для предсказания класса, к которому более вероятно принадлежит некоторый объект. Существует несколько процедур классификации, но все они сравнивают положение объекта с каждым из центроидов классов, чтобы найти «ближайший». Например, целью исследования Бардес было сформировать подпространство, определяемое канонической дискриминантной функцией, используя данные о 19 сенаторах и выделенных фракциях. Затем она, воспользовавшись результатами их голосования, вычислила значения дискриминантной функции для позиций остальных сенаторов и смогла отнести позицию каждого сенатора к одной из четырех групп. Таким образом, она определила размеры и состав фракций и выяснила, как они изменяются со временем.

КЛАССИФИЦИРУЮЩИЕ ФУНКЦИИ

Классификация — это процесс, который помогает исследователю принять решение: указанный объект «принадлежит к» или «очень похож на» данную группу (класс). Такое решение принимается на основе информации, содержащейся в дискриминантных переменных. Существует несколько способов проведения классификации. Обычно они требуют определения понятия «расстояния»

между объектом и каждым центроидом группы, чтобы можно было приписать объект к «ближайшей» группе.

Процедуры классификации могут использовать или самими дискриминантные переменные, или канонические дискриминантные функции. В первом случае дискриминантный анализ вовсе не проводится[22]. Здесь просто применяется подход максимизации различий между классами для получения функции классификации. Различение классов или размерность дискриминантного пространства на значимость не проверяется. Если же сначала определяются канонические дискриминантные функции и классификация проводится с их помощью, можно провести более глубокий анализ. К этому мы вернемся позднее, а сейчас продолжим рассмотрение классификации, когда дискриминантные переменные используются непосредственно.

Простые классифицированные функции

Фишер (1936) был первым, кто предположил, что классификация должна проводиться с помощью линейной комбинации дискриминантных переменных. Он предложил применять линейную комбинацию, которая максимизирует различия между классами, но минимизирует дисперсию внутри классов. Разработка его предложения приводит нас к определению особой линейной комбинации для каждого класса, которая называется «классифицирующая функция»*. Она имеет следующий вид:

$$h_k = b_{k0} + b_{k1}X_1 + b_{k2}X_2 + \cdots + b_{kp}X_p, \qquad (12)$$

где h_k — значение функции для класса k, а b_{ki} — коэффициенты, которые необходимо определить. Объект относится к классу с наибольшим значением (наибольшим h). Коэффициенты для классифицирующих функций определяются с помощью таких вычислений:

$$b_{ki} = (n_{\cdot} - g) \sum_{j=1}^{p} a_{ij} X_{jk_{\cdot}}, \qquad (13)$$

где b_{ki} — коэффициент для переменной i в выражении, соответствующему классу k, а a_{ij} — элемент матрицы, обратной к внутригрупповой матрице сумм попарных произведений W[23]. Постоянный член определяется так:

$$b_{k0} = -0.5 \sum_{j=1}^{p} b_{kj} X_{jk_{\cdot}} \cdot \qquad (14)$$

Мы обычно не интерпретируем эти коэффициенты классифицирующей функции, потому что они не стандартизованы и каждому классу соответствует своя функция. Точные значения функции ро-

* Во многих работах именно эти функции называются дискриминантными функциями, а функции, определяемые из соотношения (4), — каноническими переменными или каноническими дискриминаитиыми функциями (каноническими направлениями). — *Примеч. ред.*

Таблица 11

Коэффициенты простой классифицирующей функции

Переменные	Группа 1	Группа 2	Группа 3	Группа 4
CUTAID	13,040	6,283	9,064	11,941
RESTRICT	5,755	1,600	1,485	2,424
CUTASIAN	20,056	59,286	33,452	15,886
MIXED	37,016	42,909	36,761	33,253
ANTIYUGO	—2,639	7,480	2,634	0,652
ANTINEUT	8,559	—3,516	5,542	11,897
Постоянная	—77,587	—146,882	—87,329	—69,186

ли не играют: нам нужно знать лишь, для какого класса это значение наибольшее. Именно к нему объект ближе всего. Функции, описываемые соотношением (12), называются «простыми классифицирующими функциями» потому, что они предполагают лишь равенство групповых ковариационных матриц и не требуют никаких дополнительных свойств, обсуждаемых далее.

Рассмотрим табл. 11, в которой приведены коэффициенты классифицирующих функций для данных о голосовании в сенате, чтобы проиллюстрировать использование этих функций. Применив такую функцию к первичным данным по позиции сенатора Айкена, мы получим следующие значения для четырех групп: 89,742; 46,578; 78,101 и 78,221. Поскольку первое значение — наибольшее, мы отнесем позицию Айкена к первой группе (что является верным предсказанием).

Обобщенные фуикции расстояния

Более понятным способом классификации является измерение расстояний между объектом и каждым из центроидов классов, чтобы затем отнести объект в ближайший класс. Однако в тех случаях, когда переменные коррелированы, измерены в разных единицах и имеют различные стандартные отклонения, бывает трудно определить понятие «расстояния». Индийский статистик Махаланобис (1963) предложил обобщенную меру растояния, которая устраняет эти трудности. Мы можем использовать ее в следующей форме:

$$D^2(X \mid G_k) = (n_{\cdot} - g) \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}(X_i - X_{ik_{\cdot}})\,(X_j - X_{jk_{\cdot}}), \qquad (15)$$

где $D^2(X \mid G_k)$ — квадрат расстояния от точки X (данный объект) до центроида класса k. После вычисления D^2 для каждого класса классифицируем объект в группу с наименьшим D^2. Это класс, чей типичный профиль по дискриминантным переменным больше похож на профиль для этого объекта. Если расстояние до ближай-

шего класса велико, то согласие между профилями будет плохим, но по сравнению с любым другим классом — хорошим.

Соотношение (15) предполагает, что классы имеют равные ковариационные матрицы. Если это предположение не выполняется, то выражение можно модифицировать, как предлагает Татсуока (1971; 222).

Вероятность принадлежности к классу

Оказывается D^2 обладает теми же свойствами, что и статистика хи-квадрат с p степенями свободы. Таким образом, мы измеряем расстояние в «хи-квадрат единицах». Если предположить, что каждый класс является частью генеральной совокупности с многомерным нормальным распределением, то большинство объектов будет группироваться вблизи центроида, и их плотность будет убывать по мере удаления от центроида. Зная расстояние от центроида, можно сказать, какая часть класса находится ближе к центроиду, а какая — дальше от него. Следовательно, можно оценить вероятность того, что объект, настолько-то удаленный от центроида, принадлежит классу. Поскольку наши расстояния измеряются в хи-квадрат единицах, то попробуем найти значимость получения этой вероятности. Обозначим через $\Pr(X|G_k)$ вероятность того, что объект, находящийся далеко от центроида, действительно принадлежит классу k.

Относя объект к ближайшему классу в соответствии со значением D^2, мы неявно приписываем его к тому классу, для которого он имеет наибольшую вероятность принадлежности. Благодаря вероятностям, об объекте можно сказать больше простого утверждения, что он является «ближайшим» к какому-то конкретному классу. В действительности объект может с большими вероятностями принадлежать более чем одному классу или не принадлежать ни одному из них. Рассмотрим ситуацию с низким различением и высоким перекрытием классов. В этом случае объект, близкий к центроиду класса 1, будет с большой вероятностью «принадлежать» классу 2, поскольку он также «близок» к этому классу. Другая важная ситуация: объект находится на большом расстоянии от всех классов, иначе говоря — все вероятности малы. Решение приписать этот объект к ближайшему классу, может оказаться лишенным смысла, поскольку он мало похож на любой объект из этого класса. В качестве примера такой ситуации возьмем позицию сенатора Айкена. Она принадлежит к группе 1 (ближайшей группе) с вероятностью 0,1, которая очень мала. С другой стороны, позиция сенатора Бриджес с довольно высокой вероятностью (0,48) принадлежит ближайшей к нему группе (группа 4).

Ясно, что для любого объекта сумма этих вероятностей по всем классам не обязательно равна 1. Однако если мы предположим, что каждый объект должен принадлежать одной из групп, то мож-

но вычислить вероятность принадлежности для любой из групп. Вероятность того, что объект X является членом класса k, равна:

$$\Pr(G_k|X) = \frac{\Pr(X|G_k)}{\sum_{i=1}^{g} \Pr(X|G_i)}. \qquad (16)$$

Сумма этих вероятностей, часто называемых *апостериорными* вероятностями, по всем классам равна 1. Классификация наибольшей из этих величин тоже эквивалентна использованию наименьшего расстояния. Позиция сенатора Айкена с апостериорной вероятностью 1,0 принадлежит к группе 1, а позиция Бриджеса имеет апостериорную вероятность 0,99 для группы 4.

Обратите внимание на различие между этими двумя вероятностями. Апостериорная величина $\Pr(G_k|X)$ дает вероятность, что объект принадлежит классу k. А величина $\Pr(X|G_k)$ оценивает долю объектов в этом классе, которые отстоят от центроида дальше, чем X.

УЧЕТ АПРИОРНЫХ ВЕРОЯТНОСТЕЙ, ИЛИ ЦЕНА ОШИБОЧНОЙ КЛАССИФИКАЦИИ

До сих пор при обсуждении классификации предполагалось, что все классы равноправны. На практике это не всегда так. Рассмотрим, например, случай двух классов, когда 90% генеральной совокупности содержится в классе 1. Еще до вычислений ясно, что с очень большой вероятностью любой заданный объект принадлежит классу 1. Следовательно, он будет отнесен к классу 2 только при наличии очень сильных свидетельств в пользу такого решения. Это можно сделать, вычисляя апостериорные вероятности с учетом априорных знаний о вероятной принадлежности к классу.

Другая ситуация, в которой желательно использование апостериорных вероятностей, возникает, когда «стоимость» неправильной классификации существенно меняется от класса к классу. Типичный пример — применение классифицирующих функций для определения на основе различных симптомов, является ли опухоль злокачественной или доброкачественной. Вероятно, больному придется перенести много страданий при любой ошибке в диагнозе (классификации). Но больной со злокачественной опухолью, которому поставлен диагноз «доброкачественная опухоль», будет страдать больше, чем больной с доброкачественной опухолью, которому поставили диагноз «злокачественная опухоль». Если бы эти издержки неправильной классификации могли быть выражены в виде отношения, то их следовало бы использовать тем же способом, что и априорные вероятности.

В обоих примерах было бы желательно включить априорные вероятности в классифицирующую функцию, чтобы улучшить точность предположения или уменьшить «стоимость» совершения ошибок. Это можно сделать для простых классифицирующих функций

с помощью добавления натурального логарифма от априорной вероятности к постоянному члену. Или же будем модифицировать расстояние D^2, дважды вычитая натуральный логарифм от априорной вероятности. Это изменение в расстоянии математически идентично умножению $\Pr(X|G_k)$ на априорную вероятность для этого класса. Татсуока (1971; 217—232), Кули и Лохнес (1971; 262—270) дают более полное обсуждение этих модификаций.

Если классы очень различаются, то привлечение априорных вероятностей вряд ли повлияет на результат, поскольку вблизи границы между классами будет находиться очень мало объектов. Таким образом, априорные вероятности будут оказывать наибольшее воздействие, когда классы перекрываются и, следовательно, многие объекты с большой вероятностью могут принадлежать к нескольким классам. Конечно, в основе решения об использовании априорных вероятностей должны лежать теоретические соображения. Если же таких соображений нет, то лучше этого не делать. Следует также помнить, что априорные вероятности вычислены на основе *генеральной* совокупности и будут отличаться от вычисленных на основе *выборки.*

КЛАССИФИКАЦИЯ С ПОМОЩЬЮ КАНОНИЧЕСКИХ ДИСКРИМИНАНТНЫХ ФУНКЦИЙ

Классификация может быть проведена и с помощью канонических дискриминантных функций вместо использования исходных дискриминантных переменных. При этом применяются те же формулы (лишь заменяется X на f) и результаты классификации обычно бывают идентичными.

Если необходимо классифицировать большое число объектов методом расстояния и вероятностей, то, воспользовавшись дискриминантными функциями, можно значительно сократить количество работы. Вместо вычисления расстояний для p переменных нам нужны только q канонических дискриминантных функций. Для этого обычно требуется меньшее число операций (даже с учетом вычисления самих функций). Однако если мы пользовались простыми классифицирующими функциями, то применение канонических дискриминантных функций повлечет за собой увеличение объема работ.

При определенных условиях употребление канонических дискриминантных функций приведет к несовпадению результатов классификаций (имеется в виду простая классифицирующая функция. — *Примеч. ред.*). Одним из таких условий является неравенство ковариационных матриц классов. Это происходит потому, что процедура получения канонических дискриминантных функций должна использовать внутригрупповую матрицу ковариаций, являющуюся взвешенным средним матриц ковариаций для отдельных классов. В данном случае преобразование не будет точным. К сожалению, нельзя указать, как сильно должны различаться матрицы классов, чтобы применение дискриминантных функций

стало недопустимым. Татсуока (1971; 232—233) описывает случай, когда процедура, использующая канонические дискриминантные функции, давала почти такие же результаты и ее можно было повторять до тех пор, пока ковариационные матрицы классов не становились «решительно» различными.

Другая ситуация, в которой две процедуры могут давать разные результаты, возникает, когда одна или несколько канонических функций игнорируются, так как не являются статистически значимыми. Хотя в этом примере некоторые объекты могут быть классифицированы по-разному, результаты, полученные с помощью канонических дискриминантных функций, будут более точными, поскольку уменьшается влияние выборочных флуктуаций.

Бардес в своем исследовании прибегла лишь к двум из трех дискриминантных функций и не делала никаких попыток привлечь априорные вероятности. Полученные ею данные показывают, что Pr (позиция Айкена|группа 1)=0,064. Это очень маленькая вероятность, отражающая положение позиции Айкена на самом краю группы 1. Вероятности для всех других групп, по существу, равны нулю. Поэтому мы отнесем позицию Айкена к группе 1, что согласуется с результатами, найденными с помощью простых классифицирующих функций. Возвращаясь к рис. 2, мы видим, что позиция сенатора Айкена, очевидно, находится ближе всего к центроиду группы 1 (крайняя правая точка внутри группы 1).

Теперь рассмотрим подробнее два объекта из группы 1, которые находятся почти на полпути от центроида группы 1 к центроиду группы 4. Этим объектам соответствуют позиции сенаторов: справа — Кейпхарта (республиканца, штат Индиана), слева — Ноуланда (республиканца, штат Калифорния). Здесь Pr (группа 1|позиция Кейпхарта)=0,262, но Pr (группа 4|позиция Кейпхарта)=0,738. Отсюда следует, что, судя по результатам голосования, позиция Кейпхарта ближе к группе 4, несмотря на то, что первоначально на основе данных из первичного источника, Бардес отнесла его к группе 1. Для Ноуланда вероятность принадлежности его позиции к группе 1 равна 0,538, а к группе 4—0,436. Эти вероятности настолько близки, что нам трудно отдать предпочтение одной из них. Если объект находится на разграничительной линии, то иногда бывает желательным считать его неопределенным и неклассифицируемым. В действительности Бардес пересмотрела данные первичного источника о позиции Ноуланда и пришла к заключению, что они недостаточно определены, чтобы отнести позицию к какой-либо группе. Поэтому она исключила Ноуланда из дальнейшего анализа. Кроме того, были еще раз рассмотрены данные о Кейпхарте и решено, что его позиция лучше всего соответствует группе 4. Проделав эти исправления, Бардес вновь провела анализ и приступила к классификации позиций всех сенаторов уже с помощью новых дискриминантных функций.

ГРАФИЧЕСКОЕ ИЗОБРАЖЕНИЕ ОБЛАСТЕЙ

Для лучшего представления картины результатов классификации мы можем нанести разграничительные линии на график расположения объектов. На рис. 2 прямые, разделяющие, группы представляют собой эти границы (разграничительные линии). Почти горизонтальная линия справа разделяет группы 4 и 1. Объект, находящийся выше этой линии, расположен ближе к центроиду группы 4, а ниже линии — к центроиду группы 1. Подобным образом другие линии разграничивают области, где объекты будут классифицированы в замкнутые группы. Конечно, если различение слабое, многие объекты попадают вне областей их групп. В соответствии с правилами, сформулированными раньше, такие объекты будут классифицированы неверно.

Так же можно разделить одномерные графики и гистограммы. Если у нас более двух измерений, графическое изображение областей становится непрактичным из-за невозможности представления таких пространств на листе бумаги. Здесь проявляется другое преимущество классификации с помощью дискриминантных функций — в большинстве исследований требуется только одна или две функции (которые могут быть легко изображены на графике), несмотря на то, что в них используется много дискриминантных переменных. В случае одной функции разделяющая точка между двумя группами равна полусумме величин двух центроидов этих групп. Если же есть две функции, то вычисления затрудняются, но математические идеи остаются простыми. По существу, все сводится к выражению (16) с дополнительным условием:

$$D^2(X \mid G_i) = D^2(X \mid G_j).$$

Решение дает уравнение для прямой линии. Наши рассуждения предполагают, что ковариационные матрицы для отдельных классов можно считать идентичными. Если же это не так, то необходимо сделать уточнения. В случае одной функции разделяющая точка будет находиться ближе к классу с меньшим рассеянием. При двух функциях граница имеет вид кривой, которая охватывает класс с меньшей дисперсией (см. Van de Geer, 1971; 263—265).

КЛАССИФИКАЦИОННАЯ МАТРИЦА

Хотя обычно исследователи обращаются к классификации как к средству предсказания принадлежности к классу «неизвестных» объектов, мы можем использовать ее также для проверки точности процедур классификации. Для этого возьмем «известные» объекты (которыми мы пользовались при выводе классифицирующих функций) и применим к ним правила классификации. Доля правильно классифицированных объектов говорит о точности процедуры и косвенно подтверждает степень разделения классов. Можно составить таблицу, или «классификационную матрицу», описывающую результаты. Это поможет нам увидеть, какие ошибки совершаются чаще.

Таблица 12

Классификационная матрица

Исходные группы	Предполагаемые группы			
	1	2	3	4
1	8	0	0	1
2	0	2	0	0
3	0	0	5	0
4	0	0	0	3
Неизвестные	33	10	27	4

Таблица 12 представляет собой классификационную матрицу для данных о голосовании в сенате. Шесть переменных Бардес правильно предсказывают распределение по фракциям всех сенаторов (кроме Кейпхарта), чья фракционная принадлежность «известна». Точность предсказания в этом случае — 94,7% (сумма правильных предсказаний—18, поделенная на общее число «известных» объектов). Мы также видим, что ошибки в этом примере связаны с плохим разделением групп 1 и 4. В нижней строке табл. 12 дано распределение по группам «неизвестных» объектов. Это те сенаторы, чью фракционную принадлежность Бардес не смогла определить по имеющимся у нее данным. Ее главной целью было использовать дискриминантный анализ для классификации позиций этих сенаторов по результатам их голосования, после чего она продолжила исследование отношения сената к различным вариантам помощи иностранным государствам.

Процент «известных» объектов, которые были классифицированы правильно является дополнительной мерой различий между группами. Им мы воспользуемся наряду с общей Λ-статистикой Уилкса и каноническими корреляциями для указания количества дискриминантной информации, содержащейся в переменных. Как непосредственная мера точности предсказания это процентное содержание является наиболее подходящей мерой дискриминантной информации. Однако о величине процентного содержания можно судить лишь относительно ожидаемого процента правильных классификаций, когда распределение по классам производилось случайным образом. Если есть два класса, то при случайной классификации можно ожидать 50% правильных предсказаний. Для четырех классов ожидаемая точность составит только 25%. Если для двух классов процедура классификации дает 60% правильных предсказаний, то ее эффективность довольна мала, но для четырех классов такой же результат говорит о значительной эффективности, потому что случайная классификация дала бы лишь 25% правильных предсказаний. Это приводит нас к τ-статистике ошибок, которая будет стандартизованной мерой эффективности для любого количества классов:

$$\tau = \frac{n_c - \sum_{i=1}^{g} p_i n_i}{n. - \sum_{i=1}^{g} p_i n_i}, \qquad (17)$$

где n_c — число правильно классифицированных объектов, а p_i — априорная вероятность принадлежности к классу.

Выражение $\sum_{i=1}^{g} p_i n_i$ представляет собой число объектов, которые будут правильно предсказаны при случайной классификации их по классам пропорционально априорным вероятностям. Если все классы считаются равноправными, то априорные вероятности полагаются равными единице, деленной на число классов. Максимальное значение τ-статистики равно 1 и оно достигается в случае безошибочного предсказания. Нулевое значение указывает на неэффективность процедуры. τ-статистика может принимать и отрицательные значения, что свидетельствует о плохом различении или вырожденном случае. Поскольку n_c должно быть целым числом, числитель может стать отрицательным чисто случайно, когда нет различий между классами.

Для данных Бардес каждая группа имеет априорную вероятность, равную 0,25. Следовательно, сумма в τ-статистике равна $(0{,}25 \cdot 9) + (0{,}25 \cdot 2) + (0{,}25 \cdot 5) + (0{,}25 \cdot 3) = 4{,}75$. Для 18 правильных предсказаний из 19 возможных τ-статистика составит:

$$\tau = \frac{18 - 4{,}75}{19 - 4{,}75} = \frac{13{,}25}{14{,}25} = 0{,}93.$$

Это означает, что классификация с помощью дискриминантных функций делает на 93% ошибок меньше, чем ожидалось при случайной классификации (т. е. одна действительная ошибка на 14,25 ожидаемых).

ОБОСНОВАНИЕ С ПОМОЩЬЮ РАЗБИЕНИЯ ВЫБОРКИ

Как и все методы вывода, основанные на выборочных данных, процент правильных предсказаний и τ-статистика имеют тенденцию к переоценке эффективности процедуры классификации. Это происходит потому, что обоснование решения производится по той же выборке, которая применялась для получения классифицирующих функций. Выражения, использованные при создании этих функций, чувствительны к выборочным погрешностям. Таким образом, функции отражают свойства конкретной выборки более точно, чем свойства всей генеральной совокупности[24].

Если выборка достаточно велика, то мы можем при обосновании процедуры классификации взять случайное разбиение выборки на два подмножества. Одно подмножество необходимо для получения функций, а другое — только для проверки классификаций. Поскольку подмножества имеют различные выборочные ошибки, тестовое подмножество даст лучшую оценку способности предсказания свойств генеральной совокупности.

Статистики расходятся во мнениях о целесообразных размерах двух подмножеств Одни рекомендуют выбирать их равными, тогда как другие предпочитают брать бо́льшими размеры того или друго-

го подмножества. Однако главное внимание необходимо уделять тому, чтобы подмножество, используемое для вывода функций, было достаточно велико для обеспечения стабильности коэффициентов, иначе проверка будет обречена на неудачу с самого начала.

Мы рассмотрели различные процедуры классификации, которые позволяют предсказать принадлежность конкретных объектов к определенным классам, дают нам полезную информацию: 1) об отдельных объектах; 2) о различиях между классами и 3) о способности переменных как целого точно различать классы. В нашем обсуждении до сих пор предполагалось, что выбор множества дискриминантных переменных является оптимальным. Теперь перейдем к выделению некоторых подмножеств этих переменных, которые оказываются более экономичными, но столь же эффективными, как все множество.

V. ПОСЛЕДОВАТЕЛЬНЫЙ ОТБОР ПЕРЕМЕННЫХ

Исследователи часто сталкиваются с ситуациями, когда в их распоряжении оказывается несколько возможных дискриминантных переменных, а они не уверены, все ли из этих переменных полезны и необходимы. Подобные ситуации часто возникают, когда затруднительно привести точный список дискриминантных переменных. В результате собираются данные о всех переменных, которые, как «предполагается», являются хорошими дискриминаторами, или же исследование носит предварительный характер и специалисты пытаются обнаружить полезные дискриминантные переменные.

В этих ситуациях одна или больше переменных могут оказаться плохими дискриминаторами, потому что средние классов слабо различаются по этим переменным. Кроме того, две или больше переменных могут нести одинаковую информацию, хотя каждая является хорошим дискриминатором. Если некоторые из них заняты в анализе, остальные оказываются лишними. Последние не вносят никакого вклада в анализ, (хотя сами по себе они могут быть хорошими дискриминаторами), потому что в них недостаточно *новой* информации. Если нет убедительных теоретических соображений в пользу сохранения таких «избыточных» переменных, их рекомендуется исключать, поскольку они только усложняют анализ и могут даже увеличить число неправильных классификаций.

Один из способов исключения ненужных переменных состоит в использовании процедуры последовательного отбора наиболее полезных дискриминантных переменных. Прямая процедура последовательного отбора начинается с выбора переменной, обеспечивающей наилучшее одномерное различение. Затем анализируются пары, образованные отобранной и одной из оставшихся переменными, после чего находится пара, дающая наилучшее различение, из которой и отбирается переменная. Далее процедура переходит к

образованию троек из первых двух и каждой из оставшихся переменных. Наилучшая тройка определяет третью переменную. На каждом шаге этой процедуры отбирается переменная, которая в сочетании с отобранными ранее дает наилучшее различение. Процесс продолжается до тех пор, пока не будут рассмотрены все возможные переменные или пока оставшиеся переменные не перестанут улучшать различение.

Процедура последовательного отбора может работать и в обратном направлении, т. е. когда все переменные первоначально считаются «входящими» в систему, а затем на каждом шаге отбрасывается одна, самая плохая. Прямой и обратный отборы могут сочетаться, но чаще применяется прямая процедура. Если какая-либо переменная больше не дает значимого вклада в процесс различения, то она отбрасывается, но на следующем шаге может быть снова отобрана. Устранение ранее отобранной переменной происходит потому, что она в значительной степени содержит ту же дискриминантную информацию, что и другие переменные, отобранные на предыдущих шагах. В то время когда эта переменная отбиралась, она вносила существенный вклад в процесс различения. Однако переменные, отобранные на последующих шагах, в сочетании с одной или несколькими, отобранными ранее, дублируют этот вклад, таким образом переменная становится избыточной и удаляется.

Процедуры последовательного отбора порождают оптимальное множество дискриминантных переменных, которое может не быть максимальной (наилучшей) комбинацией. Чтобы получить максимальное решение, нужно проверить все возможные сочетания (пары, тройки и т. д.). Такая проверка может оказаться дорогой и требующей больших временных затрат. Процедура последовательного отбора является логичным и эффективным способом поиска лучшей комбинации, но нет гарантии, что ее конечный продукт действительно превосходит все остальные.

Последовательность, в которой отбираются переменные, не обязательно соответствует их относительной значимости. Вследствие коррелированности (что разделяет дискриминантные возможности) даже хорошие дискриминаторы могут поздно попасть или вообще не попасть в последовательность, так как их вклад в различение может оказаться меньше вклада других переменных.

КРИТЕРИИ ОТБОРА

Процедуры последовательного отбора должны использовать некоторую меру качества различения как критерий отбора. Одним из таких критериев является Λ-статистика Уилкса, но существуют и другие возможности, позволяющие расширить наше представление о различиях между классами. В этом разделе мы рассмотрим некоторые из этих возможных мер, и попытаемся определить, какая из них «лучше» соответствует цели исследования. Часто конечный результат не зависит от выбора критерия, но так бывает не всегда.

Λ-статистика Уилкса и частное F-отношение

Λ-статистика Уилкса учитывает как различия между классами, так и когезивность, или однородность, каждого класса. Под когезивностью следует понимать степень скопления объектов вокруг центроида их класса. Поэтому переменная, которая увеличивает когезивность не изменяя разделение центроидов, при отборе может оказаться предпочтительнее переменной, увеличивающей разделение без изменения когезивности.

Поскольку Λ-статистика Уилкса является «обратной» статистикой, мы будем отбирать ту переменную, для которой на этом шаге она принимает *наименьшее* значение. Как обсуждалось раньше, мы можем преобразовать Λ-статистику в полную F-статистику для проверки различий между классами. Если такое преобразование происходит, то выбор производится по *наибольшему* значению. Вместо полного F-отношения мы можем воспользоваться частным F-отношением, которое вычисляется так же, как и значение F-включения (см. ниже). Использование всех трех статистик приводит к одному и тому же результату.

V-статистика Рао

Рао (1952; 257), применяя расстояние Махаланобиса, построил статистику, которая является мерой общего разделения классов. Это обобщенная мера расстояния, известная как V-статистика Рао, допустима при любом количестве классов. Она измеряет разделение центроидов классов и не касается когезивности внутри классов. Таким образом, переменная, отобранная с помощью V-статистики, может уменьшить внутригрупповую когезию и в то же время увеличить разделение всех классов. V-статистика измеряет расстояния от каждого центроида класса до главного центроида с весами, равными размеру соответствующего класса. Следовательно, V-статистика не обеспечивает максимального разделения между всеми парами классов. (Это верно и для Λ-статистики Уилкса.) Формула для V-статистики имеет вид

$$V = (n_{\cdot} - g) \sum_{i=1}^{p'} \sum_{j=1}^{p'} a_{ij} \sum_{k=1}^{g} n_j (X_{ik\cdot} - X_{i\cdot\cdot}) (X_{jk\cdot} - X_{j\cdot\cdot}), \qquad (18)$$

где p' — число отобранных переменных (включая отобранную на текущем шаге).

Когда рассматривается большое число объектов, V-статистика имеет выборочное распределение, приблизительно совпадающее с распределением хи-квадрат с $p'(g-1)$ степенями свободы. Кроме того, изменение V-статистики, вызванное добавлением (или удалением) переменных, также имеет распределение хи-квадрат с числом степеней свободы, равным $(g-1)$, умноженное на число переменных, добавленных (удаленных) на этом шаге. Мы можем использовать это свойство при проверке статистической зна-

чимости изменения общего разделения. Если изменение не является значимым, переменную можно не включать. При добавлении переменных изменение V-статистики может оказаться отрицательным, что означает ухудшение разделения центроидов.

Квадрат расстояния Махаланобиса между ближайшими классами

Можно попытаться выделить переменную, которая порождает наибольшее разделение пары классов, являющихся ближайшими на данном шаге. Это приведет к разделению всех классов. Мы можем выбрать одну из трех статистик, чтобы оценить качество разделения. Все они используют квадрат расстояния Махаланобиса между центроидами двух классов.

Конечно, одна из этих статистик — само расстояние D^2. Это прямая непосредственная мера, в которой всем парам классов приписываются равные веса.

Межгрупповая F-статистика

F-статистика различий между двумя классами дается следующей формулой:

$$F=\frac{(n_{\cdot}-g-p'+1)n_in_j}{p'(n_{\cdot}-g)(n_i+n_j)}D^2(G_i|G_j). \qquad (19)$$

Она отличается от формулы в тесте, использующем только квадрат расстояния, тем, что здесь учитываются выборочные размеры классов. Расстояния для малых классов получат меньшие веса, чем расстояния для больших классов. Таким образом, этот критерий стремится увеличить различия между парами, содержащими большие группы.

Минимизация остаточной дисперсии

Пятый возможный критерий предназначен для минимизации остаточной дисперсии между классами. Формула имеет вид

$$R=\sum_{i=1}^{g-1}\sum_{j=i+1}^{g}\frac{4}{4+D^2(G_i|G_j)}. \qquad (20)$$

Каждый член суммы равен единице минус квадрат множественной корреляции между множеством рассматриваемых дискриминантных переменных и фиктивной переменной, идентифицирующей соответствующую пару классов. Следовательно, R является остаточной дисперсией, потому что каждый член суммы представляет собой долю дисперсии фиктивной переменной, которую *нельзя* объяснить с помощью дискриминантных переменных. Иногда число пар классов делят на $g(g-1)/2$, чтобы получить среднюю остаточную дисперсию между классами, но это не влияет на вы-

бор переменных. Кроме того, если некоторым парам нужно придать значимость, бо́льшую по сравнению с другими, каждой паре можно приписать определенный вес (см. Dixon, 1973; 243).

Учитывая одновременно все пары классов, R содействует формированию равномерного разделения классов. Этот критерий слегка отличается от первых двух, в которых два класса могут оставаться близкими друг другу, а значительное улучшение разделения получено для других классов или за счет увеличения внутригрупповой когезии. Он также отличается от третьего и четвертого критериев, в которых основное внимание обращается только на самую тесную пару.

МИНИМАЛЬНЫЕ УСЛОВИЯ ПРОВЕДЕНИЯ ОТБОРА

Большинство программ последовательного отбора требует, чтобы любая переменная удовлетворяла определенному минимуму условий, прежде чем она будет подвергнута проверке в соответствии с критерием отбора. Так, проверка толерантности позволяет обеспечить необходимую точность вычислений, а воспользовавшись частной F-статистикой, мы можем установить, что возросшее различение превосходит уровень, заданный пользователем[25]. С помощью некоторых программ также просматривается список уже отобранных переменных, чтобы проверить, не надо ли какие-либо из них отбросить.

Толерантность

Тест толерантности может обеспечить точность вычислений. Толерантность еще не отобранной переменной равна единице минус квадрат множественной корреляции между этой переменной и всеми уже отобранными переменными, когда корреляции определяются по внутригрупповой корреляционной матрице. Если проверяемая переменная является линейной комбинацией (или приблизительно равна линейной комбинации) одной или нескольких отобранных переменных, то ее толерантность равна нулю (или близка к нулю). Переменная с малой толерантностью (скажем, меньше 0,001) может привести к ошибке при вычислении матрицы, обратной W, ввиду быстрого накопления ошибок округления. Помимо вычислительных проблем, нежелательно использовать переменную, которая является линейной комбинацией отобранных переменных, потому что она не дает никакой новой информации.

Статистика F-включения

Статистика F-включения представляет собой частную F-статистику, оценивающую улучшение различения от использования рассматриваемой переменной по сравнению с различением, достигнутым с помощью других уже отобранных переменных (Dixon, 1973; 241). Если величина статистики F-включения мала, мы вряд ли

отберем такую переменную, потому что она не дает достаточно большого вклада в различение. Эта частная F-статистика с числом степеней свободы, равным $(g-1)$ и $(n_.-p'-g+1)$, в качестве теста значимости, чтобы убедиться в статистической значимости улучшения различения. Переменная должна пройти проверку толерантности и F-включения, прежде чем она будет рассмотрена в соответствии с критерием отбора.

Статистика F-удаления

Статистика F-удаления также является частной F-статистикой с числом степеней свободы, равным $(g-1)$ и $(n_.-p'-g)$. Однако она оценивает значимость ухудшения различения после удаления переменной из списка уже отобранных переменных. Эта процедура проводится в начале каждого шага, чтобы проверить, имеется ли какая-нибудь переменная, уже не вносящая достаточно большого вклада в различение, поскольку отобранные позже переменные дублируют ее вклад.

На заключительном шаге статистика F-удаления может быть использована для ранжирования дискриминантных возможностей отобранных переменных. Переменная с наибольшим значением статистики F-удаления дает наибольший вклад в различение, достигнутое благодаря другим переменным. Переменная, имеющая вторую по величине статистику F-удаления, является второй по значимости и т. д. Это ранжирование не обязательно совпадает с тем, которое можно было бы получить с помощью одномерной F-статистики, потому что она измеряет полную дискриминантную способность переменной без учета дублирования ее другими переменными.

ПРИМЕР ИСПОЛЬЗОВАНИЯ ПРОЦЕДУРЫ ПОСЛЕДОВАТЕЛЬНОГО ОТБОРА

Для того чтобы понять, как процедура последовательного отбора работает на практике, применим эту методику к данным Бардес о голосовании в сенате. Когда квадрат расстояния Махаланобиса используется в качестве критерия отбора, мы получаем результаты, приведенные в табл. 13.

На первом шаге толерантность всегда равна 1,0, потому что переменные еще не были отобраны. По той же причине здесь статистика F-включения соответствует одномерной F-статистике. В четвертом столбце даны значения D^2, среди которых мы находим наибольшее. Это значение, равное 0,492, получено для переменной CUTASIAN при сравнении групп 1 и 4. Заметьте, что самая тесная пара (пара самых близких классов) для переменной CUTASIAN не является таковой ни для какой другой переменной (для четырех групп должны быть рассмотрены шесть пар). Наш выбор статистики квадрата расстояния в качестве критерия отбора основан на предположении, что мы хотим уделить больше

Таблица 13

Статистики включения для последовательного отбора

Переменные	Толерантность	Статистика *F*-включения	Квадрат расстояния	Группы
Шаг 1				
CUTAID	1,000	2,955	0,018	3 и 4
RESTRICT	1,000	0,943	0,004	1 и 3
CUTASIAN	1,000	11,915	0,492[1]	1 и 4
MIXED	1,000	2,628	0,038	3 и 4
ANTIYUGO	1,000	4,168	0,019	2 и 3
ANTINEUT	1,000	2,900	0,194	3 и 4
Шаг 2				
CUTAID	0,521	0,748	0,820	1 и 4
RESTRICT	0,684	3,418	0,495	1 и 4
MIXED	0,305	7,981	3,014	1 и 4
ANTIYUGO	0,851	2,898	3,370	3 и 4
ANTINEUT	0,383	8,502	3,801[1]	1 и 4
Шаг 3				
CUTAID	0,507	0,700	4,590	1 и 4
RESTRICT	0,446	1,228	5,405[1]	1 и 4
MIXED	0,282	1,496	5,094	1 и 4
ANTIYUGO	0,546	1,376	4,730	1 и 4
Шаг 4				
CUTAID	0,486	0,701	5,823	1 и 4
MIXED	0,282	1,378	6,743	1 и 4
ANTIYUGO	0,488	1,887	7,519[1]	1 и 4
Шаг 5				
CUTAID	0,407	1,234	7,523	1 и 4
MIXED	0,282	1,236	8,186[1]	1 и 4
Шаг 6				
CUTAID	0,330	0,672	9,043[1]	1 и 4

[1] Обозначает переменную, отобранную на этом шаге в соответствии с квадратом расстояния Махаланобиса между двумя самыми близкими группами.

внимания влиянию рассматриваемой переменной на разделение ближайших групп. Смысл использования здесь этого критерия состоит лишь в том, чтобы проиллюстрировать работу последовательного отбора. В этом примере переменная CUTASIAN является очевидным выбором, поскольку для нее значения и квадрата расстояния, и статистики *F*-включения, намного больше, чем для любой другой переменной. Стоит отметить, однако, что на этом шаге квадрат расстояния для переменной ANTINEUT более чем в 10 раз превосходит соответствующее значение для CUTAID, в то время как значения статистики *F*-включения для них почти равны.

На втором шаге процедуры снова вычисляются все необходи-

мые статистики с учетом отобранной переменной CUTASIAN. Теперь толерантность почти наверное станет меньше единицы, поскольку она равна единице минус квадрат корреляции между CUTASIAN и другой переменной. Статистика F-включения равна частной F-статистике, отвечающей увеличению дискриминантных возможностей за счет использования соответствующей переменной после того, как переменная CUTASIAN реализовала все свои возможности. А квадрат расстояния равен наименьшей из величин, полученных для всех шести пар групп с помощью CUTASIAN и данной переменной. Здесь у ANTINEUT наибольшее из данных наименьших значений.

На шаге 3 процесс повторяется. Поскольку в качестве критерия выбора («включения») мы используем квадраты расстояний, следующей «включается» переменная RESTRICT. Однако если в качестве критерия отбора мы применяем Λ-статистику Уилкса, косвенно измеряемую статистикой F-включения, то мы выбрали бы MIXED. Расхождение вызвано тем, что каждый критерий придает особое значение собственному аспекту процесса различения.

Остающиеся шаги проводятся таким же образом до тех пор, пока не будут включены все переменные. На шаге 6 CUTAID имеет настолько малое значение статистики F-включения, что порой лучше отказаться от ее анализа.

К тому же на шаге 6 значение статистики F-удаления для ANTINEUT, падает до величины 0,996. Некоторые исследователи могут прийти к заключению, что удаление ANTINEUT оправдано, так как это значение действительно слишком мало. Тогда переходим к шагу 7, на котором будет рассматриваться включение CUTAID и ANTINEUT. Как только обнаруживается, что ни одна из этих переменных не имеет достаточно высокого значения статистики F-включения, процесс отбора будет остановлен и в дальнейшем дискриминантном анализе и классификации будут использоваться другие четыре переменные.

Этот пример специально построен так, чтобы в конечном итоге были употреблены все переменные, поскольку реальное исследование Бардес также включало все шесть переменных. В действительности у нее были причины применять все переменные, поэтому она совсем не пользовалась процедурой последовательного отбора. Если кто-то собирается работать со всеми переменными, то вряд ли применение последовательного анализа принесет ему пользу. Разумно использовать эту методику для определения переменных, которые надо исключить из-за малого вклада в процесс различения. На основе данных табл. 13 можно даже утверждать, что отбор переменных должен быть оставлен на шаге 2, поскольку ни одно из значений F-статистики не является значимым на шаге 3. Поэтому после шага 2 можно перейти к классификации. Если классифицировать только по двум переменным (CUTASIAN и ANTINEUT), ошибок будет столько же или меньше, чем при классификации по всем шести переменным. Это дает нам право

отбросить остальные четыре. В некоторых случаях использование бóльшего числа переменных приводит к ухудшению классификации.

Цель последовательного отбора — найти более экономичное подмножество, которое обладало бы такими же (если не лучшими) дискриминантными возможностями, что и полное множество. Кроме рассмотрения вопроса о возможности применения последовательного отбора, исследователь сталкивается с такими практическими проблемами, как влияние нарушений предположений, лежащих в основе дискриминантного анализа, и последствия пропуска данных. Заключительный раздел посвящен этим неприятным, но важным проблемам.

VI. ЗАКЛЮЧИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

НАРУШЕНИЕ ПРЕДПОЛОЖЕНИЙ

Мы уже немного говорили о проблемах, возникающих, когда данные не удовлетворяют математическим предположениям дискриминантного анализа. Труднее всего удовлетворить требованиям о нормальности многомерного распределения дискриминантных переменных и равенстве ковариационных матриц классов. Некоторые исследователи (см., в частности, Lachenbruch, 1975) показали, что дискриминантный анализ является достаточно устойчивым методом, допускающим некоторые отклонения от этих предположений. Кроме того, не все выводы дискриминантного анализа требуют их выполнения.

Предположение о нормальности многомерного распределения важно для проверки значимости, где сопоставляются статистики, вычисленные по выборочным данным, с теоретическим вероятностным распределением для этой статистики. Можно вычислить теоретическое распределение, сделав некоторые удобные математические предположения (например, такие, как требование, чтобы генеральная совокупность имела многомерное нормальное распределение). Если интересующая нас генеральная совокупность не удовлетворяет этому требованию, истинное выборочное распределение статистики будет отличаться от распределения, полученного теоретически. Различия между этими двумя распределениями могут быть очень малыми или очень большими в зависимости от степени нарушения предположений. Лахенбрук (1975) показал, что дискриминантный анализ не очень чувствителен к небольшим нарушениям предположения о нормальности. Это приводит лишь к некоторым потерям в эффективности и точности.

Предположение о нормальности играет важную роль в классификации, основанной на использовании вероятности принадлежности к классу. Эти вероятности вычисляются исходя из распределения хи-квадрат, что оправдано лишь, когда дискриминантные переменные имеют многомерное нормальное распределение. Если это предположение не выполняется, вычисленные вероятности будут неточными. Может оказаться, например, что вероятности

для некоторых групп будут преувеличены, в то время как вероятности для других групп — недооценены. Следовательно, эта процедура не будет оптимальной в смысле уменьшения числа неправильных классификаций.

Если ковариационные матрицы классов не равны, мы стараемся установить искажения дискриминантных функций и уравнений классификации. Один источник ошибок связан с вычислением внутригрупповой ковариационной матрицы (или других, имеющих отношение к матрице W). Внутригрупповая ковариационная матрица служит оценкой общей ковариационной матрицы классов для генеральной совокупности, образованной выборками из нескольких классов. Если матрицы для всей генеральной совокупности не равны, матрицу W все еще можно вычислить, но она уже не будет способствовать упрощению различных формул. Следовательно, канонические дискриминантные функции не дадут максимального разделения классов и вероятности принадлежности к классам будут искажены. Хотя, кажется, нет никаких процедур улучшения свойств канонических дискриминантных функций в некоторых цитированных выше работах предлагается использовать ковариационные матрицы отдельных классов для вычисления вероятности принадлежности к классу (так называемая «квадратичная дискриминация»).

Дискриминантный анализ может быть проведен и когда предположения о нормальности многомерного распределения и равенстве ковариационных матриц классов не выполняются. Задача при этом состоит в интерпретации результатов. Что они означают? И какое количество ошибок считается допустимым? В некоторых учебниках предлагаются возможные процедуры, но они приводят лишь к минимальным улучшениям, поскольку исходные отклонения не были большими. Конечно, нам трудно узнать, сколько ошибок было сделано в связи с конкретными нарушениями предположений. Однако здесь могут оказаться полезными некоторые статистики, не зависящие от этих предположений.

При определении значимости и минимального числа канонических дискриминантных функций мы не полагаемся на Λ-статистику Уилкса или связанный с ней тест значимости, основанный на хи-квадрат распределении. Вместо этого мы можем рассмотреть каноническую корреляцию и относительное процентное содержание, как было показано в разд. II. Если любая из данных величин окажется небольшой, функция будет для нас малоинтересной, даже если она — статистически значима. Тесты значимости представляют наибольший интерес в случае малых выборок. Таким образом, имея с ними дело, мы должны с бóльшим вниманием отнестись к удовлетворению предположений. Однако в случае больших выборок мы может обойтись без тестов значимости или использовать их «консервативно», когда наши данные нарушают предположения.

При классификации точность предсказания наиболее важна для объектов, расположенных вблизи границы. Если некоторый

объект с вероятностью 0,90 принадлежит к классу 1 и только с вероятностью 0,10 — к классу 2, то нам нечего беспокоиться о небольших неточностях, возникающих из-за нарушения предположений. Хотя определенная вероятность принадлежности к классу может быть неверной, наше решение приписать объект к классу 1 будет правильным, если ошибка в вычислении вероятностей не будет большой. С другой стороны, если объект имеет вероятности 0,51 для класса 1 и 0,49 для класса 2, мы должны быть очень осторожны, принимая решение. Здесь небольшая ошибка из-за нарушения предположений может привести к неправильной классификации.

Если исследователя интересует математическая модель, с помощью которой можно точно предсказывать принадлежность к классу или которая служит разумным описанием реального мира, то лучше всего воспользоваться процентом правильных классификаций. Если этот процент высок, то нарушение предположений не нанесет большого вреда. Однако, если процент правильных классификаций низок, мы не можем сказать, является ли причиной этого нарушение предположений или использование плохих дискриминантных переменных.

ДРУГИЕ ПРОБЛЕМЫ

Несколько других проблем, которые выходят за рамки этой работы, могут доставить много неприятностей пользователю дискриминантного анализа. К ним относятся: большое количество отсутствующих данных, сильно коррелированные переменные, переменная с нулевым стандартным отклонением внутри одного или нескольких классов, большие различия в размерах классов и выбросы. Хотя здесь эти проблемы не обсуждаются, читатель должен сознавать, что такие «патологии» могут оказать отрицательное влияние на точность и интерпретацию результатов дискриминантного анализа.

ЗАКЛЮЧЕНИЕ

Хотя в этом разделе внимание сфокусировано на некоторых проблемах и трудностях, возникающих при использовании дискриминантного анализа, не следует их бояться. В практических исследованиях мы часто сталкиваемся с данными, которые не согласуются с предположениями, лежащими в основе статистических методов. Зная требования, предъявляемые моделью, можно определить, когда они были нарушены, когда следует применить корректирующие меры и когда методика не соответствует целям данного исследования.

Эта работа была задумана как введение в дискриминантный анализ и включает ряд статистических процедур, предназначенных, во-первых, для изучения многомерных различий между двумя и более классами (что мы назвали «интерпретацией») и, во-

вторых, для использования нескольких переменных для предсказания принадлежности объекта к некоторому классу («классификация»). Математическая модель обычно предполагает, что переменные измеряются по интервальной шкале и имеют многомерное нормальное распределение. Мы ограничились обсуждением линейного дискриминантного анализа, который обычно требует равенства ковариационных матриц классов. Имея это в виду, исследователь может широко использовать подпрограммы дискриминантного анализа в стандартных пакетах компьютерных программ, таких, как SPSS, BMD и SAS. Читатель, желающий больше узнать обо всех особенностях дискриминантного анализа, может обратиться к работам, приведенным в списке литературы.

ПРИМЕЧАНИЯ

1. В работах (Stevens, 1946; 1951) даны определения четырех шкал измерений, принятых в статистике: наименований, порядковая, интервальная и отношений. Для интервальных измерений характерно то, что истинная разность последовательных единиц шкалы равна разности двух любых последовательных целых единиц этой шкалы. Вообще говоря, измерения по интервальной шкале соответствуют непрерывному случаю, но это ограничение совершенно не обязательно. В дискриминантном анализе требуется вычисление средних вариаций и ковариаций, поэтому измерения должны производиться на интервальном уровне. Дальнейшие сведения о шкалах измерений можно найти в работах (Blalock, 1979; Nie, 1975) и в любом вводном статистическом курсе.

2. Ковариация двух переменных является мерой их совместного изменения. Ковариация аналогична коэффициенту корреляции, но без приведения к стандартизованному виду при различных масштабах в измеряемых переменных. Соответственно ковариация может принимать любые значения и не ограничена константами —1 снизу и +1 сверху.

Часто ковариации представляются в виде матриц. Каждой переменной в матрице соответствует одна строка и один столбец. На пересечении данной строки и данного столбца находится ковариация двух переменных. На главной диагонали находятся вариации. Если данные разделены на группы, можно вычислить ковариационную матрицу для каждой группы в отдельности, используя наблюдения, принадлежащие только данной группе. Для того чтобы две ковариационные матрицы были равны, должны быть равны все соответствующие им элементы. Понятия вариации и ковариации даются во всех вводных статистических курсах, например (Blalock, 1979).

3. Большинство представленных здесь таблиц заимствовано из работы (Bardes, 1975). Везде, где это было необходимо, мы сами обработали экспериментальные данные, которые были любезно предоставлены нам Бардес. Численные значения, приведенные в работе Бардес для коэффициентов, центроидов и дискриминантных значений, не совпадают с представленными в данной работе из-за отличия способов стандартизации дискриминантных функций. Однако это не влияет на результаты интерпретации и классификации. Была использована компьютерная программа SPSS DISCRIMINANT, реализованная на ЭВМ типа IBM 360/370.

4. Можно выполнить анализ вариаций по каждой переменной в отдельности для того, чтобы выявить статистическую значимость межгрупповых отличий (см. работу Inversen, Norpoth, 1976).

Переменные, которые не дают значимых межгрупповых отличий, нужно исключить из дискриминантного анализа. Следует иметь в виду, что критерии значимости (в строгом статистическом смысле) неприменимы к данным Бардес, поскольку она изучает генеральную совокупность, а не выборку.

5. Матрица является двумерным массивом чисел. Обозначая матрицу одним символом, мы подразумеваем сразу множество чисел, объединенных в эту матрицу. Каждое число, принадлежащее матрице, называется ее элементом. Обозначением для элемента является буква с двумя индексами, первый из которых обозначает строку, где расположен элемент, а второй — столбец. Так, например, t_{ij} — элемент матрицы T, расположенный на пересечении строки i и столбца j.

Если $i=j$ в квадратной матрице, соответствующий элемент находится на главной диагонали. Матрица называется квадратной, если число столбцов равно числу строк. Для симметричной матрицы $t_{ij}=t_{ji}$, т. е. элементы над главной диагональю равны соответствующим элементам, расположенным ниже главной диагонали.

6. Дальнейшие детали, касающиеся нахождения собственных векторов и приведения дискриминантных функций к стандартному виду, можно найти в работе (Cooley, Lohnes, 1971). Кулн и Лохиес предложили приводить к стандартному виду матрицу T вместо матрицы W. Вообще говоря, эта операция является корректной, хотя дискриминантное значение некоторым образом изменяется. Как указывалось, W дает дискриминантное значение, измеренное в единицах стандартного отклонения *по каждой группе в отдельности*. Использование матрицы T приводит к дискриминантным значениям в единицах стандартного отклонения *по всему пространству*, поэтому эти значения являются меньшими числами. Выбор T и W не влияет на результаты интерпретации или классификации. Однако значения вероятностей $Pr(X|G_k)$, введенных в разд. IV, могут быть вычислены только, если используется матрица W.

Все примеры, рассмотренные в данной работе, основаны на стандартизации матрицы W.

7. Если коэффициенты дискриминантных функций приводятся к стандартному виду с использованием матрицы W, значение 1,0 соответствует одному стандартному отклонению *по данной группе*. Другими словами, если рассмотреть наблюдения, принадлежащие данной группе и вычислить их стандартное отклонение от *группового среднего* дискриминантной функции, то полученные значения будут равны единице. Подразумевается, что групповые ковариационные матрицы равны между собой и точно представляются межгрупповой ковариационной матрицей.

Если, например, нужно вычислить стандартное отклонение для *всех* наблюдений по отношению к *главному* среднему, то результирующие значения будут больше единицы (исключение — когда групповые центроиды совпадают). Причина этого заключается в том, что именно группы, а не вся система в совокупности определяют единицы измерения расстояний. Как отмечалось ранее, можно приводить к стандартному виду матрицу T, при этом стандартное отклонение от общего главного среднего по всем наблюдениям будет единичным.

8. Заметим, что направление функции является произвольным. Изменение знаков коэффициентов данной функции эквивалентно изменению направления соответствующей оси. В общем случае все направления равноправны. Но в некоторых случаях все-таки можно выделить направления, к которым «тяготеют» отдельные наблюдения. Например, для данных Бардес позиции либералов соответствует отрицательная область данных, а позиции консерваторов — положительная.

9. Отметим, что значения переменных — одни и те же для всех дискриминантных функций. Дело в том, что объекты имеют только одно значение по каждой переменной.

10. Под стандартной, или Z-формой подразумевается, что переменная должна иметь нулевое среднее и единичное стандартное отклонение.

11. Для читателей, знакомых с понятием множественной регрессии, должна быть ясна аналогия интерпретации нестандартизованных и стандартизованных дискриминантных коэффициентов с регулярным и стандартизованным коэффициентом множественной регрессии. В рассматриваемом примере стандартное отклонение шести переменных приблизительно равно. Соответственно относительная величина коэффициентов при их стандартизации изменяется незначительно. Иная ситуация наблюдается, когда стандартные отклонения отличаются друг от друга.

12. Если для приведения коэффициентов к стандартному виду используется матрица W, то дискриминантные переменные должны быть стандартизованы к общему среднему и *межгрупповым* стандартным отклонениям. Если же применяется матрица T — переменные должны быть приведены к общему среднему и к *общему* стандартному отклонению. На практике для вычисления дискриминантных значений мы обычно имеем дело с наблюдаемыми значениями переменных и нестандартизованными коэффициентами дискриминантной функции. Стандартизованные коэффициенты используются только для проведения интерпретации.

13. Структурные коэффициенты могут быть получены двумя способами. Первый применяет компьютерную программу вычислении дискриминантных значений каждой функции по соотношению (1), а затем — программу для вычислений коэффициентов корреляции Пирсона между функциями и переменными. Другой способ заключается в вычислении стандартизованных коэффициентов канонической дискриминантной функции по следующей формуле:

$$c^*_{kj} = \frac{v_{kj}\sqrt{t_{kk}}}{\sqrt{\sum_{i=1}^{p}\sum_{m=1}^{p} v_{ij}v_{mj}t_{im}}}, \qquad (21)$$

где c^*_{kj} — коэффициент j-й дискриминантной функции по k-й переменной. (Знаменатель этого равенства является константой и может быть вычислен один раз.)

Структурные коэффициенты получаются из соотношения:

$$s_{ij} = \sum_{k=1}^{p} r_{ik}c_{kj} = \sum_{k=1}^{p} \frac{t_{ik}c^*_{kj}}{\sqrt{t_{ii}t_{kk}}}, \qquad (22)$$

где s_{ij} — структурный коэффициент корреляции переменной i и функции j, а r_{ik} — корреляция между переменными i и k.

14. Эти результаты подразумевают наличие положительной корреляции между парами переменных. Если корреляция отрицательна, может наблюдаться противоположный эффект. На практике наличие множественных корреляций сильно затрудняет интеграцию стандартизованных коэффициентов.

15. Реальная значимость — это соответствие результата исследования физическому смыслу (содержанию) задачи.

16. Во многих учебниках по статистике применяются термины *каноническая переменная* для обозначения того, что мы называем «канонической дискриминантной функцией» и дискриминантная функция, которую мы в разд. IV называем «классифицирующей функцией». Другие авторы, например Кули и Лохнес (1971), применяют термин *дискриминантная функция* к «канонической дискриминантной функции». Чтобы избежать этой терминологической путаницы, мы будем пользоваться терминами «каноническая дискриминантная функция» и «классифицирующая функция».

17. Читатель должен заметить, что канонические корреляции в табл. 9 получены для небольшого числа объектов (19). Большие выборки (1000 объектов и более) затрудняют получение больших корреляций, поскольку обычно они являются более однородными.

18. Под «генеральными данными» понимается, что рассматриваемые данные об объектах исчерпывают всю генеральную совокупность. Они не являются выборкой.

19. Читатели, незнакомые с понятием статистической значимости, должны обратиться к работе (Henkel, 1976) или любому учебнику по статистике, в котором рассматриваются статистические выводы. Важно понимать, что статистическая значимость и реальная значимость — это разные понятия Статистическая значимость в первую очередь связана с проверкой, является ли выборка достаточно большой, чтобы можно было с уверенностью сказать, что рассматриваемая статистика действительно отличается от гипотетической величины (обычно нуль или «нет отличия»). Для больших выборок статистика может быть статистически значимой и не иметь реальной значимости (например, небольшое значение канонической корреляции).

20. Распределение хи-квадрат и F-распределение — теоретические вероятностные распределения, которые измеряют вероятность, что различия в групповых средних, отмеченные в выборке, вызваны случайными выборочными отклонениями, а не действительными различиями в генеральной совокупности. Каждое из этих распределений имеет свою собственную форму, зависящую от числа «степеней свободы», связанного с конкретной задачей. Необходимо знать число степеней свободы, прежде чем по таблице определить уровень вероятности с вычисленными значениями распределения хи-квадрат или F-распределения.

21. Объекты в этом примере не являются простой случайной выборкой. Следовательно, тест значимости при строгой интерпретации основных предположений неприменим.

22. Некоторые программы «дискриминантного анализа» (такие, как BMD05M и подпрограмма в SAS76) выполняют только классификацию и не вычисляют каионические дискриминантные функции.

23. Матрица, обратная квадратной, — это матрица, в которой при умножении ее на исходную диагональные элементы становятся равными единице, а все остальные — нулю. Чтобы получить более полную информацию об обратных матрицах и о том, как их вычислять, необходимо обратиться к учебнику статистики, где используется матричная алгебра (см. (Cooley and Lohnes, 1971)).

24. Эта проблема обсуждается в (Lachenbruch, 1975; 29—36).

25. Формулы для толераитности, статистик F-включения и F-удаления довольно сложны и здесь не приводятся. Заинтересовавшийся читатель может обратиться к работе (Norusis, 1979; 73—74).

ЛИТЕРАТУРА

ANDERSON, T.W. (1958) An Introduction to Multivariate Statistical Analysis. New York: John Wiley.

BARDES, B.A. (1976) "Senatorial support for foreign policy: a comparison of alternative explanations." Presented at the meeting of the Midwest Political Science Association, Chicago, April 29-May 1.

BARDES, B.A. (1975) "Senatorial realignment on foreign aid, 1953-1972: a discriminant analysis of inter-party factions." Ph.D. dissertation, University of Cincinnati.

BARR, A.J., J.H. GOODNIGHT, J.P. SALL, and J.T. HELWIG (1976) A Users's Guide to SAS-76. Raleigh, NC: Sparks Press.

BLALOCK, H.M., Jr. (1979) Social Statistics. New York: McGraw-Hill.

COOLEY, W.W. and P.R. LOHNES (1971) Multivariate Data Analysis. New York: John Wiley.

DIXON, W.J. [ed.] (1973) BMD: Biomedical Programs. Berkeley: University of California Press.

EISENSTEIN, J. and H. JACOB (1977) Felony Justice. Boston: Little, Brown.

FISHER, R.A. (1936) "The use of multiple measurements in taxonomic problems." Annals of Eugenics 7:179-188.

HENKEL, R.E. (1976) Tests of Significance. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-004. Beverly Hills and London: Sage Publications.

HEYCK, T.W. and W.R. KLECKA (1973) "British radical M.P.'s, 1874-1895: new evidence from discriminant analysis." Journal of Interdisciplinary History 4(Autumn): 161-184.

IVERSEN, G.R. and H. NORPOTH (1976) Analysis of Variance. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-001. Beverly Hills and London: Sage Publications.

KENDALL, M.G. (1968) A Course in Multivariate Analysis. New York: Hafner.

KLECKA, C.O. (1974) "The measurement of children's masculinity and femininity." Ph.D. dessertation, Northwestern University.

KLECKA, W.R. (1975) "Discriminant analysis," pp. 434-467 in N. Nie et al. SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill.

KLECKA, W.R. (1973) "The clientele of Australian parties: new perspectives through discriminant analysis." Politics 7:301-308.

KORNBERG, A. and R.C. FRASURE (1971) "Policy differences in British parliamentary parties." American Political Science Review 65:694-703.

LACHENBRUCH, P.A. (1975) Discriminant Analysis. New York: Hafner.

LEVINE, M.S. (1977) Canonical Analysis and Factor Comparison. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-006. Beverly Hills and London: Sage Publications.

MAHALANOBIS, P.C. (1963) "On the generalized distance in statistics." Proceedings of the National Institute of Science, India 12:49-55.

MORRISON, D.G. (1974) "Discriminant analysis," pp. 2.442-2.457 in R. Ferber (ed.), Handbook of Marketing Research. New York: John Willey.

MORRISON, D.G. (1969) "On the interpretation of discriminant analysis." Journal of Marketing Research 6:156-163.

NIE, N.H., C.H. HULL, J.G. JENKINS, K. STEINBRENNER, and D.H. BENT (1975) SPSS: Statistical Package for the Social Sciences. New York: McGraw-Hill.

NORUSIS, M.J. (1979) SPSS Statistical Algorithms: Release 8.0. Chicago: SPSS, Inc.
OVERALL, J.E. and C.J. KLETT (1972) Applied Multivariate Analysis. New York: McGraw-Hill.
RAO, C.R. (1965) Linear Statistical Inference and Its Applications. New York: John Wiley.
RAO, C.R. (1952) Advanced Statistical Methods in Biometric Research. New York: John Wiley.
STEVENS, S.S. (1951) "Mathematics, measurement, and psychophysics," pp. 1-49 in S.S. Stevens (ed.), The Handbook of Experimental Psychology. New York: John Wiley.
STEVENS, S.S. (1946) "On the theory of scales of measurement." Science 103:677-680.
TATSUOKA, M.M. (1971) Multivariate Analysis. New York: John Wiley.
TATSUOKA, M.M. and D.V. TIEDEMAN (1954) "Discriminant analysis." Review of Educational Research 24:402-420.
VAN DE GEER, J.P. (1971) Introduction to Multivariate Analysis for the Social Sciences. San Francisco: W.H. Freeman.
VELDMAN, D.J. (1967) Fortran Programming for the Behavioral Sciences. New York: Holt, Rinehart & Winston.

М. С. Олдендерфер, Р. К. Блэшфилд

КЛАСТЕРНЫЙ АНАЛИЗ

Mark S. Aldenderfer, Roger K. Blashfield. **Cluster Analysis** (Second Printing, 1985).

ПРЕДИСЛОВИЕ

Классификация объектов по осмысленным группам — кластеризация — является важной процедурой в области социологических исследований. Несмотря на широкое применение понятий кластеризации, кластерный анализ как формальная многомерная статистическая процедура понимается все еще плохо. Отчасти это объясняется тем, что последние десять лет техника кластеризации разрабатывалась особенно быстро, поскольку стали доступны вычислительные машины, способные выполнить большое число необходимых операций. Данный метод разрабатывается и применяется археологами, психологами, специалистами по государственному праву и социологии, поэтому часто приходится пользоваться нестандартизованной, приводящей к путанице терминологией. В связи с этим новые разработки медленно распространяются на другие дисциплины.

Настоящая работа была задумана как введение в кластерный анализ для тех, кто не имеет соответствующей подготовки и нуждается в современном и систематическом путеводителе по «лабиринту» понятий, методов и алгоритмов, связанных с идеей кластеризации. Вначале обсуждаются меры сходства — обязательная отправная точка любого анализа процесса кластеризации. Авторы отмечают расхождения в теоретических значениях этого понятия и рассматривают ряд эмпирических мер, чаще всего применяемых для измерения сходства. Затем описываются различные методы для фактической идентификации кластеров, а также процедуры обоснования и проверки адекватности результатов кластерного

ISBN 0-8039-2376-7

анализа, на что часто не обращается внимание. В работе проводятся сравнение и оценка различных понятий и методов.

Поскольку вычислительные машины почти всегда позволяют провести кластерный анализ больших множеств данных, авторы рассматривают ряд стандартных и специализированных программ. Кроме того, после каждого раздела помещены библиографические замечания. В приложении приводятся первичные данные, использованные в примерах, так что читатель может проверить, правильно ли он разобрался в описанных процедурах.

Поскольку книга сводит воедино сведения из очень обширного круга источников, читатель получит довольно полное руководство по современному применению статистических методов и вычислительных программ.

Ричард Нимн,
редактор серии

I. ВВЕДЕНИЕ

Древняя китайская классификация животных

Животные подразделяются на: (а) принадлежащих императору; б) набальзамированных; в) дрессированных; г) молочных поросят; д) сирен; е) сказочных; ж) бродячих собак; з) включенных в данную классификацию; и) дрожащих, как сумасшедшие; к) неисчислимых; л) нарисованных самой лучшей верблюжьей кисточкой; м) других; н) тех, которые только что разбили цветочную вазу и о) тех, которые издалека напоминают мух (Хорхе Луис Борхес, *Другие исследования*: 1937—1952).

Классификация является основой человеческой умозрительной деятельности. Дети очень рано начинают классифицировать объекты, окружающие их, и давать названия получаемым классам. Классификация является фундаментальным процессом научной практики, поскольку системы классификаций содержат понятия, необходимые для разработки теорий в науке.

«Кластерный анализ» — это общее название множества вычислительных процедур, используемых при создании классификации. В результате работы с процедурами образуются «кластеры» или группы очень похожих объектов. Более точно, кластерный метод — это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.

Первые работы, упоминающие о кластерных методах, появились давно, но бóльшая часть литературы по кластерному анализу была написана в течение последних двух десятилетий. Импульсом для разработки многих кластерных методов послужила книга «Начала численной таксономии», опубликованная в 1963 г. двумя биологами — Робертом Сокэлом и Петером Снитом. Сокэл и Снит утверждали, что эффективная процедура для создания биологических классификаций должна обеспечивать сбор всевозможных данных об интересующих организмах, оценивать степень сходства между этими организмами и применять некоторый метод кластеризации, чтобы поместить достаточно схожие организмы в одну и ту же группу. После этого состав каждой группы можно проанализировать, чтобы выяснить, представляют ли они разные биологические виды. Фактически Сокэл и Снит полагают, что «структура отражает процесс», т. е. структура замеченных различий и сходств между организмами может служить основой для понимания эволюционного процесса.

После выхода книги Сокэла и Снита объем литературы по кластерному анализу резко возрастает. Число публикаций о приложениях кластерного анализа во всех отраслях науки удваивается каждые три года (Blashfield and Aldenderfer, 1978 b). На наш взгляд, существуют две причины для такого возросшего интереса к кластерному анализу: 1) появление высокоскоростных компьютеров и 2) фундаментальное значение классификации как научного метода. До появления вычислительных машин применение кластерных методов для обработки больших объемов данных практически было невозможно. Для кластеризации множества данных из 200 объектов необходимо определить матрицу сходства, имеющую 19 900 уникальных значений. Определение матрицы такого размера без вычислительных машин столь утомительно и требует так много времени, что найдется мало исследователей (или их несчастных помощников), которые отважились бы на это. С широким распространением вычислительной техники стала возможной и обработка больших матриц.

Второй причиной повышенного интереса к кластеризации является то, что наука строится на классификациях, которые привносят порядок в исследования. Она содержит основные понятия, используемые наукой. Например, классификация химических элементов лежит в основе неорганической химии и атомной теории материи; классификация болезней является структурной основой медицины. Поскольку кластерные методы рассматриваются как объективные, легко воспроизводимые способы создания классификаций, то они пользуются широкой популярностью.

Ученые давно применяют кластерный анализ. Среди самых ранних из этих исследований были работы антропологов, которые определяли однородные культурные области, используя матричные методы (см. Czekanowski, 1911; Driver, 1965; Johnston, 1972). В психологии кластерный анализ рассматривался как «факторный анализ бедняка» (Tryon, 1939). Специалисты других дисциплин, особенно государственного права, также участвовали в ранних разработках методов кластеризации для общественных наук. Хотя многие теории и приложения, служившие основой кластеризации в прошлом, были отвергнуты последующими поколениями, все социальные науки и сейчас сохраняют некоторые традиции использования кластерных методов.

Несмотря на их популярность, кластерные методы все еще понимаются хуже, чем такие многомерные статистические процедуры, как факторный анализ, дискриминантный анализ и многомерное шкалирование. Литература по социальным наукам содержит ошеломляющее количество часто несовместимых терминов, методов и предпочитаемых подходов. Недостаток опубликованных руководств для начинающих в сочетании с разнобоем в терминологии и методологии затрудняют изучение кластерного анализа. Цель нашей работы — провести новичка через этот «лабиринт» кластерного анализа. Ввиду большого разнообразия методов, предложенных за последние двадцать лет, мы не сможем исчерпывающе рас-

смотреть все или даже часть методов. Поэтому мы остановимся на тех, которые сравнительно хорошо известны в области социальных наук, и, как мы полагаем, имеют достоинства, позволяющие использовать их в прикладных исследованиях.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Как мы уже отмечали, методы кластеризации конструируются для создания однородных групп объектов или единиц, которые называются кластерами. Различные приложения кластерного анализа можно свести к четырем основным задачам:

1) разработка типологии или классификации;

2) исследование полезных концептуальных схем группирования объектов;

3) порождение гипотез на основе исследования данных;

4) проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Как правило, кластерный анализ используется для создания классификаций, но в большинстве случаев прикладного анализа данных в основе исследования лежит комбинация этих задач. Чтобы лучше их понять, рассмотрим следующий пример применения кластерного анализа.

Алкоголизм — главная проблема в области психиатрии США, однако классификация больных алкоголизмом до недавнего времени не получала широкого распространения среди профессиональных психиатров. Голдстейн и Линден (1969), психологи-клиницисты, построили такую классификацию на основе кластерного анализа. Они с помощью психологического теста MMPI (миннесотское многофазовое обследование личности — Minnesota Multiphasic Personality Inventory) собрали данные о 513 больных алкоголизмом, проходивших лечение в госпитале в Индианаполисе, штат Индиана. Тест содержал 566 вопросов (типа да/нет), которые суммировались по 13 диагностическим шкалам (например, шкала шизофрении, шкала истерии).

Голдстейн и Линден разделили полученные данные на две части: основная подвыборка (239 больных) и контрольная подвыборка (251 больной). Используя основную подвыборку, они сформировали корреляционную матрицу размерностью 239×239, которая представляла сходства для MMPI-профилей этих больных, и применили кластерный метод, изобретенный Лорром (1966). Из больных основной подгруппы 114 были распределены по четырем кластерам, а оставшиеся 125 не были приписаны ни к какому кластеру. Когда такой же обработке подверглась контрольная подвыборка, снова были выделены четыре кластера, которые содержали 106 (из 251) больных алкоголизмом. Главные профили кластеров для обоих подвыборок были в основном одинаковые. Голдстейн и Линден назвали эти четыре кластера следующим образом: 1) эмоционально-неустойчивые личности; 2) психоневротики с бес-

покойством/депрессией; 3) психопатические личности и 4) больные алкоголизмом, употребляющие наркотики и обладающие параноидальными чертами.

Исследоваиие Голдстейна и Линдена сыграло важную роль, поскольку послужило моделью для более 15 последующих работ, в которых примеиялся кластерный анализ для классификации больных алкоголизмом. Большинство из этих работ в основном подтвердили обоснованиость выделения первых двух кластеров (типы I и II).

Другая работа была выполнена двумя антропологами Бертоном и Ромни (1975). Они решили исследовать, как в английском языке можио классифицировать термины, соответствующие статусу и роли индивидов в системе человеческих отношений. Данные, использованные в работе, были получены в результате классификаци 58 наиболее общих терминов, среди которых типичными были: «художник», «босс», «друг», «человек», «владелец», «поэт» и «шпион». Участвовавшим в исследовании были розданы карточки с этими словами и затем было предложено произвольным образом разложить их по группам. На число и размеры групп никаких ограничений не накладывалось. Сходство между группами терминов определялось с помощью *Z*-оценки (Muller, 1969).

Исследуя данные о сходстве с помощью неметрического многомерного шкалироваиия, авторы пытались выявить наличие некой латентной структуры для описания сходства (различия) данных терминов. Были выделеиы три шкалы: оценочная шкала, в соответствии с которой такие термины, как «игрок», «бандит» и «шпион», противопоставляются терминам «друг» и «компаньон»; шкала иерархии, по которой выявляется различие между терминами «босс» или «бригадир» и понятиями родственных и дружеских отношений, например «друг»; шкала профессиональной принадлежности, позволяющая выделить роли и соответствующие термины, связанные с работой. Затем они провели иерархический кластерный анализ, применив два различных метода к одним и тем же данным о близости. Для каждого из этих методов авторы выбрали решение, состоящее из восьми кластеров. При этом они отметили, что результаты применения методов, хотя и различаются во многих отношениях, все же имеют четыре общих кластера: 1) кластер, включающий семь терминов родства; 2) кластер дружеских отношений; 3) кластер терминов принадлежности к социальным группам и 4) кластер управленческих ролей. Бертон и Ромни пришли к выводу, что результаты, полученные с помощью обоих методов, дополняют друг друга. Они полагают, что лица, классифицировавшие термины, принимали решения на основе двух критериев отбора. Первый, упрощенный критерий подобен полученному с помощью многомерного шкалирования (оценочная шкала, шкалы иерархии и профессиональной принадлежности). Второй, более тонкий критерий подсказан результатами кластеризации и подобен четкой структуре английских терминов родственных отношений, основаиной на степеии родства между индивидами, не за-

висящей от половых различий. Результаты, полученные с помощью кластерного анализа, подтверждают неоднозначность полоролевых критериев в западном обществе, отмеченных социологами, и, кроме того, проясняют, как в английском языке классифицируются термины родства.

Последний пример — социологическое исследование Филсингера, Фолкнера и Уорленда (1969) — предназначался для создания классификации верующих. Данные были собраны с помощью шкалы религиозности (De Jong et al., 1976) в форме вопросника, который был предложен 547 старшекурсникам университета штата Пенсильвания. Было выбрано 37 вопросов, а план измерений был разработан на основе предыдущего факторного анализа этих данных (De Jong et. al., 1976). Поскольку выборка из 547 студентов слишком велика и требует для обработки больших затрат, для исследования была использована выборка из 220 студентов. Матрица сходства между парами индивидов размерностью 220×220 подверглась кластеризации. Для анализа было выбрано решение, состоявшее из семи кластеров, соответствующих семи типам верующих:

тип I — неверующие;
тип II — консерваторы;
тип III — нигилисты;
тип IV — умеренно религиозные;
тип V — крайне религиозные;
тип VI — ортодоксы;
тип VII — поклонники культа и обрядности.

Филсингер и другие сделали также попытку проверить обоснованность классификации верующих. Сначала они провели дискриминантный анализ кластеров и, как было сказано, результаты имели высокую значимость[1]. Затем сравнили объекты из разных кластеров по семи демографическим признакам. По четырем признакам (размер общины; политические взгляды; процент студентов, не связанных с церковью; принадлежность к той или иной религии) кластеры имели значимые различия. Авторы пришли к заключению, что перекрывающиеся результаты подтверждают их эмпирическую типологию верующих.

В приведенных примерах можно найти любую из основных задач кластерного анализа. Целью Голдстейна, Линдена, Филсингера и других было построение классификации, но заметную роль играет и исследование классификационных схем (MMPI и шкалы религиозности). Работа Бертона и Ромни в первую очередь была посвящена анализу данных и проверке гипотез, а построение формальной классификации было второстепенной задачей. В этом случае еще до проведения проверки гипотез авторы заметили, что их результаты подтверждаются данными, полученными с помощью более традиционных антропологических методов.

Эти примеры показывают, что, несмотря на различия в целях, типах данных и примененных методах, все исследования, исполь-

зующие кластерный анализ, характеризуют следующие пять основных шагов:

1) отбор выборки для кластеризации;

2) определение множества признаков, по которым будут оцениваться объекты в выборке;

3) вычисление значений той или иной меры сходства между объектами;

4) применение метода кластерного анализа для создания групп сходных объектов;

5) проверка достоверности результатов кластерного решения.

Каждый из перечисленных шагов играет существенную роль при использовании кластерного анализа в прикладном анализе данных.

МНОЖЕСТВА ДАННЫХ, ИСПОЛЬЗУЕМЫХ В ПРИМЕРАХ

Мы воспользуемся только одним набором данных, чтобы показать, как применяют кластерные методы. Еще один набор приводится в приложении для того, чтобы заинтересованный читатель мог поэкспериментировать с процедурами, которые мы описываем; наши результаты могут служить ориентирами для сравнения.

Первое множество данных, используемое в качестве примера, представляет собой гипотетические данные об археологических раскопках древних захоронений. Эти данные могут содержать важную для археологов информацию о социальном статусе или положении, занимаемом в обществе индивидами, найденными в могилах. Тщательно анализируя содержимое захоронений, археологи могут сделать выводы о различиях в общественном положении индивидов, что в свою очередь может помочь определить природу социального расслоения и уровень развития общества, к которому они принадлежали.

Данные изменяются в зависимости от трех измерений: возраста, пола и статуса. На участке наших археологических раскопок были «захоронены» 25 человек, которые были разделены на три возрастные группы: дети, подростки и взрослые. Представлены два статуса: элитарный и неэлитарный. Во всех захоронениях содержится до восьми различных видов объектов, найденных в них: местная керамика, наконечники стрел, обломки браслетов, обработанные камни, костяные иглы, костяные шилья, привозная керамика и металлические изделия. Каждый из этих видов объектов соответствует определенному статусу и полу; возрастные различия объектов не были включены в данные, чтобы структура набора данных оставалась относительно простой. Данные были закодированы в двоичной форме с помощью регистрации наличия или отсутствия объекта.

Второй набор данных, также искусственного происхождения, был специально создан в качестве модели классификационной задачи, с которой часто сталкиваются в психопатологии. Основное множество данных содержит информацию о 90 гипотетических

больных с тремя типами психических расстройств: психозы (П) неврозы (Н) и расстройства личности (РЛ). В каждую общую группу входило по тридцать больных. Более подробно о процессе генерации данных можно прочесть у Блэшфилда и Мори (1980). Характер заболевания больных определялся по 13 стандартным шкалам, взятым из психологического теста MMPI, описанного ранее в работе Голдстейна и Линдена (1969). Эти шкалы имеют следующие названия и аббревиатуры:

• Шкалы достоверности данных:

L — шкала лжи;
F — шкала фальсификаций;
K — шкала поправок.

• Клинические шкалы:

Hs — ипохондрия;
D — депрессия;
Hu — истерия;
Pd — психопатические отклонения;
Mf — шкала пола (мужской/женский);
Pa — паранойя;
Pt — психастения;
Sc — шизофрения;
Ma — гипомания;
Si — социальная интроверсия.

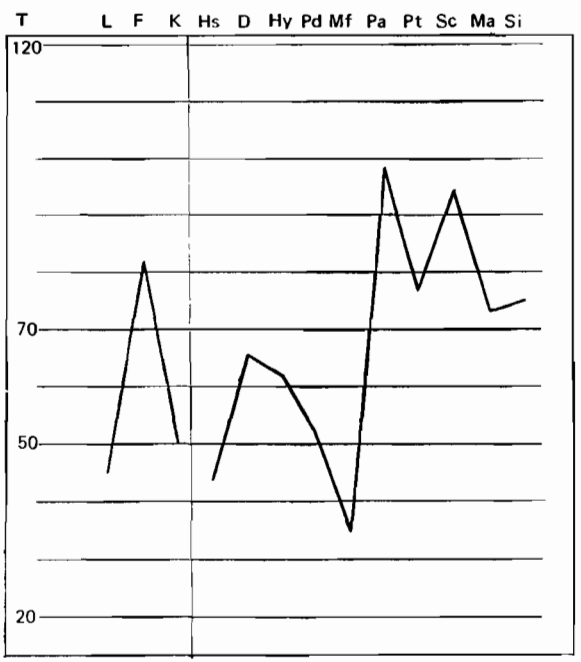

Рис. 1. Пример профиля данных MMPI-теста

Тест MMPI состоит из 566 вопросов типа да/нет, ответы на которые даются от первого лица (например, «Я люблю читать технические журналы»). Эмпирическим путем эти 566 вопросов были сгруппированы в шкалы MMPI-теста. В ходе разработки данный тест предлагался как здоровым пациентам, так и больным с психическими заболеваниями. Вопрос относился к той или иной шкале, если он позволял отделить группу больных от здоровых. Все десять «клинических» шкал были сформированы именно таким образом, а имена шкал представляют собой клинические названия групп больных, для диагностирования которых эти шкалы предназначались. Три другие стандартные шкалы являются шкалами достоверности ответов, т. е. определяют, в какой степени пациент может фальсифицировать свои симптомы.

Результаты MMPI-теста интерпретировались с помощью профилей данных о больных. На рис. 1 приведены результаты MMPI-теста для одного из 90 больных, представленных множеством данных. Значения признака изображены на профиле для каждой шкалы, при этом число 50 — нормальное значение признака, а число 70 указывает на значимое отличие от нормы. Прифили различаются в первую очередь по «пикам» или шкалам, которые имеют наивысшие значения. Для рассматриваемого пациента наивысшие значения расположены в следующем порядке Pa, Sc, F, Pt, Si и Ma. Этот профиль довольно типичен для больного с диагнозом параноидальная шизофрения.

НЕСКОЛЬКО ПРЕДОСТЕРЕЖЕНИЙ ОТНОСИТЕЛЬНО КЛАСТЕРНОГО АНАЛИЗА

Прежде чем перейти к обсуждению основных методологических этапов проведения кластерного анализа, необходимо сделать несколько предостережений общего характера.

1) *Многие методы кластерного анализа — довольно простые процедуры, которые, как правило, не имеют достаточного статистического обоснования**. Другими словами, большинство методов кластерного анализа являются эвристическими (подкрепленными лишь опытом разработчиков). Они — не более чем правдоподобные алгоритмы, используемые для создания кластеров объектов. В этом резкое отличие, например, от методов факторного анализа, который хорошо обоснован статистически. Хотя многие кластерные методы обладают важными, подробно исследованными математическими свойствами (см. Jardin and Sibson, 1971), все же важно сознавать их простоту. В этом случае маловероятно, что пользователь допустит ошибку при трактовке результата кластерного анализа.

---

* Достаточно строгая теория, охватывающая большую часть постановок задач кластер-анализа, была разработана французскими исследователями. Ее изложение можно найти в книге Э. Дидэ и др. «Методы анализа данных» (М.: Финансы и статистика, 1985). — *Примеч. ред.*

2) *Методы кластерного анализа разрабатывались для многих научных дисциплин, а потому несут на себе отпечатки специфики этих дисциплин.* Это важно отметить, потому что каждая дисциплина предъявляет свои требования к отбору данных, к форме их представления, к предполагаемой структуре классификации. Что может быть полезным в психологии, может оказаться ненужным для биологов, а так как кластерные методы порой не более чем правила для создания групп, то пользователь должен знать те особенности, которые часто сопровождают обсуждение и описание методов кластеризации.

3) *Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.* Это обычное явление в большинстве прикладных исследований. Одной из причин неодинаковых решений является то, что кластерные методы получены из разных источников, которые предопределяли использование различных правил формирования групп. Данная ситуация вносит в работу с кластерным анализом путаницу не только для начинающих, но и для опытных пользователей. Кроме того, желательно иметь специальную методику, позволяющую проверить, насколько «естественны» группы, выделенные методом кластеризации в наборе данных. Было разработано несколько процедур, способных помочь в решении этой задачи.

4) *Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные,* т. е. методы кластеризации необходимы для обнаружения структуры в данных, которую нелегко найти при визуальном обследовании или с помощью экспертов. Эта ситуация отличается от ситуации дискриминантного анализа, который более точно определяется как процедура идентификации. Последний приписывает объекты к уже существующим группам, а не создает новые группы. Хотя цель кластеризации и заключается в нахождении структуры, на деле кластерный метод привносит структуру в данные и эта структура может не совпадать с искомой, «реальной». Кластерный метод всегда размещает объекты по группам, которые могут радикально различаться по составу, если применяются различные методы кластеризации. Ключом к использованию кластерного анализа является умение отличать «реальные» группировки от навязанных методом кластеризации данных.

II. МЕРЫ СХОДСТВА

ТЕРМИНОЛОГИЯ

Для описания особенностей оценивания сходства создавалась специальная терминология. Как мы покажем позднее (см. разд. V), развитие жаргона кластерного анализа в различных отраслях науки связано с быстрым ростом и распространением самого кластерного анализа. Терминология какой-либо дисциплины образуется

таким образом, что она может перекрывать терминологию других дисциплин, даже если термины используются для описания одних и тех же предметов. Если потенциальный пользователь кластерного анализа не осведомлен о таких терминологических различиях, это может привести к большой путанице.

Термин «событие», «единица», «случай», «паттерн», «предмет», OTE (операционная таксономическая единица) обозначают объект, тогда как «переменная», «признак», «свойство», «характеристика» обозначают те черты «объектов», которые позволяют оценить их сходство. Другая группа важных терминов — «Q-анализ» и «R-анализ»; первый из них относится к связям между переменными. Кластерный анализ, например, традиционно рассматривается как «Q-техника», в то время как факторный анализ — как «R-техника».

Потенциальный пользователь кластерного анализа должен также обратить внимание на то, что матрицы данных часто формируются различными способами. В общественных науках обычно совокупность данных изображают в виде матрицы, образованной N событиями (строки матрицы), которые определяются P переменными (столбцы матрицы). В биологии имеет место обратный порядок, что приводит к матрице данных размерностью $P \times N$. В этой работе мы воспользуемся термином «первичные данные» для описания исходной матрицы событий размерностью $N \times P$ и их переменных до вычисления сходства. В соответствии с этим мы будем употреблять термины «матрица сходства» или «матрица близости» для описания матрицы сходств событий размерностью $N \times N$, вычисленной с помощью некоторой меры сходства по первичным данным.

Даже термин «сходство» не свободен от смыслового многообразия, а его синонимами являются «подобие», «близость», «связанность», «ассоциативность». Однако другие авторы ограничивают использование термина «коэффициент сходства». Например, Эверитт (1980) пользуется термином «коэффициент сходства» для обозначения тех мер, которые Сниг и Сокэл (1973) называют «коэффициентами ассоциативности». Клиффорд и Стефенсон (1975) для еще большей путаницы сводят применение термина «коэффициент ассоциативности» к значению, которое является частным случаем определений, данных Эвериттом, а также Снитом и Сокэлом. Мы будем пользоваться термином «коэффициент сходства» (или «мера сходства») и придерживаться классификации коэффициентов сходства, предложенной Снитом и Сокэлом (1973), которые подразделили эти коэффициенты на четыре группы:

1) коэффициенты корреляции;
2) меры расстояния;
3) коэффициенты ассоциативности;
4) вероятностные коэффициенты сходства.

Позже каждая из групп будет кратко описана.

ПОНЯТИЕ СХОДСТВА

То, что некоторые вещи обнаруживают между собой сходство или различие, является весьма важным моментом для процесса классификации. Несмотря на кажущуюся простоту, понятие сходства и особенно процедуры, используемые при измерении сходства, не так просты. В самом деле, понятие сходства тесно связано с такими основополагающими эпистемологическими проблемами, как: «Каким образом мы можем образовывать полезные абстрактные понятия, позволяющие внести порядок в то, что мы знаем?». Конечно, чтобы ответить на этот вопрос, нужно уметь рассортировывать вещи по классам, что требует умения объединять вещи, воспринимающиеся как схожие. Проблема сходства состоит, однако, не в простом распознавании сходных или несходных вещей, а в том, какое место эти понятия занимают в научных исследованиях. Наука для плодотворного развития должна базироваться на объективных, воспроизводимых процедурах; таким образом, разработка статистических процедур для измерения более «объективного» сходства вещей является естественным следствием необходимости в воспроизводимых и надежных классификациях.

Количественное оценивание сходства отталкивается от понятия *метрики*. При этом подходе к сходству события представляются точками координатного пространства, причем замеченные сходства и различия между точками находятся в соответствии с метрическими расстояниями между ними (Tversky, 1977). Размерность пространства определяется числом переменных, использованных для описания событий. Существует четыре стандартных критерия, которым должна удовлетворять мера сходства, чтобы быть метрикой:

1) *Симметрия.* Даны два объекта x и y; расстояние между ними удовлетворяет условию

$$d(x, y) = d(y, x) \geqslant 0.$$

2) *Неравенство треугольника.* Даны три объекта x, y, z; расстояния между ними удовлетворяют условию

$$d(x\,y) \leqslant d(x,z) + d(y,z).$$

Очевидно, это просто утверждение, что длина любой стороны треугольника меньше или равна сумме двух других сторон. Полученное выражение также называется метрическим неравенством.

3) *Различимость нетождественных объектов.* Даны два объекта x и y:

$$\text{если } d(x,y) \neq 0, \text{ то } x \neq y^*.$$

4) *Неразличимость идентичных объектов.* Для двух идентичных объектов x и x'

$$d(x,x') = 0,$$

т. е. расстояние между этими объектами равно нулю**.

* Если $x \neq y$, то $d\ (x,\ y) \neq 0$. *Примеч. ред.*

** Если $d\ (x,\ x') = 0$, то объекты x и x' идентичны. — *Примеч. ред.*

Перечисленные математические требования очень важны, поэтому многие исследователи, среди которых наиболее известны Джардин и Сибсон (1971), а также Клиффорд и Стефенсом (1975), выдвигают аргументы против механического использования коэффициентов сходства, не являющихся метриками. Не все из обсуждаемых ниже мер расстояния могут быть метриками. Ряд корреляционных мер метриками не являются. Коэффициенты, не представляющие собой метрики, могут не быть совместно монотонными; другими словами, значения различных коэффициентов на одних и тех же данных не будут согласованно изменяться. Это вызывает опасение, что коэффициенты могут указывать на наличие сильно различающихся зависимостей между объектами. Поскольку такая признанная мера сходства, как смешанный момент корреляции Пирсона, явно не удовлетворяет третьему критерию, и, как считают Клиффорд и Стефенсон (1975), во многих приложениях может не выполняться второй критерий (т. е. неравенство треугольника), то следует проверить, действительно ли некоторая мера является метрикой.

Несмотря на очевидную важность метрик, они — отнюдь не единственный способ описания сходства объектов. Конечно, исходя из философских соображений, которые начинают подтверждаться психофизиологическими исследованиями, возможно понимание сходства объектов как сравнение их характерных черт; таким образом, оценивание сходства может основываться на процессе сопоставления признаков (Tversky, 1977). Такое понятие сходства не приводит к естественной размерности для своего представления. Более того, есть большая группа социальных исследований, в которых сходство между объектами оценивается непосредственно. Например, можно брать за основу степень взаимосвязи объектов, и в исследованиях такого рода часто наблюдается асимметрия отношения сходства. Иначе говоря, объект A может соответствовать B, но B может не соответствовать A в той же степени (например, Адам может любить Бетти, хотя он Бетти вовсе не нравится). Такой тип отношений также свойствен экономике в случае, когда одно государство импортирует товаров из другой страны больше, чем оно экспортирует. Асимметрия вызывает дополнительные затруднения при вычислении коэффициентов сходства. Тверски (1977) дает хорошее введение в эти вопросы. Специалисты по кластерному анализу должны сознавать, что существует несколько видов сходства и что, хотя многие из коэффициентов и мер, обычно используемых в количественных подходах к классификации, являются метриками, все же имеются альтернативы применению этих мер, которые могут оказаться целесообразными и необходимыми в контексте исследования. Выбор меры сходства, таким образом, должен быть органической частью плана исследования, определяющегося теоретическим, практическим и философским содержанием задачи классификации.

ВЫБОР ПЕРЕМЕННЫХ

Прежде чем приступить к описанию весьма распространенных коэффициентов, используемых при оценке сходства, необходимо сделать небольшое отступление и рассказать о выборе переменных и преобразовании данных, предшествующих оцениванию. Выбор переменных в кластерном анализе является одним из наиболее важных шагов в исследовательском процессе, но, к сожалению, и одним из наименее разработанных. Основная проблема состоит в том, чтобы найти ту совокупность переменных, которая наилучшим образом отражает понятие сходства. В идеале переменные должны выбираться в соответствии с ясно сформулированной теорией, которая лежит в основе классификации. Теория является базисом для разумного выбора переменных, необходимых в исследовании. На практике, однако, теория, обосновывающая классификационные исследования, часто не сформулирована, и поэтому бывает трудно оценить, насколько выбор переменных соответствует поставленной задаче.

Важность наличия теории для руководства выбором переменных нельзя недооценивать. Искушение скатиться к наивному эмпиризму в использовании кластерного анализа очень сильно, так как метод специально создан для получения «объективной» группировки объектов. Под «наивным эмпиризмом» мы понимаем отбор и последующий анализ как можно большего количества переменных в надежде на то, что «структура» проявится, как только будет собрано достаточное количество данных. Хотя эмпирические исследования важны для любой науки, те из них, в основе которых лежит наивный эмпиризм, опасны при применении кластерного анализа ввиду эвристической природы метода и большого числа нерешенных проблем, которые компрометируют приложения (Everitt, 1979).

В большинстве видов статистического анализа данные обычно подвергаются нормировке некоторым подходящим способом. При проверке, имеет ли переменная нормальное распределение, часто производится логарифмическое или какое-нибудь другое преобразование. В том случае, если данные измерены в разных масштабах, нормировка обычно проводится таким образом, чтобы среднее равнялось нулю, а дисперсия — единице. Имеются, однако, некоторые разногласия относительно того, должна ли нормировка быть стандартной процедурой в кластерном анализе. Как указывает Эверитт (1980), нормировка к единичной дисперсии и нулевому среднему уменьшает различия между группами по тем переменным, по которым наилучшим образом обнаруживались групповые различия. Более целесообразно проводить нормировку переменных *внутри* групп (т. е. внутри кластеров), но, очевидно, этого нельзя сделать, пока объекты не разнесены по группам.

Эдельброк (1979) отметил, что переменные многомерных данных могут менять значения параметров распределения от группы к группе; таким образом, нормировка может не быть равносильным

преобразованием для этих переменных и даже может изменять соотношения между ними. Однако, исследовав методом Монте-Карло воздействие нормировки на последующий анализ с использованием коэффициента корреляции и различных иерархических кластерных методов, Эдельброк не обнаружил существенных различий в результатах классификации по нормированным и ненормированным переменным. Миллиган (1980) также показал, что нормировка, по-видимому, оказывает незначительное воздействие на результаты кластерного анализа. Другие, особенно Мэттьюз (1979), продемонстрировали, что нормировка отрицательно сказывается на адекватности результатов кластерного анализа по сравнению с «оптимальной» классификацией объектов исследования.

Ситуация относительно нормировки не совсем ясна. Пользователи, имеющие данные с существенно различными измерениями, без сомнения, захотят стандартизировать их, особенно если применяется такая мера сходства, как евклидово расстояние. Решение о проведении нормировки должно приниматься с учетом специфики решаемой задачи, при этом пользователь должен понимать, что результаты могут различаться в зависимости от принятого решения, хотя величина воздействия будет меняться от одного множества данных к другому.

Возможны и другие виды преобразования данных, многие из которых применяются одновременно с кластерным анализом. Факторный анализ и метод главных компонент часто используются в том случае, когда известно, что переменные, взятые для исследования, сильно коррелированы. Наличие сильно коррелированных переменных при вычислении меры сходства приводит, по существу, к взвешиванию этих переменных. Так, если есть три сильно коррелированные переменные, то их совместное действие эквивалентно действию лишь одной переменной, которая имеет вес, в три раза превышающий вес каждой из первоначальных переменных. Метод главных компонент и факторный анализ могут применяться для уменьшения размерности данных, тем самым создавая новые, некоррелированные переменные, которые будут употребляться в качестве первичных данных при вычислении сходства между объектами. Использование процедуры преобразования данных вызывает много споров. В факторном анализе существует тенденция к ослаблению связей между кластерами, поскольку предполагается, что факторные переменные нормально распределены. Действие факторного анализа приводит к такому преобразованию данных, при котором зависимые переменные сливаются в одну, нормально распределенную. Рольф (1970) отметил, что метод главных компонент стремится к такому преобразованию данных, при котором хорошо разделенные кластеры остаются таковыми и в редуцированном пространстве, но при этом уменьшается расстояние (и тем самым ослабляются связи) между кластерами или группами, которые были разделены слабо.

Полемика ведется и вокруг вопроса о необходимости взвеши-

вания переменных. Особенно много таких дискуссий в области биологии. Взвешивание — это манипулирование значением переменной, позволяющее ей играть бóльшую или меньшую роль в измерении сходства между объектами (Williams, 1971). Хотя идея взвешивания и проста, ее практическое применение затруднительно. Уильямс описывает пять видов взвешивания, из которых чаще всего использует выбор весов априори. Снит и Сокэл (1973) решительно возражают против априорного взвешивания и считают, что наиболее подходящий способ измерения сходства состоит в присвоении всем переменным равных весов. Однако необходимо учитывать, что Снит и Сокэл рассматривают кластеризацию как чисто эмпирический подход к созданию классификаций. Во многих случаях имеет смысл взвешивать некоторые переменные априори, если для этого есть хорошее теоретическое обоснование и процедура, позволяющая осуществить взвешивание. Поскольку вопрос взвешивания еще не стал предметом обсуждения в общественных науках, исследователи, пользующиеся кластерными методами, должны знать о существовании разногласий.

МЕРЫ СХОДСТВА

Теперь, когда задача выбора переменных и преобразования данных обсуждены, можно познакомиться с наиболее известными коэффициентами сходства. Как уже отмечалось, существует четыре их вида: коэффициенты корреляции; меры расстояния; коэффициенты ассоциативности и вероятностные коэффициенты сходства. Каждый из этих видов имеет свои достоинства и недостатки, которые следует рассматривать прежде, чем будет принято решение использовать один из них. Хотя все четыре вида мер сходства широко применялись специалистами в численной таксономии и в биологии, лишь коэффициенты корреляции и расстояния получили широкое распространение в области социальных наук. Поэтому мы уделим больше внимания этим двум типам мер.

Коэффициенты корреляции

Коэффициенты корреляции, часто называемые угловыми мерами ввиду их геометрической интерпретации, — самый распространенный тип сходства в области социальных наук. Наиболее известным является смешанный момент корреляции, предложенный Карлом Пирсоном. Первоначально использованный в качестве метода определения зависимости переменных, он был применен в количественной классификации при вычислении корреляции между объектами. В связи с этим коэффициент вычисляется следующим образом:

$$r_{jk}=\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}},$$

где x_{ij} — значение i-й переменной для j-го объекта; x_j — среднее всех значений переменных j-го объекта, а n — число переменных.

Для такого метода берутся переменные, измеренные по шкалам отношений или шкалам интервалов, а в случае бинарных данных он преобразуется в известный φ-коэффициент. Значение коэффициента корреляции изменяется от —1 до +1, причем значение нуль указывает, что между объектами нет связи. Поскольку при вычислении среднего для каждого объекта суммирование производится по всем переменным этого объекта, то стандартные критерии значимости для r здесь не имеют ясного смысла.

Часто говорят, что коэффициент корреляции оценивает *форму* в том смысле, что он нечувствителен к различиям в величине переменных, используемых для вычисления коэффициента. Как отметил Уильямс (1971), коэффициент Пирсона r чувствителен только к форме из-за неявной нормировки каждого обекта по всем переменным. Это свойство особенно важно для приложений к таким отраслям науки, как психология, социология и антропология, в которых данные часто описываются в терминах профилей. Формально профиль определяется просто как вектор значений признаков объекта, графически изображаемый в виде ломаной линии. Например, данные MMPI-теста, использованные в нашей работе, часто изображают так, чтобы для каждого индивида получилась ломаная — профиль (см. рис. 1).

Одним из главных недостатков коэффициента корреляции как меры сходства, является то, что он чувствителен к форме за счет снижения чувствительности к величине различий между переменными. Кронбах и Глезер (1953) впервые показали, что сходство между профилями определяют следующие три элемента: *форма,* т. е. спуски и подъемы ломаной линии для всех переменных; *рассеяние,* т. е. дисперсия значений переменных относительно их среднего; *поднятие* (уровень или сдвиг), т. е. среднее значение для объекта по всем переменным. Чувствительность коэффициента корреляции Пирсона лишь к форме означает, что два профиля мо-

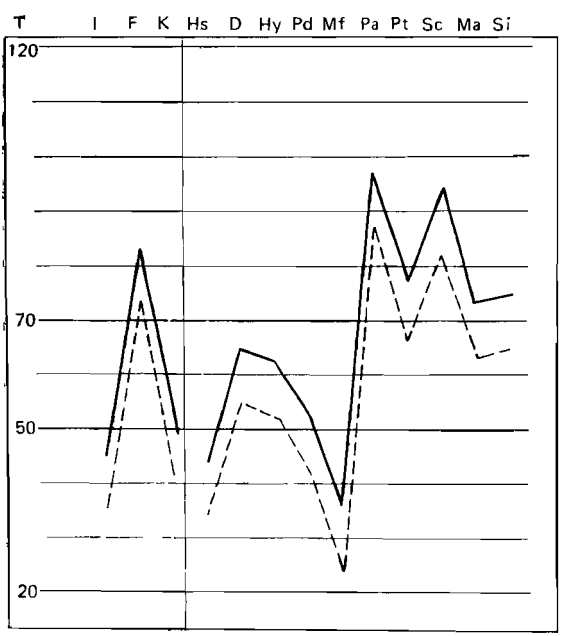

Рис. 2. Профили данных MMPI-теста

гут иметь корреляцию +1,0, и все же не быть идентичными (т. е. профили объектов не проходят через одни и те же точки). На рис. 2 показаны два профиля данных для MMPI-теста; один изображен сплошной линией, а другой — пунктирной. Формы их одинаковы. Хотя корреляция между этими двумя профилями равна +1,0, они все же не совпадают, потому что один из них приподнят. Таким образом, высокая корреляция между профилями будет наблюдаться, когда измерения одного из профилей будут линейно зависеть от измерений другого. Следовательно, при использовании коэффициента корреляции теряется некоторая информация, что может привести к неверным результатам, если не будет учтено влияние рассеяния и поднятия профиля.

Коэффициент корреляции имеет и другие недостатки. Он часто не удовлетворяет неравенству треугольника, и, как многие указывали, корреляция, вычисленная этим способом, не имеет статистического смысла, поскольку среднее значение определяется по совокупности всевозможных разнотипных переменных, а не по совокупности объектов. Смысл «среднего» по разнотипным переменным далеко не ясен.

Несмотря на эти недостатки, коэффициент широко использовался в приложениях кластерного анализа. Хаммер и Каннингхем (1981) показали, что при правильном применении кластерного метода коэффициент корреляции превосходит другие коэффициенты сходства, так как позволяет уменьшить число неверных классификаций. Парадоксально, но ценность корреляции заключается именно в том, что она не зависит от различий между переменными из-за рассеяния и сдвига. Существенную роль в успехе работы Хаммера и Каннингхема сыграло, однако, то, что исследователи смогли понять, что им нужен именно коэффициент формы, поскольку они считали, что влияние рассеяния и сдвига данных объясняется лишь субъективизмом критиков, а не недостатками, присущими этим классификациям.

Меры расстояния

Меры расстояния пользуются широкой популярностью. На практике их лучше бы называть мерами *несходства;* для большинства используемых коэффициентов большие значения соответствуют большему сходству, в то время как для мер расстояния дело обстоит наоборот. Два объекта идентичны, если описывающие их переменные принимают одинаковые значения. В этом случае расстояние между ними равно нулю. Меры расстояния обычно не ограничены сверху и зависят от выбора шкалы (масштаба) измерений. Одним из наиболее известных расстояний является евклидово расстояние, определяемое как

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

где d_{ij} — расстояние между объектами i и j, а x_{ik} — значение k-й переменной для i-го объекта. Чтобы избежать применения квадратного корня, часто величина расстояния возводится в квадрат, на что обычно указывает обозначение d^2_{ij}. Как и следовало ожидать, это выражение называют «квадратичным евклидовым расстоянием».

Можно определить и другие виды расстояния. Так, хорошо известной мерой является манхеттенское расстояние, или «расстояние городских кварталов» (city-block), которое определяется следующим образом:

$$d_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{ik}|.$$

Можно определить и другие метрики, но большинство из них являются частными формами специального класса метрических функций расстояния, известных как метрики Минковского, которые можно найти по формуле

$$d_{ij} = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \right)^{1/r}.$$

Существуют расстояния, не являющиеся метриками Минковского, и наиболее важное из них — расстояние Махаланобиса D^2, которое также носит название обобщенного расстояния (Mahalanobis, 1936). Эта метрика определяется выражением

$$d_{ij} = (x_i - x_j)' \Sigma^{-1} (X_i - X_j),$$

где Σ — общая внутригрупповая дисперсионно-ковариационная матрица, а X_i и X_j — векторы значений переменных для объектов i и j. В отличие от евклидовой и метрик Минковского, эта метрика с помощью матрицы дисперсий-ковариаций связана с корреляциями переменных. Когда корреляция между переменными равна нулю, расстояние Махаланобиса эквивалентно квадратичному евклидову расстоянию.

Несмотря на важность евклидовой и других метрик, они имеют серьезные недостатки, из которых наиболее важный состоит в том, что оценка сходства сильно зависит от различий в сдвигах данных. Переменные, у которых одновременно велики абсолютные значения и стандартные отклонения, могут подавить влияние переменных с меньшими абсолютными размерами и стандартными отклонениями. Более того, метрические расстояния изменяются под воздействием преобразований шкалы измерения переменных, при которых не сохраняется ранжирование по евклидову расстоянию. Чтобы уменьшить влияние относительных величин переменных, обычно перед вычислением расстояния нормируют переменные к единичной дисперсии и нулевому среднему. Как уже отмечалось, такое преобразование данных может вызвать затруднения.

Скиннер (1978) для вычисления сходства данных, представляемых профилями, предложил совместно использовать корреляцию и евклидово расстояние. При этом можно определить, какой

из факторов (форма, сдвиг или дисперсия) делает вклад в оценку сходства. Метод Скиннера похож на метод, предложенный Гуэртином (1966), согласно которому сначала, взяв за основу форму, с помощью корреляции создаются однородные группы объектов, а затем каждая из этих групп с помощью меры расстояния разбивается на подгруппы со схожими сдвиговыми и дисперсионными характеристиками (Skinner, 1978). Однако в методе Скиннера строится сложная функция сходства, которая объединяет расстояние и корреляцию в вычислительной процедуре, осуществляющей минимизацию ошибки измерения при оценке сходства профилей.

Поскольку в прикладном анализе данных часто возникает необходимость в нормировке, полезно рассмотреть небольшой пример, показывающий влияние нормировки на коэффициенты корреляции и расстояния. В качестве данных были взяты четыре профиля MMPI-теста. Каждому из этих профилей соответствует больной с сильной психопатологией.

В качестве исходной меры сходства для профилей был взят смешанный момент корреляции Пирсона. Результаты приведены в следующей матрице:

	A	*B*	*C*	*D*
A	XXXX	0,776	0,702	0,742
B	(3)	XXXX	0,729	0,779
C	(6)	(5)	XXXX	0,936
D	(4)	(2)	(1)	XXXX

В верхней треугольной части матрицы приведены значения корреляции, которые показывают, что все четыре профиля имеют очень схожие формы, а профили *C* и *D* даже почти идентичны (r_{CD}= =0,936). В нижней треугольной части матрицы показаны ранги, полученные в результате упорядочения по величине значений сходства от наибольшего (1) к наименьшему (6). Необходимость в ранговом упорядочении будет объяснена ниже.

После вычисления евклидовых расстояний получается матрица:

	A	*B*	*C*	*D*
A	XXXX	266	732	736
B	(2)	XXXX	532	465
C	(5)	(4)	XXXX	144
D	(6)	(3)	(1)	XXXX

Заметьте, как различаются масштабирования коэффициентов расстояния и коэффициентов корреляции. Вспомните, что абсолютные значения коэффициентов расстояния не имеют смысла. Однако и здесь обнаруживается сходство пациентов *C* и *D* (d_{CD}=144, хотя не ясно, насколько хорошим является значение 144). Общая картина сходства кажется почти одинаковой и для корреляции, и для расстояния, но существуют и различия. В частности, при использовании корреляции в качестве меры сходства наименее похожими оказались пациенты *A* и *C* (r_{AC}=0,702). Однако евклидова метрика показывает, что наименее схожими являются пациенты *A* и *D* (d_{AD}=736).

Чтобы внести еще большую путаницу, предположим, что мы решили нормировать данные. (Нормировка была действительно выполнена на основе статистики для всего множества данных, состоящего из 90 объектов). Если для оценки сходства четырех профилей после нормировки используется смешанный момент корреляции, то матрица сходства принимает вид

	A	*B*	*C*	*D*
A	XXXX	0,602	0,284	0,433
B	(2)	XXXX	0,367	0,584
C	(6)	(5)	XXXX	0,804
D	(4)	(3)	(1)	XXXX

Обратите внимание, как различаются коэффициенты корреляции для нормированных и ненормированных данных. Для ненормированных данных $r_{AC}=0{,}702$, а для нормированных $r_{AC}=0{,}284$. В обоих случаях r_{AC} — наименьшая величина в матрице, но для нормированных данных величина коэффициента корреляции показывает, что между пациентами *A* и *C* нет никакого сходства, в то время как для ненормированных данных абсолютное значение корреляции ($r=0{,}706$) свидетельствует, что пациенты *A* и *C* довольно похожи.

Наконец, в нижеприведенной матрице несходства показаны евклидовы расстояния между пациентами в случае нормированных данных:

	A	*B*	*C*	*D*
A	XXXX	0,704	2,572	2,071
B	(1)	XXXX	2,141	1,304
C	(6)	(5)	XXXX	0,870
D	(4)	(3)	(2)	XXXX

Снова величины изменяются в зависимости от того, нормированы или нет данные. Однако поскольку значение коэффициента евклидова расстояния не имеет естественного смысла, постольку эти изменения не очень важны. Что действительно важно, так это относительное изменение. Наиболее драматическим моментом является то, что коэффициент евклидова расстояния для нормированных данных показывает, что пациенты *A* и *B* — пара с наибольшим сходством, между тем, как три другие матрицы сходства указывают на то, что наиболее похожие пациенты — это *C* и *D*.

В заключение важно отметить, что все четыре матрицы порождают разные ранжирования коэффициентов сходства. Это замечание важно, так как оно показывает, что выбор коэффициента сходства и преобразования данных может плохо повлиять на соотношения, содержащиеся в итоговой матрице сходства.

Коэффициенты ассоциативности

Коэффициенты ассоциативности применяются, когда необходимо установить сходство между объектами, описываемыми бинарными переменными. Легче всего рассмотреть эти коэффициенты, обра-

тившись к 2×2-таблице ассоциативности, в которой 1 указывает на наличие переменной, а 0 — на ее отсутствие.

	1	0
1	a	b
0	c	d

Было предложено большое число (>30) таких коэффициентов, а поэтому нереально пытаться дать исчерпывающее описание всей совокупности этих мер. В основном коэффициенты ассоциативности были впервые определены в биологии, хотя, вероятно, некоторые, наиболее простые из них были найдены и в ряде других отраслей науки. Лишь небольшое число мер подверглось широкой проверке, многие вышли из употребления из-за свойств сомнительного характера. Более подробно об этом см. (Sneath and Sokal, 1973; Clifford and Stephenson, 1975; Everitt, 1980). Однако существуют три меры, которые широко используются и заслуживают специального рассмотрения. Это — простой коэффициент совстречаемости, коэффициент Жаккара и коэффициент Гауэра.

Простой коэффициент совстречаемости имеет вид

$$S=\frac{(a+d)}{(a+b+c+d)},$$

где S — сходство между двумя объектами, которое меняется в пределах от 0 до 1. Как отмечают Снит и Сокэл (1973), этот коэффициент нелегко преобразовать в метрику. Тем не менее большие усилия были направлены на то, чтобы установить приблизительные доверительные пределы. Один из небольшого числа таких методов отмечает Гудолл (1967). Этот коэффициент учитывает также и одновременное отсутствие признака у обоих объектов (как указано в клетке d матрицы ассоциативности).

Коэффициент Жаккара, определенный следующим образом

$$J=a/(a+b+c),$$

не учитывает одновременного отсутствия признака при вычислении сходства (клетка d не рассматривается). Подобно простому коэффициенту совстречаемости он изменяется от 0 до 1. Коэффициент Жаккара широко применялся в биологии при необходимости рассмотрения так называемых негативных пар (с одновременным отсутствием признака). Как заметили биологи, используя простой коэффициент совстречаемости, некоторые объекты оказываются в значительной степени схожими главным образом за счет того, что им обоим не свойственны некоторые признаки, а не за счет наличия общих характеристик. В противоположность этому коэффициент Жаккара принимает в расчет лишь те признаки, которые характерны хотя бы для одного из объектов.

Во многих областях социологических наук не стоит вопрос об учете негативных пар, но такая проблема возникает в археологии. Если предмет не был найден в захоронении, то его отсутствие может быть обусловлено либо культурными традициями, либо естест-

венными процессами распада и изнашивания. Было бы неправильно давать оценку сходства двух захоронений исходя из отсутствия в них какого-то предмета, если невозможно узнать, какое из двух возможных объяснений действительно имеет место.

Рассмотрим шесть точек из множества данных о захоронениях, чтобы кратко проиллюстрировать различия между простым коэффициентом совстречаемости и коэффициентом Жаккара:

1	Р	М	Н	1	0	0	1	0	0	0	0
5	Р	Ж	Э	0	0	1	0	0	0	1	0
8	П	М	Н	0	1	0	1	1	0	0	0
14	П	Ж	Э	1	0	0	0	1	0	1	0
18	В	М	Э	1	1	0	1	1	0	1	1
24	В	Ж	Э	1	0	0	0	1	1	1	0

Возьмем объекты 1 (ребенок, мужской пол, неэлитарное общественное положение — РМН) и 8 (подросток, мужской пол, неэлитарное общественное положение — ПМН). Матрица ассоциативности общих признаков для двух объектов размерностью 2×2 имеет вид

	ПМН	
	1	0
РМН 1	1	1
0	2	4

Другими словами, эти объекты имеют только один общий предмет. Однако четыре предмета отсутствуют в обоих захоронениях. Таким образом,

$$S=0{,}625 \quad (=5/8).$$

Тем не менее

$$J=0{,}250 \quad (=1/4).$$

Иначе говоря, в то время как простой коэффициент совстречаемости показывает, что объекты РМН и ПМН достаточно схожи, из величины коэффициента Жаккара следует, что такого сходства нет. Полная матрица сходства размерностью 6×6 в случае простого коэффициента совстречаемости имеет вид

	РМН	РЖЭ	ПМН	ПЖЭ	ВМЭ	ВЖЭ
РМН	—	0,500	0,625	0,625	0,500	0,500
РЖЭ		—	0,375	0,625	0,250	0,500
ПМН			—	0,500	0,625	0,375
ПЖЭ				—	0,625	0,875
ВМЭ					—	0,500
ВЖЭ						—

В случае коэффициента Жаккара полная матрица сходства принимает следующий вид:

	РМН	РЖЭ	ПМН	ПЖЭ	ВМЭ	ВЖЭ
РМН	—	0,000	0,250	0,250	0,333	0,200
РЖЭ		—	0,000	0,250	0,143	0,200
ПМН			—	0,200	0,500	0,166
ПЖЭ				—	0,500	0,750
ВМЭ					—	0,429
ВЖЭ						—

Как видим, эти матрицы довольно похожи. Например, они показывают, что объекты ПЖЭ, ВМЭ и ВЖЭ (недетские элитарные захоронения) имеют наибольшее сходство. Однако существуют и различия. Два детских захоронения (объекты РМН и РЖЭ) согласно коэффициенту Жаккара совсем не имеют сходства, но, судя по простому коэффициенту совстречаемости, они сравительно похожи.

Другой характерной чертой этих матриц является число «совпадений». В случае простого коэффициента совстречаемости имеется пять пар объектов, для которых $S=0{,}625$, и пять пар, для которых $S=0{,}500$. На самом деле среди пятнадцати клеток матрицы сходства размерностью 6×6 только в пяти есть неповторяющиеся значения S. Как мы позже покажем, некоторые кластерные методы не годятся для матриц сходства, у которых так много «совпадений».

Коэффициент Гауэра — единственный в своем роде, так как при оценке сходства допускает одновременное использование переменных, измеренных по различным шкалам. Коэффициент был предложен Гауэром (1971) и имеет вид

$$s_{ij}=\sum_{k=1}^{p} S_{ijk}/\sum_{k=1}^{p} W_{ijk},$$

где W_{ijk} — весовая переменная, принимающая значение 1, если сравнение объектов по признаку k следует учитывать, и 0 — в противном случае; S_{ijk} — «вклад» в сходство объектов, зависящий от того, учитывается ли признак k при сравнении объектов i и j. В случае бинарных признаков $W_{ijk}=0$, если признак k отсутствует у одного или обоих сопоставляемых объектов (Everitt, 1980). Для так называемых негативных переменных $W_{ijk}=0$. Понятно, что если все данные — двоичные, то коэффициент Гауэра сводится к коэффициенту Жаккара.

Чтобы показать, как работает этот коэффициент, расширим множество данных о захоронениях, добавив два новых признака: рост (измеренный в сантиметрах; это количественная переменная) и величину энергетических затрат, связанных с погребением (измеренных по порядковой шкале с рангами 1, 2 и 3 или соответственно низкие, средние и высокие). Матрица сходства для четырех объектов примет вид

1	Р	М	Н	1	0	0	1	0	0	0	0	69	1
7	П	М	Н	1	1	0	1	0	0	0	0	167	2
18	В	Ж	Э	1	1	0	1	1	0	1	1	179	3
25	В	М	Э	1	0	0	0	1	1	1	1	158	3

Для двоичных данных S_{ijk} вычисляется в соответствии со следующей системой подсчета:

объект i	1	1	0	0
объект j	1	0	1	0
вклад S_{ijk}	1	0	0	0
вес W_{ijk}	1	1	1	0

Для порядковых данных S_{ijk} равно 1, если сравниваемые значения равны, и 0 — в противном случае. Наконец, для количественных данных имеет место уравнение

$$S_{ijk} = 1 - |x_{ik} - x_{jk}|/R_k,$$

где x_{ik} — значение k-й переменной для объекта i, а R_k — размах значений этой переменной (разность между максимальным и минимальным значениями). В результате итоговую матрицу сходства для четырех объектов можно представить как

	РМН	ПМН	ВМЭ	ВЖЭ
РМН	—	0,527	0,285	0,170
ПМН		—	0,554	0,239
ВМЭ			—	0,726
ВЖЭ				—

Кроме возможности работать с разнотипными данными, у коэффициента есть еще несколько привлекательных особенностей. Например то, что его метрические свойства и гибкость дают возможность после простого изменения системы бинарных весов при оценке сходства учитывать и негативные пары. К сожалению, коэффициент Гауэра можно редко найти в пакетах прикладных программ по кластерному анализу, так как он практически не применяется в области социальных наук.

Вероятностные коэффициенты сходства

Радикальное отличие коэффициентов этого типа от описанных выше заключается в том, что, по сути дела, сходство между двумя объектами не вычисляется. Вместо этого мера такого типа прилагается непосредственно к исходным данным до их обработки. При образовании кластеров вычисляется информационный выигрыш (по определению Шеннона) от объединения двух объектов, а затем те объединения, которые дают минимальный выигрыш, рассматриваются как один объект. Другой особенностью вероятностных мер является то, что они пригодны лишь для бинарных данных. До сих пор не было разработано ни одной схемы использования меры этого вида для качественных и количественных переменных. Вероятностные коэффициенты сходства еще не нашли своего применения в социальных науках, но уже в течение десятилетия ими широко пользуются специалисты по численной таксономии и экологии. Более подробно об этом см. (Sneath and Sokal, 1973; Clifford and Stephenson, 1975).

БИБЛИОГРАФИЧЕСКИЕ ЗАМЕЧАНИЯ

Обсуждение коэффициентов сходства, используемых в кластерном анализе, проводится в работах Снита и Сокэла (1973), Клиффорда и Стефенсона (1975). Там же можно найти формулы для вычисления некоторых обсуждаемых мер.

Более широко теоретические вопросы, связаниые со сходством, рассматриваются в работах Хартигана (1967) и Тверски (1977). Обсуждение Скиннером (1978) формы, поднятия и рассеяния очень важно для многих применений мер сходства в социальных исследованиях. Последние три работы важны потому, что понятие сходства играет главную роль в формировании кластеров. Обычно кластеры определяются как группы *сходных* объектов. Хотя во многих приложениях кластерного анализа особое значение придается процедуре формирования кластеров, все же выбор меры сходства является решающим моментом в исследованиях, использующих кластерный анализ.

III. ОБЗОР МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА

О ПРИРОДЕ КЛАСТЕРОВ

Главная цель кластерного анализа — нахождение групп схожих объектов в выборке данных. Эти группы удобно называть кластерами. Не существует общепринятого или просто полезного определения термина «кластер», и многие исследователи считают что уже слишком поздно либо вовсе незачем пытаться найти такое определение (Bonner, 1964). Несмотря на отсутствие определения, ясно, что кластеры обладают некоторыми свойствами, наиболее важными из которых являются плотность, дисперсия, размеры, форма и отделимость. Хотя Снит и Сокэл рассматривают эти свойства для случая метрического пространства, очевидно (как они признают), что эти свойства можно логически распространить и на неметрические пространства.

Плотность — это свойство, которое позволяет определить кластер, как скопление точек в пространстве данных, относительно плотное по сравнению с другими областями пространства, содержащими либо мало точек, либо не содержащих их вовсе. Хотя четко определенной меры плотности нет, это понятие очевидно. Дисперсия характеризует степень рассеяния точек в пространстве относительно центра кластера. Несмотря на то, что между этим свойством и тем, которое используется в теории статистических выводов, есть аналогия, кластеры не всегда представляют многомерные нормальные популяции. Поэтому лучше всего рассматривать дисперсию как характеристику того, насколько близко друг к другу расположены в пространстве точки кластера. Следовательно, кластер можно назвать «плотным», если все точки находятся вблизи его центра тяжести, и «неплотным», если они разбросаны вокруг центра. Свойство кластеров — размеры — тесно связано с дисперсией; если кластер можно идентифицировать, то можно и измерить его «радиус». Это свойство полезно лишь в том случае, если рассматриваемые кластеры являются гиперсферами (т. е. имеют круглую форму) в многомерном пространстве, описываемом признаками.

Форма — это расположение точек в пространстве. Несмотря на то, что обычно кластеры изображают в форме гиперсфер или эллипсоидов, возможны кластеры и другой формы, например удлиненные кластеры. В последнем случае понятие радиуса или диаметра перестает быть полезным. Вместо этого можно вычислить «связность» точек в кластере — относительную меру расстояния между ними. Если же кластеры имеют другие, более причудливые формы (см. Everitt, 1980), то понятие связности становится менее полезным, а ценность относительных оценок диаметра и плотности, следовательно, уменьшается. Отделимость характеризует степень перекрытия кластеров и насколько далеко друг от друга они расположены в пространстве. Так, кластеры могут быть относительно близки друг к другу и не иметь четких границ, или же они могут быть разделены широкими участками пустого пространства.

С помощью этих терминов можно описать кластеры любого вида. Согласно Эверитту (1980) кластеры — это «непрерывные области (некоторого) пространства с относительно высокой плотностью точек, отделенные от других таких же областей областями с относительно низкой плотностью точек». Важность этого определения заключается в том, что оно не сводит понятие кластера к какой-то частной форме до начала анализа данных.

Разработанные кластерные методы образуют семь основных семейств:

1) иерархические агломеративные методы;
2) иерархические дивизимные методы;
3) итеративные методы группировки;
4) методы поиска модальных значений плотности;
5) факторные методы;
6) методы сгущений;
7) методы, использующие теорию графов.

Эти семейства соответствуют различным подходам к созданию групп, и применение различных методов к одним и тем же данным может привести к сильно различающимся результатам. В конкретных отраслях науки могут оказаться особенно полезными определенные семейства методов. Так, иерархические агломеративные методы чаще всего используются в биологии, тогда как факторные аналитические методы большим успехом пользуются в психологии. Когда сталкиваешься с трудной проблемой: «Какой из кластерных методов использовать?», важно помнить, что этот метод должен находиться в согласии с ожидаемым характером классификации, применяемыми признаками и мерой сходства (если она требуется для оценки подобия объектов).

Наиболее известными семействами кластерных методов, используемыми в социальных науках, являются иерархические агломеративные, иерархические дивизимные и факторные. Поэтому каждый из этих трех методов будет рассмотрен более детально на примере двух наборов данных, описанных в разд. I. Другие, менее известные семейства будут обсуждены более кратко.

ИЕРАРХИЧЕСКИЕ АГЛОМЕРАТИВНЫЕ МЕТОДЫ

Из семи семейств кластерных методов наиболее часто в приложениях употребляются иерархические агломеративные методы. Проанализировав все опубликованные в 1973 г. работы, в которых использовался кластерный анализ, Блэшфилд и Олдендерфер (1978b) нашли, что в $^2/_3$ этих статей применяется какой-либо из иерархических агломеративных методов.

Самым легким для понимания из иерархических агломеративных методов является метод одиночной связи. Рассмотрим матрицу сходства размерностью 6×6, которая была получена в разд. II с помощью коэффициента Жаккара для данных о захоронениях. *Метод одиночной связи* начинает процесс кластеризации с поиска двух наиболее похожих объектов в матрице сходства. В этом примере наиболее схожими являются объекты ПЖЭ (подросток, женский пол, элитарный) и ВЖЭ (взрослый, женский пол, элитарный) с уровнем сходства $J=0{,}750$. На следующем шаге к этой группе присоединяется объект ВМЭ, так как его коэффициент сходства с ПЖО равен 0,500. Дело в том, что по правилу объединения для метода одиночной связи новый кандидат на включение в состав кластера присоединяется к существующей группе в том случае, если он имеет наивысший уровень сходства с каким-либо из членов этой группы. Другими словами, для объединения двух объектов требуется только одна связь между ними. Третий шаг присоединяет объект ПМН к кластеру, содержащему объекты ВЖЭ, ВМЭ и ПЖЭ, потому что он тоже имеет коэффициент сходства с ВМЭ, равный 0,500. Четвертый шаг процесса кластеризации присоединяет объект РМН к группе, образованной объектами ПЖЭ, ВМЭ, ВЖЭ и ПМН с уровнем сходства $J=0{,}333$.

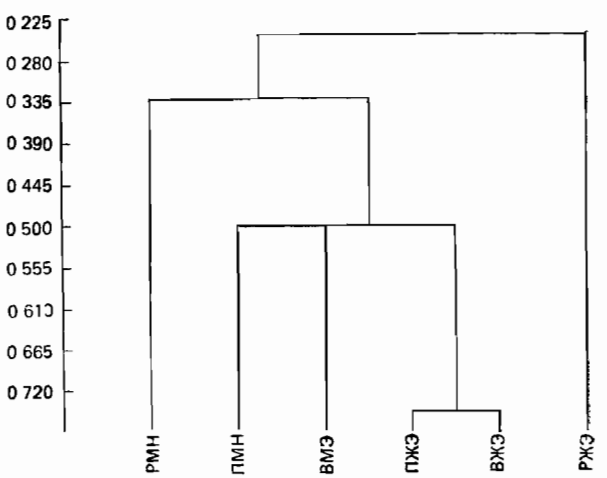

Рис 3 Дендрограмма для данных о шести захоронениях

Из этого примера можно вывести четыре важных наблюдения относительно иерархических агломеративных методов. Первое — все эти методы просматривают матрицу сходства размерностью $N \times N$ (где N — число объектов) и последовательно объединяют наиболее схожие объекты. Именно поэтому они называются агломеративными (объединяющими). Второй важный момент, на который стоит обратить внимание, состоит в том, что последовательность объединений кластеров можно представить визуально в виде древовидной диаграммы, часто называемой дендрограммой. Древовидная диаграмма, отражающая применение метода одиночной связи к данным о шести захоронениях, показана на рис. 3. Каждый шаг, на котором объединялась пара объектов, представляется ветвью этого дерева. Заметьте, что дерево изображает иерархическую организацию связей между шестью точками данных. На самом нижнем уровне все шесть точек независимы; на следующем уровне они объединяются в одну группу и три независимых объекта; наконец, на самом верхнем уровне они все объединяются в одну большую группу.

Третьим важным моментом является то, что для полной кластеризации этими методами на основе матрицы сходства размерностью $N \times N$ требуется ровно $N-1$ шагов. На первом шаге события (объекты) рассматриваются как самостоятельные кластеры. На последнем шаге все события объединяются в одну большую группу.

Наконец, для понимания иерархических агломеративных методов не нужны обширные знания. Так, метод одиночной связи не требует понимания матричной алгебры или обширной подготовки по многомерной статистике. Вместо этого дается правило, указывающее, каким образом, исходя из матрицы сходства, объекты могут объединяться в кластеры. Хотя другие иерархические агломеративные методы несколько сложнее, все они довольно просты и различаются правилами объединения объектов (которые в литературе часто называются видами связи объектов). По определению в результате работы этих кластерных методов получаются неперекрывающиеся кластеры, которые, однако, являются вложенными в том смысле, что каждый кластер может рассматриваться как элемент другого, более широкого кластера на более высоком уровне сходства. Самым распространенным способом представления результатов этих кластерных методов является дендрограмма (древновидная диаграмма), которая графически изображает иерархическую структуру, порожденную матрицей сходства и правилом объединения объектов в кластеры.

Несмотря на простоту методов, они обладают некоторыми недостатками. Если не используются специально разработанные алгоритмы, то применение иерархических алгомеративных методов может потребовать вычисления и хранения большой матрицы сходства. Необходимость в хранении такой матрицы фактически ограничивает сверху число объектов, участвующих в процессе кластеризации. Например, для набора данных из 500 объектов потре-

буются хранение и неоднократный просмотр матрицы, содержащей около 125 000 элементов. Другим недостатком кластерных методов является то, что в них объекты распределяются по кластерам лишь за один проход, а поэтому плохое начальное разбиение множества данных не может быть изменено на последующих шагах процесса кластеризации (Gower, 1967). Третий недостаток всех этих методов (за исключением метода одиночной связи) состоит в том, что они могут порождать разные решения в результате простого переупорядочения объектов в матрице сходства и, кроме того, их результаты изменяются, если некоторые объекты исключаются из рассмотрения. Устойчивость — это важное свойство любой классификации, так как устойчивые группы с бо́льшим правдоподобием представляют собой «естественные» группировки по сравнению с теми группами, которые исчезают, если некоторые объекты переупорядочены или исключены из анализа. Вопрос об устойчивости становится особенно существенным, когда мы имеем дело с малыми выборками объектов (Jardine and Sibson, 1071).

Иерархические агломеративные методы различаются главным образом по правилам построения кластеров. Некоторые авторы для обозначения способа группировки используют термин «стратегия классификации». Существует много различных правил группировки, каждое из которых порождает специфический иерархический метод. Известно по крайней мере двенадцать различных методов группировки, четыре из них наиболее распространенные: одиночной связи, полной связи, средней связи и метод Уорда.

Ланс и Уильямс (1967) получили формулу, которая позволяет описать правила группировки в общем виде для любого иерархического агломеративного метода. Формула имеет вид

$$d(h,k) = A(i) \cdot d(h,i) + A(j) \cdot d(h,j) + B \cdot d(i,j) + \\ + C \cdot ABS(d(h,i) - d(h,j)),$$

где $d(h, k)$ — различие (расстояние) между кластерами h и k, причем кластер k является результатом объединения кластеров (или объектов) i и j в ходе агломеративного шага. С помощью этой формулы можно вычислить расстояние между некоторым объектом (h) и новым кластером (k), полученным объединением объектов i и j в единый кластер. Прописными буквами обозначены параметры, которые определяют конкретный вид группировки; в методе одиночной связи, например, эти параметры принимают следующие значения: $A(i) = A(j) = 1/2$, $B = 0$ и $C = 1/2$. Полученная формула оказала большую помощь при разработке вычислительных алгоритмов для этих методов.

Чтобы проиллюстрировать работу иерархических методов и показать действие разных правил группировки, данные MMPI-теста были обработаны с помощью четырех наиболее известных методов.

Метод одиночной связи. В этом методе, описанном Снитом (1957), кластер образуется по следующему правилу: объект будет

присоединеи к уже существующему кластеру, если по крайней мере один из элементов кластера находится на том же уровне сходства, что и объект, претендующий на включение. Таким образом, присоединение определяется лишь наличием единственной связи между объектом и кластером. Главное преимущество этого метода заключается в его математических свойствах: результаты, полученные по этому методу, инвариантны к монотонным преобразованиям матрицы сходства; применению метода не мешает наличие «совпадений» в данных (Jardine and Sibson, 1971). Первое из этих свойств (инвариантность при монотонных преобразованиях) особенно важно по той причине, что все другие иерархические агломеративные методы таким свойством не обладают. Это означает, что метод одиночной связи является одним из немногих методов, результаты применения которых не изменяются при любых преобразованиях данных, оставляющих без изменения относительное упорядочение элементов матрицы сходства.

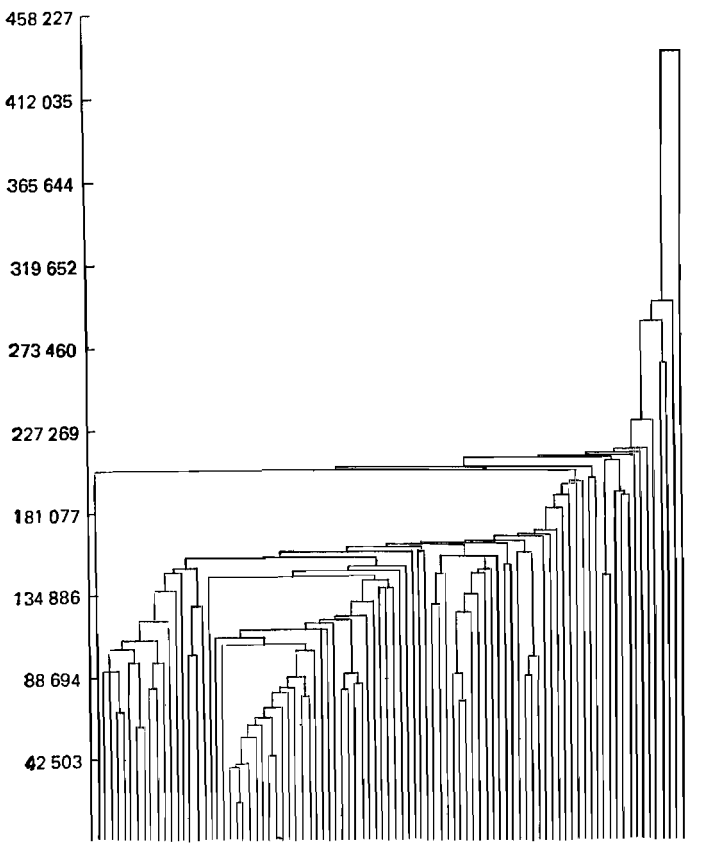

Рис. 4 Дендрограмма метода одиночной связи для данных MMPI-теста

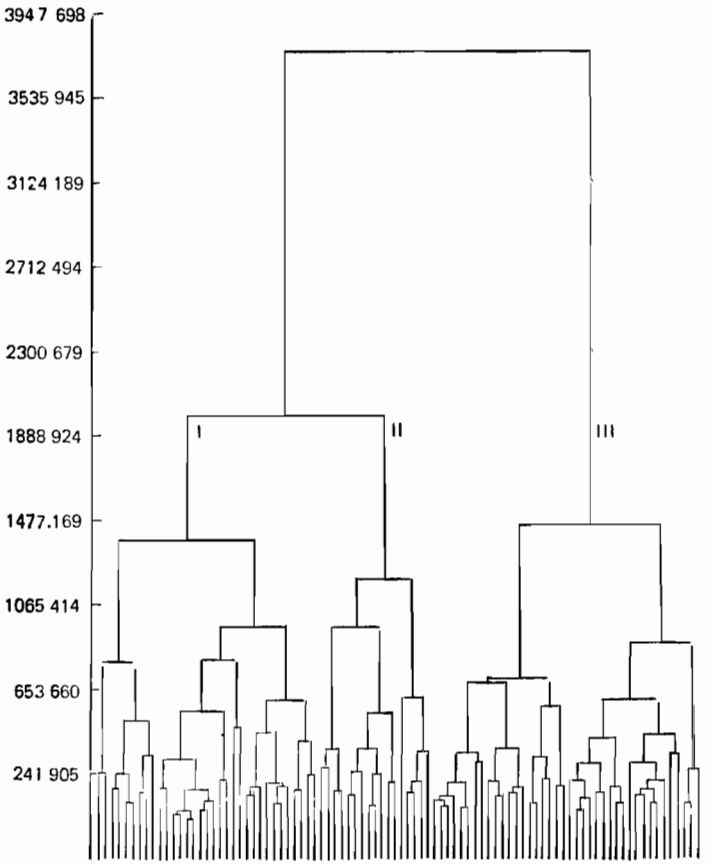

Рис 5. Дендрограмма метода полных связей для данных MMPI-теста

Главный недостаток метода одиночной связи, однако, состоит в том, что, как было показано на практике, метод приводит к появлению «цепочек» («цепной эффект»), т. е. к образованию больших продолговатых кластеров. Эффект образования цепочек можно показать на примере древовидной диаграммы для данных MMPI-теста (рис. 4). Обратите внимание, что по мере приближения к окончанию процесса кластеризации образуется один большой кластер, а все остающиеся объекты добавляются к нему один за другим. Найденное с помощью метода одиночной связи решение, состоящее из двух кластеров, является тривиальным следствием наличия одного кластера, включающего 89 объектов, и одного кластера, включающего один объект.

На рис. 4 можно отметить еще несколько интересных моментов. Во-первых, анализ рисунка не дает возможности определить,

как много кластеров содержится в данных. В противоположность этому древовидная диаграмма, полученная методом полной связи (см. рис. 5), четко указывает на наличие двух кластеров. Во-вторых, диагностируемые классы больных, тесно связанные с профилями данных MMPI-теста, не образуют четко выделенных кластеров на рисунке. В левой части дерева имеется скопление профилей больных с невротическими заболеваниями (Н), а в середине дерева скопление профилей больных с расстройствами личности (РЛ). Оставшаяся часть дерева состоит из профилей больных психозами (П), неврозами (Н) и нескольких профилей РЛ. Короче говоря, решение, порожденное методом одиночной связи, не является точным воспроизведением известной структуры данных.

Метод полных связей. В этом методе в противоположность методу одиночной связи правило объединения указывает, что сходство между кандидатами на включение в существующий кластер и любым из элементов этого кластера не должно быть меньше некоторого порогового уровня (Sokal and Michener, 1958). Настоящее правило более жесткое, чем правило для метода одиночной связи, и поэтому здесь имеется тенденция к обнаружению относительно компактных гиперсферических кластеров, образованных объектами с большим сходством. Хотя дерево, порожденное методом полных связей (рис. 5), дает наглядное представление о найденных кластерах данных, все же сравнение полученной классификации с известной не говорит о хорошем их соответствии. Приведенная ниже информация показывает распределение объектов по кластерам и диагностическим классам. Точное решение давало бы взаимооднозначное соответствие между кластерами и диагностическими классами. Однако в решении, полученном по методу полных связей, оно явно отсутствует. Как распределены объекты по классам и кластерам, показано в таблице:

		кластеры		
		I	II	III
диагнозы	П	10	0	20
	Н	15	13	2
	РЛ	8	4	18

Метод средней связи. Предложенный Сокэлом и Миченером (1958) метод средней связи разрабатывался в качестве «средства борьбы» с крайностями как метода одиночной связи, так и метода полной связи. Хотя есть несколько вариантов метода, по существу, в каждом из них вычисляется среднее сходство рассматриваемого объекта со всеми объектами в уже существующем кластере, а затем, если найденное среднее значение сходства достигает или превосходит некоторый заданный пороговый уровень сходства, объект присоединяется к этому кластеру. Чаще всего используется вариант метода средней связи, в котором вычисляется среднее арифметическое сходство между объектами кластера и кандидатом на включение. В других вариантах метода средней связи вычисляется сходство между центрами тяжести двух кластеров, подлежащих объединению. Метод средней связи широко использовался в био-

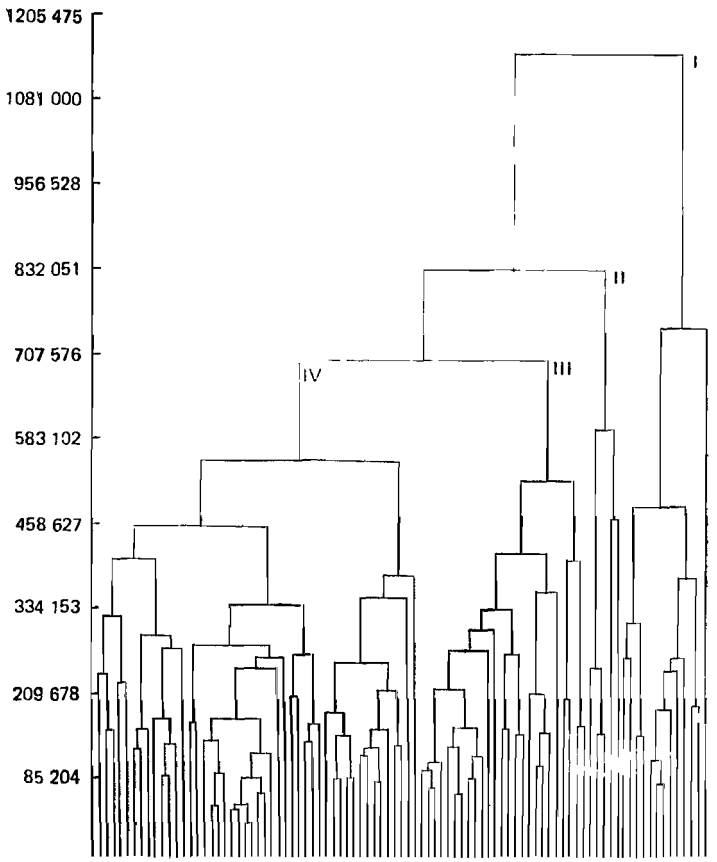

Рис 6 Дендрограмма метода средней связи для данных MMPI-теста

логии, и только недавно начал по-настоящему применяться в социальных науках

Анализ рис. 6 дает интересное соотношение между деревом, порожденным методом средней связи, и известными диагностическими классами. Первый кластер (I) содержит почти половину профилей больных психозами. Второй кластер (II) довольно мал и поровну разделен на профили больных неврозами и психозами. Третий кластер (III) содержит фактически все случаи неврозов, тогда как в четвертый (IV) самый большой кластер входят профили больных расстройствами личности и значительное число профилей больных психозами. В какой степени совпадают кластеры и диагностические классы, показывает следующая таблица:

		кластеры			
		I	II	III	IV
диагнозы	Н	0	2	26	2
	П	13	3	0	14
	РЛ	0	0	0	30

Метод Уорда. Данный метод построен таким образом, чтобы оптимизировать минимальную дисперсию внутри кластеров. Эта целевая функция известна как внутригрупповая сумма квадратов или сумма квадратов отклонений (СКО). Формула суммы квадратов отклонений имеет вид

$$\text{СКО} = x_j^2 - 1/n(\Sigma x_j)^2,$$

где x_j — значение признака j-го объекта. На первом шаге, когда каждый кластер состоит из одного объекта, СКО равна 0. По методу Уорда объединяются те группы или объекты, для которых СКО получает минимальное приращение. Метод имеет тенденцию

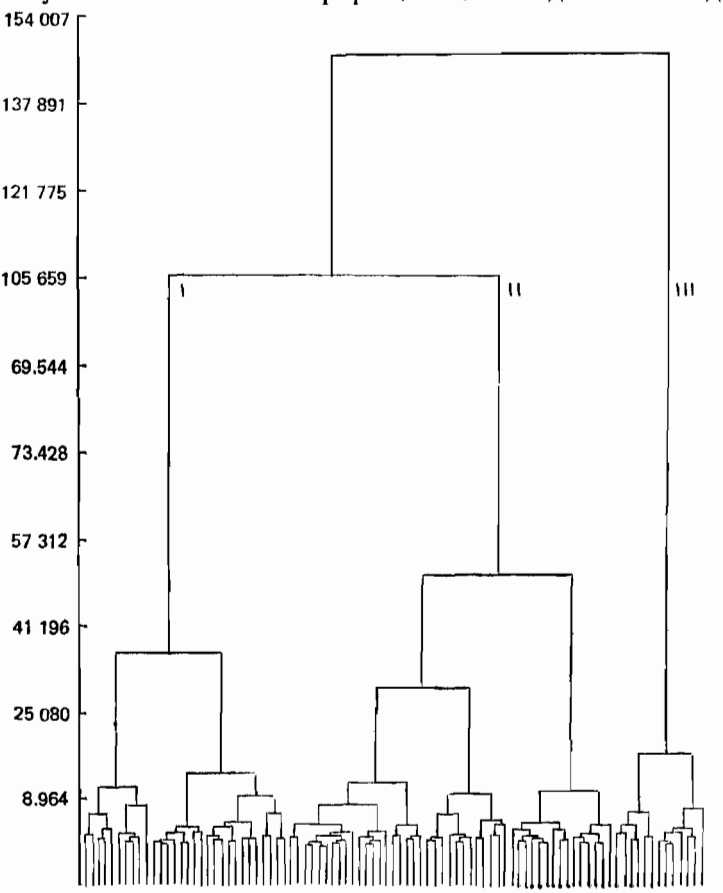

Рис. 7. Дендрограмма метода Уорда для данных MMPI-теста

к нахождению (или созданию) кластеров приблизительно равных размеров и имеющих гиперсферическую форму. Метод Уорда фактически не нашел применения в биологии, но широко используется во многих социальных науках (Blashfield, 1980).

Дерево, порожденное методом Уорда (рис. 7), ясно показывает, что найденное решение состоит из трех кластеров. Как и в случае метода средней связи, здесь имеется взаимосвязь между кластерами и диагностическими классами. Однако и метод Уорда не порождает точного решения. Ниже приводится таблица распределения объектов по кластерам и классам:

		кластер I	II	III
диагнозы	Н	29	1	0
	П	1	16	13
	РЛ	0	30	0

Обычная трудность, связанная с использованием метода Уорда, заключается в том, что найденные с его помощью кластеры можно упорядочить по величине профильного сдвига. Так, в приведенном решении профили кластера III являются наиболее приподнятыми, тогда как кластер II содержит наименее приподнятые кластеры. Практическое применение метода Уорда в социологических исследованиях показало, что он порождает решения, которые находятся под сильным воздействием величины профильного сдвига.

Имеется несколько способов сравнения различных иерархических агломеративных методов. С помощью одного из них можно проанализировать, как эти методы преобразуют соотношения между точками в многомерном пространстве. Сжимающие пространство методы изменяют эти соотношения, «уменьшая» пространство между любыми группами в данных. Когда очередная точка подвергается обработке таким методом, она скорее всего будет присоединена к уже существующей группе, а не послужит началом нового кластера. Расширяющие пространство методы действуют противоположным образом. Здесь кластеры как бы «расступаются»; таким образом в пространстве образуются мелкие, более «отчетливые» кластеры. Этот способ группировки также склонен к созданию кластеров гиперсферической формы и приблизительно равных размеров. Методы Уорда и полных связей являются методами, расширяющими пространство. И наконец, сохраняющие пространство методы, такие, как метод средней связи, оставляют без изменения свойства исходного пространства.

Уильямс и др. (1971) рассматривают свойства сужающих пространство методов как недостатки, особенно в прикладном анализе данных, тогда как, по мнению других авторов, — среди них наиболее известны Джардайн и Сибсон (1968) — эти методы предпочтительнее ввиду их хороших математических свойств, невзирая на результаты их практического использования. Эверитт (1980) уравновешивает эти две крайности замечанием, что успех применения рассматриваемых методов в анализе данных в большой степени зависит от априорных представлений об ожидаемом виде класте-

ров и действительной структуре данных. Проблема, которая будет подробно обсуждаться в одном из последующих разделов, состоит в том, чтобы определить, когда один из этих методов привносит в данные не свойственную им структуру.

ИТЕРАТИВНЫЕ МЕТОДЫ ГРУППИРОВКИ

В отличие от иерархических агломеративных методов итеративные методы группировки кластерного анализа не имели широкого применения, и специфика использования этих методов не до конца понимается их потенциальными пользователями. Большинство итеративных методов группировки работает следующим образом:

1. Начать с исходного разбиения данных на некоторое заданное число кластеров; вычислить центры тяжести этих кластеров.
2. Поместить каждую точку данных в кластер с ближайшим центром тяжести.
3. Вычислить новые центры тяжести кластеров; кластеры не заменяются на новые до тех пор, пока не будут просмотрены полностью все данные.
4. Шаги 2 и 3 повторяются до тех пор, пока не перестанут меняться кластеры.

Данные MMPI-теста были подвергнуты кластеризации с помощью процедуры k-средних процедурой CLUSTAN (Wishart, 1982) для того, чтобы продемонстрировать основные черты итеративных методов. Первый шаг состоит в формировании исходного разбиения данных. Процедура CLUSTAN произвольно распределяет 90 объектов по трем кластерам ($k=3$). Значение k задается пользователем. Затем вычисляются центры тяжести кластеров.

После этого определяются евклидовы расстояния между всеми объектами и центрами тяжести трех кластеров и объекты приписываются к ближайшему центру тяжести. Для данных MMPI-теста это означает, что 51 объект перемещается из кластера, в котором они находились первоначально, в кластер с ближайшим центром тяжести. После всех перемещений вычисляются центры тяжести новых кластеров. Эти центры тяжести уже совсем другие и приближаются к реальным центрам трех групп в данных MMPI-теста. На втором шаге все повторяется, но на этот раз производится восемь перемещений. Находятся новые центры тяжести и переходим к следующему шагу. На третьем шаге никаких перемещений не происходит. Все объекты приписываются к ближайшим центрам тяжести кластеров.

В отличие от иерархических агломеративных методов, которые требуют вычисления и хранения матрицы сходств между объектами размерностью $N \times N$, итеративные методы работают непосредственно с первичными данными. Поэтому с их помощью возможно обрабатывать довольно большие множества данных. Более того, итеративные методы делают несколько просмотров данных и могут компенсировать последствия плохого исходного разбиения дан-

ных, тем самым устраняя самый главный недостаток иерархических агломеративных методов. Эти методы порождают кластеры одного ранга, которые не являются вложенными, и поэтому не могут быть частью иерархии. Большинство итеративных методов не допускает перекрытия кластеров.

Несмотря на свои привлекательные черты, итеративные методы группировки имеют существенное ограничение. Наиболее простой способ отыскать оптимальное разбиение множества данных с помощью итеративного метода заключается в образовании всевозможных разбиений этого множества данных. Но такое, казалось бы, простое с точки зрения математических вычислений решение возможно лишь для очень небольших и тривиальных задач. Для 15 объектов и 3 кластеров этот подход требует рассмотрения 217 945 728 000 конкретных разбиений, что, очевидно, за пределами возможностей современных вычислительных машин.

Поскольку все допустимые разбиения даже для маленьких наборов данных не могут быть рассмотрены, исследователи разработали широкий круг эвристических процедур которые можно использовать для выбора небольшого подмножества из всех разбиений данных, чтобы найти или хотя бы приблизиться к оптимальному разбиению набора данных. Эта ситуация подобна той, с которой сталкиваются при эвристическом подходе к разработке правил объединения для иерархических агломеративных методов. Процедуры выбора разумны и правдоподобны, но только малая часть из них имеет достаточное статистическое обоснование.

Большинство эвристических, вычислительных и статистических свойств итеративных методов группировки могут быть описаны с помощью трех основных факторов: 1) выбора исходного разбиения; 2) типа итерации и 3) статистического критерия. Эти факторы могут сочетаться огромным количеством способов образуя алгоритмы отбора данных при определении оптимального разбиения. Не удивительно, что их различные комбинации ведут к разработке методов, порождающих разные результаты при работе с одними и теми же данными.

Исходное разбиение. Есть два основных способа начать итеративный процесс: определить начальные точки или подобрать подходящее начальное разбиение. Начальные точки определяют центры тяжести кластеров (Anderberg, 1973). Когда используются начальные точки, то при первом просмотре точки данных приписываются к ближайшим центрам тяжести кластеров. Задание начального разбиения требует детального распределения данных по кластерам. В этой процедуре центр тяжести каждого кластера определяется как многомерное среднее объектов кластера. Начальные разбиения могут выбираться случайным образом (как это было в примере с данными MMPI-теста) или же задаваться каким-либо образом самим пользователем (например, пользователь может взять в качестве исходного разбиения решение, полученное иерархической кластеризацией).

Тип итерации. Данный момент итерационного процесса связан со способом распределения объектов по кластерам. И опять имеются два основных вида итераций: по принципу k-средних и по принципу «восхождения на холм».

Итерации по принципу k-средних (они называются также «итерациями по принципу ближайшего центра» и «перемещающими итерациями») заключаются просто в перемещении объектов в кластер с ближайшим центром тяжести. Итерации по принципу k-средних могут быть либо комбинаторными, либо некомбинаторными. В первом случае перевычисление центра тяжести кластера производится после каждого изменения его состава, а во втором случае — лишь после того, как будет завершен просмотр всех данных. Кроме того, итерации по принципу k-средних подразделяются на исключающие и включающие. В итерациях исключающего типа после вычисления центра тяжести кластера рассматриваемый объект удаляется из кластера, а в итерациях включающего типа — помещается в кластер.

В итерациях, работающих по принципу «восхождения на холм», вместо присоединения объектов к кластеру в зависимости от расстояния между объектом и центром тяжести кластера, перемещение объектов производится исходя из того, будет или нет предполагаемое перемещение оптимизировать значение некоторого статистического критерия.

Статистический критерий. Методы, основанные на принципе «восхождения на холм», используют один или несколько следующих критериев (функций качества кластеризации): $\mathrm{tr}W$, $\mathrm{tr}W^{-1}B$, $\det W$ и наибольшее собственное значение матрицы $W^{-1}B$, где W — объединенная внутригрупповая ковариационная матрица, в B — объединенная межгрупповая ковариационная матрица. Каждая из этих статистик часто рассматривается в многомерном дисперсионном анализе (MANOVA); их применение выводится из статистической теории, заложенной в MANOVA. Фактически все четыре критерия связаны с обнаружением однородности кластеров в многомерном пространстве. Хотя в явном виде итерации по принципу k-средних не применяют статистический критерий при перемещении объектов, неявно они оптимизируют критерий $\mathrm{tr}W$. Таким образом, процедура k-средних минимизирует дисперсию внутри каждого кластера. Важно отметить, однако, что итерации по принципам k-средних и «восхождения на холм», используя критерий $\mathrm{tr}W$, приведут к различным результатам при одних и тех же исходных данных.

Подобно иерархическим агломеративным методам каждый из статистических критериев находит кластеры определенного вида. Критерий $\mathrm{tr}\,W$ благоприятствует образованию гиперсферических, очень однородных кластеров. Более важно, что этот критерий чувствителен к простым преобразованиям первичных данных, например, таких, как нормировка. Поскольку критерий $\det W$ не зависит от преобразований или от выбора масштаба, порождаемые им кластеры не обязаны иметь гиперсферическую форму. Его исполь-

зование, однако, предполагает, что у кластеров будет одна и та же форма, и это может вызвать некоторые затруднения в прикладном анализе данных. Скотт и Саймонс (1971) показали, что критерий det W имет тенденцию к созданию кластеров приблизительно одинаковых размеров, даже если таких кластеров нет в данных. К сожалению, характеристики других критериев известны плохо, так как они не подвергались широкому изучению и сравнению.

Одна из главных проблем, присущая всем итеративным методам, — проблема субоптимального решения. Поскольку эти методы могут выбрать лишь очень малую часть всех возможных разбиений, есть определенная вероятность, что будет выбрано субоптимальное разбиение. Такую проблему называют также проблемой локального (в противоположность глобальному) оптимума. Действительно, объективного способа определить, является ли полученное с помощью итеративного метода группировки решение глобально оптимальным, нет. Однако один подход к решению этой проблемы состоит в том, чтобы применять метод кластеризации совместно с подходящей процедурой проверки результата на достоверность (см. разд. IV).

Исследование методом Монте-Карло работы итеративных методов показало, что главная причина появления субоптимальных решений заключается в плохом исходном разбиении набора данных (Blashfield and Aldenderfer, 1978a; Milligan, 1980). Итерации по принципу k-средних чрезвычайно чувствительны к плохим начальным разбиениям и дело еще более усложняется, когда начальное приближение выбирается случайным образом (очень распространенная возможность, предоставляемая многими пакетами программного обеспечения итеративных методов). Блэшфилд и Олдендерфер (1978a) показали, что разумный выбор начального разбиения лишь ненамного улучшает положение дел, но Миллиган (1980) продемонстрировал, что итерационный процесс по принципу k-средних, использующий начальное разбиение, полученное кластеризацией по методу средней связи, приводит к лучшему восстановлению известной структуры данных по сравнению с прочими итеративными и иерархическими методами кластеризации. Другими исследователями было доказано, что итеративные методы дают оптимальное решение при любом начальном разбиении, если даиные имеют хорошую структуру (Everitt, 1980; Bayne et. al., 1980). Как видим, для решения этой задачи нужно провести больше исследований с помощью метода Монте-Карло.

ВАРИАНТЫ ФАКТОРНОГО АНАЛИЗА

Эти методы кластерного анализа весьма популярны в психологии. Они известны больше как варианты факторного анализа, обратный факторный анализ или факторизация Q-типа. Работа методов начинается с формирования корреляциоиной матрицы сходств между объектами. Обычно факторный анализ проводится с корреляционной матрицей размерностью $P \times P$, но если нужно

определить кластеры, то анализ осуществляется на основе корреляционной матрицы размерностью $N \times N$. По корреляционной матрице определяются факторы, и объекты распределяются по кластерам в зависимости от их факторных нагрузок.

Использование факторного анализа Q-типа имеет долгую историю. Самыми ревностными сторонниками этого вида кластеризации до недавнего времени были Оуверолл и Клетт (1972), а также Скиинер (1979). Предметом критики методов факторного анализа в кластеризации стали неправомерное применение линейной модели к объектам, проблема множественных факторных нагрузок (неясно, что делать с объектом, который имеет высокие нагрузки более чем для одного фактора) и двойное центрирование данных (Everitt, 1980; Fleiss et. al., 1971).

Чтобы дать читателю представление о том, как используется обратный факторный анализ, приведем пример, где рассматривается модальный профильный анализ (Skinner, 1979). В этом методе кластеризации для формирования пространства малой размерности, представляющего соотношения между объектами, взята декомпозиция Экерта — Юнга. Подход Скиннера на основе пространственной модели (концептуальные вопросы часто возникают в связи с обратным факторным анализом) обсуждается в (Skinner, 1979). Процедура состоит из трех главных шагов: 1) начальной оценки факторов; 2) увеличения числа факторов с помощью повторных выборок и 3) проверки общности факторов на новой выборке. Первый шаг этой процедуры иллюстрируется данными MMPI-теста.

Трехфакторное решение было выбрано потому, что нам заранее известно о существовании трех кластеров в данных. Первые семь собственных значений[2] решения равны:

28,07
17,16
11,49
9,39
5,39
4,60
4,22

Используя стандартные приемы факторного анализа для оценки числа факторов, можно было бы привести доводы в пользу того, что двухфакторное или четырехфакторное решение будет более приемлемо, чем трехфакторное. Тем не менее, поскольку было известно, сколько диагностических классов существует в данных, рассматривалось только трехкластерное решение.

Первым с помощью модального профильного анализа был получен биполярный фактор, который положительно коррелировал с фактором, соответствующим больным неврозами и отрицательно — с факторами, соответствующими больным расстройствами личности. Третий фактор включал в основном факторы, соответствующие больным психозами. Второй фактор состоял из всех факторов трех групп больных.

Три модальных профиля похожи на те, которые большинство психологов-клиницистов описывают как типичные MMPI-профили больных «неврозами», «расстройствами личности» и «психозами». Однако они имеют меньшие различия, чем профили для реальных трех групп в данных. Это согласуется с главной чертой обратного факторного анализа, который придает большее значение форме, а не сдвигу.

ДРУГИЕ МЕТОДЫ

Иерархические дивизимные методы являются логической противоположностью агломеративным методам. В начале процедуры (при $K=1$) все объекты принадлежат одному кластеру, а затем этот всеобъемлющий кластер разрезается на последовательно уменьшающиеся «ломтики». Есть два дивизимных вида: монотетический и политетический. Монотетический кластер — это группа, все объекты которой имеют приблизительно одно и то же значение некоторого конкретного признака. Таким образом, монотетические кластеры определяются фиксированными признаками, определенные значения которых *необходимы* для принадлежности к кластерам. В противоположность этому политетические кластеры являются группами объектов, для принадлежности к которым *достаточно* наличия определенных сочетаний из некоторого подмножества признаков. Все три метода — иерархические, агломеративные и итеративные — будут образовывать только политетические кластеры.

Монотетические дивизимные методы применяют в первую очередь к бинарным данным, а процедура деления совокупности объектов на подгруппы основана на определении признака, максимизирующего несходство между кластерами, получающимися в результате. Часто дивизимные критерии основаны на использовании статистики χ^2 или некоторых информационных статистик (Clifford and Stephenson, 1975; Everitt, 1980). Монотетический подход к дивизимной кластеризации, известный также как ассоциативный анализ, широко распространен в экологии, но применение этого метода в социальных науках ограничено археологией (Peebles, 1972; Whallon, 1971; 1972).

Методы поиска модальных значений плотности рассматривают кластер как область пространства с «высокой» плотностью точек по сравнению с окружающими областями. Они «обследуют» пространство в поисках скоплений в данных, которые и представляют собой области высокой плотности. Существуют два основных вида методов поиска модальных значений плотности: методы, основанные на кластеризации по одиночной связи, и методы разделения «смесей» многомерных вероятностных распределений.

Как отметил Эверитт (1980), методы поиска модальных значений плотности, основанные на кластеризации по одиночной связи, препятствуют образованию цепочек. В отличие от метода одиночной связи методы поиска модальных значений плотности под-

чинены строгому правилу, согласно которому предпочтение отдается образованию нового кластера, а не присоединению очередного объекта к уже существующей группе. Обычно это правило основано на измерении расстояния между существующим кластером и новым объектом или кластером (Wishart, 1969) или же на измерении среднего сходства, как в методе TAXMAP, предложенном Кармайклом и Снитом (1969). Если правило не выполняется, объединение объектов и кластеров не производится. Из этих методов широкое распространение получил модальный анализ, впервые предложенный Уишартом (1969) и позднее встроенный в пакет программ по кластерному анализу CLUSTAN (Wishart, 1982). Несмотря на привлекательность, этот метод обладает некоторыми недостатками, из которых наиболее важным является его зависимость от выбора шкал измерений. Кроме того, предполагается, что искомые в пространстве кластеры имеют сферическую форму.

Другая основная группа методов поиска модальных значений плотности — методы по определению параметров смеси распределений. Смесь определяется как совокупность выборок, представляющих различные популяции объектов. Например, множество данных MMPI-теста является смесью потому, что оно содержит выборки из трех популяций: больных неврозами, психозами и расстройствами личности. Этот подход к кластерному анализу явно основан на статистической модели, в которой элементы разных групп или классов должны иметь различные вероятностные распределения признаков. Цель кластеризации данных состоит в определении параметров, описывающих распределения для популяций.

Важные частные случаи разделения смесей реализованы в процедурах NORMIX и NORMAP, разработанных Вульфом (1970, 1971). Процедура NORMIX получает оценки максимального правдоподобия для параметров многомерных смесей нормальных распределений. Настоящий метод предполагает, что основные популяции различаются средними и ковариационными структурами Процедура NORMAP построена на более простом предположении, что структуры внутригрупповых ковариаций одинаковы. Уникальность обеих процедур NORMIX и NORMAP состоит в том, что они не распределяют объекты по кластерам, а вместо этого дают вероятность принадлежности каждого объекта к каждому из кластеров. Например, в случае перекрывающихся кластеров вероятность того, что объект принадлежит обоим кластерам, равна 0,5 (Wishart, 1982).

Методы поиска модальных значений плотности особенно чувствительны к проблеме субоптимальных решений (Everitt, 1980), поскольку уравнение максимального правдоподобия в общем случае может иметь несколько решений. Хотя в принципе можно сравнить оценки для различных неоптимальных решений, однако это нелегко сделать (или вовсе невозможно) даже для небольших задач. Другой недостаток данных методов в том, что все компоненты смеси являются многомерными нормальными распределениями. Очевидно, возможны и другие виды распределений, но

неясно, насколько устойчивы к нарушению предположения о нормальности.

Методы сгущения уникальны в том смысле, что они позволяют создавать перекрывающиеся кластеры. В отличие от иерархических методов, это семейство кластерных методов не порождает иерархические классификации; объектам разрешается быть членами нескольких кластеров. Многие ранние разработки методов сгущения относятся к лингвистическим исследованиям, поскольку именно там важно учитывать, что некоторые слова имеют различные значения.

Методы сгущения требуют вычисления матрицы сходства между объектами и определения оптимального значения статистического критерия, называемого специалистами «функцией когезии» («функция сцепления»). Затем объекты перемещаются до тех пор, пока функция не достигнет оптимального значения. Поскольку эти методы одновременно создают лишь две группы, то обычно первичные данные случайным образом разделяются на несколько начальных конфигураций, каждая из которых в дальнейшем может быть рассмотрена с точки зрения пригодности. Серьезный недостаток рассматриваемых методов состоит в том, что из-за неудачной поисковой процедуры время от времени происходит повторное обнаружение одних и тех же групп, а это не дает новой информации. Другим практическим недостатком является то, что их характеристики малоизвестны, так как эти методы не имеют широкого распространения. Джардайн и Сибсон (1968) предложили метод сгущения, основанный на теории графов, который, хотя и лишен серьезного недостатка повторного обнаружения групп, все же ограничен анализом лишь очень малых групп ($N \leqslant 25$), что обусловленно чрезвычайной вычислительной трудоемкостью (см. также Cole and Wishart, 1970).

По многим причинам методы теории графов оказались среди новых методов, доступных исследователю. Значительный интерес для теоретиков (а также для пользователей) представляет то, что кластерные методы этого семейства основаны на хорошо разработанных теоремах и аксиомах теории графов. А поскольку из теорем теории графов вытекает большое количество полезных следствий, то возможно, что эта теория станет альтернативой преимущественно эвристическому характеру других кластерных методов. Например, иерархические агломеративные методы могут быть сжато описаны в терминах теории графов (Dubes and Jain, 1980). Теория графов ведет также к созданию нуль-гипотезы, которая может быть использована при проверке наличия кластеров в матрице сходства. Она известна как «гипотеза случайного графа», утверждающая, что все ранжированные матрицы близости являются равновероятными (Ling, 1975). Кроме того, теория графов применяется при разработке более эффективных вычислительных алгоритмов для известных методов кластеризации и в некоторых случаях позволяет сделать число анализируемых объектов довольно большим.

ОПРЕДЕЛЕНИЕ ЧИСЛА КЛАСТЕРОВ

Поскольку кластерный анализ предназначен для создания однородных групп, естественно рассмотреть процедуры, позволяющие определить число полученных групп. Например, вложенная древовидная структура дендрограммы указывает на то, что в данных может находиться много различных групп, и правомерен вопрос: где нужно «обрезать» дерево, чтобы получить оптимальное число групп? Точно так же и при работе с итеративными методами пользователь должен указать число групп, присутствующих в данных, еще до создания этих групп.

К сожалению, эта проблема до сих пор находится среди нерешенных задач кластерного анализа из-за отсутствия подходящей нулевой гипотезы и сложной природы многомерных выборочных распределений.

Затруднения в создании работоспособной нулевой гипотезы вызывает отсутствие непротиворечивого и универсального определения кластерной структуры. Но, как мы уже указывали, появление такого определения маловероятно. Понятие «отсутствие структуры» в наборе данных (одна из возможных нулевых гипотез) весьма далеко от ясности, и непонятно, каким должен быть тест, позволяющий определить, есть ли в данных структура или нет. Уже созданные нулевые гипотезы (такие, как гипотеза случайного графа и гипотеза случайного положения), возможно, и полезны, но исчерпывают далеко не все возможности и должны еще найти свое место в практическом анализе данных. В любом случае «отклонение нулевой гипотезы не имеет особого значения, потому что разумные альтернативные гипотезы еще не разработаны; практичного и математически полезного определения «кластерной структуры» нет до сих пор» (Dubes and Jain, 1980).

В той же степени не поддается решению задача о разделении смеси многомерных распределений в анализе реальных данных. Хотя многие вопросы многомерных нормальных распределений хорошо разработаны, все же реальные данные не будут соответствовать этому стандарту; более того, многие выборки реальных данных являются сложными смесями, имеющими различные многомерные выборочные распределения неизвестной структуры. Поскольку не существует статистической теории и теории распределений, которые помогли бы в разделении этих смесей, также неразумно ожидать появления формальных тестов для целей кластерного анализа.

Реакция на эти ограничения была различной. В некоторых отраслях, особенно в биологии, задача определения числа кластеров не имеет первостепенной важности просто потому, что целью анализа является предварительное исследование общей картины зависимостей между объектами, представленной в виде иерархического дерева. Однако в социальных науках развиваются два основных подхода к определению числа присутствующих кластеров: эвристические процедуры и формальные тесты.

Эвристические процедуры — несомненно наиболее часто используемые методы. На самом верхнем базисном уровне иерархическое дерево «обрезается» после субъективного просмотра различных уровней дерева. Для дендрограммы (рис. 8), изображающей результаты обработки полного набора данных о захоронениях методом Уорда, применяемых евклидово расстояние, субъективная обрезка дерева приведет к выделению двух кластеров одного уровня и, возможно, трех кластеров, если рассматривать различные уровни дерева. Эту процедуру вряд ли можно назвать удовлетворительной, поскольку обычно ее результаты зависят от нужд и представлений исследователей о «правильной» структуре данных.

Более формальный, но все же эвристический подход к задаче состоит в том, чтобы графически изобразить число получаемых из иерархического дерева кластеров как функцию коэффициента слияния или смешения, равного числу способов объединения различных объектов в кластер. Значения коэффициентов слияния показаны вдоль оси *Y* древовидной диаграммы. Этот тест, вариант

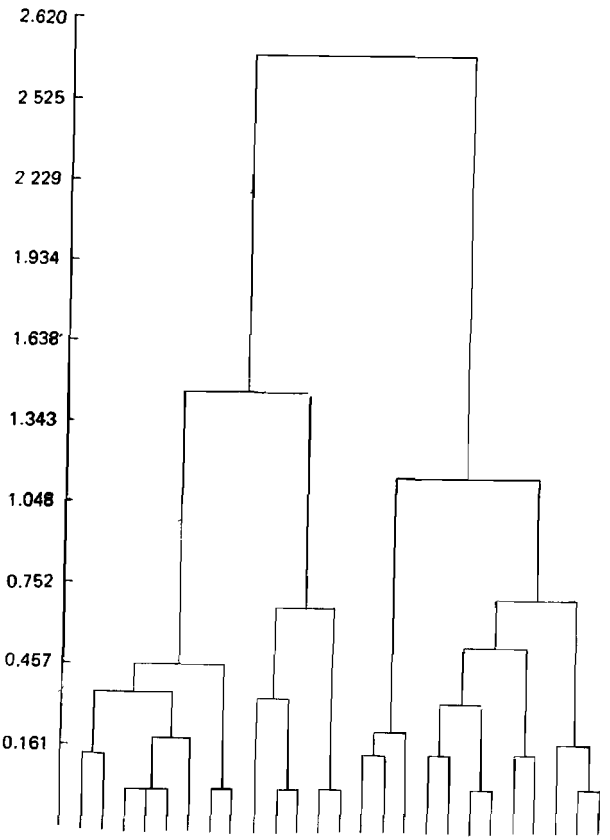

Рис. 8. Дендрограмма метода Уорда для полного набора данных о захоронениях

которого был предложен Торндайком в 1953 г., аналогичен критерию отсеивания факторного анализа. Заметное «уплощение» на этом графике говорит о том, что дальнейшее слияние кластеров не дает новой информации. На рис. 9 показан такой график для полного набора данных о захоронениях, полученный с помощью метода Уорда и евклидова расстояния. Уплощение кривой начинается вблизи решения из трех кластеров, и линия остается, по существу, плоской возле решения из двух кластеров. Отсюда следует, что в данных присутствуют три (но вероятнее всего два) кластера.

Другая субъективная процедура, несколько более формализованная, заключается в том, чтобы при новом просмотре значений коэффициента слияния найти значимые «скачки» значения коэффициента. Скачок означает, что объединяются два довольно несхожих кластера. Таким образом, число кластеров, предшествую-

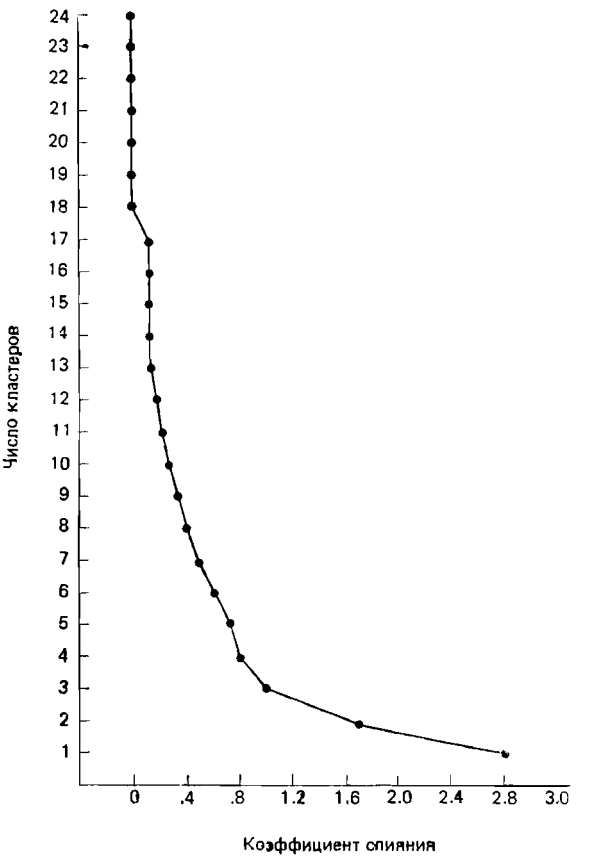

Рис. 9. График зависимости между числом кластеров и величиной коэффициента слияния, полученный с помощью метода Уорда для полного набора данных о захоронениях

щее этому объединению, является наиболее вероятным решением. Ниже показаны коэффициенты слияния, соответствующие числу кластеров, которое для полного множества данных о захоронениях принимает значения от 10 до 1.

10	кластеров	0,312	5	кластеров	0,729
9	»	0,333	4	»	0,733
8	»	0,354	3	»	1,075
7	»	0,458	2	»	1,708
6	»	0,642	1	»	2,872

Как видим, между решениями из четырех и трех кластеров есть скачок, что приводит к выводу о допустимости решения из четырех кластеров. Одна из трудностей, связанная с этой процедурой, состоит в том, что можно найти много малых скачков значения коэффициента слияния, но совершенно невозможно исходя лишь из простого визуального обследования указать, какой из этих скачков «правильный».

Этот тест был обобщен в работах (Mojena, 1977, Mojena and Wishart, 1980). Там же была разработана эвристическая процедура, позволяющая лучше определить «значимый скачок» коэффициента. «Правило остановки № 1», как его определил Мойена, предписывает, что групповой уровень или оптимальное разбиение иерархического кластерного решения получается, если удовлетворяется неравенство

$$z_{j+1} > z + ks_z,$$

где z — величина коэффициента слияния; z_{j+1} — величина коэффициента на $(j+1)$-м этапе кластерного процесса; k — стандартное отклонение, а z и s_z — среднее и стандартное отклонение коэффициентов слияния. Невыполнение неравенства говорит о том, что в данных имеется только один кластер.

На практике стандартное отклонение может быть вычислено на каждом этапе кластерного процесса, где k равно:

$$k_j = \{z_{j+1} - z\}/s_z.$$

Значения коэффициента слияния для полного набора данных о захоронениях, обработанного методом кластеризации Уорда с использованием евклидова расстояния, были рассмотрены выше. Теперь приведем значения стандартного отклонения для решений, содержащих от 1 до 4 кластеров:

	Коэффициенты слияния	Стандартное отклонение
4 кластера	0,458	0,472
3 »	0,974	1,074
2 »	1,929	1,707
1 »	3,684	2,871

В этом случае согласно правилу остановки оптимальным считается решение из трех кластеров. Уишарт (1982) отметил, что можно оценить статистическую значимость результатов, полученных с помощью этого правила, используя t-статистику с $n-2$ степе-

нями свободы, где n — число коэффициентов слияния. Процедура заключается в перемножении квадратного корня из $n - 1$ и значения стандартного отклонения k. В данном примере значения 4,79 (квадратный корень из 23) умножается на 9,74, в результате получаем 4,67. Значение значимо с уровнем 0,01 при 22 степенях свободы. Сейчас этот метод вместе с более сложным правилом встроен в процедуру CLUSTAN2.

Трудности, связанные с составными многомерными выборочными распределениями, мало сказались на разработке формальных статистических тестов, но широкое распространение получило лишь небольшое число этих тестов. Нулевая гипотеза, наиболее часто применяемая в статистических тестах, предполагает, что исследуемые данные являются случайной выборкой из генеральной совокупности с многомерным нормальным распределением. Вульф (1971), считая, что это предположение верно, предложил тест отношения правдоподобия для проверки гипотезы, что имеется r, а не r' групп. Альтернативная гипотеза, разработанная Ли (1979), заключается в следующем: данные — это выборка из генеральной совокупности с равномерным распределением. Тест, основанный на альтернативной гипотезе, использует критерий внутригрупповой суммы квадратов. Он является полезной отправной точкой в определении возможных различий между кластерами. К сожалению, тест может работать только с одним признаком. Какая бы процедура ни была выбрана, пользователь должен постоянно сознавать, что лишь малая часть этих тестов подверглась широкому изучению. Таким образом, поскольку большинство тестов плохо изучено и эвристично, то результаты их использования должны приниматься с большой осторожностью. В идеале правила определения числа имеющихся в наличии кластеров должны использоваться совместно с подходящей процедурой проверки достоверности результатов (см. разд. IV), так как может случиться, что правило остановки рекомендует такое число кластеров, которое не подтверждается результатами измерений по другим критериям.

СРАВНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Как мы уже говорили, с помощью разных методов кластеризации можно получить различные результаты для одних и тех же данных. Теперь попробуем разобраться, почему так происходит. Мы постоянно отмечали, что некоторые методы имеют присущие только им особенности и свойства. Например, метод одиночной связи имеет тенденцию к образованию длинных цепочек кластеров, в то время как метод Уорда склонен к образованию плотных гиперсферических кластеров. Понимание того, что различные методы кластеризации будут порождать заметно различающиеся результаты, имеет более чем просто академический интерес, поскольку эти методы применяются к исследованию реальных данных без особых раздумий, рутинно. Лучше еще до исчерпывающего анализа данных знать сильные и слабые стороны различных методов, чем

внезапно обнаружить, что результаты анализа во многом обязаны свойствам самого метода, а не внутренней структуре данных.

Многие сравнения кластерных методов сводятся к оценке, насколько хорошо различные методы кластеризации восстанавливают известную структуру данных. Хотя в некоторых из этих исследований были использованы реальные данные с такими же характеристикамн, в большинстве случаев применялись искусственные данные, полученные с помощью моделирования методом Монте-Карло и выборочного метода, которые специально создавались для имитации особенностей реальных данных (например, данные, имитирующие результаты MMPI-теста (Blashfield and Morey, 1980). Чаще всего наборы данных подбирались в соответствии со свойствами важных видов распределений, таких, как двумерное нормальное, многомерное нормальное и многомерное гамма-распределения. В зависимости от цели сравнения эти наборы данных изменялись в размерах (число объектов на кластер), форме кластеров, числе кластеров в данных, степени перекрытия кластеров, наличии выбросов и степени полноты классификации (должна ли классификация быть исчерпывающей). Некоторые сравнения проводились на наборах данных, удовлетворяющих ультраметрическому неравенству — более строгому варианту неравенства треугольиика, описанному в разд. II (Milligan and Issac, 1980). Внимание было уделено последствиям использования различных мер сходства.

Результаты проведенных исследований трудно свести воедино, потому что каждое из них придает особое значение своей комбинации структур данных и проверяемых методов. Поэтому не удивительно, что были получены противоречивые результаты (Milligan, 1981). Однако, по-видимому, четыре фактора оказывают на работу методов кластеризации большое влияние:

1) характеристики кластерной структуры;
2) наличие выбросов и степень полноты классификации;
3) степень перекрытия кластеров;
4) выбор меры сходства.

Наиболее важными характеристиками кластерной структуры, влияющими на работу методов кластеризации, являются форма кластеров, размеры кластеров (которые выражены в числе объектов, приходящихся на кластер, и в различиях относительных размеров кластеров) и число кластеров. Мы уже показали на нескольких примерах, что определенные методы кластеризации склонны к обнаружению определенных видов кластеров. Расширяющие пространство методы, такие, как метод Уорда, полных связей, итеративные методы группировки, использующие критерий $\mathrm{tr}\, W$, приводят к кластерам гиперсферической формы. Поэтому не надо удивляться, что в исследованиях, использующих методы Монте-Карло для создания кластеров такой формы, расширяющие пространство методы восстанавливают известную кластерную структуру лучше, чем сужающие пространство методы. Не удивительно и то, что эти методы обычно не в состоянии вос-

становить структуру кластеров, имеющих растянутую или необычную форму. Кроме того, расширяющие пространство методы имеют тенденцию находить кластеры приблизительно равных размеров. Но в этом случае, как показывают исследования по методу Монте-Карло, кластеры, состоящие из относительно небольшого числа объектов, могут слиться с кластерами больших размеров. Мойена (1977) доказал, в частности, что увеличение числа групп в данных неблагоприятно воздействует на работу метода Уорда, но этот результат не нашел подтверждения в другой работе, где применяется метод Монте-Карло. Вообще говоря, сужающие пространство методы, подобные методу одиночной связи, хорошо работают с теми кластерными структурами, для которых расширяющие пространство методы оказываются несостоятельными. Так, исследованиями, использующими метод Монте-Карло, проверено, что сужающие пространство методы действительно дают хорошее восстановление известной кластерной структуры, если кластеры хорошо определены и разделены.

Наличие выбросов и степень полноты классификации, требуемая при кластеризации, — важные факторы, влияющие на работу метода. Полная классификация является исчерпывающей: все рассматриваемые объекты должны быть размещены по группам. Основанное на методе Монте-Карло исследование влияния этого фактора показывает, что, если требуется полная классификация и данные имеют мало выбросов, то метод Уорда дает превосходное восстановление известной кластерной структуры (Kuiper and Fisher, 1975; Mojena, 1977). Однако в работах (Blashfield and Morey, 1980; Edelbrock, 1979; Edelbrock and McLaughlin, 1979; Milligan, 1980) показано, что если степень полноты классификации уменьшается, то кластеризация по методу средней связи дает восстановление такое же, что и по методу Уорда. Миллиган и Иссак (1980), воспользовавшись данными с ультраметрикой, доказали, что в действительности метод средней связи может работать лучше метода Уорда даже в случае полной классификации. Вообще может показаться, что на работе расширяющих пространство методов неблагоприятно сказывается присутствие большого числа выбросов, но это утверждение еще требует проверки. Важно помнить, что выбросы — это не просто обособленные объекты; на самом деле они могут быть представителями подгрупп, о которых в выборке содержится мало данных. Следовательно, очевидное решение проблемы выбросов (отбросить эти данные) должно быть хорошо продуманным. Независимо от их интерпретации выбросы необходимо тщательно исследовать еще до применения кластерного анализа. Для каждого выброса необходимо определить, почему он так отличен от других объектов.

Проблема перекрытия кластеров аналогична проблемам степени полноты классификации и наличия выбросов в выборке данных. Перекрытие кластеров — это просто степень, с которой кластеры занимают одно и то же пространство. Кластеры могут быть хорошо разделенными, но могут находиться и близко один к дру-

гому. Кроме того, могут присутствовать шумовые точки, т. е. точки данных, лежащие между границами кластеров. Как было показано, все эти факторы оказывают сильное влияние на работу методов кластеризации, а также, в случае перекрытия кластеров метод Уорда работает лучше большинства других методов кластеризации (Bayne et. al., 1980), тогда как метод средней связи работает плохо (Bayne et. al., 1980; Milligan, 1980). Однако при прочих равных условиях метод средней связи дает классификацию такой же полноты, что и метод Уорда для хорошо разделенных кластеров. Если же требования к полноте классификации ослаблены и допускаются перекрытия кластеров, то этот метод опять будет эквивалентен методу Уорда (Edelbrock, 1979; Edelbrock and McLaughlin, 1979; Milligan, 1980).

И наконец, на работу методов кластеризации влияет выбор меры сходства. К сожалению, были изучены только две меры: евклидово расстояние и коэффициент смешанного момента корреляции. Фактически во всех описанных выше исследованиях по методу Монте-Карло сравнивалась работа метода Уорда, использовавшего евклидово расстояние, и метода средней связи, применявшего коэффициент смешанного момента корреляции. Хотя, кажется, выбор меры все же приводит к некоторым различиям в результатах, его воздействие скрадывается воздействием характеристик кластерной структуры, требуемой степени полноты классификации и перекрытия кластеров. Надо еще очень много поработать с другими коэффициентами, прежде чем можно будет оценить влияние выбора меры сходства на работу методов кластеризации.

БИБЛИОГРАФИЧЕСКИЕ ЗАМЕЧАНИЯ

В этом разделе было рассмотрено лишь несколько различных методов кластеризации. Конечно, здесь не дается исчерпывающего обсуждения всего того, что известно о рассмотренных методах. Для новичков, желающих ознакомиться с другими введениями в методы кластеризации, мы рекомендуем обратиться к работам: (Bailey, 1975; Everitt, 1980; и Lorr, 1983). Все три работы вполне доступны и различаются лишь своей направленностью. Опытному пользователю мы посоветовали бы ознакомиться с обзором Кормака (1971), хотя он немного и устарел, а также с работой Эвритта (1979), являющейся сводом нерешенных задач кластерного анализа. Кроме того, имеется пять книг, которые содержат большое количество детальной информации о различных методах кластеризации и их работе: (Anderberg, 1973; Clifford and Stephenson, 1975; Hartigan, 1975; Mezzich and Solomon, 1980; Sneath and Socal, 1973).

IV. МЕТОДЫ ПРОВЕРКИ ОБОСНОВАННОСТИ РЕШЕНИЙ

В этом разделе обсуждаются пять методов проверки достоверности (обоснованности) решений кластерного анализа: 1) кофенетическая корреляция; 2) тесты значимости для признаков, используемых при создании кластеров; 3) повторная выборка; 4) тесты значимости для независимых признаков и 5) методы Монте-Карло.

КОФЕНЕТИЧЕСКАЯ КОРРЕЛЯЦИЯ

Кофенетическая корреляция была впервые предложена Сокэлом и Рольфом в 1962 г. Она является главной мерой обоснованности решения, предлагаемой специалистами по численной таксономии (Sneath and Sokal, 1973). Эта мера используется только вместе с иерархическим агломеративным методом. Кофенетическая корреляция необходима для определения, насколько хорошо характер отношений (сходство/несходство) между объектами представляется деревом или деидрограммой, полученными с помощью иерархического метода кластеризации.

Решение для данных о шести захоронениях, полученное методом одиночной связи с использованием коэффициента Жаккарда, представлено в виде иерархического дерева (см. рис. 3). Просмотрев дерево, можно получить представление о сходствах для любой пары объектов. Например, объект ПЖЭ (подросток, женский пол, элитарный) и ВЖЭ (взрослый, женский пол, элитарный) довольно похожи, поскольку они объединяются относительно «высокой» ветвью дерева. С другой стороны, объекты РЖЭ и ПЖЭ мало похожи, так как они не объединяются в единый кластер до самого последнего шага (т. е. они объединяются лишь у основания дерева).

С помощью дерева, приведенного на рис. 3, можно построить вторичную матрицу сходства между всеми парами объектов, соответствующую рассматриваемому иерархическому решению:

	РМН	РЖЭ	ПМН	ПЖЭ	ВМЭ	ВЖЭ
РМН	—	0,250	0,333	0,333	0,333	0,333
РЖЭ		—	0,250	0,250	0,250	0,250
ПМН			—	0,500	0,500	0,500
ПЖЭ				—	0,500	0,750
ВМЭ					—	0,500
ВЖЭ						—

Каждый элемент матрицы представляет собой значение сходства для уровня, на котором определенная пара объектов была объединена в общий кластер. Важно отметить, что эта матрица сходства имеет не более $N-1$ различных элементов, так как для иерархического агломеративного метода всегда требуется $N-1$

шагов объединения. Исходная матрица содержит до $N(N-1)/2$ различных элементов и имеет вид

	РМН	РЖЭ	ПМН	ПЖЭ	ВМЭ	ВЖЭ
РМН	—	0,000	0,250	0,250	0,333	0,200
РЖЭ		—	0,000	0,250	0,143	0,200
ПМН			—	0,200	0,500	0,167
ПЖЭ				—	0,500	0,750
ВМЭ					—	0,429
ВЖЭ						—

Кофенетическая корреляция является корреляцией между значениями исходной матрицы сходства и вторичной матрицы сходства. Таким образом, кофенетическая корреляция для решения, полученного методом одиночной связи и показанного на рис. 3, равна $C=0{,}810$.

Несмотря на довольно частое применение, кофенетическая корреляция имеет и явные недостатки. Во-первых, использование смешанного момента корреляции предполагает, что нормально распределенные значения в двух матрицах коррелированы. Это предположение обычно не выполняется для значений вторичной матрицы сходства, так как кластерные методы в значительной степени определяют распределение значений сходства в этой матрице. Таким образом, применение коэффициента корреляции для оценки степени сходства между значениями двух матриц не является оптимальным. Во-вторых, поскольку число различных значений во вторичной матрице сходства меньше, чем в исходной матрице, то и количество информации, содержащейся в каждой из двух матриц, весьма различно. Холгерссон (1978) провел исследование с помощью метода Монте-Карло для того, чтобы проанализировать характеристики кофенетической корреляции, и обнаружил, что она является плохим индикатором качества кластерного решения.

ТЕСТЫ ЗНАЧИМОСТИ ДЛЯ ПРИЗНАКОВ, НЕОБХОДИМЫЕ ПРИ СОЗДАНИИ КЛАСТЕРОВ

Другой процедурой, которая часто используется в прикладных исследованиях с применением кластерного анализа, является многомерный дисперсионный анализ (MANOVA) признаков, необходимых для получения решения. Цель анализа — выяснить с помощью тестов для проверки гипотезы однородности, значимо ли разбиение данных на кластеры[3]. В отличие от кофенетической корреляции, применяемой для анализа правильности иерархического дерева, выполнение стандартных тестов значимости связано с качеством кластерного решения, представляющего собой разбиение множества данных. Таким образом, процедуру MANOVA можно применять к решениям, полученным любым методом кластеризации, лишь бы он порождал разбиения (например, итеративные методы группировки, иерархические методы, варианты факторного анализа).

Ясно, что использование MANOVA для проверки гипотезы однородности кластеров вполне разумно. Более того, она становится весьма популярной процедурой, потому что ее результаты всегда имеют высокую значимость. Так, при исследовании типов верующих (Filsinger et. al., 1979), обсуждавшемся в разд. I, были обнаружены значимые различия между кластерами после проведения дискриминантного анализа признаков, необходимых при создании кластеров. В действительности дискриминантный анализ правильно классифицировал 96% субъектов. Эти результаты свидетельствуют, что кластерное решение, полученное Филсингером и другими, хорошо описывает типы верующих людей. Однако *такое использование дискриминантного анализа (или MANOVA, или многократно ANOVA) оказывается статистически неправомерным.*

Чтобы понять это, рассмотрим следующий пример. Предположим, что группа исследователей проводит *IQ*-тест среди случайно отобранных детей из одного класса по всей школьной системе. Далее предположим, что в этом наборе данных на самом деле нет кластеров. После того как будет построена диаграмма вдоль оси *IQ*-оценок, исследователи получат нормальное распределение со средним, равным 100 (именно такое значение можно было предсказать исходя из нормативных данных по этому признаку). Тем не менее допустим, что они все же решили провести кластерный анализ полученных данных, несмотря на унимодальное распределение по *IQ*-признакам. Найденное кластериое решение делит выборку на две группы: с коэффициентом *IQ*, превышающим 100, и с коэффициентом *IQ* не выше 100. Если затем исследователи проведут дисперсионный анализ для сравнения групп по величине их *IQ*-оценок, применение *F*-теста покажет высокую значимость! Этот «высокозначимый» результат будет иметь место, несмотря на то, что в данных не существует ни одного кластера. С помощью методов кластерного анализа (по определению) объекты разделяются на кластеры, которые фактически не перекрываются по признакам, применявшимся при создании кластеров. Проверки значимости различий между кластерами по этим признакам будут всегда давать положительные результаты, поскольку результаты таких проверок всегда положительны независимо от того, есть в данных кластеры или нет. Описанное использование тестов в лучшем случае бесполезно, в худшем — ведет к заблуждениям.

ПОВТОРНАЯ ВЫБОРКА

Первые два метода проверки достоверности результатов применяются часто, но они имеют серьезные недостатки. Специалисты по методологии кластерного анализа не рассматривают эти методы в качестве методики обоснования результатов (Hartigan, 1975a).

Третий метод позволяет оценить степень повторяемости кластерного решения в серии наборов данных. Если для различных выборок из одной и той же генеральной совокупности получается одинаковое кластерное решение, то напрашивается вывод, что это решение присуще всей совокупности. Маловероятно, что неустойчивое кластерное решение может отражать свойства генеральной совокупности. Эта методика уже рассматривалась в качестве примера в разд. I. Голдстейн и Линден (1969), проводя исследование больных алкоголизмом, разделили всю полученную выборку на две части, к которым затем применили один и тот же метод кластерного анализа. В результате в обоих решениях были обнаружены те же самые четыре кластера.

Методика повторных выборок фактически предназначена для проверки соответствия кластерного решения внутренней структуре генеральной совокупности. То, что одни и те же кластеры обнаруживаются в различных подмножествах, когда к ним применяются одинаковые кластерные методы, *не* доказывают обоснованность решения. Другими словами, при неудачной попытке повторить кластерное решение оно отвергается, но успешное повторение не дает гарантии достоверности этого решения.

ТЕСТЫ ЗНАЧИМОСТИ ДЛЯ ВНЕШНИХ ПРИЗНАКОВ

Процедуры, включенные в данную категорию, вероятно, считаются лучшими среди способов обоснования достоверности кластерного решения, но, к сожалению, этот подход мало использовался, несмотря на его потенциальные возможности. По существу, процедура заключается в проведении теста значимости, с помощью которого сравниваются кластеры по признакам, *не* применявшимся при получении кластерного решения. Этой методикой пользовались при исследовании верующих Филзингер и др. (1979). Они сопоставили полученные кластеры по семи демографическим признакам, не участвовавшим при формировании кластеров. В результате были обнаружены значимые различия по четырем из этих семи признаков.

Интересное исследование, в котором использовался более сложный тест внешней обоснованности кластерного решения, было проведено Финни и Мусом (1979). Эти исследователи, подобно Голдстейну и Линдену (1979) (см. разд. I), хотели определить, возможно ли выделить подтипы больных алкоголизмом. Проанализировав данные вопросников о 429 больных, они нашли восемь кластеров. Для этих же больных в течение шести месяцев были собраны данные по пяти признакам: 1) самоотчет о потреблении алкоголя; 2) отказ от спиртного; 3) физические повреждения; 4) реабилитация и 5) выполнение профессиональных функций. Финни и Мус нашли, что восемь кластеров различались по этим пяти признакам. Это было показано F-тестом значимости с помощью процедуры ANOVA. Кроме того, исследователи обнаружили значимые

связи между участием больных в лечебных программах и их принадлежностью к кластерам.

Сила внешнего обоснования заключается в том, что оно непосредственно проверяет достоверность кластерного решения по отношению к подходящему критерию. Одной из причин, почему этот подход к проверке достоверности решения редко используется в исследованиях с кластерным анализом, является высокая стоимость методологического планирования сбора данных для рассматриваемого критерия. Другая вероятная причина заключается в чисто исследовательском (поисковом) характере работ, где необходим кластерный анализ. Отсутствие разработанной теории, сопровождающей весь процесс создания классификации, не позволяет выделить группу внешних критериев, соответствующих целям исследования. Однако кластерные решения, успешно прошедшие проверку на достоверность, по сравнению с прочими решениями намного ценнее.

ПРОЦЕДУРЫ МОНТЕ-КАРЛО

Последний подход к обоснованию достоверности решений используется сравнительно мало и в некоторой степени труден для изложения. В сущности, этот подход заключается в применении процедур Монте-Карло, применяющих генераторы случайных чисел, для создания наборов данных с осиовными характеристиками, соответствующими характеристикам реальных данных, но не содержащих кластеров. Одни и те же методы кластеризации употребляются как к реальным данным, так и к искусственным, а полученные решения сравниваются с помощью подходящих методов. Пример такого процесса, использующий данные MMPI-теста, возможно лучший способ проиллюстрировать этот подход.

Шаг 1. Создание рандомизированного набора данных. С помощью генератора случайных чисел создается множество искусственных данных, которое не имеет кластеров, но обладает теми же характеристиками, что и реальный набор данных. Чтобы сделать это, мы вычислили общие средние, стандартные отклонения и матрицу корреляций между признаками для исходного множества данных MMPI-теста о 90 больных. Далее для создания рандомизированного набора данных мы написали короткую программу на Фортране, которая использует генератор случайных чисел из пакета программ IMSL. Этот генератор порождает данные, являющиеся выборкой из генеральной совокупности с многомерным нормальным распределением с заданным вектором средних и заданной ковариационной матрицей. Первый шаг может показаться труднопреодолимым для пользователя, но в действительности такую программу довольно легко написать: требуется лишь 36 операторов Фортрана. В результате получаем рандомизированное множество данных о 90 гипотетических больных, которое не содержит кластеров.

Шаг 2. Применение одного и того же метода кластерного ана-

лиза к обоим наборам данных. Для сравнения результатов кластерного анализа каждый из наборов данных подвергся обработке по итерационному методу *k*-средних (мы воспользовались процедурой BMDPKM). Программа начала свою работу с создания начального разбиения, а затем последовательно применяла метод *k*-средних, описанный в разд. III, для формирования заданного числа кластеров. Поскольку известно, что реальные данные состоят из трех групп, то мы решили рассмотреть только решение, в которое входят три кластера.

Средние, найденные по рандомизированным данным, сильно отличаются от средних, найденных по реальным данным. Кроме того, отметим, что средние этих групп можно упорядочить по возрастанию. Другими словами, один кластер содержит сильно приподнятые профили, другой — умеренно приподнятые, а средние третьего кластера довольно малы. Наш опыт применения кластерного анализа к рандомизированным данным свидетельствует, что многие методы кластеризации формируют такие кластеры из случайных данных, которые можно упорядочить по возрастанию их средних.

Шаг 3. Сравнение кластерных решений. Последний шаг заключается в сравнении выходных статистик кластерных решений, полученных по реальному и искусственному наборам данных. В этом случае мы воспользуемся мерой достоверности, основанной на *F*-отношении, которая имеется в пакете программ BMDPKM. Значения F-отношения, вычисленные с помощью однофакторной ANOVA по кластерам для всех 13 признаков, приводятся ниже:

L	9,4	*Mf*	1,5
F	69,7	*Pa*	63,7
K	10,6	*Pt*	26,4
Hs	47,7	*Sc*	59,3
D	27,6	*Ma*	27,7
Hy	21,1	*Si*	27,9
Pd	38,5		

Обратите внимание, что большинство значений довольно велико. Действительно, за исключением значения признака *Mf*, *F*-отношение принимает значения от 9,4 до 69,7. Если применить тесты значимости к этим 13 признакам, то 12 из них окажутся значимыми. Однако, как было показано выше, такое использование тестов значимости неправомерно.

Следующее множество значений представляет собой соответствующие *F*-отношения трехкластерного решения в случае рандомизированных данных. Поскольку в рандомизированных данных кластеров нет, то эти значения являются одноточечными оценками нулевых значений *F*-отношений. Вообще говоря, значения *F*-отношений трехкластерного решения не меньше значений *F*-отношений реальных данных. Действительно, эти *F*-отношения имеют значения от 11,9 до 77,4 (опять, исключая признак *Mf*):

L	13,7	*K*	52,4
F	22,6	*Hs*	18,8
D	55,7	*Pt*	77,4
Hy	11,9	*Sc*	67,4
Pd	14,9	*Ma*	19,8
Mf	0,1	*Si*	31,2
Pa	36,4		

О чем же говорит результат сравнения? *F*-отношение, вычисленное с помощью программы BMDPKM, дает пользователю представление об однородности кластеров. Когда рассматриваются абсолютные значения первого множества *F*-отношений, они кажутся разумно большими и, по всей видимости, говорят о том, что кластеры в какой-то степени однородны. Однако *F*-отношения для данных, не имеющих кластеров, столь же велики. Это доказывает, что первое множество *F*-отношений недостаточно велико для того, чтобы пользователь мог отвергать нулевую гипотезу об отсутствии кластеров.

Графический вывод программы BMDPKM можно использовать для наглядного представления структуры результатов. На рис. 10 показана схема расположения трех кластеров, представленных в двумерном пространстве основных компонент. На этой схеме очень хорошо видны три кластера. Однако если также изобразить кластеры рандомизированных данных (рис. 11), то три «кластера» кажутся непересекающимися, но не столь плотными, как реальные кластеры. Заметьте, что на схемах между кластерами нет очевидных границ. Вместо этого графическое отображение обоих решений показывает, что кластеры могут быть просто произвольным разбиением полного набора данных. Сравнивая графическое изображение реальных данных с изображением рандомизированных данных, видно, что пользователю будет трудно отбросить нулевую гипотезу об отсутствии кластеров.

Следовательно, по результатам работы программы можно заключить, что решение из трех кластеров соответствует структуре реальных данных и что сформированные кластеры однородны и хорошо разделены. Использование метода Монте-Карло позволяет формализовать проведение сравнительного анализа результатов вычислительных программ кластеризации.

Рассмотрим еще один набор данных MMPI-теста для 90 больных. Этот набор данных был выбран таким образом, чтобы имеющиеся в нем три группы больных (с психозами, неврозами и расстройствами личности) были очень плотными и хорошо выраженными. Вновь выполним три шага: 1) формирование рандомизированного множества данных; 2) проведение кластерного анализа реальных и рандомизированных данных; 3) сравнение результатов. Применялся тот же самый метод кластеризации BMDPKM. Результирующие выходные статистики *F*-отношения показаны ниже:

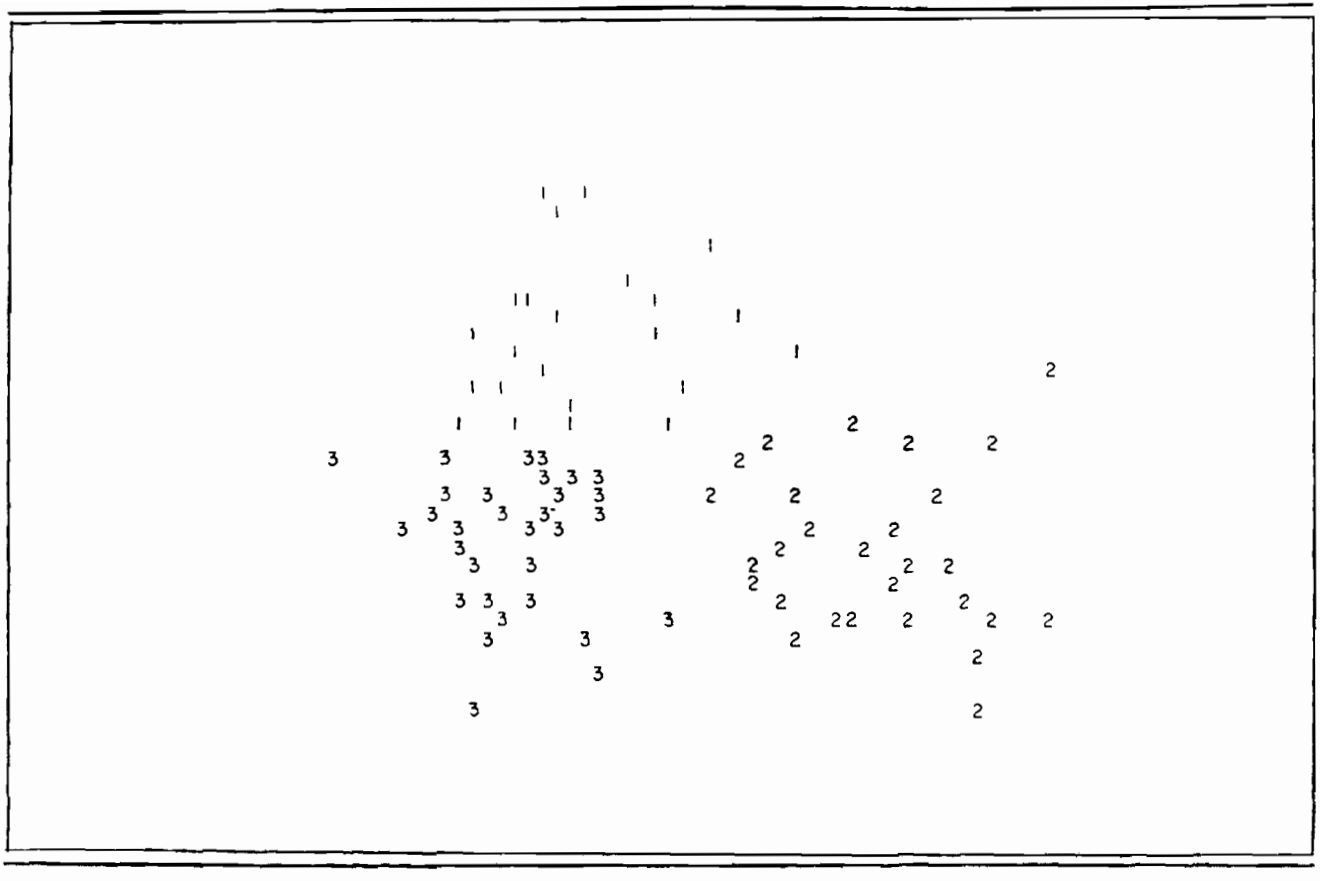

Рис. 10. Схема расположения трех кластеров решения для реальных данных MMPI-теста

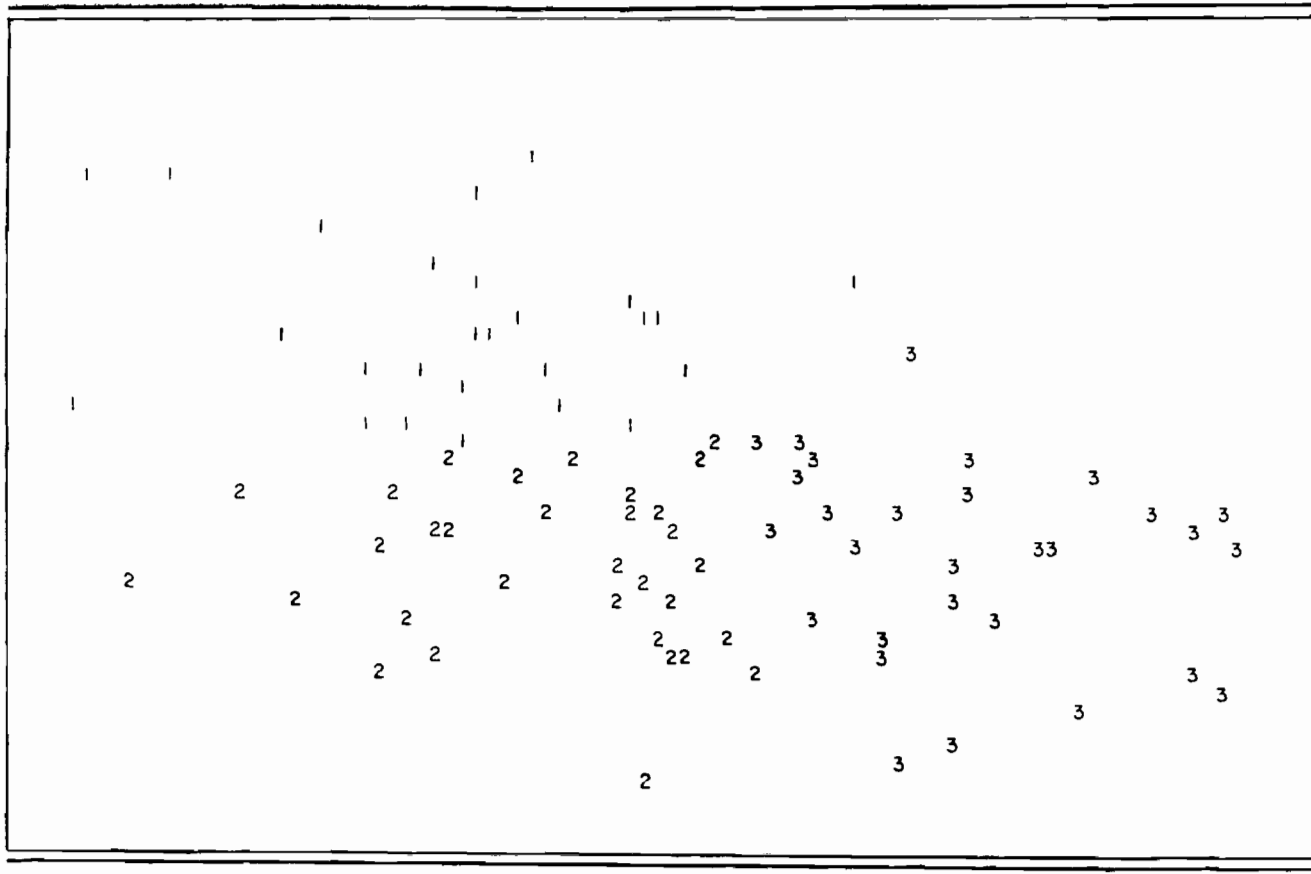

Рис. 11. Схема расположения трех кластеров решения для рандомизированных данных MMPI-теста

	Реальные данные	Рандомизированные данные
L	55,1	19,1
F	895,2	91,4
K	70,2	37,6
Hs	250,9	39,4
D	115,1	37,4
Hy	151,4	25,1
Pd	414,3	28,5
Mf	4,1	3,9
Pa	497,5	75,9
Pt	129,5	72,1
Sc	365,0	91,8
Ma	370,9	30,0
Si	243,7	59,4

Заметьте насколько больше *F*-отношения для реальных данных, чем для рандомизированных. Почти все они являются трехзначными числами, и по любым стандартам они будут казаться очень большими величинами. На рис. 12 и 13 приведены схемы расположения кластеров для реальных и рандомизированных данных соответственно. Обратите внимание, что для реальных данных кластеры очень плотные и между ними существуют четкие границы. У рандомизированных данных такой структуры не отмечается. Хотя большинство процедур обоснования достоверности решений изучены плохо и требуют осторожного обращения, некоторые из них необходимо использовать во всех исследованиях, где применяется кластерный анализ. Читателю, желающему продолжить изучение затронутой здесь темы, предлагаем следующие работы: (Dubes and Jain, 1980; Rohlf, 1974; Skinner and Blashfield, 1982; Chambers and Kleiner, 1982).

V. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ КЛАСТЕРНОГО АНАЛИЗА И ЛИТЕРАТУРА ПО КЛАСТЕРИЗАЦИИ

Программное обеспечение кластерного анализа можно разделить на четыре основные категории: 1) библиотеки подпрограмм и алгоритмов; 2) общие пакеты программ по статистике, содержащие и методы кластеризации; 3) пакеты программ по кластерному анализу и 4) простые программы, реализующие какой-либо вид кластеризации (Blashfield et al., 1982). Поскольку исчерпывающий обзор программного обеспечения кластерного анализа выходит за рамки нашей работы, мы сосредоточим внимание лишь на тех программах и пакетах, которые получили широкое распространение.

БИБЛИОТЕКИ ПОДПРОГРАММ И АЛГОРИТМОВ

В настоящее время доступны три основные библиотеки программ и алгоритмов: книги (Anderberg, 1973; Hartigan, 1975) и

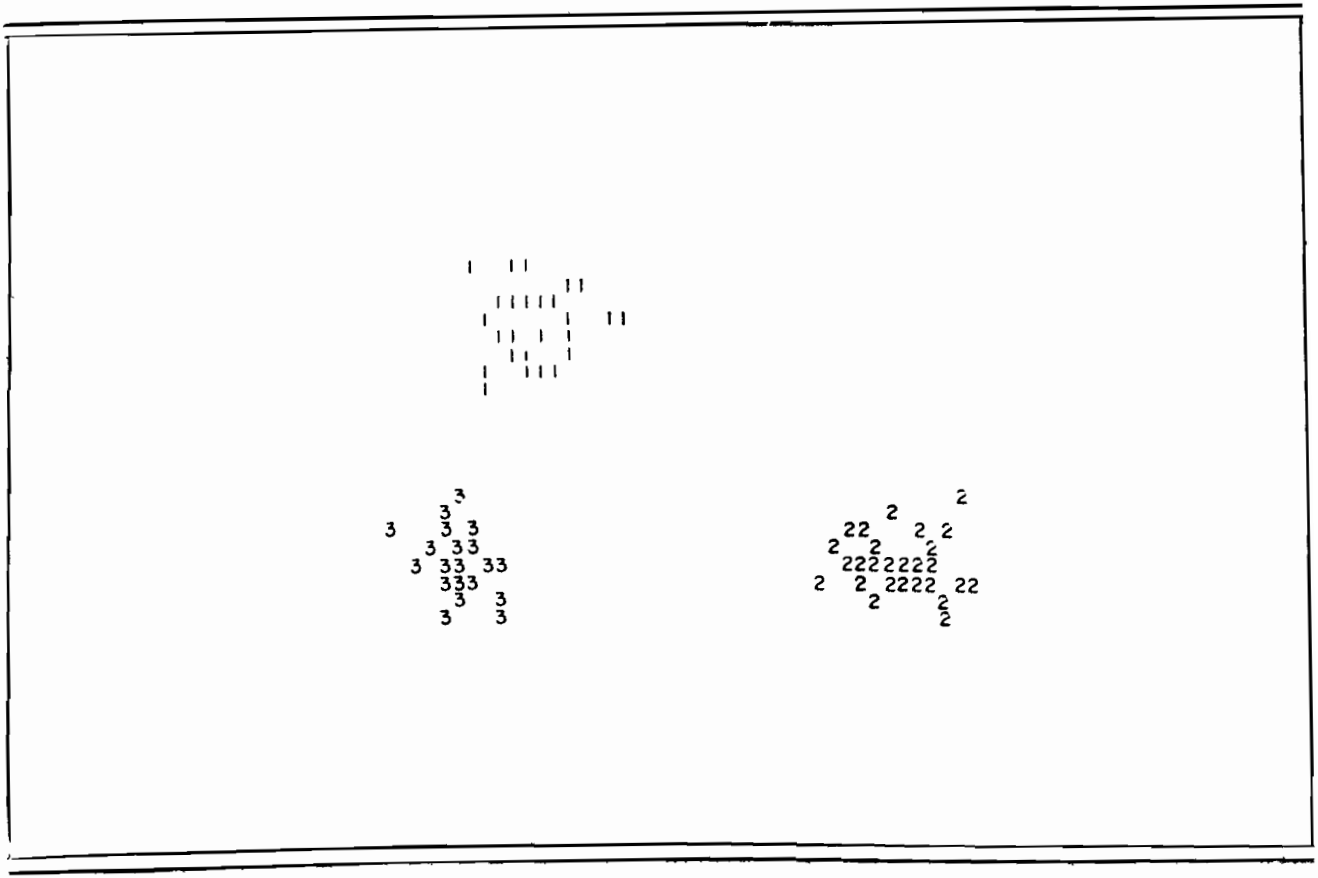

Рис. 12. Схема трехкластерного решения для данных MMPI-теста с плотными кластерами

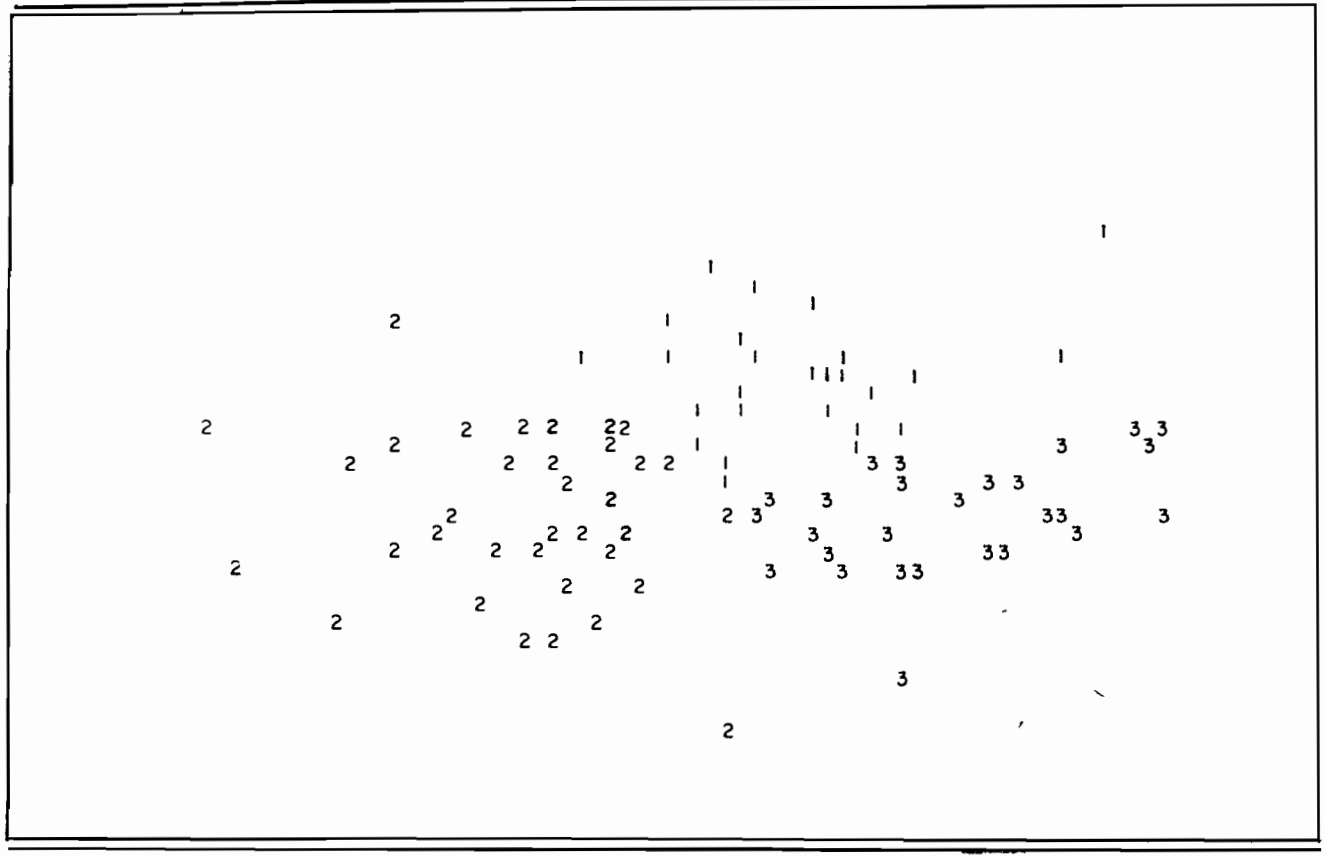

Рис. 13. Схема для рандомизнрованных данных

программы из Международной математической и статистической библиотеки (IMSL, 1980). Поскольку большая часть этого программного обеспечения довольно запутана, пользователь должен применять все возможности языка управления заданиями для редактирования и последующего выполнения программ. Если воспользоваться современным программистским жаргоном, то можно сказать, что это программное обеспечение не очень «дружественно к пользователю». Прежде чем начать работу с программами, пользователь должен быть знаком как с языком управления заданиями вычислительной системы, так и с языком Фортран, который использовался при разработке этих программ. В общем, уровень программной поддержки пользователя очень низок. Алгоритмы Хартигана описаны в отдельном руководстве пользователя (Dallal, 1975), тогда как алгоритмы Андерберга можно найти лишь в его книге. Хотя в документацию собрания IMSL-подпрограмм входят и описания алгоритмов кластеризации, это не облегчает пользование алгоритмами. Несмотря на широкий выбор методов и вспомогательных программ, новичку не рекомендуется пользоваться алгоритмами этой категории до тех пор, пока не появятся обстоятельные руководства.

ПАКЕТЫ СТАТИСТИЧЕСКИХ ПРОГРАММ, СОДЕРЖАЩИЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ

Возможно, наиболее удобным и общедоступным программным обеспечением кластерного анализа являются подпрограммы, содержащиеся в таких пакетах статистических программ, как: BMDP (Dixon, 1983), SAS (SAS Institute, 1982), SPSS (SPSS, 1984). Концепции, заложенные в основу этих пакетов, хорошо известны. Они открывают непрограммистам сравнительно легкий доступ к сложным статистическим методам решения широкого круга исследовательских задач. Помогает пользователю и то, что в пакетах программ используется язык управления заданиями вычислительной системы, позволяющий с минимальными усилиями передавать вычислительной системе запросы пользователя. Эти пакеты программ содержат также разнообразные методы отбора и обработки данных, позволяющие сделать сложный анализ простым и выполнимым. Если же пакет программ содержит метод, представляющий интерес для пользователя, то преимущества применения уже существующих пакетов статистических программ становятся значительными.

За исключением программы BMDP, число различных дополнительных программ кластеризации, содержащихся в большинстве статистических пакетов, очень мало. Например, ранняя версия SAS содержала только один метод кластеризации, а SPSS — ни одного. Однако такое положение существенно изменилось. Для BMDP были разработаны четыре процедуры кластерного анализа: 1) методы одиночной, полной и средней связей для группировки

признаков; 2) методы средней связи (центроидная группировка), одиночной связи и k-ближайших соседей для группировки объектов; 3) блочный метод кластеризации (Hartigan, 1975) для одновременной группировки объектов и признаков; 4) итеративный метод k-средних, образующий разбиения объектов. (Последняя процедура, BMDPKM, была использована в примере, иллюстрирующем применение методики повторных выборок; см. разд. IV.) Процедуры BMDP снабжены хорошими описаниями, имеют понятные распечатки и ими довольно легко пользоваться. Наиболее серьезными недостатками этого пакета программ являются небольшое число иерархических агломеративных методов кластеризации объектов и возможность выбора лишь четырех мер сходства (евклидово расстояние, метрика Миковского, расстояние хи-квадрат и φ-коэффициент).

Во втором пакете статистических программ, SAS, до недавнего времени был лишь один метод кластерного анализа — метод полной связи. Однако в недавнюю версию этого пакета (SAS, 1982) включены существенные добавления, хотя, как ни странно, в нем уже нет метода полной связи. Пакет программ (в рамках процедуры CLUSTER) сейчас содержит метод центра тяжести, метод Уорда и иерархический агломеративный метод средней связи. Евклидово расстояние — все еще единственная используемая мера сходства. В процедуре FASTCLUS был добавлен метод k-средних (центроидный метод группировки Андерберга). И наконец, в пакет был включен факторный метод кластеризации признаков (процедура VARCLUS) В пакете было увеличено число диагностических программ, аналогичных имеющимся в пакете CLUSTAN. Значительный интерес представляет новая процедура остановки при определении числа кластеров — кубический критерий кластеризации. Эта процедура была добавлена в последнюю версию пакета программ, но авторы SAS не опубликовали никаких работ, которые могли бы продемонстрировать ее обоснованность или практичность в прикладных исследованиях.

В программе SPSS в настоящее время нет ни одного метода кластерного анализа. Однако есть новая процедура CLUSTER (Bajaj, 1979), которая, возможно, будет включена в SPSS. Новая процедура содержит 27 мер сходства, большинство из них — коэффициенты ассоциативности. В пакете имеется семь иерархических агломеративных методов (включая методы одиночной, полной и средней связей и метод Уорда), которые может применять пользователь.

ПАКЕТЫ ПРОГРАММ КЛАСТЕРНОГО АНАЛИЗА

С точки зрения серьезного исследователя, пакеты программ кластерного анализа обладают максимальной гибкостью и большими удобствами для пользователя. Они сочетают преимущества общих пакетов статистических программ (интегрированный язык управления, процедуры отбора и обработки данных) с чертами,

представляющими особый интерес для пользователя кластерного анализа (разнообразие методов кластеризации, спецнальные диагностические программы и улучшенная графика). Огромное значение имеет то, что многие из этих пакетов программ содержат малодоступные или даже уникальные методы кластеризации и аналитические процедуры, связанные со специальными задачами и структурами данных.

Наиболее известным из пакетов программ кластерного анализа является программа CLUSTAN. Новейшая редакция CLUSTAN (версия 2.1; Wishart, 1982) включает одиннадцать процедур, которые содержат все семейства методов кластеризации, определенные в разд. III, за исключением вариантов факторного анализа. Это следующие процедуры:

- HIERARCHY — восемь иерархических агломеративных методов
- CENTROID — центроидный иерархический метод
- RELOCATE — итеративный метод k-средних
- MODE — модальный анализ
- DENSITY — улучшенный метод модального анализа
- DIVIDE — монотетический дивизимный метод
- KDEND — метод теории графов Джардайна и Сибсона
- DNDRITE — метод минимального покрывающего дерева
- EUCLID — итеративный метод, использующий нелинейное программирование
- NORMIX — метод разделения многомерных нормальных смесей
- INVARIANT — итеративный метод оптимизации многомерных индикаторов однородности кластеров

Среди других особенностей этой программы — кластерная диагностика и средства проверки обоснованности решений, включающие процедуры RULES и COMPARE, в которых реализованы правила остановки Мойена (1977) и кофенетический коэффициент корреляции Мойена и Уишарта (1980). Процедура CORREL содержит в общей сложности 38 мер сходства, а пакет программ имеет вспомогательную процедуру, позволяющую пользователю определить коэффициент сходства любого вида (DEFINE). С помощью других вспомогательных программ можно управлять кластерной диагностикой или графическим выводом информации.

Есть еще три пакета программ, посвященных кластерному анализу: BCTRY (Tryon and Bailey, 1970), CLUS (Friedman and Rubin, 1967), NTSYS (Rohlf et al., 1974). Из этих трех пакетов NTSYS является наиболее важным, поскольку в него включены методы и идеи, обсуждавшиеся в книге по кластерному анализу и численной таксономии (Sneath and Sokal, 1973). Помимо кластерного анализа, пакет NTSYS содержит несколько многомерных статистических процедур, в том числе многомерное шкалирование и факторный анализ. Пакет программ BCTRY создан на основе книги (Tryon and Bailey, 1970) и включает ряд методов кластеризации, отражающих подход Трайона к факторному анализу.

Последний пакет программ CLUS в настоящее время используется редко, а заинтересованный читатель может найти современную версию этой программы в новейшей редакции CLUSTAN.

ПРОСТЫЕ ПРОГРАММЫ КЛАСТЕРНОГО АНАЛИЗА

Простые программы кластерного анализа именно таковыми и являются. Эти программы написаны главным образом на Фортране. Они обычно реализуют один или два алгоритма кластеризации. Некоторым образом они напоминают подпрограммы первой категории, определенные выше, в том, что они требуют от пользователя знания языка управления заданиями вычислительной системы и языка, на котором написаны программы. Вообще говоря, эти программы почти не имеют средств отладки программ, плохо документированы и выводят мало информации. Однако простые программы важны, потому что они часто используются в определенных отраслях науки, а также лежат в основе алгоритмов, входящих в такие пакеты программ, как SAS, IMSL и OSIRIS. Наиболее известной из этих программ является HGROUP, реализующая метод, объединивший методы одиночной и полной связей (Johnston, 1967) и ISODATA, реализующая гибкий итеративный метод группировки, широко применяемый в технике (Hall and Khanna, 1977).

ЛИТЕРАТУРА ПО КЛАСТЕРНОМУ АНАЛИЗУ

Классификация является основным этапом научного исследования, но характер, методы и цели исследования в любой области науки определяются стоящими перед ней задачами и изучаемыми ею явлениями. Поэтому неудивительно, что кластерный анализ — метод, рекламируемый как «объективное» средство классификации, принимает различные формы и определяется многими, часто противоречащими друг другу способами. Также естественно, что литературу по кластерному анализу можно найти в самых различных журналах (по электротехнике, биологии, библиотечному делу, по психиатрии и т. д.). Необходимо отдавать себе отчет в том, что методы кластерного анализа разрабатываются широким кругом научных дисциплин и что под этим названием собрано большое количество совершенно различных методов.

В разд. I мы описали причины возросшего интереса к кластерному анализу. Одновременно с разработкой новых методов и алгоритмов кластеризации отмечался быстрый рост числа статей, связанных с кластеризацией, во многих областях науки. Но если в начале 60-х годов быстрый рост публикаций был ограничен, в какой-то степени, рамками биологических наук, в конце 60-х — начале 70-х годов кластерный анализ распространился фактически на все области научных исследований. Например, в 1973 г. в 162 журналах, включая *Acta Psyhologica, American Antiguity, Computer Journal, Journal of Biochemistry, Quarterly Journal of Medicine, Journal of Marketing Research, Systematic Zoology* и *Journal*

of Ecology, было опубликовано 292 статьи, которые либо цитировали хотя бы одну из основных работ по кластерному анализу, либо использовали в своих названиях термины «кластерный анализ» или «численная таксономия» (Blashfield and Aldenderfer, 1978). Тематика исследований варьирует от анализа морфологии мумифицированных грызунов в Новой Гвинее до изучения результатов голосования сенаторов США, от анализа поведенческих функций замороженных тараканов при их размораживании до исследования географического распределения некоторых видов лишая в Саскачеване.

Такой взрыв публикаций оказал огромное влияние на развитие и применение кластерного анализа. Но, к сожалению, имеются и негативные стороны. Быстрый рост публикаций по кластерному анализу повлек за собой образование группировок пользователей и как следствие — создание жаргона, используемого лишь группировками, его создавшими (Blashfield and Aldenderfer, 1978; Blashfield, 1980).

О формировании жаргона специалистами в области социальных наук свидетельствует, например, разнообразная терминология, относящаяся к методу Уорда. «Метод Уорда» в литературе называется по-разному. Известны по крайней мере еще четыре его названия: «метод минимальной дисперсии», «метод суммы квадратов ошибок», «иерархическая группировка, минимизирующая trW» и «HGROUP». Первые два названия указывают просто на критерий, оптимум которого определяется в методе Уорда, тогда как третье связано с суммой квадратов ошибок, являющейся монотонным преобразованием следа матрицы W, внутригрупповой ковариационной матрицы. Наконец, широко применяемое название «HGROUP» — это название популярной компьютерной программы, которая реализует метод Уорда (Veldman, 1967).

Образование жаргона мешает развитию междисциплинарных связей, препятствует эффективному сравнению методологии и результатов применения кластерного анализа в различных областях науки, ведет к ненужным затратам усилий (повторное изобретение одних и тех же алгоритмов) и, наконец, не дает новым пользователям глубоко понять выбранные ими методы (Blashfield and aldenderfer, 1978). Например, авторы одного исследования в области социальных наук (Rogers and Linden, 1973) сравнили три различных метода кластеризации, применяя одни и те же данные. Они называли эти методы следующим образом: «иерархическая группировка», «иерархическая кластеризация или HCG» и «кластерный анализ». И ни одно из этих названий не было привычным для методов кластеризации. Начинающий пользователь программ кластерного анализа будет сбит с толку всеми существующими названиями и не сможет связать их с другими описаниями методов кластеризации. Опытные пользователи окажутся в трудном положении при сравнении своих исследований с аналогичными работами. Возможно, мы впадаем в крайность, но жаргон представляет собой серьезную проблему.

В последние годы развитие кластерного анализа несколько замедлилось, судя и по числу публикаций, и по числу дисциплин, где этот метод применяется. Можно сказать, что в настоящее время психология, социология, биология, статистика и некоторые технические дисциплины выходят на стадию консолидации в отношении кластерного анализа.

Количество статей, воспевающих достоинства кластерного анализа, постепенно уменьшается. При этом все чаще появляются работы, в которых на контрольных данных проводится сравнение применимости различных методов кластеризации. В литературе стало уделяться больше внимания и приложениям. Многие исследования направлены на разработку практических мер для проверки обоснованности результатов, полученных с помощью кластерного анализа. Все это свидетельствует о серьезных попытках создать разумную статистическую теорию методов кластеризации.

РЕКОМЕНДАЦИИ ПО СОСТАВЛЕНИЮ ОТЧЕТОВ ОБ ИССЛЕДОВАНИЯХ, ИСПОЛЬЗУЮЩИХ КЛАСТЕРНЫЙ АНАЛИЗ

Цель нашей работы — помочь потенциальному пользователю освоить кластерный анализ и познакомить его с многообразием методов, литературой, программным обеспечением и терминологией. Надеемся, что после ее прочтения, специалисты смогут оценить достоинства и недостатки различных подходов и методов классификации. Здесь рассматриваются те проблемы и способы их устранения, о которых должен знать каждый пользователь программ кластерного анализа. Хотелось бы, однако, предложить несколько рекомендаций, которые вряд ли улучшат качество научных исследований, но все же дадут возможность сравнить результаты исследований, проведенных с помощью кластеризации.

1. *Необходимо давать четкое описание метода кластеризации.* Это поможет избавиться от жаргона в публикациях. Несомненно, одним из стандартов может служить книга Снита и Сокэла (1973). В нашей работе мы следуем их терминологии и рекомендуем ее другим. Название метода должно сопровождаться соответствующими ссылками.

2. *Необходимо четко указывать, какая мера сходства была выбрана (или статистический критерий, если используется итеративный метод).* Как было показано в разд. II и III, выбор меры сходства может сильно повлиять на результат, полученный с помощью кластерного анализа. Если выбор меры сходства не указывается в сообщении, читатель не сможет определить, какое влияние оказал этот выбор на результат кластерного анализа.

3. *Необходимо указывать, какой программой пользовался исследователь.* Блэшфилд (1977) показал, что с помощью разных программ, работавших в полном соответствии с одним и тем же методом кластеризации и идентичными мерами сходства, были получены сильно различающиеся результаты. В данном случае существенно отличались друг от друга формулы, по которым вычислялось евклидово расстояние. Хотя обе программы были вполне корректны, в одной из них не извлекается квадратный корень из выражения. К сожалению, при описании меры сходства в каждой программе употреблялся термин «евклидово расстояние».

4. *Необходимо указать процедуры, которые применялись при определении числа кластеров.* Эта рекомендация важна при попытке повторить исследование. Кроме того, очевидно, что простое утверждение: «Для анализа было выбрано решение, содержащее десять кластеров», никого не удовлетворит, пока не будут указаны причины такого выбора.

5. *Необходимо привести убедительные свидетельства обоснованности решения кластерного анализа.* Это, возможно, основной шаг при использовании кластерного анализа и все же он редко применяется и пользователями программ метода, и потребителями отчетов. Различные методы кластеризации приводят

к разным результатам для одних и тех же данных, но следует доказать их достоверность.

Приложение.

Пример множества данных (данные о захоронениях).

1	C	M	N	1	0	0	1	0	0	0	0
2	C	M	N	0	0	0	1	0	0	0	0
3	C	M	E	1	0	0	1	0	0	1	1
4	C	F	N	1	0	1	0	0	0	0	0
5	C	F	E	0	0	1	0	0	0	1	0
6	C	F	E	1	0	1	0	0	0	1	0
7	T	M	N	1	1	0	1	0	0	0	0
8	T	M	N	0	1	0	1	1	0	0	0
9	T	M	N	1	0	0	1	1	0	0	0
10	T	M	N	1	1	0	1	1	0	0	0
11	T	M	E	1	1	0	1	1	0	1	1
12	T	F	N	0	0	0	0	1	0	0	0
13	T	F	N	1	0	0	0	1	0	0	0
14	T	F	E	1	0	0	0	1	0	1	0
15	A	M	N	1	1	0	1	1	0	0	0
16	A	M	N	0	1	0	1	1	0	0	0
17	A	M	N	1	1	0	1	0	0	0	0
18	A	M	E	1	1	0	1	1	0	1	1
19	A	M	E	1	0	0	1	0	0	1	0
20	A	F	N	0	0	0	0	0	1	0	0
21	A	F	N	1	0	0	0	0	1	0	0
22	A	F	N	0	0	0	0	1	1	0	0
23	A	F	N	1	0	0	0	0	0	0	0
24	A	F	E	1	0	0	0	1	1	1	0
25	A	F	E	1	0	0	0	1	1	1	1

C — ребенок; T — подросток; A — взрослый; M — мужской пол; F — женский пол; N — не элитарный; E — элитарный; 1 — присутствует; 0 — отсутствует.

ПРИМЕЧАНИЯ

1. Как показано в разд. IV, эта «процедура обоснования» на самом деле не обосновывает результаты кластеризации.

2 Собственные значения часто используются в стандартном факторном анализе для того, чтобы дать представление о степени важности факторов и помочь определить количество факторов в данных.

3. Здесь можно считать, что выполнение MANOVA эквивалентно проведению однофакторного дисперсионного анализа для каждого из признаков, использовавшихся при создании кластеров, или проведению дискриминантного анализа

ЛИТЕРАТУРА

ANDERBERG, M. (1973) Cluster Analysis for Applications. New York: Academic Press.

BAILEY, K. (1975) "Cluster Analysis," in D. Heise (ed) Sociological Methodology. San Francisco: Jossey-Bass.

BAJAJ, S.R. (1979) "A preliminary version of subprogram CLUSTER." Applications Division, Northwestern University. (unpublished)

BAYNE, R., J. BEAUCHAMP, C. BEGOVICH, and V. KANE (1980) "Monte Carlo comparisons of selected clustering procedures." Pattern Recognition 12:51-62.

BLASHFIELD, R.K. (1980)"The growth of cluster analysis:Tryon,Ward,and Johnson." Multivariate Behavioral Research 15:439-458.

BLASHFIELD, R.K. (1977) "On the equivalence of four software programs for performing hierarchical cluster analysis." Psychometrika 42:429-431.

BLASHFIELD, R.K. (1976) "Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods." Psychological Bulletin 83:377-388.

BLASHFIELD, R.K. and M. ALDENDERFER (1978a) "Computer programs for performing iterative partitioning cluster analysis." Applied Psychological Measurement 2:533-541.

BLASHFIELD, R.K. (1978b) "The literature on cluster analysis." Multivariate Behavioral Research 13:271-295.

BLASHFIELD, R.K. and L. MOREY (1982) "Cluster analysis software," pp. 245-266 in P. Krishnaiah and L. Kanal (eds.) Handbook of Statistics, vol 2 Amsterdam: North-Holland.

BLASHFIELD, R. and L. MOREY (1980) "A comparison of four clustering methods using MMPI Monte Carlo data." Applied Psychological Measurement 4(1):57-64.

BONNER, R.E. (1964) "On some clustering techniques." I.B.M. Journal of Research and Development 8:22-32.

BURTON, M.L. and A.K. ROMNEY (1975) "A multidimensional representation of role terms." American Ethnologist 2:397-408.

CARMICHAEL, J.W. and P.H.A. SNEATH (1969) "Taxometric maps." Systematic Zoology 18:267-276.

CLIFFORD, H. and W. STEPHENSON (1975) An Introduction to Numerical Taxonomy. New York: Academic Press.

COLE, A.J. and D. WISHART (1970) "An improved algorithm for the Jardine-Sibson method of generating overlapping clusters." Computer Journal 13:156-163.

CORMACK, R. (1971) "A review of classification " Journal of the Royal Statistical Society (Series A) 134:321-367.

CRONBACH, L. and G. GLESER (1953) "Assessing similarity between profiles." Psychological Bulletin 50:456-473.

CZEKANOWSKI, J. (1911) "Objectiv kriterien in der ethnologie." Korrespondenzblatt der Deutschen Geselischaft für Anthropologie, Ethnologie, und Urgeschichte 42:1-5.

DALLAL, G.E. (1975) "A users' guide to J.A. Hartigan's clustering algorithms." Yale University. (unpublished)

DeJONG, G., J. FAULKNER, and R. WARLAND (1976) "Dimensions of religiosity reconsidered: evidence from a cross-cultural study." Social Forces 54:866-889.

DIXON, W. (1983) BMDP Statistical Software. Berkeley: University of California Press.

DRIVER, H.E. (1965) "Survey of numerical classification in anthropology," pp. 304-344 in D.Hymes (ed.) The Use of Computers in Anthropology. The Hague: Mouton.
DUBES, R. and A. JAIN (1980) "Clustering methodologies in exploratory data analysis." Advances in Computers 19:113-228.
EDELBROCK, C. (1979) "Comparing the accuracy of hierarchical clustering algorithms: the problem of classifying everybody." Multivariate Behavioral Research 14:367-384.
EDELBROCK, C. and B. McLAUGHLIN (1979) "Intraclass correlations as metrics for hierarchical cluster analysis: parametric comparisons using the mixture model." Paper presented at the Seventh Annual Meeting, Classification Society.
EVERITT, B. (1980) Cluster Analysis. New York: Halsted.
EVERITT, B. (1979) "Unresolved problems in cluster analysis." Biometrics 35:169-181.
FILSINGER, E., J. FAULKNER, and R. WARLAND (1979) "Empirical taxonomy of religious individuals: an investigation among college students." Sociological Analysis 40:136-146.
FINNEY, J.W. and R.H. MOOS (1979) "Treatment and outcome for empirical subtypes of alcoholic patients." Journal of Consulting and Clinical Psychology 47: 25-38.
FLEISS, J., W. LAWLOR, S. PLATMAN, and R. FIEVE (1971) "On the use of inverted factor analysis for generating typologies." Journal of Abnormal Psychology 77:127-132.
FRIEDMAN, H.P. and J. RUBIN (1967) "On some invariant criteria for grouping data." Journal of the American Statistical Association 62:1159-1178.
GOLDSTEIN, S.G. and J.D. LINDEN (1969) "Multivariate classification of alcoholics by means of the MMPI." Journal of Abnormal Psychology 74:661-669.
GOODALL, D. (1967) "The distribution of the matching coefficient." Biometrics 23:647-656.
GOWER, J.C. (1971) "A general coefficient of similarity and some of its properties." Biometrics 27:857-872.
GOWER, J.C. (1967) "A comparison of some methods of cluster analysis." Biometrics 23:623-637.
GUERTIN, W (1966) "The search for recurring patterns among individual profiles." Educational and Psychological Measurement 26:151-165.
HALL, D.J. and D. KHANNA (1977) "The ISODATA method of computation for the relative perception of similarities and differences in complex and real data," in K. Enslein, A. Ralston, and H.W. Wilf (eds.) Statistical Methods for Digital Computers, vol. 3. New York: John Wiley.
HAMER, R. and J. CUNNINGHAM (1981) "Cluster analyzing profile data confounded with interrater differences: a comparison of profile association measures." Applied Psychological Measurement 5:63-72.
HARTIGAN, J. (1975) Clustering Algorithms. New York: John Wiley.
HARTIGAN, J. (1967) "Representation of similarity matrices by trees." Journal of Statistical Computing and Computer Simulation 4:187-213.
HOLGERSON, M. (1978) "The limited value of cophenetic correlation as a clustering criterion." Pattern Recognition 10:287-295.
IMSL (1980) IMSL Reference Manual Library 1, Ed. 6, vol. 1 and 2. Houston, TX: Author.
JARDINE, N. and R. SIBSON (1971) Mathematical Taxonomy. New York: John Wiley.

JARDINE, N. and R. SIBSON (1968) "The construction of hierarchic and non-hierarchic classifications." Computer Journal 11:117-184.
JOHNSON, S. (1967) "Hierarchical clustering schemes." Psychometrika 38:241-254.
KUIPER, F. and L. FISHER (1975) "A Monte Carlo comparison of six clustering procedures." Biometrics 31:777-783.
LANCE, G. and W. WILLIAMS (1967) "A general theory of classificatory sorting strategies." Computer Journal 9:373-380.
LEE, K. (1979) "Multivariate tests for clusters." Journal of the American Statistical Association 74:708-714.
LING, R. (1975) "An exact probability distribution of the connectivity of random graphs." Journal of Mathematical Psychology 12:90-98.
LORR, M. (1983) Cluster Analysis for Social Sciences. San Francisco: Jossey-Bass.
LORR, M. (1966) Explorations in Typing Psychotics. New York: Pergamon.
LORR, M and B. RADHAKRISHNAN (1967) "A comparison of two methods of cluster analysis." Educational and Psychological Measurement 27:47-53.
MAHALANOBIS, P. (1936) "On the generalized distance in statistics." Proceedings of the National Institute of Science, Calcutta 12:49-55.
MATTHEWS, A. (1979) "Standardization of measures prior to clustering." Biometrics 35:892.
MEZZICH, J. and H. SOLOMON (1980) Taxonomy and Behavioral Science: Comparative Performance of Grouping Methods. New York: Academic Press.
MILLER, G.A. (1969) "A psychological method to investigate verbal concepts." Journal of Mathematical Psychology 6:169-191.
MILLIGAN, G.W. (1981) "A Monte Carlo study of thirty internal criterion measures for cluster analysis." Psychometrika 46:187-199.
MILLIGAN, G.W. (1980) "An examination of the effect of six types of error perturbation of fifteen clustering algorithms." Psychometrika 45:325-342.
MILLIGAN, G.W. and P.O. ISSAC (1980) "The validation of four ultrametric clustering algorithms." Pattern Recognition 12:41-50.
MOJENA, R. (1977) "Hierarchical grouping methods and stopping rules–an evaluation." Computer Journal 20:359-363.
MOJENA, R. and D. WISHART (1980) "Stopping rules for Ward's clustering method," pp. 426-432 in Proceedings of COMPSTAT 1980. Würzburg, West Germany, Physika-Verlag.
OVERALL, J. and C. KLETT (1972) Applied Multivariate Analysis. New York: McGraw-Hill.
PEEBLES, C. (1972) "Monothetic-divisive analysis of Moundville burials." Newsletter of Computer Archaeology 7:1-11.
ROGERS, G. and J.D. LINDEN (1973) "Use of multiple discriminant function analysis in the evaluation of three multivariate grouping techniques." Education and Psychological Measurement 33:787-802.
ROHLF, F.J. (1977) "Computational efficiency of agglomerative clustering algorithms." Technical Report RC-6831. IBM Watson Research Center.
ROHLF, F.J. (1974) "Methods of comparing classifications." Annual Review of Ecology and Systematics 5:101-113.
ROHLF, F.J. (1970) "Adaptive hierarchical clustering schemes." Systematic Zoology 19:58-82.
ROHLF, F.J., J. KISHPAUGH, and D. KIRK (1974) "NT-SYS users manual." State University of New York at Stony Brook

SAS Institute (1982) SAS User's Guide: Statistics. New York: Author.
SCOTT, A.J. and M.J. SYMONS (1971) "Clustering methods based on the likelihood ratio criteria. Biometrics 27:387-397.
SKINNER, H. (1979) "Dimensions and clusters: a hybrid approach to classification." Applied Psychological Measurement 3:327-341.
SKINNER, H. (1978) "Differentiating the contribution of elevation, scatter, and shape in profile similarity." Educational and Psychological Measurement 38:297-308.
SKINNER, H. and R. BLASHFIELD (1982) "Increasing the impact of cluster analysis research: the case of psychiatric classification." Journal of Consulting and Clinical Psychology 50:727-734.
SNEATH, P. (1957) "The application of computers to taxonomy." Journal of General Microbiology 17:201-226.
SNEATH, P. and R. SOKAL (1973) Numerical Taxonomy. San Francisco: W.H. Freeman.
SOKAL, R. and C.D. MICHENER (1958) "A statistical method for evaluating systematic relationships." University of Kansas Scientific Bulletin 38:1409-1438.
SOKAL, R. and F. ROHLF (1962) "The comparison of dendrograms by objective methods." Taxon 11:33-40.
SOKAL, R. and P. SNEATH (1963) Principles of Numerical Taxonomy. San Francisco: W.H. Freeman.
SPSS, Inc. (1984) SPSSX. New York: McGraw-Hill.
TRYON, R. (1939) Cluster Analysis. New York: McGraw-Hill.
TRYON, R. and D.E. BAILEY (1970) Cluster Analysis. New York: McGraw-Hill.
TVERSKY, A. (1977) "Features of similarity." Psychological Review 84:327-352.
VELDMAN, D.J. (1967) "FORTRAN programming for the behavioral sciences." New York: Holt, Rinehart and Winston.
WARD, J. (1963) "Hierarchical grouping to optimize an objective function." Journal of the American Statistical Association 58:236-244.
WHALLON, R. (1972) "A new approach to pottery typology." American Antiquity 37:13-34.
WHALLON, R. (1971) "A computer program for monothetic subdivisive classification in archaeology." Technical Report 1, University of Michigan Museum of Anthropology, Ann Arbor.
WILLIAMS, W. (1971) "Principles of clustering." Annual Review of Ecology and Systematics 2:303-326.
WILLIAMS, W., G.N. LANCE, M.B. DALE, and H.T. CLIFFORD (1971) "Controversy concerning the criteria for taxometric strategies." Computer Journal 14:162-165.
WISHART, D. (1982) "Supplement, CLUSTAN user manual, Third Edition." Program Library Unit, Edinburgh University.
WISHART, D. (1969) "Mode analysis: a generalization of nearest neighbor which reduces chaining effects," pp. 282-311 in A. Cole (ed.) Numerical Taxonomy. London: Academic Press.
WOLFE, J.H. (1971) "A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions." Naval Personnel and Training Research Laboratory Technical Bulletin STB 72-2. San Diego, California.
WOLFE, J.H. (1970) "Pattern clustering by multivariate mixture analysis." Multivariate Behavioral Research 5:329-350.

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

1. Айвазян С. А., Бежаева З. И., Староверов О. В. Классификация многомерных наблюдений. — М.: Статистика, 1974.

2. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1988.

3. Аренс Х., Лёйтер Ю. Многомерный дисперсионный анализ. — М.: Финансы и статистика, 1985.

4. Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. — М.: Мир, 1982.

5. Благуш П. Факторный анализ с обобщениями. — М.: Финансы и статистика, 1988.

6. Болч Б. У., Хуань К. Д. Многомерные статистические методы для экономики.— М.: Статистика, 1979.

7. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. — М.: Прогресс, 1976.

8. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке: Методы планирования эксперимента. — М.: Мир, 1981.

9. Дуда Р., Харт П. Распознавание образов и анализ сцен. — М.: Мир, 1976.

10. Дюран Б., Оделл П. Кластерный анализ. — М.: Статистика, 1977.

11. Елисеева И. М., Рукавишников В. О. Группировка, корреляция, распознавание образов. — М.: Статистика, 1977.

12. Енюков И. С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА.—М.: Финансы и статистика, 1986.

13. Жамбю М. Иерархический кластер-анализ и соответствия. — М.: Финансы и статистика, 1988.

14. Загоруйко Н. Г., Елкина В. Н., Емельянов С. В., Лбов Г. С. Пакет прикладных программ ОТЭКС (для анализа данных). — М.: Финансы и статистика, 1986.

15. Иберла К. Факторный анализ. — М.: Статистика, 1980.

16. Кендалл М. Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. — М.: Наука, 1976.

17. Кильдишев Г. С., Аболенцев Ю. И. Многомерные группировки.— М.: Статистика, 1978.

18. Классификация и кластер/Под ред. Дж. Райзина. — М.: Мир, 1980.

19. Крамер Г. Математические методы статистики. — М.: ИИЛ, 1948.

20. Мандель И. Д. Кластерный анализ — М.: Финансы и статистика, 1988.

21. Математика, в социологии: Моделирование и обработка информации/ А. Г. Аганбегян, Х. Блейлок и др. — М.: Мир, 1977.

22. Миркин Б. Г. Анализ качественных признаков. — М.: Статистика, 1976.

23. Миркин Б. Г. Анализ качественных признаков и структур. — М.: Статистика, 1980.

24. Миркин Б. Г. Группировки в социально-экономических исследованиях. — М.: Финансы и статистика, 1985.

25. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1.— М.: Финансы и статистика, 1981.

26. Мостеллер Ф., Тьюки Дж. Анализ даиных и регрессия. Вып. 2.— М.· Фиинансы и статистика, 1982.

27. Окунь Я. Факторный анализ. — М.: Статистика, 1974.

28. Плюта В. Многомерный сравнительный анализ в экономических исследованиях (методы таксономии и факторного анализа). — М.: Статистика, 1980.

29. Рао С. Р. Линейные статистические методы и их применения. — М.: Наука, 1968.

30. Статистические методы для ЭВМ/Под ред. К. Энслейна, Э. Рэлстона, Г. С. Уилфа. — М.: Наука, 1986.

31. Уилкс С. Математическая статистика. — М.: Наука, 1967.

32. Харман Г. Современный факторный анализ. — М.: Статистика, 1972.

Научное издание

**Джейн-Он Ким, Чарльз У. Мьюллер, Уильм Р. Клекка,
Марк С. Олдендерфер, Роджер К. Блэшфилд**

**ФАКТОРНЫЙ, ДИСКРИМИНАНТНЫЙ
И КЛАСТЕРНЫЙ АНАЛИЗ**

Книга одобрена на объединенном заседании редколлегии серий «Математико-статистические методы за рубежом» и «Библиотечка иностранных книг для экономистов и статистиков» 28.05.87 г.

Зав. редакцией *К. В. Коробов*
Редактор *О. А. Ермилина*
Мл. редакторы *Т. Т. Гришкова, Н. Е. Мендрова*
Худож. редактор *Ю. И. Артюхов*
Техн. редактор *Л. Г. Челышева*
Корректоры *Г. А. Башарина, Г. В. Хлопцева, Н. П. Сперанская*

ИБ № 2303

Сдано в набор 10.01.89. Подписано в печать 06.04.89. Формат $60\times90^{1}/_{16}$. Бум. кн.-журн. Гарнитура «Литературная». Печать высокая. Усл. печ. л. 13,5. Усл. кр.-отт. 13,5. Уч.-изд. л. 15,1. Тираж 7000 экз. Заказ 1385. Цена 2 руб.

Издательство «Финансы и статистика», 101000, Москва, ул Чернышевского, 7.

Областная типография управления издательств, полиграфии и книжной торговли Ивановского облисполкома. 153628. Иваново, ул. Типографская, 6.