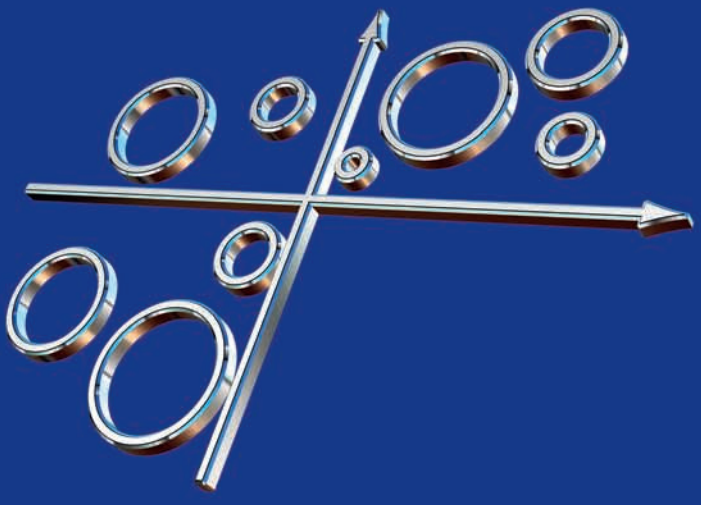
WILEY

Cluster Analysis

5th Edition



Brian S. Everitt • Sabine Landau
Morven Leese • Daniel Stahl

WILEY SERIES IN PROBABILITY AND STATISTICS

Cluster Analysis

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series can be found on http://www.wiley.com/WileyCDA/Section/id-300611.html.

Cluster Analysis

5th Edition

Brian S. Everitt • Sabine Landau
Morven Leese • Daniel Stahl
King's College London, UK

WILEY
A John Wiley and Sons, Ltd., Publication

This edition first published 2011

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Everitt, Brian.
Cluster Analysis / Brian S. Everitt. – 5th ed.
p. cm. – (Wiley series in probability and statistics ; 848)
Summary: "This edition provides a thorough revision of the fourth edition which focuses on the practical aspects of cluster analysis and covers new methodology in terms of longitudinal data and provides examples from bioinformatics. Real life examples are used throughout to demonstrate the application of the theory, and figures are used extensively to illustrate graphical techniques. This book includes an appendix of getting started on cluster analysis using R, as well as a comprehensive and up-to-date bibliography."– Provided by publisher.
Summary: "This edition provides a thorough revision of the fourth edition which focuses on the practical aspects of cluster analysis and covers new methodology in terms of longitudinal data and provides examples from bioinformatics"– Provided by publisher.
Includes bibliographical references and index.
ISBN 978-0-470-74991-3 (hardback)
1. Cluster analysis. I. Title.
QA278.E9 2011
519.5'3–dc22

2010037932

A catalogue record for this book is available from the British Library.

Print ISBN: 978-0-470-74991-3
ePDF ISBN: 978-0-470-97780-4
oBook ISBN: 978-0-470-97781-1
ePub ISBN: 978-0-470-97844-3

Set in 10.25/12pt Times Roman by Thomson Digital, Noida, India

To Joanna, Rachel, Hywel and Dafydd
Brian Everitt

To Premjit
Sabine Landau

To Peter
Morven Leese

To Charmen
Daniel Stahl

Contents

Preface

It is now over 35 years since the first edition of *Cluster Analysis* was published. During this lengthy time period the topic has been in, and occasionally out, of fashion, but the book itself has remained a popular and hopefully useful account of a wide range of numerical methods for exploring multivariate data with a view to uncovering or discovering groups or clusters of homogeneous observations. Such clustering techniques have been employed in a remarkable number of different disciplines. In psychiatry the techniques have been used to refine existing diagnostic categories. In archaeology clustering has been used to investigate the relationship between various types of artefacts. In market research, methods of cluster analysis have been applied to produce groups of consumers with different purchasing patterns. And in the first decade of the 21st century cluster analysis is of considerable interest and importance in the new field of bioinformatics, where it has been used to identify groups of genes with similar patterns of expression with the aim of helping to answer questions of how gene expression is affected by various diseases and which genes are responsible for specific hereditary diseases.

In this fifth edition of *Cluster Analysis,* new material dealing with recent developments and applications, particularly in bioinformatics, has been added to each chapter. Chapter 6, dealing with finite mixture models as the basis of clustering, has been completely rewritten to take account of new work in the area, and a new chapter, Chapter 7, deals with the application of mixture models to structured data, for example repeated measures data. And, of course, a very important difference between this fifth edition and the previous edition is the addition of an extra author, Dr Daniel Stahl.

Like the previous four editions we hope that this book will continue to provide a readable, relatively low-tech introduction to clustering and its possibilities and limitations for research workers in a variety of disciplines, for applied statisticians and for graduate students in statistics and related subjects.

Brian S. Everitt
Sabine Landau
Morven Leese
Daniel Stahl
London

Acknowledgement

We owe a great debt to our copy editor, Clare Lendrem, for her excellent work on the manuscript of our book.

1

An introduction to classification and clustering

1.1 Introduction

> An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand.
>
> Steven Pinker, *How the Mind Works*, 1997.

One of the most basic abilities of living creatures involves the grouping of similar objects to produce a classification. The idea of sorting similar things into categories is clearly a primitive one since early man, for example, must have been able to realize that many individual objects shared certain properties such as being edible, or poisonous, or ferocious and so on.

Classification, in its widest sense, is needed for the development of language, which consists of words which help us to recognize and discuss the different types of events, objects and people we encounter. Each noun in a language, for example, is essentially a label used to describe a class of things which have striking features in common; thus animals are named as cats, dogs, horses, etc., and such a name collects individuals into groups. Naming and classifying are essentially synonymous.

As well as being a basic human conceptual activity, classification is also fundamental to most branches of science. In biology for example, classification of organisms has been a preoccupation since the very first biological investigations. Aristotle built up an elaborate system for classifying the species of the animal

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

kingdom, which began by dividing animals into two main groups, those having red blood (corresponding roughly to our own vertebrates), and those lacking it (the invertebrates). He further subdivided these two groups according to the way in which the young are produced, whether alive, in eggs, as pupae and so on.

Following Aristotle, Theophrastos wrote the first fundamental accounts of the structure and classification of plants. The resulting books were so fully documented, so profound and so all-embracing in their scope that they provided the groundwork of biological research for many centuries. They were superseded only in the 17th and 18th centuries, when the great European explorers, by opening the rest of the world to inquiring travellers, created the occasion for a second, similar programme of research and collection, under the direction of the Swedish naturalist, Linnaeus. In 1737, Carl von Linné published his work *Genera Plantarum*, from which the following quotation is taken:

> All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar. The greater the number of natural distinctions this method comprehends the clearer becomes our idea of things. The more numerous the objects which employ our attention the more difficult it becomes to form such a method and the more necessary.
>
> For we must not join in the same genus the horse and the swine, though both species had been one hoof'd nor separate in different genera the goat, the reindeer and the elk, tho' they differ in the form of their horns. We ought therefore by attentive and diligent observation to determine the limits of the genera, since they cannot be determined *a priori*. This is the great work, the important labour, for should the genera be confused, all would be confusion.

In biology, the theory and practice of classifying organisms is generally known as *taxonomy*. Initially, taxonomy in its widest sense was perhaps more of an art than a scientific method, but eventually less subjective techniques were developed largely by Adanson (1727–1806), who is credited by Sokal and Sneath (1963) with the introduction of the *polythetic* type of system into biology, in which classifications are based on many characteristics of the objects being studied, as opposed to *monothetic* systems, which use a single characteristic to produce a classification.

The classification of animals and plants has clearly played an important role in the fields of biology and zoology, particularly as a basis for Darwin's theory of evolution. But classification has also played a central role in the developments of theories in other fields of science. The classification of the elements in the periodic table for example, produced by Mendeleyev in the 1860s, has had a profound impact on the understanding of the structure of the atom. Again, in astronomy, the classification of stars into *dwarf* stars and *giant* stars using the Hertzsprung–Russell plot of temperature against luminosity (Figure 1.1) has strongly affected theories of stellar evolution.

Classification may involve people, animals, chemical elements, stars, etc., as the entities to be grouped. In this text we shall generally use the term *object* to cover all such possibilities.

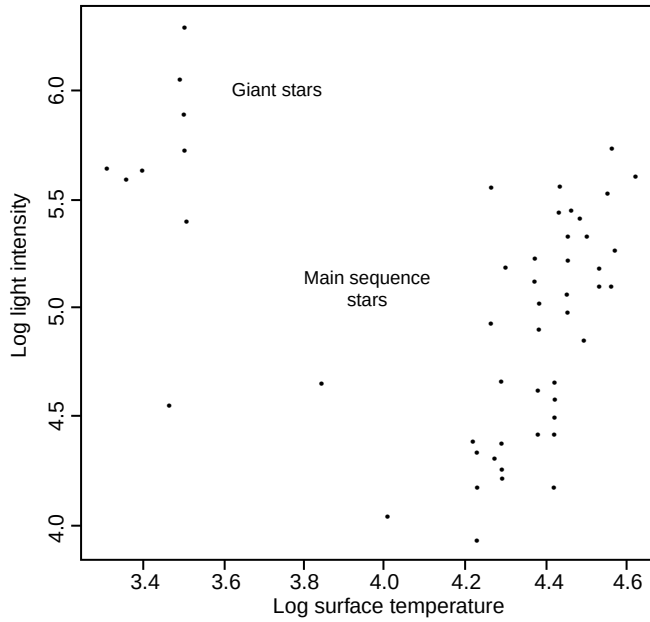


Figure 1.1 *Hertzsprung–Russell plot of temperature against luminosity.*

1.2 Reasons for classifying

At one level, a classification scheme may simply represent a convenient method for organizing a large data set so that it can be understood more easily and information retrieved more efficiently. If the data can validly be summarized by a small number of groups of objects, then the group labels may provide a very concise description of patterns of similarities and differences in the data. In market research, for example, it may be useful to group a large number of respondents according to their preferences for particular products. This may help to identify a 'niche product' for a particular type of consumer. The need to summarize data sets in this way is increasingly important because of the growing number of large databases now available in many areas of science, and the exploration of such databases using cluster analysis and other multivariate analysis techniques is now often called *data mining*. In the 21st century, data mining has become of particular interest for investigating material on the World Wide Web, where the aim is to extract useful information or knowledge from web page contents (see, Liu, 2007 for more details).

In many applications, however, investigators may be looking for a classification which, in addition to providing a useful summary of the data, also serves some more fundamental purpose. Medicine provides a good example. To understand and treat disease it has to be classified, and in general the classification will have two main aims. The first will be *prediction* – separating diseases that require different

treatments. The second will be to provide a basis for research into *aetiology* – the causes of different types of disease. It is these two aims that a clinician has in mind when she makes a diagnosis.

It is almost always the case that a variety of alternative classifications exist for the same set of objects. Human beings, for example, may be classified with respect to *economic status* into groups such as *lower class*, *middle class* and *upper class*; alternatively they might be classified by annual consumption of alcohol into *low*, *medium* and *high*. Clearly such different classifications may not collect the same individuals into groups. Some classifications are, however, more likely to be of general use than others, a point well-made by Needham (1965) in discussing the classification of humans into men and women:

> The usefulness of this classification does not begin and end with all that can, in one sense, be strictly inferred from it – namely a statement about sexual organs. It is a very useful classification because classing a person as a man or woman conveys a great deal more information, about probable relative size, strength, certain types of dexterity and so on. When we say that persons in class *man* are more suitable than persons in class *woman* for certain tasks and conversely, we are only incidentally making a remark about sex, our primary concern being with strength, endurance etc. The point is that we have been able to use a classification of persons which conveys information on many properties. On the contrary a classification of persons into those with hair on their forearms between $\frac{3}{16}$ and $\frac{1}{4}$ inch long and those without, though it may serve some particular use, is certainly of no general use, for imputing membership in the former class to a person conveys information in this property alone. Put another way, there are no known properties which divide up a set of people in a similar manner.

A similar point can be made in respect of the classification of books based on subject matter and their classification based on the colour of the book's binding. The former, with classes such as *dictionaries*, *novels*, *biographies,* etc., will be of far wider use than the latter with classes such as *green*, *blue*, *red,* etc. The reason why the first is more useful than the second is clear; the subject matter classification indicates more of a book's characteristics than the latter.

So it should be remembered that in general a classification of a set of objects is not like a scientific theory and should perhaps be judged largely on its usefulness, rather than in terms of whether it is 'true' or 'false'.

1.3 Numerical methods of classification – cluster analysis

Numerical techniques for deriving classifications originated largely in the natural sciences such as biology and zoology in an effort to rid taxonomy of its traditionally subjective nature. The aim was to provide *objective* and *stable* classifications. Objective in the sense that the analysis of the same set of organisms by the same sequence of numerical methods produces the same classification; stable in that the

classification remains the same under a wide variety of additions of organisms or of new characteristics describing them.

A number of names have been applied to these numerical methods depending largely on the area of application. *Numerical taxonomy* is generally used in biology. In psychology the term *Q analysis* is sometimes employed. In the artificial intelligence literature *unsupervised pattern recognition* is the favoured label, and market researchers often talk about *segmentation*. But nowadays *cluster analysis* is probably the preferred generic term for procedures which seek to uncover groups in data.

In most applications of cluster analysis a *partition* of the data is sought, in which each individual or object belongs to a single cluster, and the complete set of clusters contains all individuals. In some circumstances, however, overlapping clusters may provide a more acceptable solution. It must also be remembered that one acceptable answer from a cluster analysis is that no grouping of the data is justified.

The basic data for most applications of cluster analysis is the usual $n \times p$ multivariate data matrix, $\mathbf{X}$, containing the variable values describing each object to be clustered; that is,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}.$$

The entry x_{ij} in $\mathbf{X}$ gives the value of the jth variable on object i. Such a matrix is often termed 'two-mode', indicating that the rows and columns correspond to different things.

The variables in $\mathbf{X}$ may often be a mixture of continuous, ordinal and/or categorical, and often some entries will be missing. Mixed variables and missing values may complicate the clustering of data, as we shall see in later chapters. And in some applications, the rows of the matrix $\mathbf{X}$ may contain *repeated measures* of the *same* variable but under, for example, different conditions, or at different times, or at a number of spatial positions, etc. A simple example in the time domain is provided by measurements of, say, the heights of children each month for several years. Such *structured data* are of a special nature in that all variables are measured on the same scale, and the cluster analysis of structured data may require different approaches from the clustering of unstructured data, as we will see in Chapter 3 and in Chapter 7.

Some cluster analysis techniques begin by converting the matrix $\mathbf{X}$ into an $n \times n$ matrix of inter-object *similarities*, *dissimilarities* or *distances* (a general term is *proximity*), a procedure to be discussed in detail in Chapter 3. (Such matrices may be designated 'one-mode', indicating that their rows and columns index the same thing.) But in some applications the inter-object similarity or dissimilarity matrix may arise directly, particularly in experiments where people are asked to judge the perceived similarity or dissimilarity of a set of stimuli or objects of interest. As an

Table 1.1 Dissimilarity data for all pairs of 10 colas for 2 subjects.

Subject 1

	Cola number									
	1	2	3	4	5	6	7	8	9	10
1	0									
2	16	0								
3	81	47	0							
4	56	32	71	0						
5	87	68	44	71	0					
6	60	35	21	98	34	0				
7	84	94	98	57	99	99	0			
8	50	87	79	73	19	92	45	0		
9	99	25	53	98	52	17	99	84	0	
10	16	92	90	83	79	44	24	18	98	0

Subject 2

	Cola number									
	1	2	3	4	5	6	7	8	9	10
1	0									
2	20	0								
3	75	35	0							
4	60	31	80	0						
5	80	70	37	70	0					
6	55	40	20	89	30	0				
7	80	90	90	55	87	88	0			
8	45	80	77	75	25	86	40	0		
9	87	35	50	88	60	10	98	83	0	
10	12	90	96	89	75	40	27	14	90	0

example, Table 1.1 shows judgements about various brands of cola made by two subjects, using a visual analogue scale with anchor points 'some' (having a score of 0) and 'different' (having a score of 100). In this example the resulting rating for a pair of colas is a dissimilarity – low values indicate that the two colas are regarded as alike and vice versa. A similarity measure would have been obtained had the anchor points been reversed, although similarities are usually scaled to lie in the interval [0,1], as we shall see in Chapter 3.

In this text our main interest will centre on clustering the objects which define the rows of the data matrix **X**. There is, however, no fundamental reason why some clustering techniques could not be applied to the columns of **X** to cluster the variables, perhaps as an alternative to some form of *factor analysis* (see Everitt and Dunn, 2001). This issue of clustering variables will be taken up briefly in Chapter 8.

Cluster analysis is essentially about *discovering* groups in data, and clustering methods should not be confused with *discrimination* and *assignment* methods (in the artificial intelligence world the term *supervised learning* is used), where the groups are known *a priori* and the aim of the analysis is to construct rules for classifying new individuals into one or other of the known groups. A readable account of such methods is given in Hand (1981). More details of recently developed techniques are available in McLachlan (2004).

1.4 What is a cluster?

Up to this point the terms cluster, group and class have been used in an entirely intuitive manner without any attempt at formal definition. In fact it turns out that formal definition is not only difficult but may even be misplaced. Bonner (1964), for example, has suggested that the ultimate criterion for evaluating the meaning of such terms is the value judgement of the user. If using a term such as 'cluster' produces an answer of value to the investigator, that is all that is required.

Bonner has a point, but his argument is not entirely convincing, and many authors, for example Cormack (1971) and Gordon (1999), attempt to define just what a cluster is in terms of internal cohesion – *homogeneity* – and external isolation – *separation*. Such properties can be illustrated, informally at least, with a diagram such as Figure 1.2. The 'clusters' present in this figure will be clear to most observers without attempting an explicit formal definition of the term. Indeed, the example indicates that no single definition is likely to be sufficient for all situations. This may explain why attempts to make the concepts of homogeneity and separation mathematically precise in terms of explicit numerical indices have led to numerous and diverse criteria.

It is not entirely clear how a 'cluster' is recognized when displayed in the plane, but one feature of the recognition process would appear to involve assessment of the relative distances between points. How human observers draw perceptually coherent clusters out of fields of 'dots' will be considered briefly in Chapter 2.

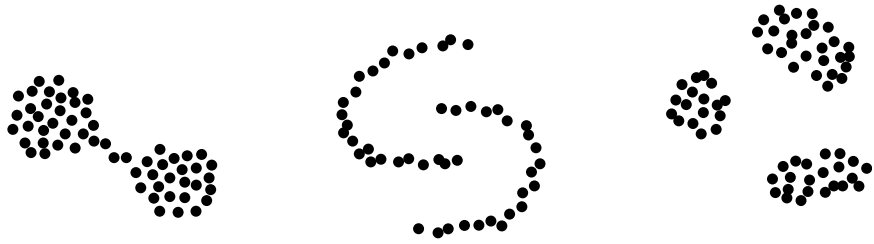


Figure 1.2 *Clusters with internal cohesion and/or external solution. (Reproduced with permission of CRC Press from Gordon, 1980.)*

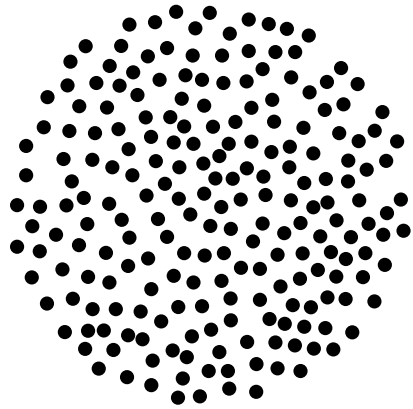


Figure 1.3 *Data containing no 'natural' clusters. (Reproduced with permission of CRC Press from Gordon, 1980.)*

A further set of two-dimensional data is plotted in Figure 1.3. Here most observers would conclude that there is no 'natural' cluster structure, simply a single homogeneous collection of points. Ideally, then, one might expect a method of cluster analysis applied to such data to come to a similar conclusion. As will be seen later, this may not be the case, and many (most) methods of cluster analysis *will* divide the type of data seen in Figure 1.3 into 'groups'. Often the process of dividing a homogeneous data set into different parts is referred to as *dissection*, and such a procedure may be useful in specific circumstances. If, for example, the points in Figure 1.3 represented the geographical locations of houses in a town, dissection might be a useful way of dividing the town up into compact postal districts which contain comparable numbers of houses – see Figure 1.4. (This example was suggested by Gordon, 1980.) The problem is, of course, that since in most cases

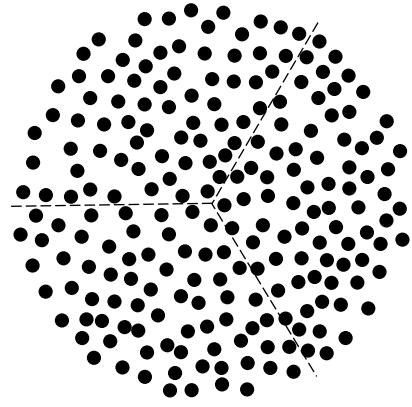


Figure 1.4 *Dissection of data in Figure 1.3 (Reproduced with permission of CRC Press from Gordon, 1980.)*

the investigator does not know *a priori* the structure of the data (cluster analysis is, after all, intended to help to uncover any structure), there is a danger of interpreting *all* clustering solutions in terms of the existence of distinct (natural) clusters. The investigator may then conveniently 'ignore' the possibility that the classification produced by a cluster analysis is an artefact of the method and that actually she is *imposing* a structure on her data rather than discovering something about the actual structure. This is a very real problem in the application of clustering techniques, and one which will be the subject of further discussion in later chapters.

1.5 Examples of the use of clustering

The general problem which cluster analysis addresses appears in many disciplines: biology, botany, medicine, psychology, geography, marketing, image processing, psychiatry, archaeology, etc. Here we describe briefly a number of applications of cluster analysis reported in some of these disciplines. Several of these applications will be described more fully in later chapters, as will a variety of other applications not mentioned below.

1.5.1 Market research

Dividing customers into homogeneous groups is one of the basic strategies of marketing. A market researcher may, for example, ask how to group consumers who seek similar benefits from a product so he or she can communicate with them better. Or a market analyst may be interested in grouping financial characteristics of companies so as to be able to relate them to their stock market performance.

An early specific example of the use of cluster analysis in market research is given in Green *et al.* (1967). A large number of cities were available that could be used as test markets but, due to economic factors, testing had to be restricted to only a small number of these. Cluster analysis was used to classify the cities into a small number of groups on the basis of 14 variables including city size, newspaper circulation and per capita income. Because cities within a group could be expected to be very similar to each other, choosing one city from each group was used as a means of selecting the test markets.

Another application of cluster analysis in market research is described in Chakrapani (2004). A car manufacturer believes that buying a sports car is not solely based on one's means or on one's age but it is more a lifestyle decision, with sports car buyers having a pattern of lifestyle that is different from those who do not buy sports cars. Consequently, the manufacturer employs cluster analysis to try to identify people with a lifestyle most associated with buying sports cars, to create a focused marketing campaign.

1.5.2 Astronomy

Large multivariate astronomical data bases are frequently suspected of containing relatively distinct groups of objects which must be distinguished from each other.

Astronomers want to know how many distinct classes of, for example, stars there are on the basis of some statistical criterion. The typical scientific questions posed are 'How many statistically distinct classes of objects are in this data set and which objects are to be assigned to which classes? Are previously unknown classes of objects present?' Cluster analysis can be used to classify astronomical objects, and can often help astronomers find unusual objects within a flood of data. Examples include discoveries of high-redshift quasars, type 2 quasars (highly luminous, active galactic nuclei, whose centres are obscured by gas and dust), and brown dwarfs.

One specific example is the study reported by Faúndez-Abans *et al.* (1996), who applied a clustering technique due to Ward (1963) (see Chapter 4) to data on the chemical composition of 192 planetary nebulae. Six groups were identified which were similar in many respects to a previously used classification of such objects, but which also showed interesting differences.

A second astronomical example comes from Celeux and Govaert (1992), who apply normal mixture models (see Chapter 6) to stellar data consisting of a population of 2370 stars described by their velocities towards the galactic centre and towards the galactic rotation. Using a three-cluster model, they find a large-size, small-volume cluster, and two small-size, large-volume clusters.

For a fuller account of the use of cluster analysis in astronomy see Babu and Feigelson (1996).

1.5.3 Psychiatry

Diseases of the mind are more elusive than diseases of the body, and there has been much interest in psychiatry in using cluster analysis techniques to refine or even redefine current diagnostic categories. Much of this work has involved depressed patients, where interest primarily centres on the question of the existence of *endogenous* and *neurotic* subtypes. Pilowsky *et al.* (1969), for example, using a method described in Wallace and Boulton (1968), clustered 200 patients on the basis of their responses to a depression questionnaire, together with information about their mental state, sex, age and length of illness. (Notice once again the different types of variable involved.) One of the clusters produced was identified with endogenous depression. A similar study by Paykel (1971), using 165 patients and a clustering method due to Friedman and Rubin (1967) (see Chapter 5), indicated four groups, one of which was clearly psychotic depression. A general review of the classification of depression is given in Farmer *et al.* (1983).

Cluster analysis has also been used to find a classification of individuals who attempt suicide, which might form the basis for studies into the causes and treatment of the problem. Paykel and Rassaby (1978), for example, studied 236 suicide attempters presenting at the main emergency service of a city in the USA. From the pool of available variables, 14 were selected as particularly relevant to classification and used in the analysis. These included age, number of previous suicide attempts, severity of depression and hostility, plus a number of demographic characteristics. A number of cluster methods, for example Ward's method, were applied to the data,

and a classification with three groups was considered the most useful. The general characteristics of the groups found were as follows:

- Group 1: Patients take overdoses, on the whole showing less risk to life, less psychiatric disturbance, and more evidence of interpersonal rather than self-destructive motivation.
- Group 2: Patients in this group made more severe attempts, with more self-destructive motivation, by more violent methods than overdoses.
- Group 3: Patients in this group had a previous history of many attempts and gestures, their recent attempt was relatively mild, and they were overly hostile, engendering reciprocal hostility in the psychiatrist treating them.

A further application of cluster analysis to parasuicide is described in Kurtz *et al.* (1987), and Ellis *et al.* (1996) also investigated the use of cluster analysis on suicidal psychotic outpatients, using *average linkage clustering* (see Chapter 4). They identified four groups which were labelled as follows:

- negativistic/avoidant/schizoid
- avoidant/dependent/negativistic
- antisocial
- histrionic/narcissistic.

And yet another psychiatric example is provided by the controversy over how best to classify eating disorders in which there is recurrent binge eating. Hay *et al.* (1996) investigated the problem by applying Ward's method of cluster analysis to 250 young women each described by five sub-scales derived from the 12th edition of the Eating Disorder Examination (Fairburn and Cooper, 1993). Four subgroups were found:

- objective or subjective bulimic episodes and vomiting or laxative misuse;
- objective bulimic episodes and low levels of vomiting or laxative misuse;
- subjective bulimic episodes and low levels of vomiting or laxative misuse;
- heterogeneous in nature.

1.5.4 Weather classification

Vast amounts of data are collected on the weather worldwide. Exploring such data using cluster analysis may provide new insights into climatological and environmental trends that have both scientific and practical significance. Littmann (2000), for example, applies cluster analysis to the daily occurrences of several surface pressures for weather in the Mediterranean basin, and finds 20 groups that explain rainfall variance in the core Mediterranean regions. And Liu and George (2005) use fuzzy k-means clustering (see Chapter 8) to account for the spatiotemporal nature of weather data in the South Central USA. One further example is provided by Huth

et al. (1993), who analyse daily weather data in winter months (December–February) at Prague Clementinum. Daily weather was characterized by eight variables such as daily mean temperature, relative humidity and wind speed. Average linkage (see Chapter 4) was used to group the data into days with similar weather conditions.

1.5.5 Archaeology

In archaeology, the classification of artefacts can help in uncovering their different uses, the periods they were in use and which populations they were used by. Similarly, the study of fossilized material can help to reveal how prehistoric societies lived. An early example of the cluster analysis of artefacts is given in Hodson *et al.* (1966), who applied single linkage and average linkage clustering (see Chapter 4) to brooches from the Iron Age and found classifications of demonstrable archaeological significance. Another example is given in Hodson (1971), who used a *k-means* clustering technique (see Chapter 5) to construct a taxonomy of hand axes found in the British Isles. Variables used to describe each of the axes included length, breadth and pointedness at the tip. The analysis resulted in two clusters, one of which contained thin, small axes and the other thick, large axes, with axes in the two groups probably being used for different purposes. A third example of clustering artefacts is that given in Mallory-Greenough and Greenough (1998), who again use single linkage and average linkage clustering on trace-element concentrations determined by inductively coupled plasma mass spectrometry in Ancient Egyptian pottery. They find that three groups of Nile pottery from Mendes and Karnak (Akhenatan Temple Project excavations) can be distinguished using lead, lithium, ytterbium and hafnium data.

An example of the clustering of fossilized material is given in Sutton and Reinhard (1995), who report a cluster analysis of 155 coprolites from Antelope House, a prehistoric Anasazi site in Canyon de Chelly, Arizona. The analysis revealed three primary clusters: whole kernel maize, milled maize, and nonmaize, which the authors interpreted as representing seasonal- and preference-related cuisine.

1.5.6 Bioinformatics and genetics

The past decade has been witness to a tremendous growth in *Bioinformatics,* which is the coming together of molecular biology, computer science, mathematics and statistics. Such growth has been accelerated by the ever-expanding genomic and proteomic databases, which are themselves the result of rapid technological advances in DNA sequencing, gene expression measurement and macromolecular structure determination. Statistics and statisticians have played their most important role in this scientific revolution in the study of gene expression. Genes within each cell's DNA provide the templates for building the proteins necessary for many of the structural and biochemical process that take place in each and every one of us. But although most cells in human beings contain the full complement of genes that

make up the entire human genome, genes are selectively expressed in each cell depending on the type of cell and tissue and general conditions both within and outside the cell. Molecular biology techniques have made it clear that major events in the life of a cell are regulated by factors that alter the expression of the gene. Attempting to understand how expression of genes is selectively controlled is now a major activity in modern biological research. DNA microarrays (Cortese, 2000) are a revolutionary breakthrough in experimental molecular biology that have the ability to simultaneously study thousands of genes under a multitude of conditions and provide a mass of data for the researcher. These new types of data share a common characteristic, namely that the number of variables (p) greatly exceeds the number of observations (n); such data is generally labelled *high dimensional*. Many classical statistical methods cannot be applied to high-dimensional data without substantial modifications. But cluster analysis can be used to identify groups of genes with similar patterns of expression, and this can help provide answers to questions of how gene expression is affected by various diseases and which genes are responsible for specific hereditary diseases. For example, Selinski and Ickstadt (2008) use cluster analysis of single-nucleotide polymorphisms to detect differences between diseased and control individuals in case-control studies, and Eisen *et al*. (1998) use clustering of genome-wide expression data to identify cancer subtypes associated with survival; Witten and Tibshirani (2010) describe a similar application of clustering to renal cell carcinoma data. And Kerr and Churchill (2001) investigate the problem of making statistical inferences from clustering tools applied to gene expression data.

1.6 Summary

Cluster analysis techniques are concerned with exploring data sets to assess whether or not they can be summarized meaningfully in terms of a relatively small number of groups or clusters of objects or individuals which resemble each other and which are different in some respects from individuals in other clusters. A vast variety of clustering methods have been developed over the last four decades or so, and to make discussion of them simpler we have devoted later chapters to describing particular classes of techniques – cluster analysis clustered, so-to-speak! But before looking at these formal methods of cluster analysis, we will, in Chapter 2, examine some graphical approaches which may help in uncovering cluster structure, and then in Chapter 3 consider the measurement of similarity, dissimilarity and distance, which is central to many clustering techniques. Finally, in Chapter 9 we will confront the difficult problem of cluster validation, and try to give potential users of cluster analysis some useful hints as to how to avoid being misled by artefactual solutions.

2

Detecting clusters graphically

2.1 Introduction

Graphical views of multivariate data are important in all aspects of their analysis. In general terms, graphical displays of multivariate data can provide insights into the structure of the data, and in particular, from the point of view of this book, they can be useful for suggesting that the data may contain clusters and consequently that some formal method of cluster analysis might usefully be applied to the data. The usefulness of graphical displays in this context arises from the power of the human visual system in detecting patterns, and a fascinating account of how human observers draw perceptually coherent clusters out of fields of dots is given in Feldman (1995). However, the following caveat from the late Carl Sagan should be kept in mind.

> Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent.

In this chapter we describe a number of relatively simple, *static* graphical techniques that are often useful for providing evidence for or against possible cluster structure in the data. Most of the methods are based on an examination of either *direct* univariate or bivariate marginal plots of the multivariate data (i.e. plots obtained using the original variables), or *indirect* one- or two-dimensional 'views' of the data obtained from the application to the data of a suitable dimension-reduction technique, for example principal components analysis. For an account of *dynamic* graphical methods that may help in uncovering clusters in high-dimensional data, see Cook and Swayne (2007).

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

2.2 Detecting clusters with univariate and bivariate plots of data

It is generally argued that a unimodal distribution corresponds to a homogeneous, unclustered population and, in contrast, that the existence of several distinct modes indicates a heterogeneous, clustered population, with each mode corresponding to a cluster of observations. Although this is well known not to be universally true (see, for example, Behboodian, 1970), the general thrust of the methods to be discussed in this section is that the presence of some degree of multimodality in the data is relatively strong evidence in favour of some type of cluster structure. There are a number of formal tests for the presence of distinct modes in data that may sometimes be useful; for example, those suggested by Hartigan and Hartigan (1985), Good and Gaskins (1980), Silverman (1981, 1983) and Cheng and Hall (1998). But here we shall not give details of these methods, preferring to concentrate on a rather more informal 'eye-balling' approach to the problem of mode detection using suitable one- and two-dimensional plots of the data.

2.2.1 Histograms

The humble histogram is often a useful first step in the search for modes in data, particularly, of course, if the data are univariate. Figure 2.1, for example, shows a histogram of myelinated lumbosacral ventral root fibre sizes from a kitten of a particular age. The distribution is distinctly bimodal, suggesting the presence of two relatively distinct groups of observations in the data. Here it is known that the first mode is associated with axons of gamma neurones and the second with alpha neurones.

Now consider the data shown in Table 2.1, which gives the velocities of 82 galaxies from six well-separated conic sections of space. The data are intended to shed light on whether or not the observable universe contains superclusters of galaxies surrounded by large voids. The evidence for the existence of the superclusters would be in the multimodality of the distribution of velocities. A histogram of the data does give some evidence of such multimodality – see Figure 2.2.

With multivariate data, histograms can be constructed for each separate variable, as we shall in a later example in this chapter (see Section 2.2.4), but in general this will not be of great help in uncovering cluster structure in the data because the marginal distribution of each variable may not reflect accurately the multivariate distribution of the complete set of variables. It is, for example, possible for clustered multivariate data to give rise to unimodal histograms for some or all of the separate variables.

2.2.2 Scatterplots

The basis of two-dimensional views of the data is the simple *xy*-scatterplot, which has been in use since at least the 18th century (see Tufte, 1983). To begin, consider the scatterplot shown in Figure 2.3. To a human observer (such as the reader), it is

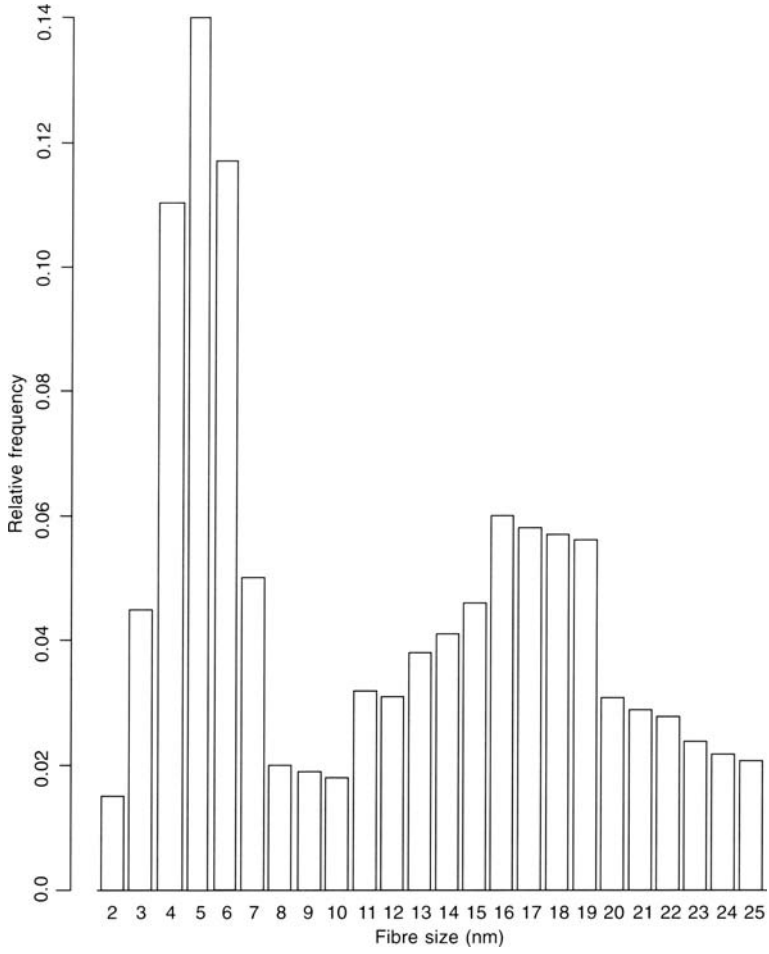


Figure 2.1 *Histogram of myelinated lumbosacral ventral root fibre sizes from a kitten of a particular age.*

Table 2.1 Velocities of 82 galaxies (km/s).

9172	9558	10406	18419	18927	19330	19440	19541	19846	19914	19989
20179	20221	20795	20875	21492	21921	22209	22314	22746	22914	23263
23542	23711	24289	24990	26995	34279	9350	9775	16084	18552	19052
19343	19473	19547	19856	19918	20166	20196	20415	20821	20986	21701
21960	22242	22374	22747	23206	23484	23666	24129	24366	25633	32065
9483	10227	16170	18600	19070	19349	19529	19663	19863	19973	20175
10215	20629	20846	21137	21814	22185	22249	22495	22888	23241	23538
23706	24285	24717	26960	32789						

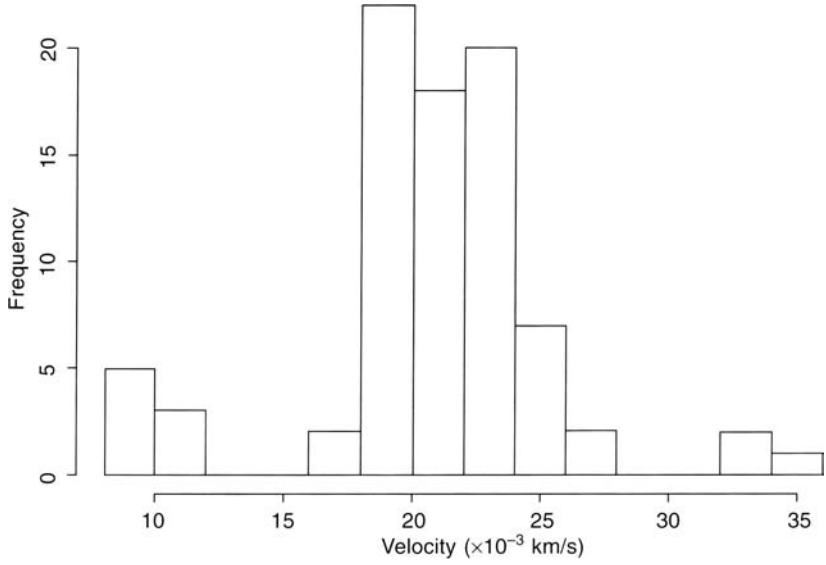


Figure 2.2 *Histogram of velocities of 82 galaxies.*

obvious that there are two natural groups or clusters of dots. (It is not entirely clear how a 'cluster' is recognized when displayed in the plane, although one feature of the recognition process probably involves the assessment of the relative distances between points.) This conclusion is reached with no conscious effort of thought. The relative homogeneity of each cluster of points and the degree of their separation makes the task of identifying them simple.

But now consider the example provided by Figure 2.4, which is a scatterplot of two measures of intensity, PD and T_2, for 2836 voxels from a functional magnetic resonance imaging (fMRI) study reported in Bullmore *et al*. (1995). *Some* cluster structure is apparent from this plot, but in parts of the diagram the 'overplotting'

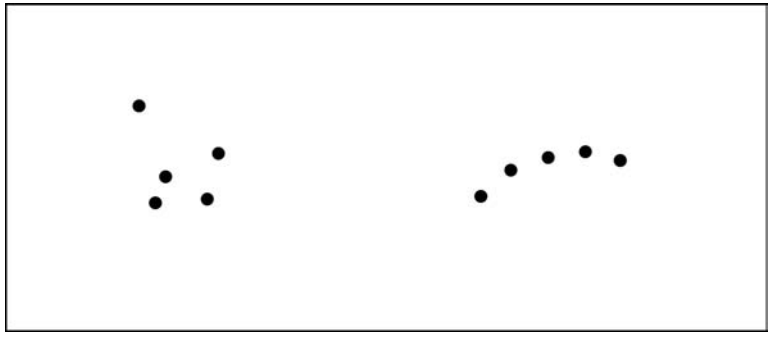


Figure 2.3 *Scatterplot showing two distinct groups of points.*

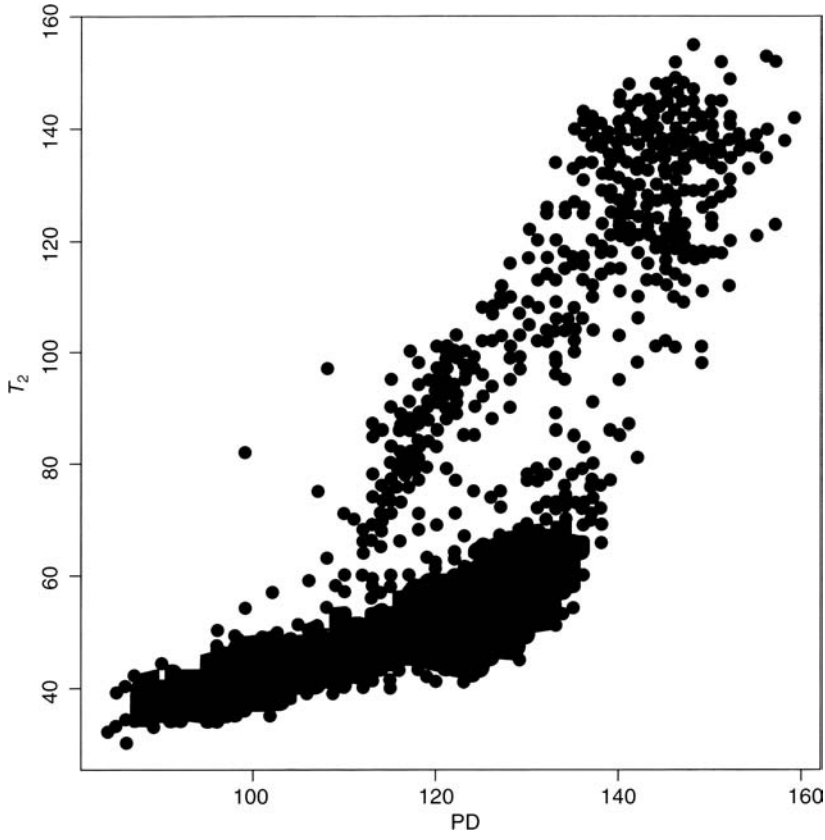


Figure 2.4 *Scatterplot of fMRI data.*

may hide the presence of other clusters. And such will be the case for the scatterplots of many data sets found in practice. The basic scatterplot might, it appears, need a little help if it is to reveal *all* the structure in the data.

2.2.3 Density estimation

Scatterplots (and histograms) can be made to be more useful by adding a numerical estimate of the bivariate (univariate) density of the data. If we are willing to assume a particular form for the distribution of the data, for example bivariate (univariate) Gaussian, density estimation would be reduced to simply estimating the values of the density's parameters. (A particular type of parameter density function useful for modelling clustered data will be described in Chapter 6.) More commonly, however, we wish to allow the data to speak for themselves and so use one of a variety of *nonparametric estimation* procedures now available. Density estimation is now a large topic and is covered in several books including Silverman (1986), Scott (1992), Wand and Jones (1995), Simonoff (1996) and Bowman and Azzalini

(1997). Here we content ourselves with a brief account of *kernel density estimates*, beginning for simplicity with estimation for the univariate situation and then moving on to the usually more interesting (at least from a clustering perspective) case of bivariate data.

Univariate density estimation

From the definition of a probability density, if the random variable X has density f,

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x-h < X < x+h). \tag{2.1}$$

For any given h, a naïve estimator of $P(x-h < X < x+h)$ is the proportion of the observations $X_1, X_2, \ldots, X_n$ falling in the interval $(x-h,\ x+h)$; that is

$$\hat{f}(x) = \frac{1}{2hn} [\text{no. of } X_1, X_2, \ldots, X_n \text{ falling in } (x-h,\ x+h)]. \tag{2.2}$$

If we introduce a weight function W given by

$$W(x) = \left\{ \begin{array}{ll} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{array} \right\}, \tag{2.3}$$

then the naïve estimator can be rewritten as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x-X_i}{h} \right). \tag{2.4}$$

Unfortunately, this estimator is not a continuous function and is not satisfactory for practical density estimation. It does, however, lead naturally to the kernel estimator defined by

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h} \right), \tag{2.5}$$

where K is known as the *kernel function* and h as the *bandwidth* or *smoothing parameter*. The kernel function must satisfy the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{2.6}$$

Usually, but not always, the kernel function will be a symmetric density function, for example, the normal.

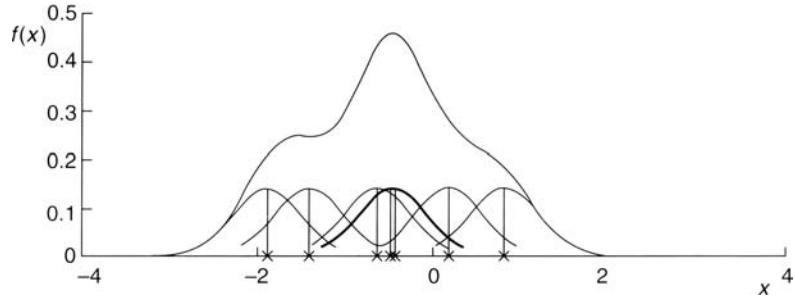


Figure 2.5 *Kernel estimate showing individual kernels.*

The kernel estimator is a sum of 'bumps' placed at the observations. The kernel function determines the shape of the bumps, while the window width h determines their width. Figure 2.5 (taken from Silverman, 1986) shows the individual bumps $n^{-1}h^{-1}K[(x-X_i)/h]$, as well as the estimate $\hat{f}$ obtained by adding them up.

Three commonly used kernel functions are *rectangular*, *triangular* and *Gaussian*:

- rectangular

$$K(x) = \frac{1}{2} \text{ for } |x| < 1,\ 0 \text{ otherwise} \tag{2.7}$$

- triangular

$$K(x) = 1-|x| \text{ for } |x| < 1,\ 0 \text{ otherwise} \tag{2.8}$$

- Gaussian

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}. \tag{2.9}$$

The three kernel functions are illustrated in Figure 2.6.

As an illustration of applying univariate kernel density estimators, Figure 2.7 shows a number of estimates for the velocities of galaxies data in Table 2.1, obtained by using different kernel functions and/or different values for the bandwidth. The suggestion of multimodality and hence 'clustering' in the galaxy velocities is very strong.

Bivariate density estimation

The kernel density estimator considered as a sum of 'bumps' centred at the observations has a simple extension to two dimensions (and similarly for

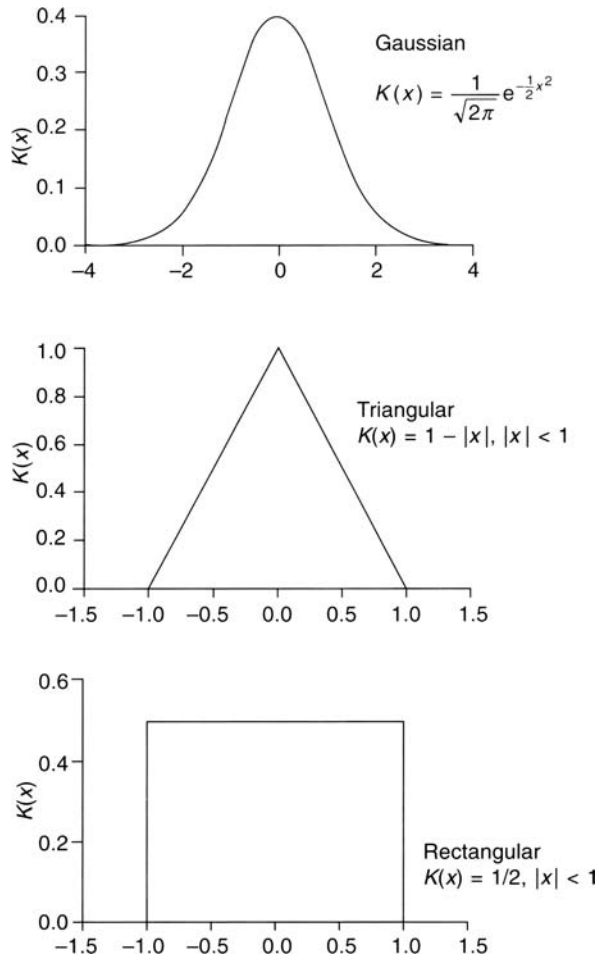


Figure 2.6 *Three kernel functions.*

more than two dimensions). The bivariate estimator for data (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ is defined as

$$\hat{f}(x, y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right). \tag{2.10}$$

In this estimator each coordinate direction has its own smoothing parameter, h_x and h_y. An alternative is to scale the data equally for both dimensions and use a single smoothing parameter.

For bivariate density estimation a commonly used kernel function is the standard bivariate normal density

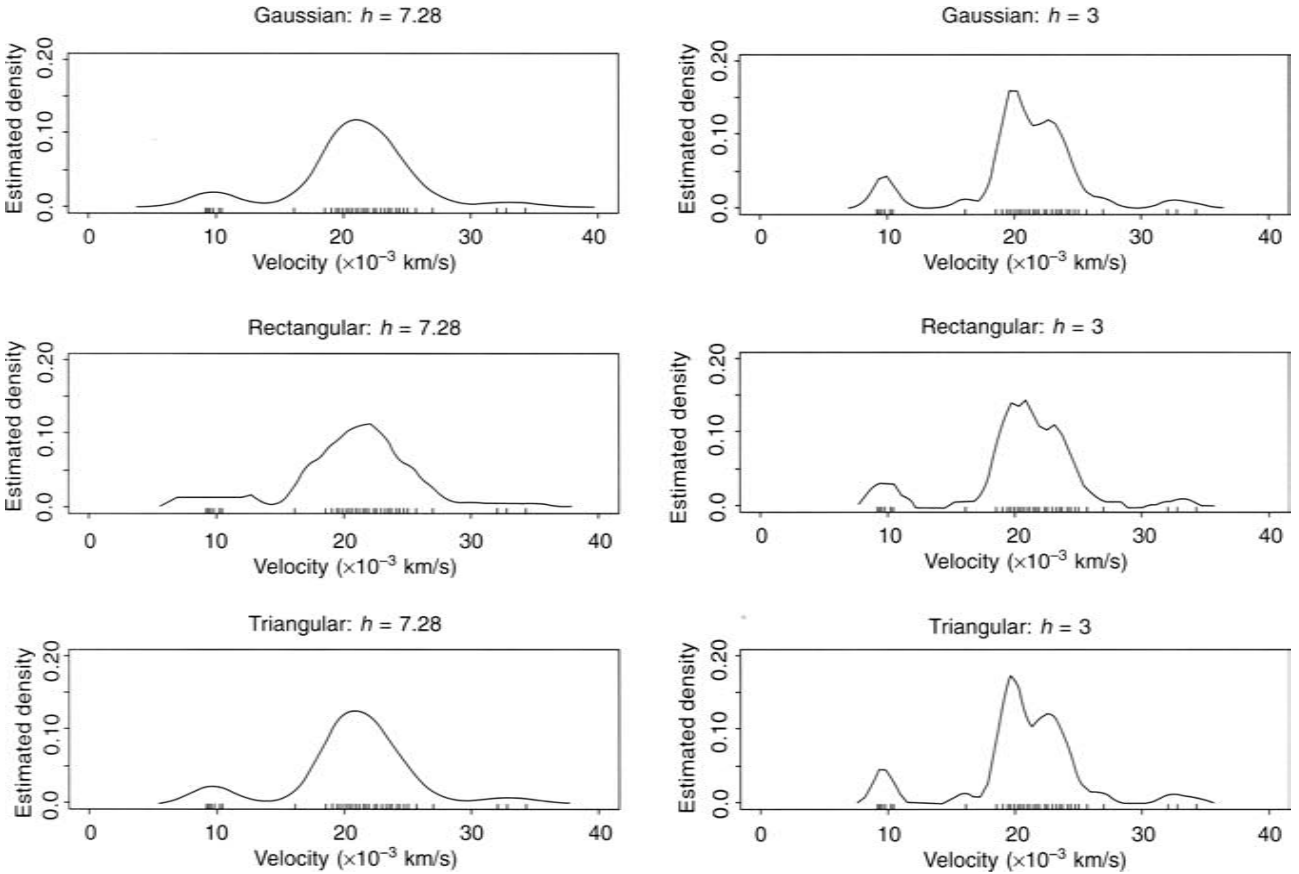


Figure 2.7 *Kernel density estimates for galaxy velocity data.*

$$K(x,y) = \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(x^2+y^2\right)\right]. \tag{2.11}$$

Another possibility is the bivariate Epanechikov kernel given by

$$K(x,y)\begin{cases} = \dfrac{2}{\pi}\left(1-x^2-y^2\right) & \text{if } x^2+y^2 < 1 \\ = 0 \text{ otherwise} \end{cases}. \tag{2.12}$$

According to Venables and Ripley (1999), the bandwidth should be chosen to be proportional to $n^{-\frac{1}{5}}$; unfortunately the constant of proportionality depends on the unknown density. The tricky problem of bandwidth estimation is considered in detail in Silverman (1986).

Returning to the fMRI data introduced in the previous subsection and plotted in Figure 2.4, we will now add the contours of the estimated bivariate density of the data to this scatterplot, giving the diagram shown in Figure 2.8. The enhanced scatterplot gives evidence of at least three clusters of voxels, a structure that could not be seen in Figure 2.4. Here, three clusters makes good sense because they correspond largely to grey matter voxels, white matter voxels and cerebrospinal fluid (CSF) voxels. (A more formal analysis of these data is presented in Chapter 6.)

2.2.4 Scatterplot matrices

When we have multivariate data with three or more variables, the scatterplots of each pair of variables can still be used as the basis of an initial examination of the data for informal evidence that the data have some cluster structure, particularly if the scatterplots are arranged as a *scatterplot matrix*. A scatterplot matrix is defined as a square, symmetric grid of bivariate scatterplots (Cleveland, 1994). This grid has p rows and p columns, each one corresponding to a different one of the p variables. Each of the grid's cells show a scatterplot of two variables. Because the scatterplot matrix is symmetric about its diagonal, variable j is plotted against variable i in the ijth cell, and the same variables appear in cell ji with the x and y axes of the scatterplots interchanged. The reason for including both the upper and lower triangles of the grid, despite its seeming redundancy, is that it enables a row and a column to be visually scanned to see one variable against all others, with the scales for the one lined up along the horizontal or the vertical.

To illustrate the use of the scatterplot matrix combined with univariate and bivariate density estimation, we shall use the data on body measurements shown in Table 2.2. (There are, of course, too few *observations* here for an entirely satisfactory or convincing demonstration, but they will serve as a useful example.) Figure 2.9 is a scatterplot matrix of the data in which each component scatterplot is enhanced with the estimated bivariate density of the two variables, and the diagonal panels display the histograms for each separate variable. There is clear evidence of the presence of two separate groups, particularly in the waist–hips measurements scatterplot. In this case the explanation is simple – the data contain measurements on both males and females.

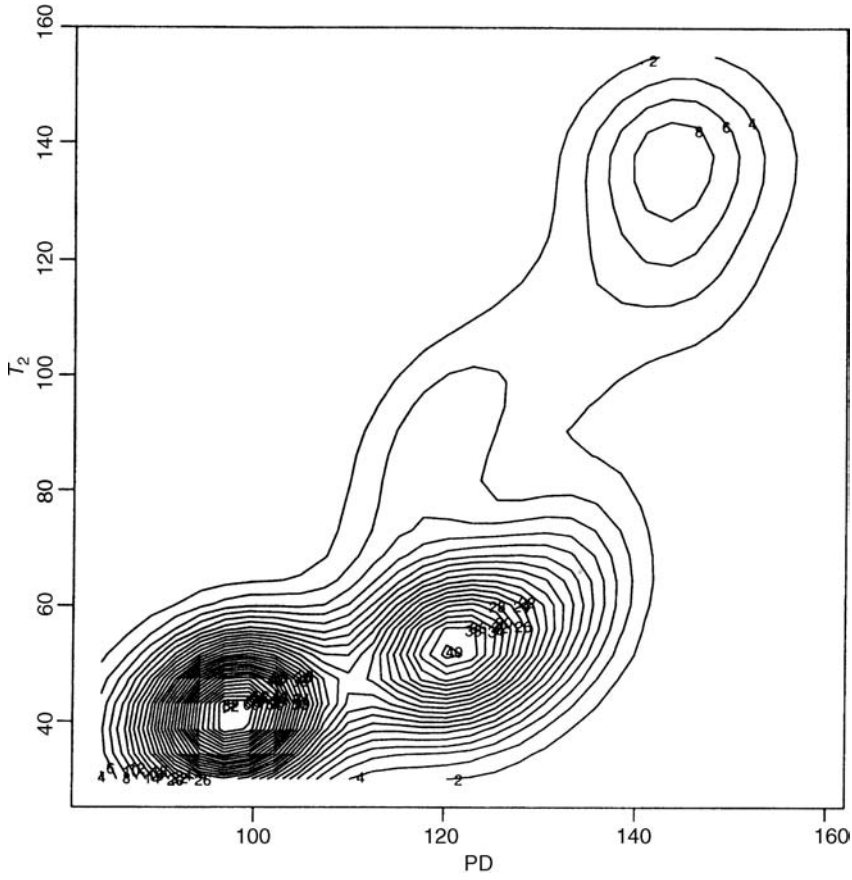


Figure 2.8 *Scatterplot of fMRI data enhanced with estimated bivariate density.*

As a second example of using the scatterplot matrix combined with bivariate density estimation, we shall use the data shown in Table 2.3 collected in a study of air pollution in 41 US cities. The variables recorded are as follows:

SO_2	Sulphur dioxide content of air in micrograms per cubic metre;
TEMP	Average annual temperature (°F);
MANUF	Number of manufacturing enterprises employing 20 or more workers;
POP	Population size (1970 census) in thousands;
WIND	Average wind speed in miles per hour;
PRECIP	Average annual precipitation in inches;
DAYS	Average number of days with precipitation per year.

Here we shall use the six climate and human ecology variables to assess whether there is any evidence that there are groups of cities with similar profiles. (The sulphur dioxide content of the air will be used in Chapter 5 to aid in

Table 2.2 Body measurements data (inches).

Subject	Chest	Waist	Hips
1	34	30	32
2	37	32	37
3	38	30	36
4	36	33	39
5	38	29	33
6	43	32	38
7	40	33	42
8	38	30	40
9	40	30	37
10	41	32	39
11	36	24	35
12	36	25	37
13	34	24	37
14	33	22	34
15	36	26	38
16	37	26	37
17	34	25	38
18	36	26	37
19	38	28	40
20	35	23	35

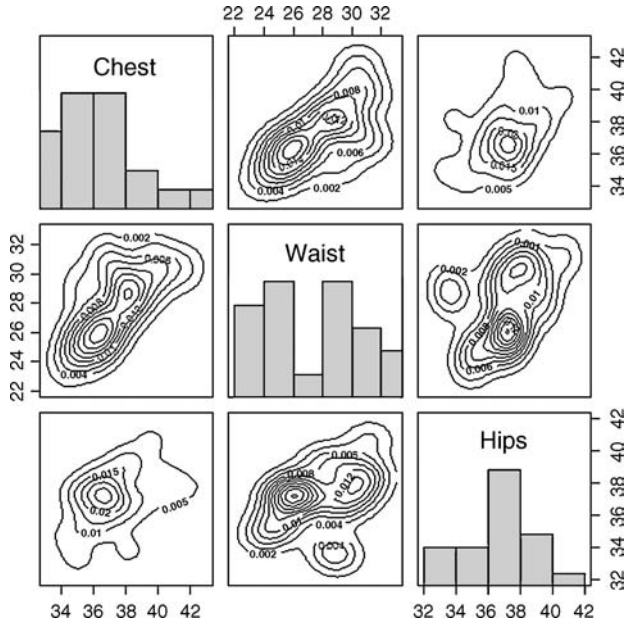


Figure 2.9 *Scatterplot matrix of body measurements enhanced by bivariate density estimates and histograms.*

Table 2.3 Air pollution in US cities.

City	SO2	TEMP	MANUF	POP	WIND	PRECIP	DAYS
Phoenix	10	70.3	213	582	6.0	7.05	36
Little Rock	13	61.0	91	132	8.2	48.52	100
San Francisco	12	56.7	453	716	8.7	20.66	67
Denver	17	51.9	454	515	9.0	12.95	86
Hartford	56	49.1	412	158	9.0	43.37	127
Wilmington	36	54.0	80	80	9.0	40.25	114
Washington	29	57.3	434	757	9.3	38.89	111
Jacksonville	14	68.4	136	529	8.8	54.47	116
Miami	10	75.5	207	335	9.0	59.80	128
Atlanta	24	61.5	368	497	9.1	48.34	115
Chicago	110	50.6	3344	3369	10.4	34.44	122
Indianapolis	28	52.3	361	746	9.7	38.74	121
Des Moines	17	49.0	104	201	11.2	30.85	103
Wichita	8	56.6	125	277	12.7	30.58	82
Louisville	30	55.6	291	593	8.3	43.11	123
New Orleans	9	68.3	204	361	8.4	56.77	113
Baltimore	47	55.0	625	905	9.6	41.31	111
Detroit	35	49.9	1064	1513	10.1	30.96	129
Minneapolis	29	43.5	699	744	10.6	25.94	137
Kansas	14	54.5	381	507	10.0	37.00	99
St Louis	56	55.9	775	622	9.5	35.89	105
Omaha	14	51.5	181	347	10.9	30.18	98
Albuquerque	11	56.8	46	244	8.9	7.77	58
Albany	46	47.6	44	116	8.8	33.36	135
Buffalo	11	47.1	391	463	12.4	36.11	166
Cincinnati	23	54.0	462	453	7.1	39.04	132
Cleveland	65	49.7	1007	751	10.9	34.99	155
Columbus	26	51.5	266	540	8.6	37.01	134
Philadelphia	69	54.6	1692	1950	9.6	39.93	115
Pittsburgh	61	50.4	347	520	9.4	36.22	147
Providence	94	50.0	343	179	10.6	42.75	125
Memphis	10	61.6	337	624	9.2	49.10	105
Nashville	18	59.4	275	448	7.9	46.00	119
Dallas	9	66.2	641	844	10.9	35.94	78
Houston	10	68.9	721	1233	10.8	48.19	103
Salt Lake City	28	51.0	137	176	8.7	15.17	89
Norfolk	31	59.3	96	308	10.6	44.68	116
Richmond	26	57.8	197	299	7.6	42.59	115
Seattle	29	51.1	379	531	9.4	38.79	164
Charleston	31	55.2	35	71	6.5	40.75	148
Milwaukee	16	45.7	569	717	11.8	29.07	123

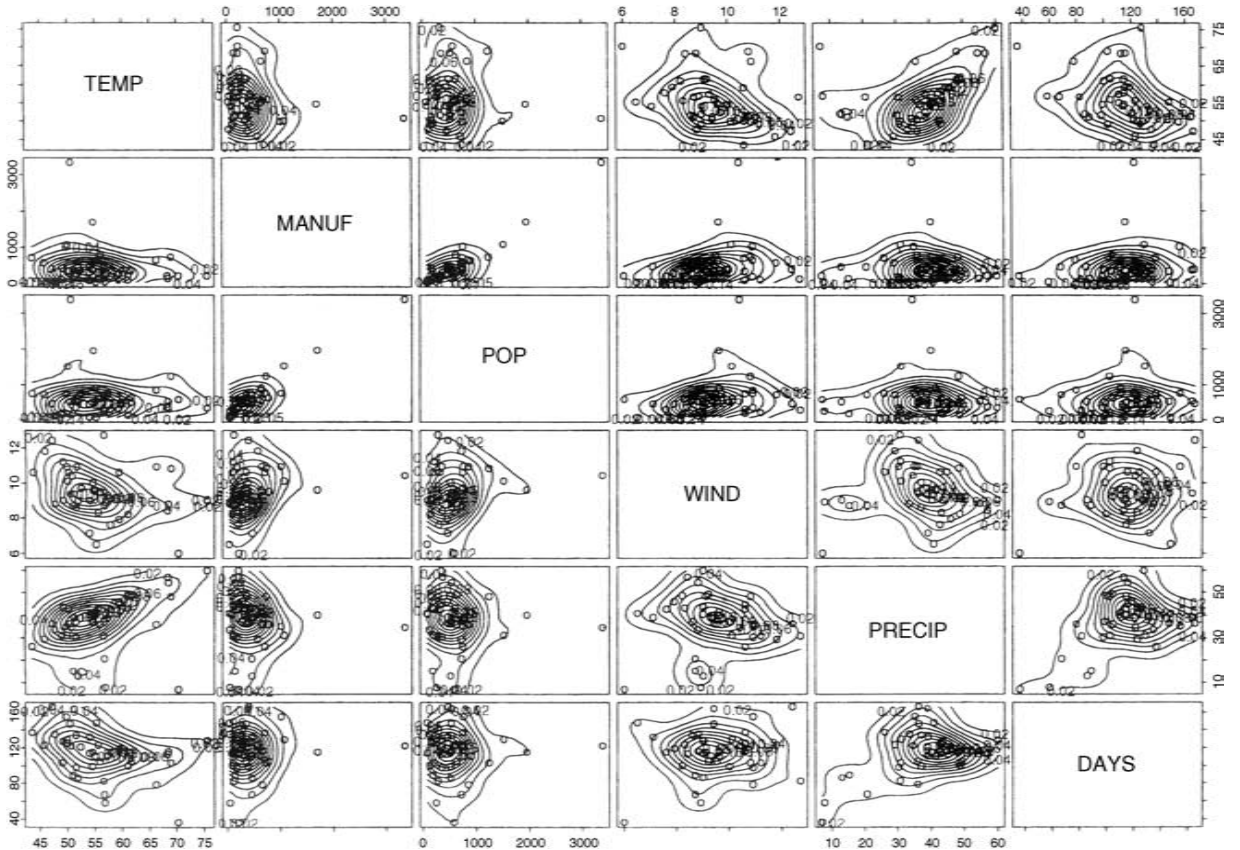


Figure 2.10 *Scatterplot matrix of air pollution data in which each panel is enhanced with an estimated bivariate density.*

evaluating a more detailed cluster analysis of these data.) A scatterplot matrix of the data in which each component scatterplot is enhanced with the estimated bivariate density of the two variables is shown in Figure 2.10. The plot gives little convincing evidence of any group structure in the data. The most obvious feature of the data is the presence of several 'outlier' observations. As such observations can be a problem for some methods of cluster analysis, identification of them prior to applying these methods is often very useful.

2.3 Using lower-dimensional projections of multivariate data for graphical representations

Scatterplots and, to some extent, scatterplot matrices are more useful for exploring multivariate data for the presence of clusters when there are only a relatively small number of variables. When the number of variables, p, is moderate to large (as it will be for many multivariate data sets encountered in practice), the two-dimensional marginal views of the data suffer more and more from the fact that they do not necessarily reflect the true nature of the structure present in p dimensions. But all is not necessarily lost, and scatterplots etc. may still be useful after the data have been projected into a small number of dimensions in some way that preserves their multivariate structure as fully as possible. There are a number of possibilities, of which the most common (although not necessarily the most useful) is *principal components analysis.*

2.3.1 Principal components analysis of multivariate data

Principal components analysis is a method for transforming the variables in a multivariate data set into new variables which are uncorrelated with each other and which account for decreasing proportions of the total variance of the original variables. Each new variable is defined as a particular linear combination of the original variables. Full accounts of the method are given in Everitt and Dunn (2001) and Jackson (2003), but in brief:

- The first principal component, y_1, is defined as the linear combination of the original variables, $x_1, x_2, \ldots, x_p$, that accounts for the maximal amount of the variance of the x variables amongst all such linear combinations.
- The second principal component, y_2, is defined as the linear combination of the original variables that accounts for a maximal amount of the remaining variance subject to being uncorrelated with y_1. Subsequent components are defined similarly.
- So principal components analysis finds new variables $y_1, y_2, \ldots, y_p$ defined as follows:

$$\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ &\vdots \\ y_p &= a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p, \end{aligned} \tag{2.13}$$

with the coefficients being chosen so that $y_1, y_2, \ldots, y_p$ account for decreasing proportions of the variance of the x variables and are uncorrelated. (Because the variance can be scaled by rescaling the combination defining a principal component, the coefficients defining each principal component are such that their sums of squares equal one.)

- The coefficients are found as the eigenvectors of the sample covariance matrix **S** or, more commonly, the sample correlation matrix **R** when the variables are on very different scales. It should be noted that there is, in general, no simple relationship between the components found from these two matrices. (A similar scaling decision is also often a problem in cluster analysis, as we shall see in Chapter 3).
- The variances of $y_1, y_2, \ldots, y_p$ are given by the eigenvalues of **S** or **R**. Where the first few components account for a large proportion of the variance of the observed variables, it is generally assumed that they provide a parsimonious summary of the original data, useful perhaps in later analysis.

In a clustering context, principal components analysis provides a means of projecting the data into a lower dimensional space, so as to make visual inspection hopefully more informative. The points of the q-dimensional projection ($q < p$) onto the first q principal components lie in a q-dimensional space, and this is the best-fitting q-dimensional space as measured by the sum of the squares of the distances from the data points to their projections into this space.

The use of principal components analysis in the context of clustering will be illustrated on the data shown in Table 2.4, which gives the chemical composition in terms of nine oxides as determined by atomic absorption spectrophotometry, of 46 examples of Romano-British pottery (Tubb *et al.*, 1980). Here there are nine variables, recorded for each piece of pottery. The scatterplot matrix of the data shown in Figure 2.11 does show some evidence of clustering, but it is quite difficult to see clearly any detail of the possible structure in the data.

The results of a principal components analysis on the correlation matrix of the pottery data are shown in Table 2.5. The first two components have eigenvalues greater than one, and the third component has an eigenvalue very close to one, and so we will use these three components to represent the data graphically. (Here we shall not spend time trying to interpret the components.) A scatterplot matrix of the first three component scores for each piece of pottery is shown in Figure 2.12. Each separate scatterplot is enhanced with a contour plot of the appropriate estimated bivariate density. The diagram gives strong evidence that there are at least three clusters in the data. Here an explanation for the cluster structure is found by looking at the regions from which the pots arise; the three clusters correspond to pots from three different regions – see Everitt (2005).

(A last word of caution is perhaps needed about using principal components in the way described above, because the principal components analysis may attribute sample correlations due to clusters as due to components. The problem is taken up in detail in Chapters 3, 6 and 7.)

Table 2.4 Results of chemical analysis of Roman-British pottery (Tubb *et al.*, 1980).

Sample number	Chemical component								
	Al_2O_3	Fe_2O_3	MgO	CaO	Na_2O	K_2O	TiO_2	MnO	BaO
1	18.8	9.52	2.00	0.79	0.40	3.20	1.01	0.077	0.015
2	16.9	7.33	1.65	0.84	0.40	3.05	0.99	0.067	0.018
3	18.2	7.64	1.82	0.77	0.40	3.07	0.98	0.087	0.014
4	16.9	7.29	1.56	0.76	0.40	3.05	1.00	0.063	0.019
5	17.8	7.24	1.83	0.92	0.43	3.12	0.93	0.061	0.019
6	18.8	7.45	2.06	0.87	0.25	3.26	0.98	0.072	0.017
7	16.5	7.05	1.81	1.73	0.33	3.20	0.95	0.066	0.019
8	18.0	7.42	2.06	1.00	0.28	3.37	0.96	0.072	0.017
9	15.8	7.15	1.62	0.71	0.38	3.25	0.93	0.062	0.017
10	14.6	6.87	1.67	0.76	0.33	3.06	0.91	0.055	0.012
11	13.7	5.83	1.50	0.66	0.13	2.25	0.75	0.034	0.012
12	14.6	6.76	1.63	1.48	0.20	3.02	0.87	0.055	0.016
13	14.8	7.07	1.62	1.44	0.24	3.03	0.86	0.080	0.016
14	17.1	7.79	1.99	0.83	0.46	3.13	0.93	0.090	0.020
15	16.8	7.86	1.86	0.84	0.46	2.93	0.94	0.094	0.020
16	15.8	7.65	1.94	0.81	0.83	3.33	0.96	0.112	0.019
17	18.6	7.85	2.33	0.87	0.38	3.17	0.98	0.081	0.018
18	16.9	7.87	1.83	1.31	0.53	3.09	0.95	0.092	0.023
19	18.9	7.58	2.05	0.83	0.13	3.29	0.98	0.072	0.015
20	18.0	7.50	1.94	0.69	0.12	3.14	0.93	0.035	0.017
21	17.8	7.28	1.92	0.81	0.18	3.15	0.90	0.067	0.017
22	14.4	7.00	4.30	0.15	0.51	4.25	0.79	0.160	0.019
23	13.8	7.08	3.43	0.12	0.17	4.14	0.77	0.144	0.020
24	14.6	7.09	3.88	0.13	0.20	4.36	0.81	0.124	0.019
25	11.5	6.37	5.64	0.16	0.14	3.89	0.69	0.087	0.009
26	13.8	7.06	5.34	0.20	0.20	4.31	0.71	0.101	0.021
27	10.9	6.26	3.47	0.17	0.22	3.40	0.66	0.109	0.010
28	10.1	4.26	4.26	0.20	0.18	3.32	0.59	0.149	0.017
29	11.6	5.78	5.91	0.18	0.16	3.70	0.65	0.082	0.015
30	11.1	5.49	4.52	0.29	0.30	4.03	0.63	0.080	0.016
31	13.4	6.92	7.23	0.28	0.20	4.54	0.69	0.163	0.017
32	12.4	6.13	5.69	0.22	0.54	4.65	0.70	0.159	0.015
33	13.1	6.64	5.51	0.31	0.24	4.89	0.72	0.094	0.017
34	11.6	5.39	3.77	0.29	0.06	4.51	0.56	0.110	0.015
35	11.8	5.44	3.94	0.30	0.04	4.64	0.59	0.085	0.013
36	18.3	1.28	0.67	0.03	0.03	1.96	0.65	0.001	0.014
37	15.8	2.39	0.63	0.01	0.04	1.94	1.29	0.001	0.014
38	18.0	1.50	0.67	0.01	0.06	2.11	0.92	0.001	0.016
39	18.0	1.88	0.68	0.01	0.04	2.00	1.11	0.006	0.022
40	20.8	1.51	0.72	0.07	0.10	2.37	1.26	0.002	0.016
41	17.7	1.12	0.56	0.06	0.06	2.06	0.79	0.001	0.013
42	18.3	1.14	0.67	0.06	0.05	2.11	0.89	0.006	0.019
43	16.7	0.92	0.53	0.01	0.05	1.76	0.91	0.004	0.013
44	14.8	2.74	0.67	0.03	0.05	2.15	1.34	0.003	0.015
45	19.1	1.64	0.60	0.10	0.03	1.75	1.04	0.007	0.018

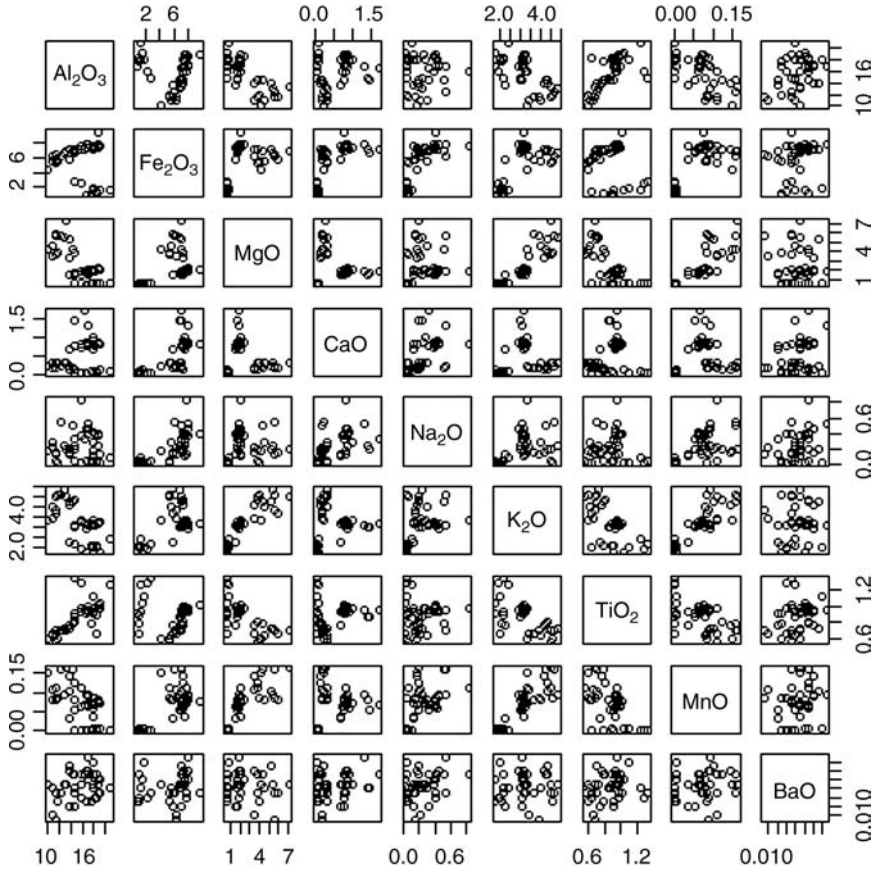


Figure 2.11 *Scatterplot matrix of data on Romano-British pottery.*

2.3.2 Exploratory projection pursuit

According to Jones and Sibson (1987),

> The term 'projection pursuit' was first used by Friedman and Tukey (1974) to name a technique for the exploratory analysis of reasonably large and reasonably multivariate data sets. Projection pursuit reveals structure in the original data by offering selected low-dimensional orthogonal projections of it for inspection.

In essence, projection pursuit methods seek a q-dimensional projection of the data that maximizes some measure of 'interestingness', usually for $q = 1$ or $q = 2$ so that they can be visualized. Principal components analysis is, for example, a projection pursuit method in which the index of interestingness is the proportion of total variance accounted for by the projected data. It relies for its success on the tendency for large variation to also be interestingly structured variation, a connection which is

Table 2.5 Results of a principal components analysis on the Romano-British pottery data.

Oxide	Component 1	Component 2	Component 3
Al_2O_3	0.35	0.33	−0.12
Fe_2O_3	−0.33	0.40	0.26
MgO	−0.44	−0.19	−0.15
CaO	-----	0.50	0.48
Na_2O	−0.22	0.46	-----
K_2O	−0.46	-----	−0.10
TiO_2	0.34	0.30	-----
MnO	−0.46	-----	−0.14
BaO	-----	0.38	−0.79
Component standard deviation	2.05	1.59	0.94

----- = the loading is very close to zero.

not necessary and which often fails to hold in practice. The essence of projection pursuit can be described most simply by considering only one-dimensional solutions.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a p-dimensional data set of size n; then the projection is specified by a p-dimensional vector α such that $\alpha\alpha' = 1$. (A two-dimensional projection would involve a $p \times 2$ matrix). The projected data in this case are one-dimensional data points $z_1, \ldots, z_n$, where $z_i = \alpha\mathbf{x}_i$ for $i = 1, \ldots, n$. The merit of a particular configuration $\{z_1, \ldots, z_n\}$ in expressing important structure in the data will be quantified by means of a projection index, $I_{(z_1,\ldots,z_n)}$. (Principal components

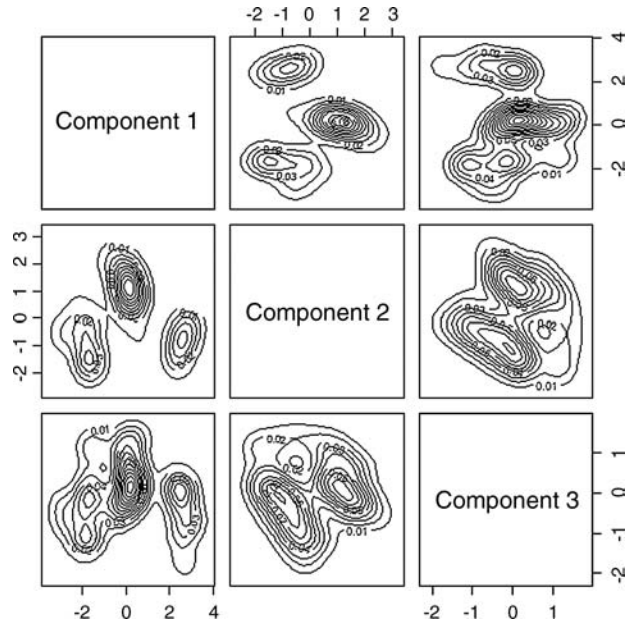


Figure 2.12 *Scatterplot of the first three principal components of the Romano-British pottery data enhanced by bivariate density estimates.*

reduction effectively uses sample variance as a projection index.) Most frequently the index used is a measure of the discrepancy between the density of the projected data and the density of an 'uninteresting' projection, for example the standard normal, but many other indices of 'interestingness' have been suggested. Details are given in Jones and Sibson (1987), Ripley (1996) and Sun (1998).

Once an index is chosen, a projection is found by numerically maximizing the index over the choice of projection; that is, a q-dimensional projection is determined by a $p \times q$ orthogonal matrix. With q small (usually one or two), this may appear to imply a relatively simple optimization task, but Ripley (1996) points out why this is not always the case:

- The index is often very sensitive to the projection directions, and good views may occur with sharp and well-separated peaks in the optimization space.
- The index may be very sensitive to small changes in the data configuration and so have very many local maxima.

As a result of such problems, Friedman, in the discussion of Jones and Sibson (1987), reflects thus:

> It has been my experience that finding the substantive minima of a projection index is a difficult problem, and that simple gradient-guided methods (such as steepest descent) are generally inadequate. The power of a projection pursuit procedure depends crucially on the reliability and thoroughness of the numerical optimizer.

Ripley (1996) supports this view and suggests that it may be necessary to try many different starting points, some of which may reveal projections with large values of the projection index. Posse (1990, 1995), for example, considers an almost random search which he finds to be superior to the optimization methods of Jones and Sibson (1987) and Friedman (1987).

As an illustration of the application of projection pursuit, we will describe an example given in Ripley (1996). This involves data from a study reported in Campbell and Mahon (1974) on rock crabs of the genus *Leptograpsus*. Each specimen has measurements on the following five variables:

FL	width of frontal lip
RW	rear width
CL	length along midline
CW	maximum width of the carapace
BD	body depth.

(All measurements were in mm).

Two hundred crabs were measured and four groups suspected in advance. Four projections of the data, three found by using projection pursuit with different projection indices, are shown in Figure 2.13. Only the projection in Figure 2.13(b) matches up with the *a priori* expectation of four groups.

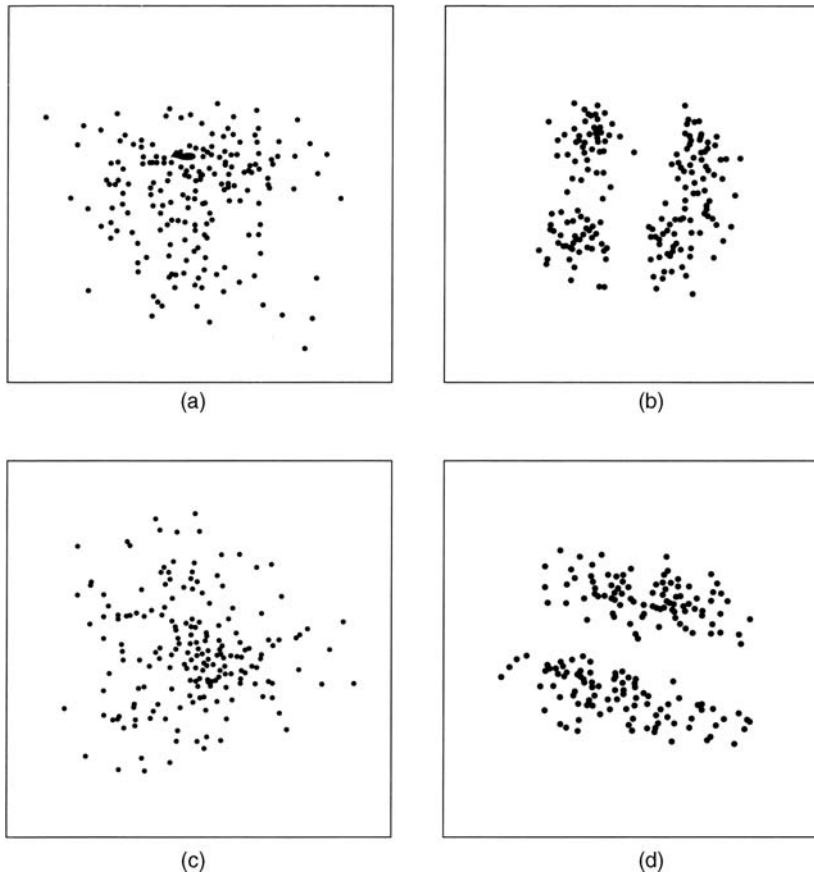


Figure 2.13 *Results from projection pursuit applied to 200 specimens of the rock crab genus* Leptograpsus: (a) random projection; (b) from projection pursuit with the Hermite index; (c) from projection pursuit with the Friedman–Tukey index; (d) from projection pursuit with the Friedman (1987) index. (Reproduced with permission of Cambridge University Press, from Pattern Recognition and Neural Networks, 1996, B. D. Ripley.)

A further example is provided by Friedman and Tukey (1974). The data consists of 500 seven-dimensional observations taken in a particle physics experiment. Full details are available in the original paper. The scatterplots in Figure 2.14 show projections of the data onto the first two principal components and onto a plane found by projection pursuit. The projection pursuit solution shows structure not apparent in the principal components plot.

Other interesting examples are given in Posse (1995). An account of software for projection pursuit is given in Swayne *et al.* (2003), and a more recent account of the method is given in Su *et al.* (2008).

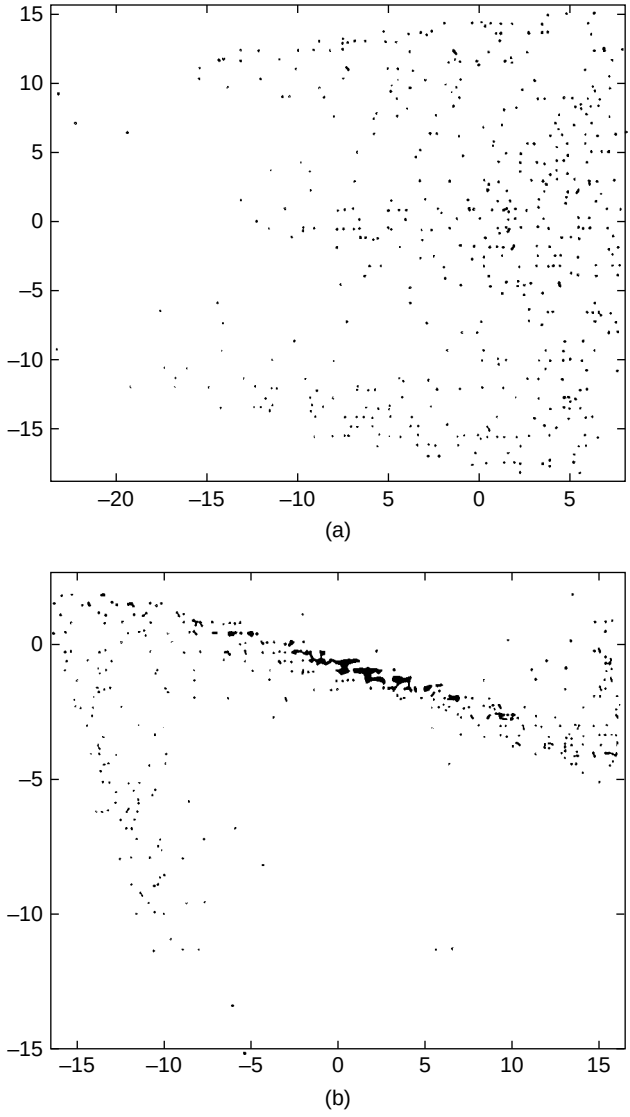


Figure 2.14 *Two-dimensional projections of high-energy physics data: (a) first two principal components; (b) projection pursuit. (Reproduced from Friedman and Tukey (1974) with the permission of the Institute of Electrical and Electronic Engineers, Inc.)*

2.3.3 Multidimensional scaling

It was mentioned in Chapter 1 that the first step in many clustering techniques is to transform the $n \times p$ multivariate data matrix $\mathbf{X}$ into an $n \times n$ matrix, the entries of which give measures of distance, dissimilarity or similarity between pairs of

individuals. (Chapter 3 will be concerned with details of these concepts.) It was also mentioned that such distance or dissimilarity matrices can arise directly, particularly from psychological experiments in which human observers are asked to judge the similarity of various stimuli of interest. For the latter in particular, a well established method of analysis is some form of *multidimensional scaling* (MDS). The essence of such methods is an attempt to represent the observed similarities or dissimilarities in the form of a geometrical model by embedding the stimuli of interest in some coordinate space so that a specified measure of distance, for example *Euclidean* (see Chapter 3), between the points in the space represents the observed proximities. Indeed, a narrow definition of multidimensional scaling is the search for a low-dimensional space in which points in the space represent the stimuli, one point representing each stimulus, such that the distances between the points in the space d_{ij} match as well as possible, in some sense, the original dissimilarities δ_{ij} or similarities s_{ij}. In a very general sense this simply means that the larger the observed dissimilarity value (or the smaller the similarity value) between two stimuli, the further apart should be the points representing them in the geometrical model. More formally, distances in the derived space are specified to be related to the corresponding dissimilarities in some simple way, for example linearly, so that the proposed model can be written as follows:

$$\begin{aligned} \delta_{ij} &= f(d_{ij}) \\ d_{ij} &= h(\mathbf{x}_i, \mathbf{x}_j), \end{aligned} \tag{2.14}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are q-dimensional vectors ($q < p$) containing the coordinate values representing stimuli i and j, f represents the assumed functional relationship between the observed dissimilarities and the derived distances, and h represents the chosen distance function. In the majority of applications of MDS, h is taken, often implicitly, to be Euclidean. The problem now becomes one of estimating the coordinate values to represent the stimuli. In general this is achieved by optimizing some goodness-of-fit index measuring how well the fitted distances match the observed proximities – for details see Everitt and Rabe-Hesketh (1997).

The general aim of MDS is most often to uncover the dimensions on which human subjects make similarity judgements. But it can also be helpful in clustering applications for displaying clusters in a two-dimensional space so that they can be visualized. An example from Shepard (1974) will illustrate how. In investigating the strengths of mental associations among 16 familiar kinds of animals, Shepard started with the quantitative information to form a MDS solution and subsequently obtained typical information by performing a hierarchical cluster analysis (see Chapter 4). Shepard gained substantially increased insight into the data structure after superimposing the typical information on the MDS 'map' of the data by enclosing cluster members within closed boundaries – see Figure 2.15.

Another example of the use of MDS for the clustering of fragrances is given in Lawless (1989), and a further recent example is described in Adachi (2002).

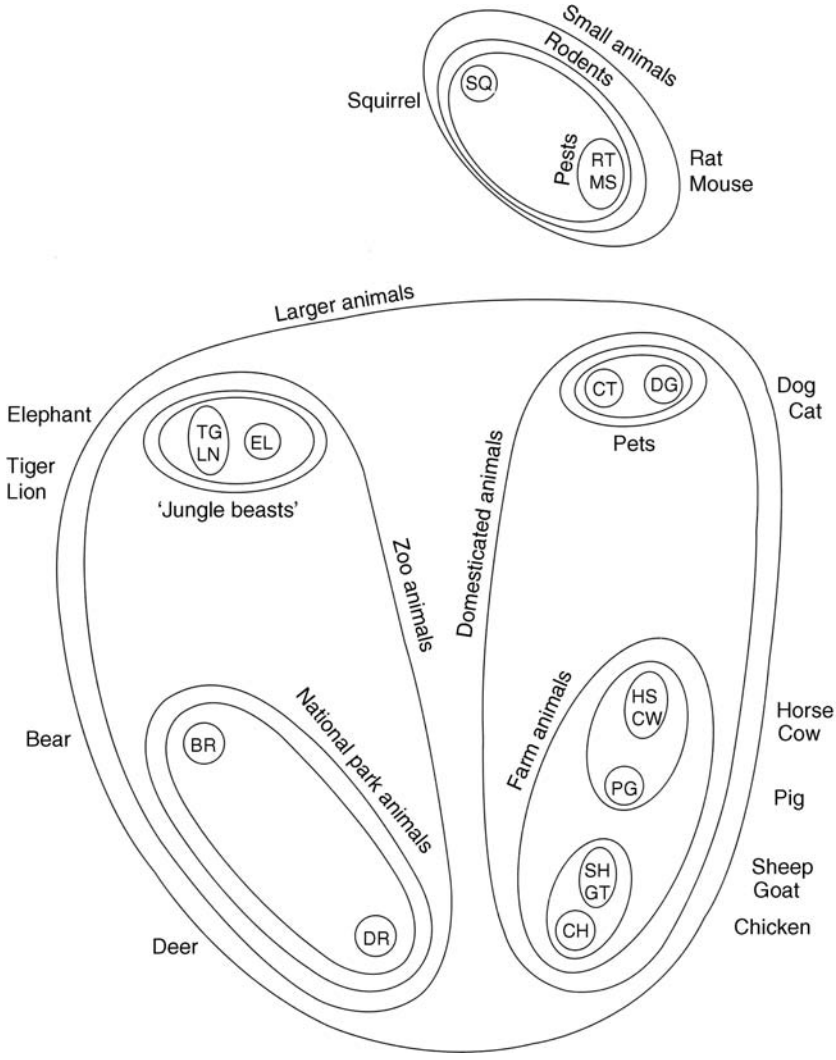


Figure 2.15 *Mental associations among 16 kinds of animals. MDS solution plus hierarchical clustering solution (reproduced with permission from Shepard, 1974).*

2.4 Three-dimensional plots and trellis graphics

The previous sections have primarily dealt with plotting data in some two-dimensional space using either the original variables, or derived variables constructed in some way so that a low-dimensional projection of the data is informative. But it is possible to plot more than two dimensions directly, as we will demonstrate in this section by way of a number of examples.

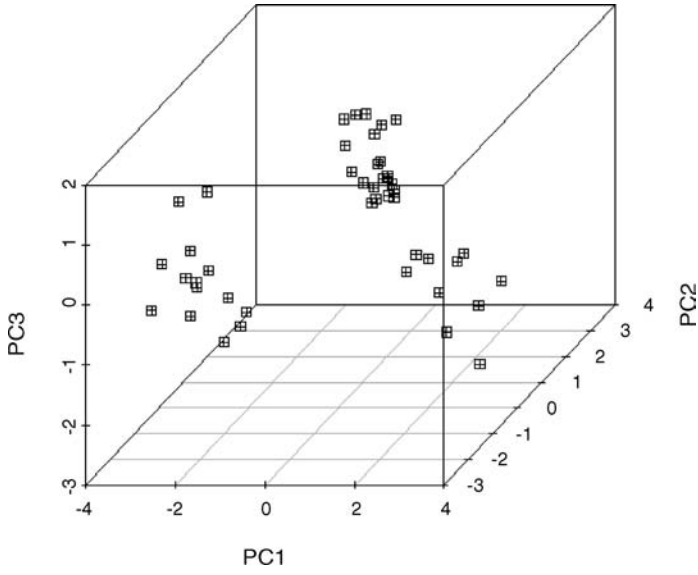


Figure 2.16 *Three-dimensional plot of the Romano-British pottery data in the space of their first three principal component scores.*

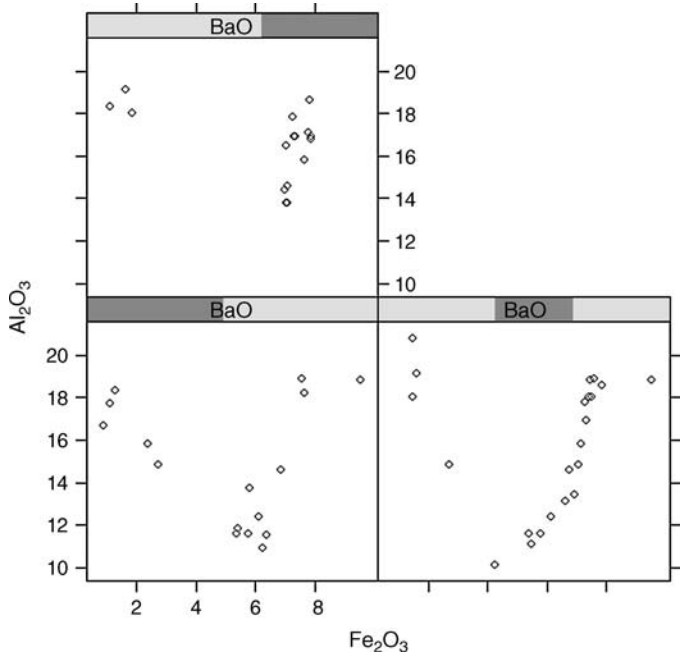


Figure 2.17 *Trellis graphic of three variables from the Romano-British pottery data.*

First, three-dimensional plots can now be obtained directly from most statistical software and these can, on occasions, be helpful in looking for evidence of cluster structure. As an example, Figure 2.16 shows a plot of the Romano-British pottery data in the space of the first three principal components. The plot very clearly demonstrates the presence of three separate clusters, confirming what was found previously in Section 2.3.1 using scatterplot matrices and estimated bivariate densities.

Trellis graphics (Becker and Cleveland, 1994) and the associated lattice graphics (Sarkar, 2008) give a way of examining high-dimensional structure in data by means of one-, two- and three-dimensional graphs. The quintessential feature of the approach is *multiple conditioning,* which allows some type of plot to be displayed for different values of a given variable or variables. Although such graphics are not specifically designed to give evidence of clustering in multivariate data, they might, nevertheless, be helpful for this is some situations. As a simple example of a trellis graphic we will look at scatterplots of Al_2O_3 values against Fe_2O_3 values conditioned on the values of BaO, after dividing these values in to three groups as follows:

Intervals	Min	Max	Count
1	0.0085	0.0155	16
2	0.0145	0.0185	22
3	0.0175	0.0236	16

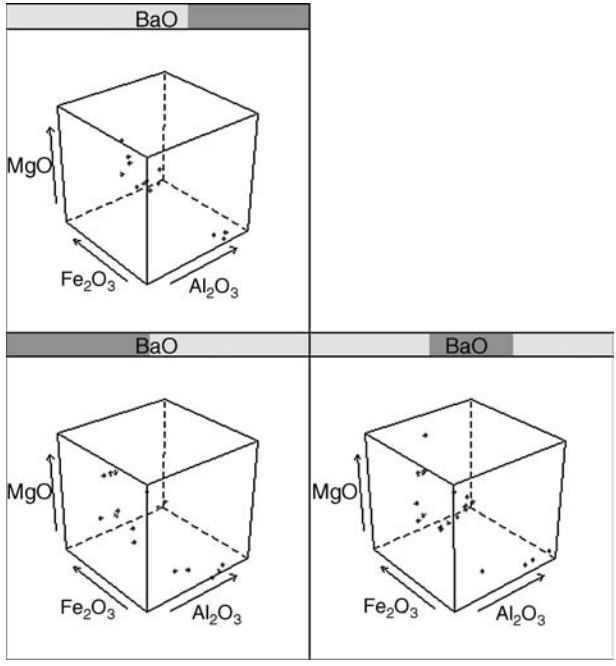


Figure 2.18 *Trellis graphic of four variables from the Romano-British pottery data.*

The resulting trellis diagram is shown in Figure 2.17. The evidence of cluster structure in the scatter plots of Al_2O_3 against Fe_2O_3 appears to hold for all values of BaO.

A more complex example of a trellis graphic is shown in Figure 2.18. Here three-dimensional plots of MgO against Al_2O_3 and Fe_2O_3 are conditioned on the three levels of BaO used previously. For the lowest values of BaO, the separation of the points into clusters appears to be less distinct than for the medium and higher values.

Many other exciting examples of the use of trellis graphics, although not specifically for evidence of clustering, are given in Sarkar (2008).

2.5 Summary

Evidence for the presence of clusters of observations can often be found by some relatively simple plots of data, enhanced perhaps by appropriate density estimates. In one or two dimensions, for example, clusters can sometimes be identified by looking for separate modes in the estimated density function of the data. Such an approach can be used on data sets where the number of variables is greater than two, by first projecting the data into a lower dimensional space using perhaps principal components or a projection pursuit approach. Although in this chapter nonparametric density estimation has been used in an informal manner to identify modes in univariate and bivariate data, Silverman (1986) describes how it can be used as the basis of a more formal approach to clustering. When only a proximity matrix is available, multidimensional scaling can be used in an effort to detect cluster structure visually. And three-dimensional plots and trellis graphics may also, on occasions, offer some graphical help in deciding whether a formal cluster analysis procedure might usefully be applied to the data. (In this chapter, our concern has been graphical displays that may help in deciding whether or not the application of some formal method of cluster analysis to a data set might be worthwhile; in Chapter 9 we will look at some graphics that may be helpful in displaying the clusters obtained from particular clustering techniques.)

3

Measurement of proximity

3.1 Introduction

Of central importance in attempting to identify clusters of observations which may be present in data is knowledge on how 'close' individuals are to each other, or how far apart they are. Many clustering investigations have as their starting point an $n \times n$ one-mode matrix, the elements of which reflect, in some sense, a quantitative measure of closeness, more commonly referred to in this context as *dissimilarity*, *distance* or *similarity*, with a general term being *proximity*. Two individuals are 'close' when their dissimilarity or distance is small or their similarity large. Proximities can be determined either directly or indirectly, although the latter is more common in most applications. Directly determined proximities are illustrated by the cola tasting experiment described in Chapter 1; they occur most often in areas such as psychology and market research.

Indirect proximities are usually derived from the $n \times p$ multivariate (two-mode) matrix, $\mathbf{X}$, introduced in Chapter 1. There is a vast range of possible proximity measures, many of which we will meet in this chapter. But as an initial example, Table 3.1 shows data concerning crime rates of seven offences ($p = 7$) for 16 cities in the USA ($n = 16$), with an accompanying dissimilarity matrix, the elements of which are calculated as the Euclidean distances between cities (see Section 3.3) after scaling each crime variable to unit variance (a technique that will be discussed in Section 3.8). We see that Washington and Detroit are judged to be the two cities most alike with respect to their crime profiles, and Los Angeles and Hartford the least alike.

To discuss indirect proximity measures in general, we shall first consider measures suitable for categorical variables, then those useful for continuous variables and finally look at measures suitable for data sets containing both

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

Table 3.1 (a) City crime data per 100 000 population (reproduced with permission from Hartigan, 1975).

	Murder/ manslaughter	Rape	Robbery	Assault	Burglary	Larceny	Auto theft
Atlanta (AT)	16.50	24.80	106.00	147.00	1112.00	905.00	494.00
Boston (BO)	4.20	13.30	122.00	90.00	982.00	669.00	954.00
Chicago (CH)	11.60	24.70	340.00	242.00	808.00	609.00	645.00
Dallas (DA)	18.10	34.20	184.00	293.00	1668.00	901.00	602.00
Denver (DE)	6.90	41.50	173.00	191.00	1534.00	1368.00	780.00
Detroit (DT)	13.00	35.70	477.00	220.00	1566.00	1183.00	788.00
Hartford (HA)	2.50	8.80	68.00	103.00	1017.00	724.00	468.00
Honolulu (HO)	3.60	12.70	42.00	28.00	1457.00	1102.00	637.00
Houston (HS)	16.80	26.60	289.00	186.00	1509.00	787.00	697.00
Kansas City (KC)	10.80	43.20	255.00	226.00	1494.00	955.00	765.00
Los Angeles (LA)	9.70	51.80	286.00	355.00	1902.00	1386.00	862.00
New Orleans (NO)	10.30	39.70	266.00	283.00	1056.00	1036.00	776.00
New York (NY)	9.40	19.40	522.00	267.00	1674.00	1392.00	848.00
Portland (PO)	5.00	23.00	157.00	144.00	1530.00	1281.00	488.00
Tucson (TU)	5.10	22.90	85.00	148.00	1206.00	756.00	483.00
Washington (WA)	12.50	27.60	524.00	217.00	1496.00	1003.00	739.00

Data from the United States Statistical Abstract (1970).

Table 3.1 (b) Dissimilarity matrix calculated from Table 3.1 (a).

	AT	BO	CH	DA	DE	DT	HA	HO	HS	KC	LA	NO	NY	PO	TU	WA
AT	0.00	4.24	2.78	2.79	3.85	3.84	3.29	3.58	2.30	3.21	5.51	3.24	4.87	3.09	2.42	3.58
BO	4.24	0.00	3.59	5.31	4.36	4.78	3.29	3.22	4.04	4.10	6.27	3.98	5.05	4.40	3.40	4.42
CH	2.78	3.59	0.00	3.61	4.39	3.69	3.59	4.66	2.75	3.19	5.56	2.48	4.54	4.22	2.97	3.05
DA	2.79	5.31	3.61	0.00	3.44	2.85	5.09	4.87	1.84	2.27	3.61	2.94	3.94	3.74	3.80	2.90
DE	3.85	4.36	4.39	3.44	0.00	2.48	4.79	3.55	3.37	1.90	2.66	2.47	3.13	2.58	3.69	3.12
DT	3.84	4.78	3.69	2.85	2.48	0.00	5.39	4.62	2.33	1.85	2.88	2.43	1.92	3.58	4.34	1.09
HA	3.29	3.29	3.59	5.09	4.79	5.39	0.00	2.53	4.31	4.65	6.88	4.56	5.69	3.10	1.53	4.86
HO	3.58	3.22	4.66	4.87	3.55	4.62	2.53	0.00	4.02	4.11	5.92	4.55	4.77	2.18	2.52	4.45
HS	2.30	4.04	2.75	1.84	3.37	2.33	4.31	4.02	0.00	2.07	4.31	2.77	3.52	3.51	3.27	1.98
KC	3.21	4.10	3.19	2.27	1.90	1.85	4.65	4.11	2.07	0.00	2.80	1.65	3.25	3.24	3.34	2.19
LA	5.51	6.27	5.56	3.61	2.66	2.88	6.88	5.92	4.31	2.80	0.00	3.40	3.34	4.62	5.62	3.73
NO	3.24	3.98	2.48	2.94	2.47	2.43	4.56	4.55	2.77	1.65	3.40	0.00	3.43	3.63	3.48	2.58
NY	4.87	5.05	4.54	3.94	3.13	1.92	5.69	4.77	3.52	3.25	3.34	3.43	0.00	3.81	4.97	2.07
PO	3.09	4.40	4.22	3.74	2.58	3.58	3.10	2.18	3.51	3.24	4.62	3.63	3.81	0.00	2.32	3.55
TU	2.42	3.40	2.97	3.80	3.69	4.34	1.53	2.52	3.27	3.34	5.62	3.48	4.97	2.32	0.00	3.95
WA	3.58	4.42	3.05	2.90	3.12	1.09	4.86	4.45	1.98	2.19	3.73	2.58	2.07	3.55	3.95	0.00

categorical and continuous variables. Special attention will be paid to proximity measures suitable for data consisting of repeated measures of the same variable, for example taken at different time points. In a clustering context, one important question about an observed proximity matrix is whether it gives any *direct* evidence that the data are, in fact, clustered. This question will be addressed in Chapter 9.

3.2 Similarity measures for categorical data

With data in which all the variables are categorical, measures of similarity are most commonly used. The measures are generally scaled to be in the interval [0, 1], although occasionally they are expressed as percentages in the range 0–100%. Two individuals i and j have a similarity coefficient s_{ij} of unity if both have identical values for all variables. A similarity value of zero indicates that the two individuals differ maximally for all variables. (It would of course be a simple matter to convert a similarity s_{ij} into a dissimilarity δ_{ij} by taking, for example $\delta_{ij} = 1 - s_{ij}$.)

3.2.1 Similarity measures for binary data

The most common type of multivariate categorical data is where all the variables are binary, and a large number of similarity measures have been proposed for such data. All the measures are defined in terms of the entries in a cross-classification of the counts of matches and mismatches in the p variables for two individuals; the general version of this cross-classification is shown in Table 3.2.

A list of some of the similarity measures that have been suggested for binary data is shown in Table 3.3; a more extensive list can be found in Gower and Legendre (1986). The reason for such a large number of possible measures has to do with the apparent uncertainty as to how to deal with the count of zero–zero matches (d in Table 3.2). In some cases, of course, zero–zero matches are completely equivalent to one–one matches, and therefore should be included in the calculated similarity measure. An example is gender, where there is no preference as to which of the two categories should be coded zero or one. But in other cases the inclusion or otherwise of d is more problematic; for example, when the zero category corresponds to the genuine absence of some property, such as wings in a study of insects. The question that then needs to be asked is do the co-absences contain

Table 3.2 Counts of binary outcomes for two individuals.

		Individual i		
	Outcome	1	0	Total
Individual j	1	a	b	$a + b$
	0	c	d	$c + d$
	Total	$a + c$	$b + d$	$p = a + b + c + d$

Table 3.3 Similarity measures for binary data.

Measure	Formula
S1: Matching coefficient	$s_{ij} = (a+d)/(a+b+c+d)$
S2: Jaccard coefficient (Jaccard, 1908)	$s_{ij} = a/(a+b+c)$
S3: Rogers and Tanimoto (1960)	$s_{ij} = (a+d)/[a+2(b+c)+d]$
S4: Sneath and Sokal (1973)	$s_{ij} = a/[a+2(b+c)]$
S5: Gower and Legendre (1986)	$s_{ij} = (a+d)\Big/\left[a+\frac{1}{2}(b+c)+d\right]$
S6: Gower and Legendre (1986)	$s_{ij} = a\Big/\left[a+\frac{1}{2}(b+c)\right]$

useful information about the similarity of two objects? Attributing a large degree of similarity to a pair of individuals simply because they both lack a large number of attributes may not be sensible in many situations. In such cases, measures that ignore the co-absence count d in Table 3.2, for example Jaccard's coefficient (S2) or the coefficient proposed by Sneath and Sokal (S4), might be used (see Table 3.3). If, for example, the presence or absence of a relatively rare attribute such as blood type AB negative is of interest, two individuals with that blood type clearly have something in common, but it is not clear whether the same can be said about two people who do not have the blood type. When co-absences are considered informative, the simple matching coefficient (S1) is usually employed. Measures S3 and S5 are further examples of symmetric coefficients that treat positive matches (a) and negative matches (d) in the same way. The coefficients differ in the weights that they assign to matches and nonmatches. (The question of the weights of variables will be discussed in detail in Section 3.7.)

Sneath and Sokal (1973) point out that there are no hard and fast rules regarding the inclusion or otherwise of negative or positive matches. Each set of data must be considered on its merits by the investigator most familiar with the material involved. The choice of similarity measure on that basis is particularly important, since the use of different similarity coefficients can result in widely different values. While some coefficients can be shown to lead to the same ordering (Gower and Legendre (1986) point out that S2, S4 and S6 are monotonically related, as are S1, S3 and S5), others, for example the matching coefficient and Jaccard's coefficient, can lead to different assessments of the relative similarities of a set of objects.

3.2.2 Similarity measures for categorical data with more than two levels

Categorical data where the variables have more than two levels – eye colour, for example – could be dealt with in a similar way to binary data, with each level of a variable being regarded as a single binary variable. This is not an attractive approach, however, simply because of the large number of 'negative' matches

which will inevitably be involved. A superior method is to allocate a score s_{ijk} of zero or one to each variable k, depending on whether the two individuals i and j are the same on that variable. These scores are then simply averaged over all p variables to give the required similarity coefficient as

$$s_{ij} = \frac{1}{p}\sum_{k=1}^{p} s_{ijk}. \tag{3.1}$$

An interesting modification of this coefficient is found in genetics when evaluating the similarity of DNA sequences. The variable values available for each sequence are the nucleotides (four possible categories: adenine (A), guanine (G), thymine (T) and cytosine (C)) found at each of p positions. An intuitive measure of the dissimilarity, s_{ij}, between two sequences i and j would be the proportions of positions at which both sequences have the same nucleotides, or, in dissimilarity terms, the proportions of positions at which both sequences have different nucleotides. However, genetic similarity between two species is intended to reflect the time lapsed since both species had the last common ancestor. It can be shown that the intuitive measure of dissimilarity increases exponentially in time and reaches an asymptote. Thus, there is not a simple, linear relationship between the dissimilarity measure and time since last common ancestor; a certain change in dissimilarity will reflect varying changes in the genetic proximity, dependent on the value of the measure. To correct for this problem, Jukes and Cantor (1969) first suggested the logarithmic transformed genetic dissimilarity measure

$$\delta_{ij}^{\mathrm{JC}} = -\left(\frac{3}{4}\right)\ln\left[1-\left(\frac{4}{3}\right)\delta_{ij}\right]; \tag{3.2}$$

a different formulation of which is given by Tajima (1993). It is also desirable that a genetic dissimilarity measure weights down transitions (e.g. a mutation of A to G) which occur far more frequently than tranversions (e.g. a mutation from A to T). Modifications of the Jukes–Cantor dissimilarity to this effect have been suggested by Kimura (1980).

An alternative definition of similarity for categorical variables is to divide all possible outcomes of the kth variable into mutually exclusive subsets of categories, allocate s_{ijk} to zero or one depending on whether the two categories for individuals i and j are members of the same subset, and then determine the proportion of shared subsets across variables. This similarity measure has been applied in the study of how languages are related. Linguists have identified sets of core-meaning word categories which are widely understood across cultures – such as 'water', 'child' or 'to give' – assembled in the so-called Swadesh 200-word list. For each language under study and each meaning on the Swadesh list, words can be gathered, providing an n languages $\times$ p Swadesh meanings matrix of words. Linguists can further divide the words from different languages for the same Swadesh meaning into mutually exclusive cognate classes, where two words are considered

'cognate' if they have the same meaning when narrowly defined and have been generated by sound changes from a common ancestor. Two words from different languages are then assigned $s_{ijk} = 1$ if they are members of the same cognate class, and the proportion of cognate classes shared by two languages is a measure of their relatedness. For more details see Dyen *et al.* (1992) and Atkinson *et al.* (2005).

3.3 Dissimilarity and distance measures for continuous data

When all the recorded variables are continuous, proximities between individuals are typically quantified by dissimilarity measures or distance measures, where a dissimilarity measure, δ_{ij}, is termed a *distance measure* if it fulfils the *metric (triangular) inequality*

$$\delta_{ij} + \delta_{im} \geq \delta_{jm} \tag{3.3}$$

for pairs of individuals *ij*, *im* and *jm*. An $n \times n$ matrix of dissimilarities, $\mathbf{\Delta}$, with elements δ_{ij}, where $\delta_{ii} = 0$ for all i, is said to be *metric*, if the inequality (3.3) holds for all triplets (i, j, m). From the metric inequality follows that the dissimilarity between individuals i and j is the same as that between j and i, and that if two points are close together then a third point has a similar relation to both of them. Metric dissimilarities are by definition nonnegative. (In the remainder of the text we refer to metric dissimilarity measures specifically as distance measures and denote the $n \times n$ matrix of distances $\mathbf{D}$, with elements, d_{ij}.)

A variety of measures have been proposed for deriving a dissimilarity matrix from a set of continuous multivariate observations. Commonly used dissimilarity measures are summarized in Table 3.4. More extensive lists can be found in Gower (1985), Gower and Legendre (1986) or Jajuga *et al.* (2003). All distance measures are formulated so as to allow for differential weighting of the quantitative variables (in Table 3.4, the w_k, $k = 1, \ldots, p$ denote the nonnegative weights of the p variables). We defer the issue of how these weights should be chosen until Section 3.7 and simply assume at this stage that the variables are weighted equally (all $w_k = 1$).

Proposed dissimilarity measures can be broadly divided into distance measures and correlation-type measures. The distance measure most commonly used is *Euclidean distance* (D1)

$$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right]^{1/2}, \tag{3.4}$$

where x_{ik} and x_{jk} are, respectively, the kth variable value of the p-dimensional observations for individuals i and j. This distance measure has the appealing property that the d_{ij} can be interpreted as physical distances between two

Table 3.4 Dissimilarity measures for continuous data.

Measure	Formula
D1: Euclidean distance	$d_{ij} = \left[\sum_{k=1}^{p} w_k^2 \left(x_{ik} - x_{jk}\right)^2\right]^{1/2}$
D2: City block distance	$d_{ij} = \sum_{k=1}^{p} w_k \left\lvert x_{ik} - x_{jk}\right\rvert$
D3: Minkowski distance	$d_{ij} = \left(\sum_{k=1}^{p} w_k^r \left\lvert x_{ik} - x_{jk}\right\rvert^r\right)^{1/r} \quad (r \geq 1)$
D4: Canberra distance (Lance and Williams, 1966)	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^{p} w_k \left\lvert x_{ik} - x_{jk}\right\rvert / \left(\left\lvert x_{ik}\right\rvert + \left\lvert x_{jk}\right\rvert\right) & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$
D5: Pearson correlation	$\delta_{ij} = \left(1 - \phi_{ij}\right)/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k \left(x_{ik} - \bar{x}_{i\cdot}\right)\left(x_{jk} - \bar{x}_{j\cdot}\right) \Big/ \left[\sum_{k=1}^{p} w_k \left(x_{ik} - \bar{x}_{i\cdot}\right)^2 \sum_{k=1}^{p} w_k \left(x_{jk} - \bar{x}_{j\cdot}\right)^2\right]^{1/2}$ where $\bar{x}_{i\cdot} = \sum_{k=1}^{p} w_k x_{ik} \Big/ \sum_{k=1}^{p} w_k$
D6: Angular separation	$\delta_{ij} = \left(1 - \phi_{ij}\right)/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} \Big/ \left(\sum_{k=1}^{p} w_k x_{ik}^2 \sum_{k=1}^{p} w_k x_{jk}^2\right)^{1/2}$

p-dimensional points $\mathbf{x}_i' = (x_{i1}, \ldots x_{ip})$ and $\mathbf{x}_j' = (x_{j1}, \ldots x_{jp})$ in Euclidean space. Formally this distance is also known as the l_2 norm. The city block distance or l_1 norm (D2) describes distances on a rectilinear configuration. It is also referred to as the *taxicab* (Krause, 1975), *rectilinear* (Brandeau and Chiu, 1988) or *Manhattan* (Larson and Sadiq, 1983) distance, because it measures distances travelled in street configuration. Both the Euclidean ($r = 2$) or the city block ($r = 1$) distance are special cases of the general *Minkowski distance* (D3) or l_r norm.

The *Canberra distance measure* (D4) is very sensitive to small changes close to $x_{ik} = x_{jk} = 0$. It is often regarded as a generalization of the dissimilarity measure for binary data. In this context D4 can be divided by the number of variables, p, to ensure a dissimilarity coefficient in the interval [0, 1], and it can then be shown that D4 for binary variables is just one minus the matching coefficient S1 in Table 3.3 (Gower and Legendre, 1986).

It has often been suggested (e.g. Strauss *et al.*, 1973; Cliff *et al.*, 1995) that the correlation, ϕ_{ij}, between the p-dimensional observations of the ith and jth subject can be used to quantify the similarity between them. Measures D5 and D6 are

examples of the derivation of dissimilarity measures from correlation coefficients. Measure D5 employs the Pearson correlation coefficient, and measure D6 the cross-product index. Since for correlation coefficients we have that

$$-1 \leq \phi_{ij} \leq 1, \tag{3.5}$$

with the value '1' reflecting the strongest possible positive relationship and the value '-1' the strongest possible negative relationship, these coefficients can be transformed into dissimilarities, δ_{ij}, within the interval [0, 1] as shown in Table 3.4. The correlation coefficient, ϕ_{ij}, used to construct D6 is the cosine of the angle between two vectors connecting the origin to the ith and jth p-dimensional observation respectively. A similar interpretation exists for D5, except now the vectors start from the 'mean' of the p-dimensional observation.

The use of correlation coefficients in this context is far more contentious than its noncontroversial role in assessing the linear relationship between two variables based on a sample of n observations on the variables. When correlations between two individuals are used to quantify their similarity the rows of the data matrix are standardized, not its columns. Clearly, when variables are measured on different scales the notion of a difference between variable values, and consequently that of a mean variable value or a variance, is meaningless (for further critiques see Fleiss and Zubin, 1969 or Jardine and Sibson, 1971). In addition, the correlation coefficient is unable to measure the difference in size between two observations. For example, the three-dimensional data points $\mathbf{x}_i' = (1, 2, 3)$ and $\mathbf{x}_j' = (3, 6, 9)$ have correlation $\phi_{ij} = 1$, while $\mathbf{x}_j$ is three times the size of $\mathbf{x}_i$. However, the use of a correlation coefficient can be justified for situations where all the variables have been measured on the same scale and the precise values taken are important only to the extent that they provide information about the subject's relative profile. For example, in classifying animals or plants the absolute sizes of the organisms or their parts are often less important than their shapes. In such studies the investigator requires a dissimilarity coefficient that takes the value zero if and only if two individuals' profiles are multiples of each other. The angular separation dissimilarity measure has this property.

Of the dissimilarity measures introduced here, as the name suggests, measures D1 to D4 can be shown to be general distance measures (Gower and Legendre, 1986). Correlation-type measures D5 and D6 can also be shown to result in metric dissimilarity matrices (Anderberg, 1973).

Large data sets of continuous variables measured on the same scale are currently being generated in genetic research due to recent advances in microarray-based genomic surveys and other high-throughput approaches. These techniques produce masses of continuous expression numbers for thousands or tens of thousands of genes under hundreds of experimental conditions, which are increasingly collected in reference databases or 'compendia' of expression profiles. The conditions may refer to times in a cell growth experiment or simply a collection of a number of experimental interventions. The pattern of expression of a gene across a range of experimental conditions informs on the status of cellular processes, and

researchers are often interested in identifying genes with similar function. In a clustering context this means that we are dealing with a (very large) matrix of continuous expression values for n genes (rows) under p conditions (columns). Since the intuitive biological notion of 'coexpression' between two patterns seems to focus on shape rather than magnitude, the correlation coefficient has been revived as a similarity measure in this area, although other measures for continuous outcomes have also been put forward (Eisen *et al.*, 1998). However, the Pearson correlation is sensitive to outliers, and expression data is notoriously noisy. This has prompted a number of suggestions from this area for modifying correlation coefficients when used as similarity measures; for example, robust versions of correlation coefficients (Hardin *et al.*, 2007) such as the jacknife correlation (Heyer *et al.*, 1999), or altogether more general association coefficients such as the *mutual information distance measure* (Priness *et al.*, 2007). (For more on proximity measures for data with variables measured on the same scale see Section 3.5.)

A desirable property of dissimilarity matrices is that they are *Euclidean*, where an $n \times n$ dissimilarity matrix, $\mathbf{\Delta}$, with elements δ_{ij}, is said to be Euclidean if the n individuals can be represented as points in space such that the Euclidean distance between points i and j is δ_{ij}. The Euclidean property is appealing since, like the Euclidean distance measure, it allows the interpretation of dissimilarities as physical distances. If a dissimilarity matrix is Euclidean then it is also metric, but the converse does not follow. As an example, Gower and Legendre (1986) presented the following dissimilarity matrix

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 2 & 2 & 1.1 \\ 2 & 0 & 2 & 1.1 \\ 2 & 2 & 0 & 1.1 \\ 1.1 & 1.1 & 1.1 & 0 \end{pmatrix} \end{array}.$$

The matrix arises from a situation in which the observations of individuals 1, 2 and 3 form an equilateral triangle of side length 2 units, with the position of individual 4 being equidistant (1.1 units) from each of the other three positions (Figure 3.1). It can be shown that $\mathbf{D}$ is metric simply by verifying the triangular inequality for all possible triplets of individuals. If this configuration is to be Euclidean then the smallest distance that the position of individual 4 can be from the other points is when it is coplanar with them and at their centroid. But this corresponds to a minimal distance of $2\sqrt{3}/3 = 1.15$, which is greater than 1.1. Thus $\mathbf{D}$ is metric but not Euclidean.

Gower and Legendre (1986) also show that, out of the distance measures D1–D4, only the Euclidean distance itself (D1) produces Euclidean dissimilarity matrices.

In many cases, similarity and dissimilarity matrices can be transformed to be Euclidean. Gower (1966) showed that if a similarity matrix, $\mathbf{S}$, with elements s_{ij}, is nonnegative definite, then the matrix $\mathbf{D}$, with elements d_{ij} defined as

$$d_{ij} = \sqrt{(1 - s_{ij})} \tag{3.6}$$

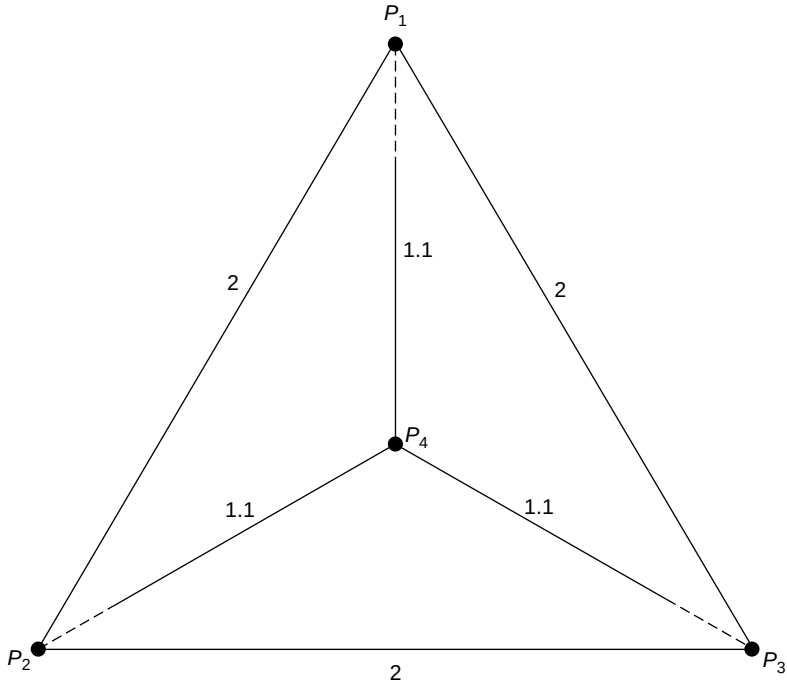


Figure 3.1 *An example of a set of distances that satisfy the metric inequality but which have no Euclidean representation.* (Reproduced with permission from Gower and Legendre, 1986.)

is Euclidean. All the similarity coefficients given in Table 3.3 (except S6) can be shown to have a positive semidefinite similarity matrix (Gower and Legendre, 1986); hence the corresponding distances defined according to Equation (3.6) are Euclidean. Furthermore, for any dissimilarity matrix, $\boldsymbol{\Delta}$, with elements δ_{ij}, constants c_1 and c_2 exist, such that the matrix $\mathbf{D}$, with elements

$$d_{ij} = \sqrt{\left(\delta_{ij}{}^2 + c_1\right)} \quad \text{(Lingoes, 1971)} \tag{3.7}$$

or

$$d_{ij} = \delta_{ij} + c_2 \quad \text{(Cailliez, 1983)} \tag{3.8}$$

is Euclidean (both Lingoes and Cailliez comment on how the relevant constants can be found). Further investigations of the relationships between dissimilarity matrices, distance matrices and Euclidean matrices are carried out in Gower and Legendre (1986) and Cailliez and Kuntz (1996).

3.4 Similarity measures for data containing both continuous and categorical variables

There are a number of approaches to constructing proximities for mixed-mode data, that is, data in which some variables are continuous and some categorical. One possibility would be to dichotomize all variables and use a similarity measure for binary data; a second to rescale all the variables so that they are on the same scale by replacing variable values by their ranks among the objects and then using a measure for continuous data (e.g. Wright *et al.*, 2003); and a third to construct a dissimilarity measure for each type of variable and combine these, either with or without differential weighting into a single coefficient (e.g. Bushel *et al.*, 2007). More complex suggestions are made in Estabrook and Rodgers (1966), Gower (1971), Legendre and Chodorowski (1977), Lerman (1987) and Ichino and Yaguchi (1994). Here we shall concentrate on the similarity measure proposed by Gower (1971). Gower's general similarity measure is given by

$$s_{ij} = \sum_{k=1}^{p} w_{ijk} s_{ijk} \Big/ \sum_{k=1}^{p} w_{ijk}, \tag{3.9}$$

where s_{ijk} is the similarity between the ith and jth individual as measured by the kth variable, and w_{ijk} is typically one or zero depending on whether or not the comparison is considered valid. The value of w_{ijk} is set to zero if the outcome of the kth variable is missing for either or both of individuals i and j. In addition, w_{ijk} can be set to zero if the kth variable is binary and it is thought appropriate to exclude negative matches. For binary variables and categorical variables with more than two categories, the component similarities, s_{ijk}, take the value one when the two individuals have the same value and zero otherwise. For continuous variables, Gower suggests using the similarity measure

$$s_{ij} = 1 - \left|x_{ik} - x_{jk}\right| / R_k, \tag{3.10}$$

where $R_k$is the range of observations for the kth variable. (In other words, the city block distance is employed after scaling the kth variable to unit range.)

To illustrate the use of Gower's coefficient we consider data from a survey about students' preferences and attitudes towards video and computer games provided by Nolan and Speed (1999). The target population was university students who would be taking part in statistics labs as part of their learning. Ninety-one students took part in the survey in 1994. The records of 11 students including one that had not yet experienced video games (subject 82) are shown in Table 3.5. We will use Gower's general similarity to measure the resemblance between subject's attitudes towards video games. We did not wish to exclude any matches on our binary variables (*ever*, *busy*, *educ*), and therefore in the absence of missing values set $w_{ijk} = 1$ for all i, j, k. For example, the similarity between subjects 1 and 2 was calculated as

Table 3.5 Results from video games survey (reproduced with permission from Nolan and Speed, 1999).

Subject	Preferences and attitudes towards video games							Sex	Age	Grade
	ever	*time*	*like*	*where*	*freq*	*busy*	*educ*			
1	Yes	2	Somewhat	H. com.	Weekly	No	Yes	Female	19	A
2	Yes	0	Somewhat	H. com.	Monthly	No	No	Female	18	C
3	Yes	0	Somewhat	Arcade	Monthly	No	No	Male	19	B
4	Yes	0.5	Somewhat	H. com.	Monthly	No	Yes	Female	19	B
5	Yes	0	Somewhat	H. com.	Semesterly	No	Yes	Female	19	B
6	Yes	0	Somewhat	H. sys.	Semesterly	No	No	Male	19	B
7	Yes	0	Not really	H. com.	Semesterly	No	No	Male	20	B
8	Yes	0	Somewhat	H. com.	Semesterly	No	No	Female	19	B
9	Yes	2	Somewhat	H. sys.	Daily	Yes	Yes	Male	19	A
10	Yes	0	Somewhat	H. com.	Semesterly	No	Yes	Male	19	A
82	No	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Male	19	A

ever = Have you ever played video games before?
time = Time spent playing video games in week prior to survey, in hours.
like = Do you like to play video games?
where = Where do you play video games? (H. com. = home computer; H. sys. = home system).
freq = How often do you play video games?
busy = Do you play when you are busy?
educ = Is playing video games educational?
Age = age in years.
Grade = grade expected in course.
n.a. = not applicable.

$$s_{1,2} = \frac{1\times 1+1\times(1-|2-0|/30)+1\times(1-|2-2|/3)+1\times 1+1\times(1-|2-3|/3)+1\times 1+1\times 0}{1+1+1+1+1+1+1} = 0.8. \tag{3.11}$$

(Note that here we have treated ordinal variables (*like*, *freq*) as if their ranks were on an interval scale – this is not ideal but seemed preferable to treating them as nominal variables and only declaring two values as similar when they have exactly the same rank). In contrast, when comparing any subject to subject 82 who had never experienced video games and therefore could not comment on his liking etc. of them, weights were set to invalid ($w_{i82k} = 0$ for all i and $k = 2, \ldots, 7$). For example, the similarity between subjects 1 and 82 is

$$s_{1,82} = \frac{1\times 0+0+0+0+0+0+0}{1+0+0+0+0+0+0} = 0. \tag{3.12}$$

Part of the dissimilarity matrix is shown in Table 3.6.

Gower (1971) shows that, when there are no missing values, the similarity matrix resulting from using his suggested similarity coefficient is positive semi-definite, and hence the dissimilarity matrix defined according to Equation (3.6) is Euclidean.

Table 3.6 Part of Gower's dissimilarity matrix for video games survey data.

	1	2	3	4	5	6	7	8	9	10	82
1	0.00										
2	0.20	0.00									
3	0.34	0.14	0.00								
4	0.05	0.15	0.29	0.00							
5	0.10	0.19	0.33	0.05	0.00						
6	0.39	0.19	0.19	0.34	0.29	0.00					
7	0.30	0.10	0.24	0.24	0.19	0.19	0.00				
8	0.25	0.05	0.19	0.19	0.14	0.14	0.05	0.00			
9	0.33	0.53	0.53	0.39	0.44	0.44	0.63	0.58	0.00		
10	0.10	0.19	0.33	0.05	0.00	0.29	0.19	0.14	0.44	0.00	
82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Most general-purpose statistical software such as `Stata` (www.stata.com/), `SAS` (http://support.sas.com/), `SPSS` (www.spss.com/) or `R` (www.r-project.org/; see packages `cluster`, `clusterSim` and `proxy`) implement a number of measures for converting a two-mode data matrix into a one-mode (dis)similarity matrix. Most of the (dis)similarity measures listed in Tables 3.3 and 3.4 for binary and continuous data are available. Gower's similarity measure is provided in `Stata` and `R` (function `daisy` in package `cluster`).

3.5 Proximity measures for structured data

In some applications, the multivariate $n \times p$ matrix, $\mathbf{X}$, consists of *repeated measures* of the same outcome variable but under different conditions, measuring different concepts, at different times, at a number of spatial positions, etc., and the appearance of such data in a genetics context has been mentioned previously in Section 3.3. A simple example in the time domain is given by measurements of, say, the heights of children obtained each month over several years. Such data are of a special nature in that all variables are measured on the same scale and the individual data vectors are referenced by another p-dimensional variable such as condition, time or space. For example, for repeated measures made at times $t_1, t_2, \ldots, t_p$, the reference variable is simply the vector of associated times $(t_1, t_2, \ldots, t_p)'$ – for the children's height example it would contain the months at which heights were measured. For repeated measures obtained under different experimental conditions, say A, B or C, the reference vector would be of the form $(\mathrm{A}, \ldots, \mathrm{A}, \mathrm{B}, \ldots, \mathrm{B}, \mathrm{C}, \ldots, \mathrm{C})'$. In this section, we discuss proximity measures that are particularly suited to such *structured data*; that is, measures that make use of the fact that all the variable values arise from the same data space *and* acknowledge and exploit the existence of a reference variable. Conceptualizing repeated measures as a data matrix with an accompanying reference vector is useful when it comes to determining appropriate summaries of an object's variable values and resulting

measures of dissimilarities between objects, as we shall see below. It also helps to model the means and covariances of the repeated measures by a reduced set of parameters, as we shall see in Chapter 7 which gives a detailed account of model-based cluster analysis of structured data.

The simplest and perhaps most commonly used approach to exploiting the reference variable is in the construction of a reduced set of relevant summaries per object which are then used as the basis for defining object similarity. What constitutes an appropriate summary measure will depend on the context of the substantive research. Here we look at some approaches for choosing summary measures and resulting proximity measures for the most frequently encountered reference vectors – 'time', 'experimental condition' and 'underlying factor'.

When the reference variable is time and the functional form of individual time curves is known, then parameter estimates obtained by fitting linear or nonlinear regression models to individual time courses may represent such a set of summaries (see, e.g., Bansal and Sharma, 2003). Returning to the children example, if it were reasonable to assume linear growth over the study period, then a child's growth curve could be described fully by only two summary statistics – the intercept and slope of the curve. We could estimate these two parameters by regressing a child's height measurements against the reference variable (assessment months). The proximity between the growth curves of two children could then be captured by a suitable measure of (dis)similarity between children's regression coefficients, for example the Euclidean distance between the children's standardized regression coefficients. (We will consider the issue of variable standardization in Section 3.8.)

When the reference variable allocates the repeated measures into a number of classes, as is for example the case when gene expressions are obtained over a range of experimental conditions, then a typical choice of summary measure is simply an object's (gene's) mean variable value (mean gene expression) per class (condition). The summary approach can be expanded by using not only the summary measures of interest but also the precision of these estimates in the construction of proximities. For example, Hughes *et al.* (2000) construct a similarity measure for genes by summarizing expression levels across microarrays using mean levels per experimental condition. These authors then measure the similarity between two such sets of means by a weighted correlation coefficient, with weights chosen inversely proportional to the respective standard errors of the means. (We will consider the issue of variable weighting in Section 3.7.)

Often, structured data arise when the variables can be assumed to follow a known *factor model*. (Factor models are described in, for example, Bollen (1989) or Loehlin (2004); and a more comprehensive account will be given in Chapter 7). Briefly, under a so-called *confirmatory factor analysis model,* each variable or item can be allocated to one of a set of underlying factors or concepts. The factors cannot be observed directly but are 'indicated' by a number of items, each of which is measured on the same scale. Many questionnaires employed in the behaviour and social sciences produce multivariate data of this type. For example, the well-known Hospital Anxiety and Depression Scale (HADS; Zigmond and Snaith, 1983) assesses patients on 14 items, each of which is scored

on a four-point Likert scale (1 = 'most of the time', 2 = 'a lot of the time', 3 = 'occasionally' and 4 = 'not at all'). Seven of the items have been shown to capture the unobserved concept 'depression', while the remaining seven items target the factor 'anxiety'. Thus the data generated under such models are structured, as all items are measured on the same scale and the structure can be identified by a reference vector whose entries are the factors that have been targeted. For example, the $n \times 14$-dimensional structured data generated by administering the HADS to a sample of n subjects would have a reference vector of the form $(\text{ANX}, \text{DEP}, \text{ANX}, \ldots, \text{DEP})'$. A categorical factor reference variable can be used in the same way as a categorical condition reference variable to construct appropriate summaries per factor level. Returning to the HADS data, we could summarize the data by an $n \times 2$-dimensional matrix consisting of patients' means (or medians or totals) over anxiety and depression items, and then measure the proximity between two subjects by the distance between their bivariate summaries.

Finally, note that the summary approach, while typically used with continuous variables, is not limited to variables on an interval scale. The same principles can be applied to deal with categorical data. The difference is that summary measures now need to capture relevant aspects of the distribution of the categorical variables over repeated measures. Summaries such as quantiles (ordinal only), proportions of particular categories or the mode of the distribution would be obvious choices. For example, consider a data matrix consisting of variables indicating the absence/presence of a number of symptoms of psychiatric patients. If the symptoms can be grouped into domains such as 'cognition', 'executive functioning', etc., then one approach for measuring patient dissimilarities would be to determine for each patient and symptom domain the proportion of possible symptoms that are present, and then to measure the distances between the patients' domain proportions.

Rows of $\mathbf{X}$ which represent ordered lists of elements – that is all the variables provide a categorical outcome and these outcomes can be aligned in one dimension – are more generally referred to as *sequences*. Sequences occur in many contexts: in genetics, DNA or protein sequences may need to be aligned (Sjőlander, 2004); letters in a word form a sequence; or events such as jobs or criminal careers may need to be considered in a temporal context. Sequences produce categorical repeated measures data, with their structure being captured by a reference vector which indicates a variable's position in the dimension in which alignment takes place (e.g. position in word, order in time). Some of the approaches described above for repeated measures in the temporal domain could be used to construct proximities, but recent interest in sequences, in particular in the field of genetics, has prompted the development of algorithms for determining dissimilarity measures which specifically exploit the aligned nature of the categorical data. We will introduce some of the more popular ones below.

An example of a set of sequences is given by Brinsky-Fay *et al.* (2006). These authors consider artificial sequences of quarterly employment states over a period of up to 36 months for 500 graduates. Here we look at the first 10 months only. An individual's monthly employment state was classed as one of five possible categories: 'higher education', 'vocational education', 'employment',

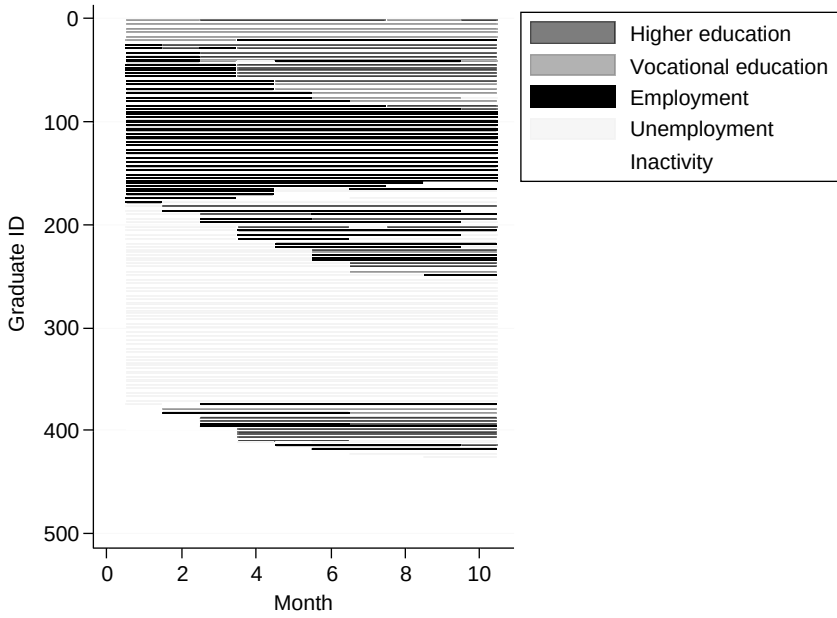


Figure 3.2 *Sequence index plot for employment.* (Reproduced with permission from Brinsky-Fay *et al.*, 2006.)

'unemployment' and 'inactivity', and it was deemed important to take account of the temporal ordering when judging the resemblance between two sequences. For a small number of categories, sequence data is readily summarized by means of a sequence index plot (Scherer, 2001), in which the x-axis represents the order dimension (here month), the y-axis refers to individual sequences, and the categories observed for each sequence are colour coded. Figure 3.2 shows the sequence index plot for the employment data. We can see that the majority of individuals in our sample start off being 'unemployed' or 'inactive', with more than half of these individuals never changing their status over the 10-month period.

So-called *sequence analysis* in an area of research in sociology and psychology that centres on problems of events and actions in their temporal context and includes the measurements of similarities between sequences (see, e.g., Abbott, 1995; Abbott and Tsay, 2000). Perhaps the most popular measure of dissimilarity between two sequences is the *Levenshtein distance* (Levenshtein, 1966), which has received a lot of interest in information theory and computer science, and counts the minimum number of operations needed to transform one sequence of categories into another, where an operation is an insertion, a deletion or a substitution of a single category. Such operations are only applicable on aligned sets of categories, and so counting the number of operations leads to a dissimilarity measure for sequences. Each operation can be assigned a penalty weight (a typical choice would be to give double the penalty to a substitution than to an insertion or

deletion.) The measure is also sometimes called the 'edit distance', due to its application in determining the similarity between words for spell checkers. Care needs to be taken to deal with gaps in the sequence or sequences of variable lengths. (For suitable approaches under these scenarios see Abbott and Tsay, 2000.)

Optimal matching algorithms (OMAs) need to be employed to find the minimum number of operations required to match one sequence to another. One such algorithm for aligning sequences is the Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), which is commonly used in bioinformatics applications to align protein or nucleotide sequences. `Stata`'s implementation of this algorithm (`sqom` command) was used to generate a 500×500 distance matrix for the employment sequences; an extract for 10 sequences is shown in Table 3.7. The number of operations needed to convert one sequence into another varies widely. For example, while only 2 operations are required to convert the sequence for graduate 57 (vocational education during the whole 10-month period) into the sequence for graduate 397 (vocational education in all months except 1), 20 operations are needed to change sequence 57 into sequence 1 (mix of higher education and employment).

The *Jaro similarity measure* (Jaro, 1995) is a related measure of similarity between sequences of categories often used to delete duplicates in the area of record linkage. It makes use of the alignment information by counting the number, m, of matching characters and the number, t, of transpositions. Two categories are considered matching if they are no further than $p/2 - 1$ positions away from each other on the alignment scale (e.g. letter number). A transition is a swap of two categories within a sequence. Then the Jaro similarity is defined as

$$s^{\text{Jaro}} = \frac{1}{3}\left(\frac{2m}{p} + \frac{m-t}{m}\right), \tag{3.13}$$

with the Jaro–Winkler measure rescaling this index to give more favourable ratings to sequences that match from the beginning for a set prefix length (see Winkler, 1999).

Table 3.7 Part of Levenshtein distance matrix for employment history sequences.

Graduate ID	397	112	381	57	247	442	269	97	50	1
397	0									
112	12	0								
381	14	4	0							
57	2	12	14	0						
247	8	12	14	8	0					
442	14	14	14	14	14	0				
269	16	16	16	16	16	2	0			
97	18	18	18	18	10	18	20	0		
50	18	8	6	20	20	18	18	20	0	
1	18	8	6	20	20	16	16	20	2	0

3.6 Inter-group proximity measures

So far we have been concerned with measuring the proximity between two individuals. As we will see in the following chapters, in clustering applications it also becomes necessary to consider how to measure the proximity between groups of individuals. There are two basic approaches to defining such inter-group proximities. Firstly, the proximity between two groups might be defined by a suitable summary of the proximities between individuals from either group. Secondly, each group might be described by a representative observation by choosing a suitable summary statistic for each variable, and the inter-group proximity defined as the proximity between the representative observations.

3.6.1 Inter-group proximity derived from the proximity matrix

For deriving inter-group proximities from the matrix of inter-individual proximities, there are a variety of possibilities. We could, for example, take the smallest dissimilarity between any two individuals, one from each group. In the context of distances, this would be referred to as *nearest-neighbour distance* and is the basis of the clustering technique known as *single linkage* (see Chapter 4). The opposite of nearest-neighbour distance is to define the inter-group distances as the largest distance between any two individuals, one from each group. This is known as *furthest-neighbour distance* and constitutes the basis of the *complete linkage* cluster method (again see Chapter 4). Instead of employing the extremes, the inter-group dissimilarity can also be defined as the average dissimilarity between individuals from both groups. Such a measure is used in *group average clustering* (see Chapter 4).

3.6.2 Inter-group proximity based on group summaries for continuous data

One obvious method for constructing inter-group dissimilarity measures for continuous data is to simply substitute group means (also known as the *centroid*) for the variable values in the formulae for inter-individual measures such as the Euclidean distance or the city block distance (Table 3.4). If, for example, group A has mean vector $\bar{\mathbf{x}}'_{\mathrm{A}} = \left(\bar{x}_{\mathrm{A}1}, \ldots, \bar{x}_{\mathrm{A}p}\right)$ and group B mean vector $\bar{\mathbf{x}}'_{\mathrm{B}} = (\bar{x}_{\mathrm{B}1}, \ldots, \bar{x}_{\mathrm{B}p})$, then the Euclidean inter-group distance would be defined as

$$d_{\mathrm{AB}} = \left[\sum_{k=1}^{p} \left(\bar{x}_{\mathrm{A}k} - \bar{x}_{\mathrm{B}k}\right)^2\right]^{1/2}. \tag{3.14}$$

More appropriate, however, might be measures which incorporate, in one way or another, knowledge of within-group variation. One possibility is to use

Mahalanobis's (1936) *generalized distance*, D^2, given by

$$D^2 = (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)' \mathbf{W}^{-1} (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B), \tag{3.15}$$

where $\mathbf{W}$ is the pooled within-group covariance matrix for the two groups. When correlations between variables within groups are slight, D^2 will be similar to the squared Euclidean distance calculated on variables standardized by dividing by their within-group standard deviation. Thus, the Mahalanobis distance increases with increasing distances between the two group centres and with decreasing within-group variation. By also employing within-group correlations, the Mahalanobis distance takes account of the (possibly nonspherical) shape of the groups.

The use of Mahalanobis D^2 implies that the investigator is willing to assume that the covariance matrices are at least approximately the same in the two groups. When this is not so, D^2 is an inappropriate inter-group measure, and for such cases several alternatives have been proposed. Three such distance measures were assessed by Chaddha and Marcus (1968), who concluded that a measure suggested by Anderson and Bahadur (1962) had some advantage. This inter-group distance measure is defined by

$$\delta_{AB} = \max_t \frac{2\mathbf{b}_t'\mathbf{d}}{\left(\mathbf{b}_t'\mathbf{W}_A\mathbf{b}_t\right)^{1/2} + \left(\mathbf{b}_t'\mathbf{W}_B\mathbf{b}_t\right)^{1/2}}, \tag{3.16}$$

where $\mathbf{W}_A$ and $\mathbf{W}_B$ are the $p \times p$ sample covariance matrices in group A and B respectively, $\mathbf{d} = \bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B$ and $\mathbf{b}_t = (t\mathbf{W}_A + (1-t)\mathbf{W}_B)^{-1}\mathbf{d}$.

Another alternative is the *normal information radius* (NIR) suggested by Jardine and Sibson (1971). This distance is defined as

$$\text{NIR} = \frac{1}{2}\log_2\left\{\frac{\det\left[\frac{1}{2}(\mathbf{W}_A + \mathbf{W}_B)\right] + \frac{1}{4}(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)'(\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)}{\det(\mathbf{W}_A)^{1/2}\det(\mathbf{W}_B)^{1/2}}\right\}. \tag{3.17}$$

When $\mathbf{W}_A = \mathbf{W}_B = \mathbf{W}$ this is reduced to

$$\text{NIR} = \frac{1}{2}\log_2\left(1 + \frac{1}{4}D^2\right), \tag{3.18}$$

where D^2 is the Mahalanobis distance. The NIR can therefore be regarded as providing a generalization of D^2 to the heterogeneous covariance matrices case.t

3.6.3 Inter-group proximity based on group summaries for categorical data

Approaches for measuring inter-group dissimilarities between groups of individuals for which categorical variables have been observed have been considered by a number of authors. Balakrishnan and Sanghvi (1968), for example, proposed a

dissimilarity index of the form

$$G^2 = \sum_{k=1}^{p} \sum_{l=1}^{c_k+1} \frac{(p_{\mathrm{A}kl} - p_{\mathrm{B}kl})^2}{p_{kl}}, \tag{3.19}$$

where $p_{\mathrm{A}kl}$ and $p_{\mathrm{B}kl}$ are the proportions of the lth category of the kth variable in group A and B respectively, $p_{kl} = \frac{1}{2}(p_{\mathrm{A}kl} + p_{\mathrm{B}kl})$, $c_k + 1$ is the number of categories for the kth variable and p is the number of variables.

Kurczynski (1969) suggested adapting the generalized Mahalanobis distance, with categorical variables replacing quantitative variables. In its most general form, this measure for inter-group distance is given by

$$D_{\mathrm{p}}^2 = (\mathbf{p}_{\mathrm{A}} - \mathbf{p}_{\mathrm{B}})' \, \mathbf{W}_{\mathrm{p}}^{-1} (\mathbf{p}_{\mathrm{A}} - \mathbf{p}_{\mathrm{B}}), \tag{3.20}$$

where $\mathbf{p}_{\mathrm{A}} = (p_{\mathrm{A}11}, p_{\mathrm{A}12}, \ldots, p_{\mathrm{A}1c_1}, p_{\mathrm{A}21}, p_{\mathrm{A}22}, \ldots, p_{\mathrm{A}2c_2}, \ldots, p_{\mathrm{A}k1}, p_{\mathrm{A}k2}, \ldots, p_{\mathrm{A}kc_k})'$ contains sample proportions in group A and $\mathbf{p}_{\mathrm{B}}$ is defined in a similar manner, and $\mathbf{W}_{\mathrm{p}}$ is the $m \times m$ common sample covariance matrix, where $m = \sum_{k=1}^{p} c_k$. Various alternative forms of this dissimilarity measure may be derived, depending on how the elements of $\mathbf{W}_{\mathrm{p}}$ are calculated. Kurczynski (1970), for example, shows that if each variable has a multinomial distribution, and the variables are independent of one another, then the dissimilarity measure in (3.20) is equal to the dissimilarity measure defined in (3.19). Kurczynski (1969) also demonstrated some important applications of inter-group distance measures for categorical data where the variables are gene frequencies.

3.7 Weighting variables

To weight a variable means to give it greater or lesser *importance* than other variables in determining the proximity between two objects. All of the distance measures in Table 3.4 are, in fact, defined in such a way as to allow for differential weighting of the quantitative variables. The question is 'How should the weights be chosen?' Before we discuss this question, it is important to realize that the selection of variables for inclusion into the study already presents a form of weighting, since the variables not included are effectively being given the weight zero. Similarly, the common practice of standardization, which we shall look at in detail in the next section, can be viewed as a special case of weighting the variables.

The weights chosen for the variables reflect the importance that the investigator assigns to the variables for the classification task. This assignment might either be the result of a judgement on behalf of the investigator or of the consideration of some aspect of the data matrix, **X**, itself. In the former case, when the investigator determines the weights, this can be done by specifying the weights directly or indirectly. The methods proposed by Sokal and Rohlf (1980) and Gordon (1990) are examples of indirect weight assignment. These authors obtain perceived

dissimilarities between selected (possibly hypothetical) objects and also observe variable values for those objects. They then model the dissimilarities using the underlying variables and weights that indicate their relative importance. The weights that best fit the perceived dissimilarities are then chosen.

A common approach to determining the weights from the data matrix, $\mathbf{X}$, is to define the weights w_k of the kth variable to be inversely proportional to some measure of variability in this variable. This choice of weights implies that the importance of a variable decreases when its variability increases. Several measures of variability have been used to define the weights. For a continuous variable, the most commonly employed weight is either the reciprocal of its standard deviation or the reciprocal of its range. Milligan and Cooper (1988) studied eight approaches to variability weighting for continuous data, and concluded that weights based on the sample range of each variable are the most effective. Employing variability weights is equivalent to what is commonly referred to as *standardizing* the variables. We will therefore revisit this approach in the next section on standardizing variables.

The previous approach assumed the importance of a variable to be inversely proportional to the total variability of that variable. The total variability of a variable comprises variation both within and between groups which may exist within the set of individuals. The aim of cluster analysis is typically to identify such groups. Hence it can be argued that the importance of a variable should not be reduced because of between-group variation (on the contrary, one might wish to assign more importance to a variable that shows larger between-group variation). As Fleiss and Zubin (1969) show, defining variable weights inversely proportional to a measure of total variability can have the serious disadvantage of diluting differences between groups on the variables which are the best discriminators. Figure 3.3 illustrates this problem.

Of course, if we knew the groups, using the within-group standard deviation of the kth variable to define weights would largely overcome this problem. Or, more generally, for equal covariances, Mahalanobis's generalized distance could be used to define the distance between two objects i and j with vectors of measurements $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$D_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)' \mathbf{W}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right), \tag{3.21}$$

where $\mathbf{W}$ is the pooled within-group covariance matrix. But in the clustering context group membership is not available prior to the analysis. Nevertheless, attempts have been made to estimate the within-group variation without knowing the cluster structure. Art *et al.* (1982) proposed an approach for determining a Mahalanobis-type distance matrix, using an iterative algorithm to identify pairs of observations that are likely to be within the same cluster and use these 'likely clusters' to calculate a within-cluster covariance matrix, $\mathbf{W}^*$. Gnanadesikan *et al.* (1995) suggested extending their approach by also estimating the between-cluster covariance matrix, $\mathbf{B}^*$, and calculating Mahalanobis-type distances based on $\operatorname{diag}\left(\mathbf{B}^*\right)\left[\operatorname{diag}\left(\mathbf{W}^*\right)\right]^{-1}$ instead of $\left(\mathbf{W}^*\right)^{-1}$. They argued that, this way, the data could be used 'to suggest weights which would emphasize the variables with the most promise for revealing clusters'.

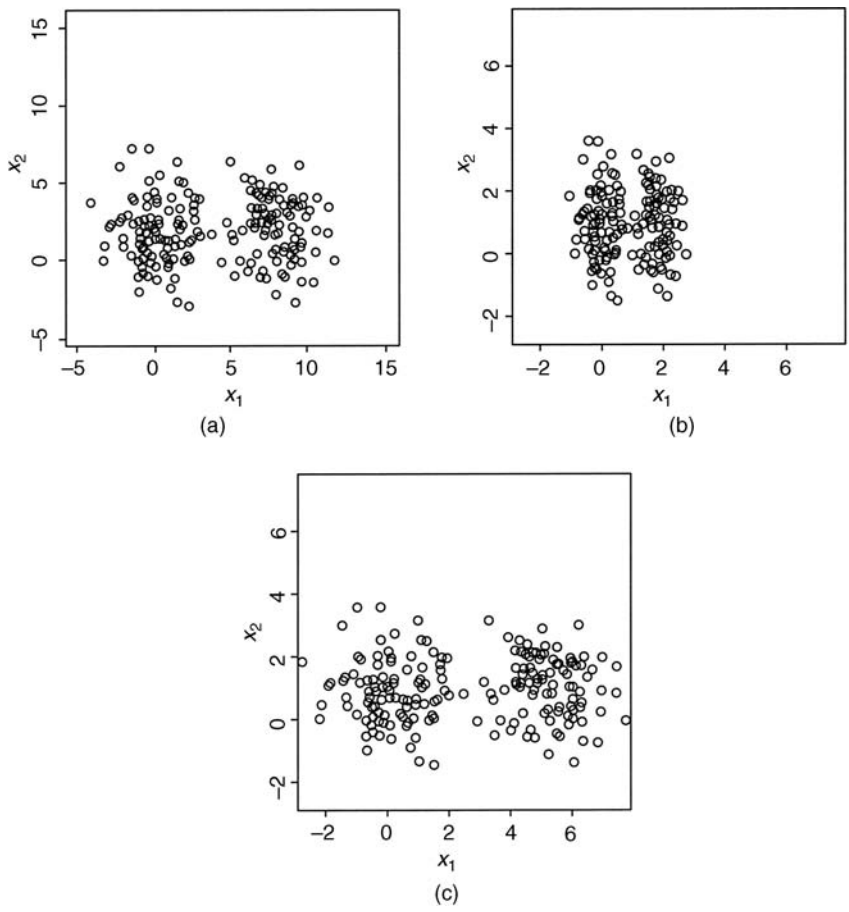


Figure 3.3 *Illustration of standardization problem. (a) Data on original scale. (b) Undesirable standardization: weights based on total standard deviations. (c) Desirable standardization: weights based on within-group standard deviations.*

An alternative criterion for determining the importance of a variable from the data has been proposed by De Soete (1986), who suggests finding weights, one for each variable, which yield weighted Euclidean distances that minimize a criterion for departure from *ultrametricity* (a term defined and discussed in the next chapter; see Section 4.4.3). This is motivated by a well-known relationship between distances that satisfy the ultrametric inequality and the existence of a unique hierarchical tree (see Chapter 4). In a simulation study, Milligan (1989) found this algorithm useful as a means of identifying variables that are important for the clustering of objects.

We have already mentioned precision weights in the context of defining proximities for structured data in Section 3.5. For variables on the same scale,

the original data matrix can be converted into a smaller data matrix of individual summaries on the basis of the underlying reference variable. Often not only summaries, such as the means within conditions, but also their precision can be estimated from the original data matrix. When relatively more or less weight is to be given to the summaries to acknowledge measurement error, weighting by the inverse of the standard error of the new summary variables is a possibility.

A further method of constructing weights from the data matrix is *variable selection*. Here, the idea is that, as in multiple regression modelling, an empirical selection procedure can be employed to identify a subset of the initial variables for inclusion in cluster analysis. The procedure results in weights of value one for selected variables and value zero for excluded variables. Examples of such selection procedures are the computer-intensive approaches proposed by Fowlkes *et al.* (1988); Carmone *et al.* (1999); Brusco and Cradit (2001) and Steinley and Brusco (2008a). In essence, such procedures proceed in an iterative fashion to identify variables which, when contributing to a cluster algorithm, lead to internally cohesive and externally isolated clusters and, when clustered singly, produce reasonable agreement with cluster solutions provided by other subsets of variables. In a simulation study, Steinley and Brusco (2008b) show that their latest algorithm (Steinley and Brusco, 2008a), which involves a screening step for 'clusterability' and evaluates all feasible subsets rather than forward selection, outperforms a number of competitors. We will take another look at the issue of variable selection in the context of model-based cluster analysis in Chapter 6.

Gnanadesikan *et al.* (1995) assessed the ability of squared distance functions based on data-determined weights, both those described above and others, to recover groups in eight simulated and real continuous data sets in a subsequent cluster analysis. Their main findings were:

(i) Equal weights, (total) standard deviation weights, and range weights were generally ineffective, but range weights were preferable to standard deviation weights.
(ii) Weighting based on estimates of within-cluster variability worked well overall.
(iii) Weighting aimed at emphasizing variables with the most potential for identifying clusters did enhance clustering when some variables had a strong cluster structure.
(iv) Weighting to optimize the fitting of a hierarchical tree was often even less effective than equal weighting or weighting based on (total) standard deviations.
(v) Forward variable selection was often among the better performers. (Note that all-subsets variable selection was not assessed at the time.)

Giving unambiguous advice as to how the variables should be weighted in the construction of dissimilarity measures is difficult; nevertheless, some points can be made. Firstly, as Sneath and Sokal (1973) point out, weights based on subjective judgements of what is important might simply reflect an existing

classification of the data. This is not what is generally required in cluster analysis. More commonly, methods of cluster analysis are applied to the data in the hope that previously unnoticed groups will emerge. Thus it is generally advisable to reduce subjective importance judgements to the initial selection of variables to be recorded, with this selection reflecting the investigator's judgement of relevance for the purpose of classification. Secondly, as the study by Gnanadesikan *et al.* (1995) shows, an overall optimal criterion for determining importance weights empirically (from the data matrix) has not been identified so far; the clustering performance of distance measures based on such weights appears to depend on the (in practice unknown) cluster structure. However, weights derived by measuring non-importance by estimated within-group variability appear to have the most potential for recovering groups in subsequent cluster analysis. And two of the most popular strategies, throwing lots of variables into a standard distance-based clustering algorithm (equal weighting) in the hope that no important ones will be omitted, and employing weights based on the standard deviations of the variables, appear to be ineffective.

3.8 Standardization

In many clustering applications the variables describing the objects to be clustered will not be measured in the same units. Indeed they may often be variables of different types, as we have already seen in Section 3.4. It is clear that it would not be sensible to treat, say, weight measured in kilograms, height measured in metres, and anxiety rated on a four-point scale as equivalent in any sense in determining a measure of similarity or distance. When all the variables have been measured on a continuous scale, the solution most often suggested to deal with the problem of different units of measurement is to simply standardize each variable to unit variance prior to any analysis. A number of variability measures have been used for this purpose. When the standard deviations calculated from the complete set of objects to be clustered are used, the technique is often referred to as *autoscaling, standard scoring* or *z-scoring*. Alternatives are division by the median absolute deviations, or by the ranges, with the latter shown to outperform autoscaling in many clustering applications (Milligan and Cooper, 1988; Gnanadesikan *et al*., 1995; Jajuga and Walesiak, 2000).

As pointed out in the previous section, standardization of variables to unit variance can be viewed as a special case of weighting. Here the weights are simply the reciprocals of the measures chosen to quantify the variance of the variables – typically the sample standard deviation or sample range of continuous variables. Thus, when standardizing variables prior to analysis the investigator assumes that the importance of a variable decreases with increasing variability. As a result of standardization being a special case of weighting, some of the recommendations made with respect to the choice of weights carry over to standardization: if the investigator cannot determine an appropriate unit of measurement and standardizing variables becomes necessary, it is preferable to standardize variables using a measure of within-group variability rather than one of total variability. In a

clustering context, the methods suggested by Art *et al.* (1982) and Gnanadesikan *et al.* (1995) for determining weights from the data matrix look promising, and implementations of their $\mathbf{W}^*$ algorithm are available in some software (Gnanadesikan, 1997). In the end, the best way of dealing with the problem of the appropriate unit of measurement might be to employ a cluster method which is invariant under scaling, thus avoiding the issue of standardization altogether. Cluster methods whose grouping solutions are not affected by changes in a variable's unit of measurement will be discussed in later chapters (see, for example, Section 5.3).

3.9 Choice of proximity measure

An almost endless number of similarity or dissimilarity coefficients exist. Several authors have provided categorizations of the various coefficients (Cheetham and Hazel, 1969; Hubálek, 1982; Gower and Legendre, 1986; Baulieu, 1989) in terms of what are generally considered their important properties (e.g. scale of data, metric and Euclidean properties of dissimilarity matrices). Unfortunately, the properties are not conclusive for choosing between coefficients. As Gower and Legendre (1986) pointed out, 'a coefficient has to be considered in the context of the descriptive statistical study of which it is a part, including the nature of the data, and the intended type of analysis'. But they did suggest some criteria which might help in making a choice.

Firstly, the nature of the data should strongly influence the choice of the proximity measure. Under certain circumstances, for example, quantitative data might be best regarded as binary, as when dichotomizing 'noisy' quantitative variables (Legendre and Legendre, 1983), or when the relevant purpose that the investigator has in mind depends on a known threshold value. As an example, Gower and Legendre (1986) considered classifying river areas according to their suitability for growing edible fish as judged by threshold levels of pesticides and heavy metals. Here the data were dichotomized according to whether measurements were above or below some toxicity level.

Next the choice of measure should depend on the scale of the data. Similarity coefficients based on Table 3.2 should be used when the data is binary. As mentioned before, the choice of proximity measure then centres around the treatment of co-absences. For continuous data, distance or correlation-type dissimilarity measures should be used according to whether 'size' or 'shape' of the objects is of interest (see previous discussion). For data that involve a mixture of continuous and binary variables, a number of coefficients have been suggested. Further mixed coefficients are easily constructed by combining proximity measures for categorical and continuous data.

Finally, the clustering method to be used might have some implications for the choice of the coefficient. For example, making a choice between several proximity coefficients with similar properties, which are also known to be monotonically related, such as S1, S3 and S5 in Table 3.3, can be avoided by employing a cluster

method that depends only on the ranking of the proximities, not their absolute values. (More details on cluster methods that are invariant under monotonic transformations of the proximity matrix, such single and complete linkage, are given in the next Chapter.) Similarly, as mentioned before, if a scale-invariant cluster analysis method is to be employed to group continuous data, the issue of weighting variables/standardization becomes irrelevant. (For more details on such methods see Chapters 5 and 6.)

Gower and Legendre (1986) present a detailed discussion of the choice of similarity or dissimilarity measure and give a decision-making table that may often be helpful in the process. However, they conclude that it is not possible in all circumstances to give a definite answer as to what measure is best to use.

3.10 Summary

Different measures of similarity or dissimilarity calculated from the same set of individuals can, and often will, lead to different solutions when used as the basis of a cluster analysis. Consequently, it would be extremely useful to know which particular measures are 'optimal' in some sense. Unfortunately, and despite a number of comparative studies (see Cheetham and Hazel, 1969; Boyce, 1969; Williams *et al.*, 1966), the question cannot be answered in any absolute sense, and the choice of measure will be guided largely by the type of variables being used and the intuition of the investigator. One recommendation which appears sensible, however, is that of Sneath and Sokal (1973), who suggest that the simplest coefficient applicable to a data set be chosen, since this is likely to ease the possibly difficult task of interpretation of final results.

4

Hierarchical clustering

4.1 Introduction

In a hierarchical classification the data are not partitioned into a particular number of classes or clusters at a single step. Instead the classification consists of a series of partitions, which may run from a single cluster containing all individuals, to n clusters each containing a single individual. Hierarchical clustering techniques may be subdivided into *agglomerative* methods, which proceed by a series of successive fusions of the n individuals into groups, and *divisive* methods, which separate the n individuals successively into finer groupings. Both types of hierarchical clustering can be viewed as attempting to find the optimal step, in some defined sense (see later), at each stage in the progressive subdivision or synthesis of the data, and each operates on a proximity matrix of some kind (see Chapter 3). A useful review of the standard methods has been given by Gordon (1987).

With hierarchical methods, divisions or fusions, once made, are irrevocable so that when an agglomerative algorithm has joined two individuals they cannot subsequently be separated, and when a divisive algorithm has made a split it cannot be undone. As Kaufman and Rousseeuw (1990) colourfully comment: 'A hierarchical method suffers from the defect that it can never repair what was done in previous steps'. Hawkins *et al.* (1982) illustrate the problem in the following way. Suppose a single variable is measured on eight objects, giving the results (−2.2, −2, −1.8, −0.1, 0.1, 1.8, 2, 2.2). The data contain three obvious 'clusters'. If the first split was into two clusters on the basis of the size of the usual t-statistic, the middle cluster (−0.1, 0.1) would be divided to produce the two clusters (−2.2, −2, −1.8, −0.1) and (0.1, 1.8, 2, 2.2). To recover them would

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

entail the nonhierarchical approach of continuing to a four-cluster solution and then merging these two.

Since all agglomerative hierarchical techniques ultimately reduce the data to a single cluster containing all the individuals, and the divisive techniques will finally split the entire set of data into n groups each containing a single individual, the investigator wishing to have a solution with an 'optimal' number of clusters will need to decide when to stop. The tricky problem of deciding on the correct number of clusters is discussed in Section 4.4.4.

Hierarchical classifications produced by either the agglomerative or divisive route may be represented by a two-dimensional diagram known as a *dendrogram,* which illustrates the fusions or divisions made at each stage of the analysis. An example of such a diagram is given in Figure 4.1. Some further properties of dendrograms and how they may be used in interpreting the results of hierarchical clustering techniques are discussed in Section 4.4.

The structure in Figure 4.1 resembles an evolutionary tree, and it is in biological applications that hierarchical classifications are perhaps most relevant. According to Rohlf (1970), a biologist, 'all other things being equal', aims for a system of nested clusters. Other areas where hierarchical classifications might be particularly appropriate are studies of social systems, and in museology and

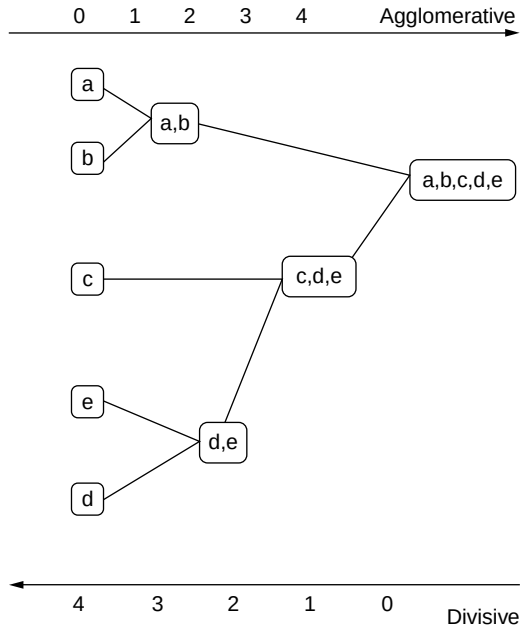


Figure 4.1 *Example of a hierarchical tree structure. (* Taken from *Finding Groups in Data,* 1990, Kaufman and Rousseeuw, with permission of the publisher, John Wiley & Sons, Inc.).

librarianship, where hierarchies are implicit in the subject matter. As will be seen later, hierarchical clustering methods have been applied in many other areas where there is not necessarily an underlying hierarchical structure. Although they may still often be usefully applied in these areas, if only to provide a starting point for a more complex clustering procedure, the following caveat of Hawkins *et al.* (1982) should be borne in mind: 'users should be very wary of using hierarchic methods if they are not clearly necessary'.

The following sections describe commonly used agglomerative techniques, and their properties. These properties are potentially applicable to divisive techniques also, but they are discussed here in relation to agglomerative techniques, since this is where most technical research has been concentrated. A description of some divisive techniques in Section 4.3 will be followed by a discussion of issues common to both agglomerative and divisive techniques. Several applications of hierarchical clustering techniques will be described in Section 4.5.

4.2 Agglomerative methods

Agglomerative procedures are probably the most widely used of the hierarchical methods. They produce a series of partitions of the data: the first consists of n single-member 'clusters'; the last consists of a single group containing all n individuals. The basic operation of all such methods is similar, and will be illustrated for two specific examples, *single linkage* and *centroid linkage.* At each stage the methods fuse individuals or groups of individuals which are closest (or most similar). Differences between the methods arise because of the different ways of defining distance (or similarity) between an individual and a group containing several individuals, or between two groups of individuals (see Chapter 3 for further details). Before giving a summary of the most widely used methods, we illustrate the general approach using two examples.

4.2.1 Illustrative examples of agglomerative methods

In this section, two hierarchical techniques are illustrated, the first requiring solely a proximity matrix, the second requiring access to a data matrix. The first illustration is of one of the simplest hierarchical clustering methods, *single linkage,* also known as the nearest-neighbour technique. It was first described by Florek *et al.* (1951) and later by Sneath (1957) and Johnson (1967). The defining feature of the method is that the distance between groups is defined as that of the closest pair of individuals, where only pairs consisting of one individual from each group are considered (nearest-neighbour distance; see Section 3.6). Single linkage serves to illustrate the general procedure of a hierarchical method, and in the example below it is applied as an agglomerative method. However, it could equally well be applied as a divisive method, by starting with a cluster containing all objects and then splitting into two clusters whose nearest-neighbour distance is a maximum.

Consider the following distance matrix:

$$\mathbf{D}_1 = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}.$$

The smallest nonzero entry in the matrix is that for individuals 1 and 2, so these are joined to form a two-member cluster. Distances between this cluster and the other three individuals are obtained as

$$\begin{aligned} d_{(12)3} &= \min(d_{13}, d_{23}) = d_{23} = 5.0 \\ d_{(12)4} &= \min(d_{14}, d_{24}) = d_{24} = 9.0 \\ d_{(12)5} &= \min(d_{15}, d_{25}) = d_{25} = 8.0. \end{aligned}$$

A new matrix may now be constructed whose entries are inter-individual and cluster-individual distance values:

$$\mathbf{D}_2 = \begin{array}{c} (12) \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}.$$

The smallest entry in $\mathbf{D}_2$ is that for individuals 4 and 5, so these now form a second two-member cluster and a new set of distances are found:

$$\begin{aligned} d_{(12)3} &= 5.0 \text{ as before} \\ d_{(12)(45)} &= \min(d_{14}, d_{15}, d_{24}, d_{25}) = d_{25} = 8.0 \\ d_{(45)3} &= \min(d_{34}, d_{35}) = d_{34} = 4.0. \end{aligned}$$

These may be arranged in a matrix $\mathbf{D}_3$:

$$\mathbf{D}_3 = \begin{array}{c} (12) \\ 3 \\ (45) \end{array} \begin{pmatrix} 0.0 & & \\ 5.0 & 0.0 & \\ 8.0 & 4.0 & 0.0 \end{pmatrix}.$$

The smallest entry is now $d_{(45)3}$, and individual 3 is added to the cluster containing individuals 4 and 5. Finally the groups containing individuals 1, 2 and 3, 4, 5 are combined into a single cluster.

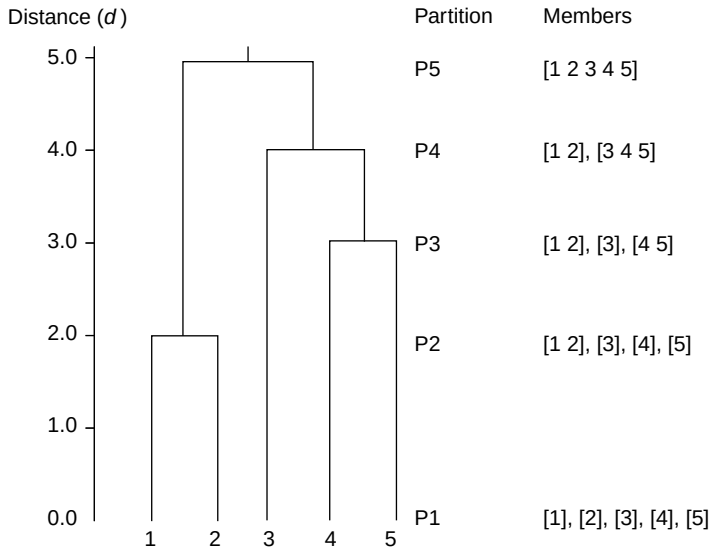


Figure 4.2 *Dendrogram for worked example of single linkage, showing partitions at each step.*

The dendrogram illustrating the process, and the partitions produced at each stage are shown in Figure 4.2; the *height* in this diagram represents the distance at which each fusion is made. Dendrograms and their features are described in more detail in Section 4.4.1.

Single linkage operates directly on a proximity matrix. Another type of clustering, *centroid* clustering, however, requires access to the original data. To illustrate this type of method, it will be applied to the following set of bivariate data:

Object	Variable 1	Variable 2
1	1.0	1.0
2	1.0	2.0
3	6.0	3.0
4	8.0	2.0
5	8.0	0.0

Suppose Euclidean distance is chosen as the inter-object distance measure, giving the following distance matrix:

$$\mathbf{C}_1 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0.00 & & & & \\ 1.00 & 0.00 & & & \\ 5.39 & 5.10 & 0.00 & & \\ 7.07 & 7.00 & 2.24 & 0.00 & \\ 7.07 & 7.28 & 3.61 & 2.00 & 0.00 \end{pmatrix}.$$

Examination of $\mathbf{C}_1$ shows that c_{12} is the smallest entry, and objects 1 and 2 are fused to form a group. The mean vector (centroid) of the group is calculated (1, 1.5) and a new Euclidean distance matrix is calculated:

$$\mathbf{C}_2 = \begin{matrix} (12) \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 0.00 & & & \\ 5.22 & 0.00 & & \\ 7.02 & 2.24 & 0.00 & \\ 7.16 & 3.61 & 2.00 & 0.00 \end{pmatrix}.$$

The smallest entry in $\mathbf{C}_2$ is c_{45}, and objects 4 and 5 are therefore fused to form a second group, the mean vector of which is (8.0, 1.0). A further distance matrix $\mathbf{C}_3$ is now calculated:

$$\mathbf{C}_3 = \begin{matrix} (12) \\ 3 \\ (45) \end{matrix} \begin{pmatrix} 0.00 & & \\ 5.22 & 0.00 & \\ 7.02 & 2.83 & 0.00 \end{pmatrix}.$$

In $\mathbf{C}_3$ the smallest entry is $c_{(45)3}$, and so objects 3, 4 and 5 are merged into a three-member cluster. The final stage consists of the fusion of the two remaining groups into one.

An important point to note about the two methods mentioned above (and all the methods discussed in this chapter) is that the clusterings proceed hierarchically, each being obtained by the merger of clusters from the previous level. So, for example, in neither of the examples above could clusters (1, 2, 4) and (3, 5) have been formed, since neither is obtainable by merging existing clusters.

4.2.2 The standard agglomerative methods

In addition to those introduced in the previous section, there are several other possible inter-group proximity measures (see Section 3.6), each giving rise to a different agglomerative method. For example, *complete linkage* (or furthest neighbour) is opposite to single linkage, in the sense that distance between groups is now defined as that of the most distant pair of individuals. In *group average linkage* – also known as the unweighted pair-group method using the average approach (UPGMA) – the distance between two clusters is the average of the distance between all pairs of individuals that are made up of one individual from each group. All these three methods (single, complete and average) use a proximity matrix as input, and the inter-cluster distances they use are each illustrated graphically in Figure 4.3.

Another agglomerative hierarchical method is *centroid clustering* – also known as the unweighted pair-group method using the centroid approach (UPGMC) – which uses a data matrix rather than a proximity matrix and involves merging clusters with the most similar mean vectors. *Median linkage* – the weighted

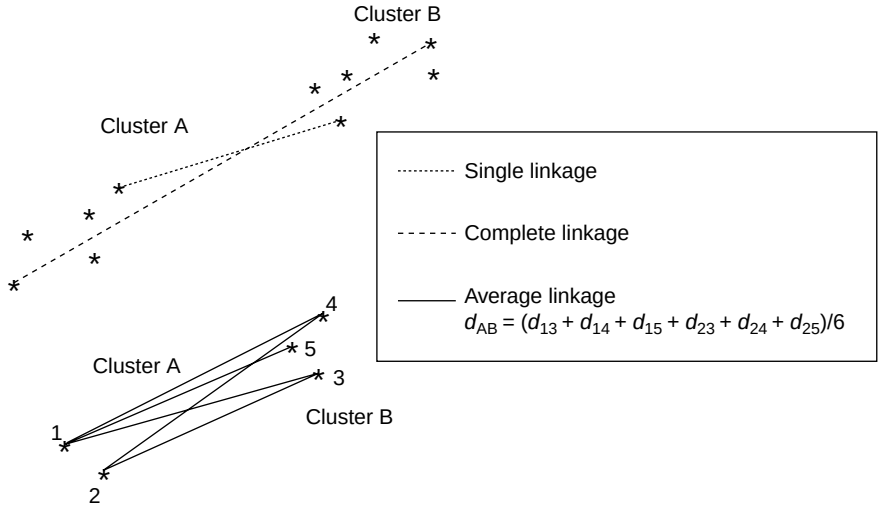


Figure 4.3 *Examples of three inter-cluster distance measures: single, complete and average.*

pair-group method using the centroid approach (WPGMC) – is similar, except that the centroids of the constituent clusters are weighted equally to produce the new centroid of the merged cluster. This is to avoid the objects in the more numerous of the pair of clusters to be merged dominating those in the smaller cluster. The new centroid is thus intermediate between the two constituent clusters.

In the numerical illustration of centroid linkage shown in Section 4.2.1, Euclidean distance was used, as is most common. While other proximity measures are possible with centroid or median linkage, they would lack interpretation in terms of the raw data (see Anderberg, 1973).

Ward (1963) introduced a third type of method, in which the fusion of two clusters is based on the size of an error sum-of-squares criterion. The objective at each stage is to minimize the increase in the total within-cluster error sum of squares, E, given by

$$E = \sum_{m=1}^{g} E_m,$$

where

$$E_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p_k} \left(x_{ml,k} - \bar{x}_{m,k} \right)^2, \tag{4.1}$$

in which $\bar{x}_{m,k} = (1/n_m) \sum_{l=1}^{n_m} x_{ml,k}$ (the mean of the mth cluster for the kth variable), $x_{ml,k}$ being the score on the kth variable ($k = 1, \ldots, p$) for the lth object

($l = 1, \ldots, n_m$) in the mth cluster ($m = 1, \ldots, g$). This increase is proportional to the squared Euclidean distance between the centroids of the merged clusters, but the method differs from centroid clustering in that centroids are weighted by $n_m n_q/(n_m + n_q)$ when computing distances between centroids, where n_m and n_q are the numbers of objects in the two clusters m and q.

Weighted average linkage (McQuitty, 1966), also known as WPGMA, is similar to (group) average linkage but weights inter-cluster distances according to the inverse of the number of objects in each class, as in the case of median compared to centroid linkage.

The seven methods introduced so far are summarized in Table 4.1, along with some remarks about some of their typical characteristics, which will be amplified below.

Some other hierarchical methods, related to the techniques described above, should also be mentioned. The *sum-of-squares* method (Jambu, 1978; Podani, 1989) is similar to Ward's method but is based on the sum of squares within each cluster rather than the increase in sum of squares in the merged cluster.

Lance and Williams (1967) also introduced a new *flexible* method defined by values of the parameters of a general recurrence formula, outlined in the next subsection. Many of the mathematical properties of the standard hierarchical methods can be defined in terms of the parameters of the Lance and Williams formulation, and in Section 4.4.3 some of these are introduced.

4.2.3 Recurrence formula for agglomerative methods

The Lance and Williams recurrence formula gives the distance between a group k and a group (ij) formed by the fusion of two groups (i and j) as

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|, \tag{4.2}$$

where d_{ij} is the distance between groups i and j. Lance and Williams used the formula to define a new 'flexible' scheme, with parameter values $\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, $\beta < 1$, $\gamma = 0$. By allowing β to vary, clustering schemes with various characteristics can be obtained. They suggest small negative values for β, such as -0.25, although Scheibler and Schneider (1985) suggest -0.50.

The inter-group distance measures used by many standard hierarchical clustering techniques can, by suitable choice of the parameters α_i, α_j, β and γ, be contained within this formula, as shown in Table 4.2, which also shows additional properties of these methods, to be discussed in Section 4.4.3. Single linkage, for example, corresponds to the parameter values $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = 0$ and $\gamma = -\frac{1}{2}$, and (4.2) is

$$d_{k(ij)} = \frac{1}{2} d_{ki} + \frac{1}{2} d_{ki} d_{kj} - \frac{1}{2} \left| d_{ki} - d_{kj} \right| \tag{4.3}$$

If $d_{ki} > d_{kj}$, then $|d_{ki} - d_{kj}| = d_{ki} - d_{kj}$ and $d_{k(ij)} = d_{kj}$. Similarly, if $d_{ki} < d_{kj}$, $|d_{ki} - d_{kj}| = d_{kj} - d_{ki}$, and consequently the recurrence formula gives the

Table 4.1 Standard agglomerative hierarchical clustering methods.

Method	Alternative name[a]	Usually used with:	Distance between clusters defined as:	Remarks
Single linkage Sneath (1957)	Nearest neighbour	Similarity or distance	Minimum distance between pair of objects, one in one cluster, one in the other	Tends to produce unbalanced and straggly clusters ('chaining'), especially in large data sets. Does not take account of cluster structure.
Complete linkage Sorensen (1948)	Furthest neighbour	Similarity or distance	Maximum distance between pair of objects, one in one cluster, one in the other	Tends to find compact clusters with equal diameters (maximum distance between objects). Does not take account of cluster structure.
(Group) Average linkage Sokal and Michener (1958)	UPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	Tends to join clusters with small variances. Intermediate between single and complete linkage. Takes account of cluster structure. Relatively robust.
Centroid linkage Sokal and Michener (1958)	UPGMC	Distance (requires raw data)	Squared Euclidean distance between mean vectors (centroids)	Assumes points can be represented in Euclidean space (for geometrical interpretation). The more numerous of the two groups clustered dominates the merged cluster. Subject to reversals.
Weighted average linkage McQuitty (1966)	WPGMA	Similarity or distance	Average distance between pair of objects, one in one cluster, one in the other	As for UPGMA, but points in small clusters weighted more highly than points in large clusters (useful if cluster sizes are likely to be uneven).
Median linkage Gower (1967)	WPGMC	Distance (requires raw data)	Squared Euclidean distance between weighted centroids	Assumes points can be represented in Euclidean space for geometrical interpretation. New group is intermediate in position between merged groups. Subject to reversals.
Ward's method Ward (1963)	Minimum sum of squares	Distance (requires raw data)	Increase in sum of squares within clusters, after fusion, summed over all variables	Assumes points can be represented in Euclidean space for geometrical interpretation. Tends to find same-size, spherical clusters. Sensitive to outliers.

[a]U = unweighted; W = weighted; PG = pair group; A = average; C = centroid.

Table 4.2 Hierarchical agglomerative clustering methods: admissibility conditions and Lance–Williams parameters.

	Admissibility conditions[a]				Lance–Williams parameters[b]		
Method	U	C	P	M	α_i	β	γ
Single linkage	N	N	Y	Y	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage	N	N	Y	Y	$\frac{1}{2}$	0	$\frac{1}{2}$
Average linkage	N	N	N	N	$n_i/(n_i + n_j)$	0	0
Centroid linkage	Y	N	N	N	$n_i/(n_i + n_j)$	$-n_i n_j/(n_i + n_j)^2$	0
Median linkage	Y	N	Y	N	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward's method	N	Y	N	N	$(n_k + n_i)/(n_k + n_i + n_j)$	$-n_k/(n_k + n_i + n_j)$	0

[a] U = no reversals; C = convex; P = point proportional; M = monotone.
[b] n_k, n_i and n_j are the respective cluster sizes when cluster k is joined to the fusion of clusters i and j (see Equation (4.2)).

required

$$d_{k(ij)} = \min\left(d_{ki}, d_{kj}\right) \tag{4.4}$$

The Lance and Williams recursive formula can be used to program many hierarchical agglomerative methods, and such algorithms use computer time of the order of $n^2\log(n)$. Improvements can be made on this, as discussed by Hansen and Jaumard (1997). For example, algorithms based on merging edges in the minimum spanning tree representation of single linkage are proportional to n^2. (Divisive algorithms are intrinsically more difficult to program efficiently, and this is partly why they are less widely used.)

4.2.4 Problems of agglomerative hierarchical methods

To illustrate some of the potential problems of these agglomerative methods, a set of simulated data will be clustered using single, complete and average linkage. The data consist of 50 points simulated from two bivariate normal distributions with mean vectors (0, 0) and (4, 4), and common covariance matrix

$$\Sigma = \begin{pmatrix} 16.0 & 1.5 \\ 1.5 & 0.25 \end{pmatrix}.$$

Two intermediate points have been added for the first analysis, in order to illustrate a problem known as *chaining* often found when using single linkage. Figure 4.4 gives the single linkage dendrogram and Figure 4.5 shows some of the results of the cluster analyses.

Figure 4.4 shows a typical single linkage dendrogram. There is little clear structure, with the two intermediate points (51 and 52) linking the two main clusters, which are gradually pulled together into one large cluster, isolating two singletons until the final step. Note that although the outlying points 8 and 29 are

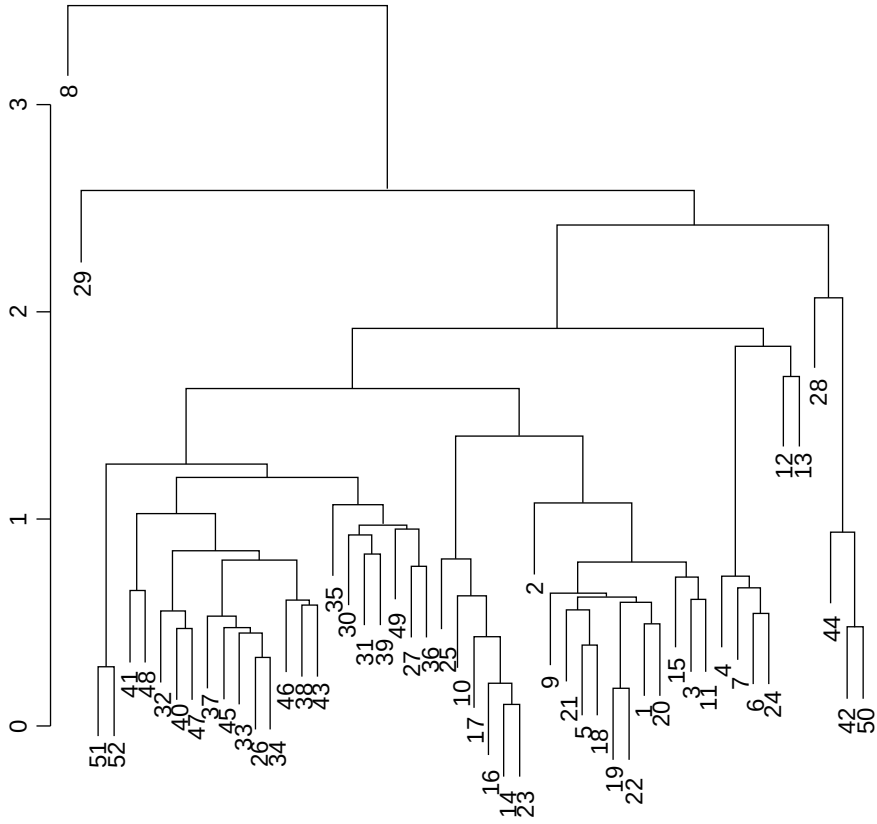


Figure 4.4 *Dendrogram showing single linkage clustering of simulated data set (see also Figure 4.5(a)).*

close together on the dendrogram, they are those at the extreme opposite ends of the main clusters. Note also that this form of dendrogram places the labels for the points just underneath the place where they first join a cluster. This makes the order of joining evident. Some other software would place all labels along the zero line at the bottom.

Figure 4.5(a) shows the chaining of the two main groups together in single linkage, and the isolation of one outlier, if two groups are specified. (If three groups are specified, the other outlier is hived off, still leaving one large cluster.) Despite the obvious lack of success in recovering the two groups, this example does illustrate a potential benefit of applying single linkage, namely that it can be used to identify outliers, since these are left as singletons if they are sufficiently far from their nearest neighbour.

Complete (Figure 4.5(c)) and average linkage (Figure 4.5(d)) techniques were equally unsuccessful in cluster recovery, with or without intermediate points, and whatever number of clusters was specified (from two to five). They tended to

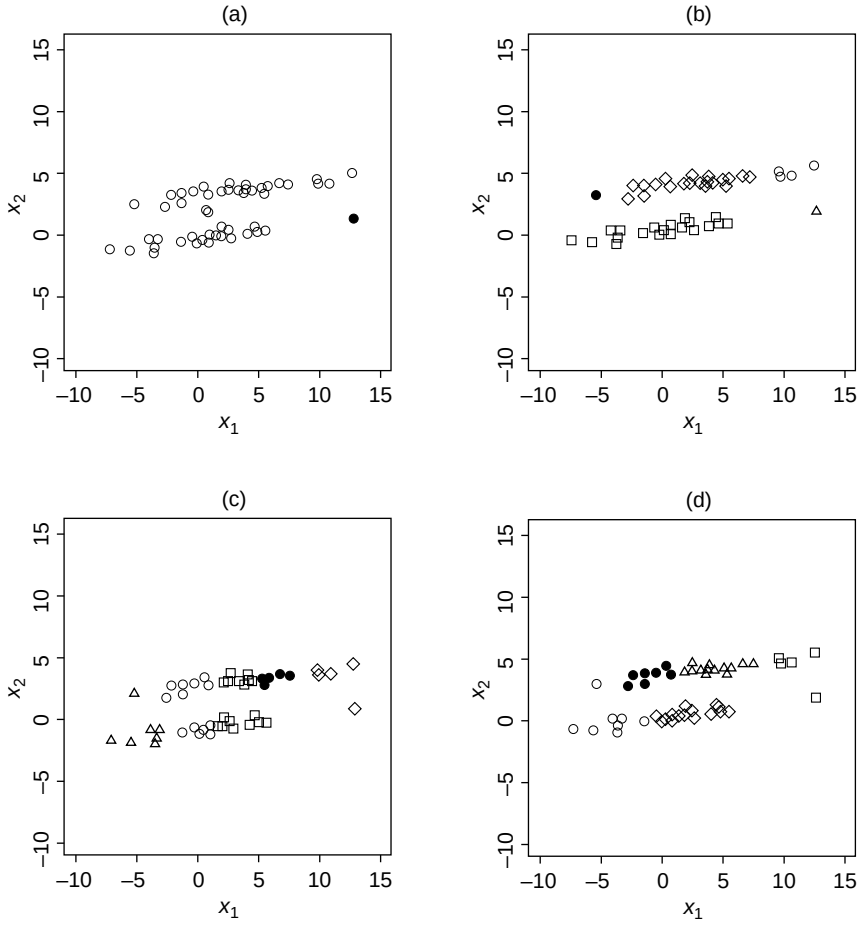


Figure 4.5 *Clusters obtained by four different methods from simulated data: (a) single linkage, with intermediate points, two-cluster solution; (b) single linkage, no intermediate points, five-cluster solution; (c) complete linkage, no intermediate points, five-cluster solution; (d) average linkage, no intermediate points, five-cluster solution.*

impose spherical clusters, forming a cluster in the middle, part from group 1 and part from group 2. The five-cluster solution for single linkage (Figure 4.5(b)) was relatively more successful, since the five clusters could be amalgamated into two, to form the correct groups. (Of course, such an amalgamation would destroy the hierarchy, just as in the Hawkins *et al.* example given in Section 4.1.)

These small examples show some of the problems of agglomerative methods and their relative failure to recover nonspherical clusters. (Similar problems were found in the more general empirical studies to be discussed in the next subsection.) It also shows how crucial it is to make the correct choice as to the number of clusters present (see Section 4.4.4), and the advisability of plotting the

raw data where feasible. A practical computational consideration is non-uniqueness due to ties (for single linkage and some other methods). In the case of non-uniqueness, a decision as to which clusters to fuse needs to be made, and this is usually a default choice determined by software. It is generally recommended to run analyses with different choices to check for robustness. The more sophisticated model-based clustering techniques to be discussed in Chapters 6 and 7 have the potential to overcome some of these problems.

4.2.5 Empirical studies of hierarchical agglomerative methods

Empirical studies of hierarchical methods are of two main types. One type simulates clusters in data of a particular type and then assesses the characteristics and recovery of clusters. The other is based on real data from a particular subject matter, the criterion in the latter usually being the interpretability of clusters. Examples of the former include a review by Milligan (1981) and a study reported by Hands and Everitt (1987). The latter concluded that Ward's method performed very well when the data contained clusters with approximately the same numbers of points, but poorly when the clusters were of different sizes. In that situation, centroid clustering appeared to give the most satisfactory results. Cunningham and Ogilvie (1972) and Blashfield (1976) also concluded that for clusters with equal numbers of points Ward's method is successful, otherwise centroid group average and complete linkage are preferable.

Studies that focus on the stability of clustering in the presence of outliers or noise include that by Hubert (1974), who found that complete linkage is less sensitive to observational errors than single linkage. (A related point is the observation of Hartigan (1975), that single linkage is dependent on the smallest distances, and they need to be measured with low error for single linkage to be successful.)

An empirical study based on the subject-matter approach is that of Duflou and Maenhaut (1990). These authors compared seven standard methods (those in Table 4.1 and one other) on data involving chemical concentrations in the brain. They rejected centroid and median linkage because of reversals (a type of inconsistency in the hierarchy; see Section 4.4.3), and concluded that, of the remainder, Ward's method and complete linkage gave interpretable results and correctly distinguished grey and white matter areas in the brain. A further example is provided by Baxter (1994), who summarizes the position in archaeology, where empirical studies generally favour Ward's method and average linkage.

It has to be recognized that hierarchical clustering methods may give very different results on the same data, and empirical studies are rarely conclusive. What is most clear is that no one method can be recommended above all others and, as Gordon (1998) points out, hierarchical methods are in any case only stepwise optimal. A few general observations can, however, be made. Single linkage, which has satisfactory mathematical properties and is also easy to program and apply to large data sets, tends to be less satisfactory than other methods because of

'chaining'; this is the phenomenon in which separated clusters with 'noise' points in between them tend to be joined together. Ward's method often appears to work well but may impose a spherical structure where none exists.

4.3 Divisive methods

Divisive methods operate in the opposite direction to agglomerative methods, starting with one large cluster and successively splitting clusters. They are computationally demanding if all $2^{k-1} - 1$ possible divisions into two subclusters of a cluster of k objects are considered at each stage. However, for data consisting of p binary variables, relatively simple and computationally efficient methods, known as *monothetic divisive methods*, are available. These generally divide clusters according to the presence or absence of each of the p variables, so that at each stage clusters contain members with certain attributes either all present or all absent. The data for these methods thus need to be in the form of a two-mode (binary) matrix. The term 'monothetic' refers to the use of a single variable on which to base the split at a given stage; *polythetic* methods, to be described in Section 4.3.2, use all the variables at each stage. While less commonly used than agglomerative methods, divisive methods have the advantage, pointed out by Kaufman and Rousseeuw (1990), that most users are interested in the main structure in their data, and this is revealed from the outset of a divisive method.

4.3.1 Monothetic divisive methods

The choice of the variable in monothetic divisive methods on which a split is made depends on optimizing a criterion reflecting either cluster homogeneity or association with other variables. This tends to minimize the number of splits that have to be made. An example of the homogeneity criterion is the *information content, C* (which in this case signifies disorder or chaos), defined by p variables and n objects (Lance and Williams, 1968):

$$C = pn\log n - \sum_{k=1}^{p} [f_k \log f_k - (n - f_k)\log(n - f_k)], \tag{4.5}$$

where f_k is the number of individuals having the kth attribute. If a group X is to be split into two groups A and B, the reduction in C is $C_X - C_A - C_B$. The ideal set of clusters would have members with identical attributes and C equal to zero; hence clusters are split at each stage according to possession of the attribute which leads to the greatest reduction in C.

Instead of cluster homogeneity, the attribute used at each step can be chosen according to its overall association with all attributes remaining at this step: this is sometimes termed *association analysis* (Williams and Lambert, 1959), especially in ecology. For example, for one pair of variables V_i and V_j with values 0 and 1, the frequencies observed might be:

V_j	V_i	
	1	0
1	a	b
0	c	d

Common measures of association (summed over all pairs of variables) are the following:

$$|ad-bc| \tag{4.6}$$

$$(ad-bc)^2 \tag{4.7}$$

$$(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)] \tag{4.8}$$

$$\sqrt{(ad-bc)^2 n/[(a+b)(a+c)(b+d)(c+d)]} \tag{4.9}$$

$$(ad-bc)^2/[(a+b)(a+c)(b+d)(c+d)] \tag{4.10}$$

The split at each stage is made according to the presence or absence of the attribute whose association with the others (i.e. the summed criterion above) is a maximum. The first two criteria, (4.6) and (4.7), have the advantage that there is no danger of computational problems if any of the marginal totals are zero (Kaufman and Rousseeuw, 1990). The last three, (4.8), (4.9) and (4.10), are all related to the usual chi-squared statistic, its square root, and the Pearson correlation coefficient, respectively. Hubálek (1982) gives a review of 43 such coefficients.

Appealing features of monothetic divisive methods are the easy classification of new members, and the inclusion of cases with missing values. The latter can be dealt with as follows. If there are missing values for a particular variable, V_1 say, the nonmissing variable with the largest absolute association with it is determined, V_2, say. The missing value for V_1 is replaced by the value of V_2 for the same observation (positive association between V_1 and V_2) or $1 - V_2$ (negative association).

A further advantage of monothetic divisive methods is that it is obvious which variables produce the split at any stage of the process. However, a general problem with these methods is that the possession of a particular attribute, which is either rare or rarely found in combination with others, may take an individual down the 'wrong' path. Typical uses of the method are in medicine (as diagnostic keys; see, for example, Payne and Preece, 1980) and in mortuary studies in archaeology, where it can be argued that social stratum in life might be reflected by the possession of a common set of grave goods (see O'Shea, 1985, for example).

A new method of divisive clustering has been proposed by Piccarreta and Billari (2007), which can be used for sequence data such as life-course histories. The method uses the logic of classification and regression tree (CART) analysis

(Breiman et al., 1984), thus enabling some of the most useful features of CART analysis to be employed, such as tree pruning by cross-validation to identify the appropriate number of clusters. Piccarreta and Billari define two new types of data derived from the original sequences: *auxiliary variables* and *state permanence sequences*. This means that, rather than having completely different dependent and independent variables (as in CART), the variables defining the splits, the criterion for assessing the homogeneity of the clusters, and the data characterizing the clusters are all derived from the sequence data.

The splits in this new method are made with the aim of producing 'pure' clusters. However, whereas, in CART, purity is defined in terms of a dependent or outcome variable, here 'impurity' is defined as the summed OMA distance between all pairs of units (see Chapter 3, Section 3.5 for a description of the OMA distance measure). The auxiliary variables, which are used to split the sample, can be defined in various ways according to the subject matter; for example, they might be the times at which a particular state is reached for the first time, second time, etc. The divisive procedure operates by choosing the auxiliary variables that lead to the greatest improvement in the within-cluster purity at each stage of the splitting process. This leads to a tree with nodes (clusters) which can then be simplified by pruning – cutting back branches. As in CART, this is at the expense of within-cluster purity, which has to be balanced against increased simplicity. Once a satisfactory solution has been found, the exemplar (see Section 4.4.1) – here the medoid of each cluster – can be used to summarize the clusters using a representation that retains some key features of the original sequence: the *state permanence sequence*, which indicates the length of time in each state. Section 4.5.4 describes the authors' application of this method. `SAS` routines are available from the authors of the paper.

4.3.2 Polythetic divisive methods

Polythetic divisive methods are more akin to the agglomerative methods discussed above, since they use all variables simultaneously, and can work with a proximity matrix. The procedure of MacNaughton-Smith *et al.* (1964) avoids considering all possible splits, a potential problem of polythetic divisive methods. It proceeds by finding the object that is furthest away from the others within a group, and using that as the seed for a splinter group. Each object is then considered for entry to the splinter group: any that are closer to the splinter group are moved into it. The step is repeated, the next cluster for splitting being chosen as the largest in diameter (defined by the largest dissimilarity between any two objects).

The process has been described as follows by Kaufman and Rousseeuw (1990), who have developed a program, `diana` (DIvisive ANAlysis clustering), which is implemented in `S-plus` and `R`, and this is often used as the method's appellation.

The mechanism somewhat resembles the way a political party might split up due to inner conflicts: firstly the most discontented member leaves the party and starts a new one, and then some others follow him until a kind of equilibrium is attained. So we first need to know which member disagrees most with the others.

To illustrate this, consider the following distance matrix for seven individuals:

$$\mathbf{D} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{pmatrix} 0 & & & & & & \\ 10 & 0 & & & & & \\ 7 & 7 & 0 & & & & \\ 30 & 23 & 21 & 0 & & & \\ 29 & 25 & 22 & 7 & 0 & & \\ 38 & 34 & 31 & 10 & 11 & 0 & \\ 42 & 36 & 36 & 13 & 17 & 9 & 0 \end{pmatrix}.$$

The individual used to initiate the splinter group is the one whose average distance from the other individuals is a maximum. This is found to be individual 1, giving the initial groups as (1) and (2, 3, 4, 5, 6, 7). Next the average distance of each individual in the main group to the individuals in the splinter group is found, followed by the average distances of each individual in the main group to the other individuals in this group. The difference between these two averages is then found. In this example this leads to:

Individual in main group	Average distance to splinter group (A)	Average distance to main group (B)	$B-A$
2	10.0	25.0	15.0
3	7.0	23.4	16.4
4	30.0	14.8	−15.2
5	29.0	16.4	−12.6
6	38.0	19.0	−19.0
7	42.0	22.2	−19.8

The maximum difference is 16.4, for individual 3, which is therefore added into the splinter group, giving the two groups (1, 3) and (2, 4, 5, 6, 7). Repeating the process gives the following:

Individual in main group	Average distance to splinter group (A)	Average distance to main group (B)	$B-A$
2	8.5	29.5	21.0
4	25.5	13.2	−12.3
5	25.5	15.0	−10.5
6	34.5	16.0	−18.5
7	39.0	18.7	−20.3

So now individual 2 joins the splinter group to give groups (1, 3, 2) and (4, 5, 6, 7), and the process is repeated to give:

Individual in main group	Average distance to splinter group (A)	Average distance to main group (B)	$B-A$
4	24.3	10.0	−14.3
5	25.3	11.7	−13.6
6	34.3	10.0	−24.3
7	38.0	13.0	−25.0

As all the differences are now negative, the process would continue (if desired) on each subgroup separately.

4.4 Applying the hierarchical clustering process

To make best use of hierarchical techniques, both agglomerative and divisive, the user often needs to consider the following points (in addition to the choice of proximity measure):

- graphical display of the clustering process
- comparison of dendrograms
- mathematical properties of methods
- choice of partition
- hierarchical algorithms.

These will be discussed briefly in this section. (Further general information on some of these points will be given in Chapter 9.)

4.4.1 Dendrograms and other tree representations

The dendrogram, or tree diagram, is a mathematical and pictorial representation of the complete clustering procedure, as already illustrated. Here some terminology is given (see Figure 4.6). The *nodes* of the dendrogram represent clusters, and the lengths of the stems *(heights)* represent the distances at which clusters are joined, as already defined in Section 4.2.1. As noted in Section 4.2.4, the stems may be drawn so that they do not extend to the zero line of the diagram, in order to indicate the order in which objects first join clusters. Dendrograms which do not have numerical information attached to the stems are termed *unweighted* or *ranked.* Most dendrograms have two edges emanating from each node *(binary trees).* The arrangement of nodes and stems is the *topology* of the tree.

The names of objects attached to the terminal nodes are known as *labels.* Internal nodes are not usually labelled, although Carroll and Chang (1973) give an example of this where, for instance, the internal node ‘arm’ is above the terminal nodes ‘elbow’ and ‘hand’. Typical or representative members of the clusters can be associated with the internal nodes, called *exemplars* or *centrotypes,* and are defined as the objects having the maximum within-cluster average similarity

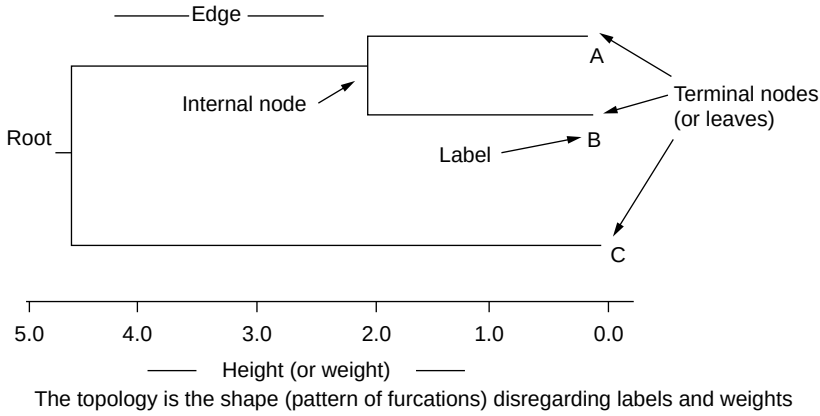


Figure 4.6 *Some terminology used in describing dendrograms.*

(or minimum dissimilarity). A particular type of centrotype is the *medoid* (the object with the minimum *absolute* distance to the other members of the cluster). The dendrogram itself describes the process by which the hierarchy has been obtained, whereas the exemplar and internal node labels describe particular partitions, once these have been chosen.

It is important to realize that the same data and clustering procedure can give rise to 2^{n-1} dendrograms with different appearances, depending on the order in which the nodes are displayed. This can be envisaged by imagining the dendrogram as a mobile in three-dimensional space: the stems from each node can swing around through 180 degrees without changing inter-cluster relationships. Most software packages choose the algorithm for drawing dendrograms automatically, but algorithms for optimizing the appearance of dendrograms have been developed, for example by using internal (Gale *et al.*, 1984) or external (Degerman, 1982) evidence. Wishart (1999) has proposed a robust method that optimizes the rank order of the proximities. The method involves considering each cluster fusion in turn, by reversing the order of the cases within the cluster but without affecting the topology of the tree so as to optimize an objective function.

A number of extensions to dendrograms have been developed. *Espaliers*, for example, are generalized dendrograms in which the length of the horizontal line conveys information about the relative homogeneity and separation of clusters. Hansen *et al.* (1996) discuss the details of these, and give a number of examples and an algorithm for converting a standard dendrogram into an espalier. The *pyramid* is a further specialized type of dendrogram for representing overlapping clusters (see Chapter 8). De Soete and Carroll (1996) give examples of these and other types of tree representation.

The additive tree (or *path length tree*) is a generalization of the dendrogram in which the lengths of the paths between the nodes represent proximities between objects, and in which the *additive inequality* (or *four-point condition*) holds. This

generalization of the ultrametric inequality (see Section 4.4.3) is a necessary and sufficient condition for a set of proximities to be represented in the form of an additive tree. The additive inequality is as follows:

$$d_{xy} + d_{uv} \leq \max|d_{xu} + d_{yv} + d_{yu}| \text{ for all } x,\ y,\ u,\ v. \tag{4.11}$$

Further details are given in Everitt and Rabe-Hesketh (1997), and an example of an additive tree showing genetic associations between various ethnic groups has kindly been provided by Kenneth Kidd (see Figure 4.7). This is a representation of pairwise genetic distances among 30 human populations, generated by a searching routine (Kidd and Sgaramella-Zonta, 1971) that makes topological changes around small or negative branches starting from the neighbour-joining tree produced by the PHYLIP package (Felsenstein, 1989). These branch lengths are the least-squares solution to the complete set of linear equations that relate each pairwise distance to the sum of the branch lengths connecting those populations. For these 30 populations there are $n(n-1)/2 = 435$ pairwise distances to be explained by addition of different combinations of $2n - 3 = 57$ branch lengths. Each tree topology is represented by a different set of equations. Of the 8.69×10^{36} possible trees (sets of linear equations), only about 100 were actually evaluated, and the tree in Figure 4.7 had the smallest $\sum e^2$ (the quantity minimized by least squares

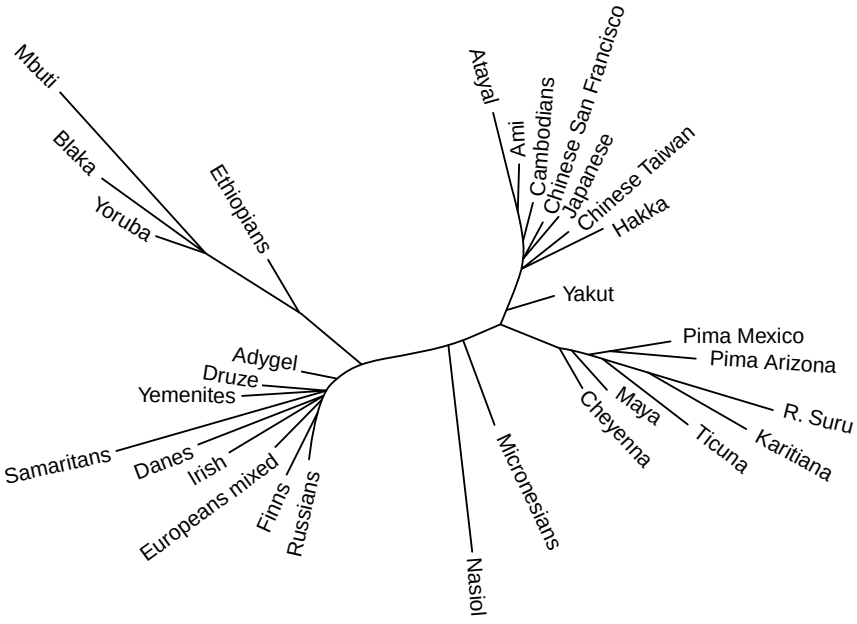


Figure 4.7 *An additive tree representation of pairwise genetic distances among 30 human populations; descriptions of these populations can be found in the ALFRED database at http://info.med.yale.edu/genetics/kkidd.* (Reproduced with permission from Kenneth K. Kidd.)

for each set of linear equations); several others were almost as good as this tree, with only small differences around the very small branches.

Unlike the dendrogram, where each terminal node is equidistant from a single node at the top of the hierarchy, this type of tree is not so rooted. As a concrete example of this, consider the case in which the distances represent genetic differences between species, based on the total number of mutations from a common origin. The additive tree would allow the estimation of the evolutionary time between the appearance of two species from the total path length between them, but it would not be possible to say for certain which species was earlier, since their common origin could be placed at any internal node in the tree. This example is typical of applications of additive trees, which are commonly employed in evolutionary studies to reconstruct phylogenies.

4.4.2 Comparing dendrograms and measuring their distortion

It may be required to compare two dendrograms without making a particular choice as to the particular partition corresponding to a specific number of clusters. Furthermore, hierarchical clustering techniques impose a hierarchical structure on data and it is usually necessary to consider whether this type of structure is acceptable or whether it introduces unacceptable distortion of the original relationships amongst the objects as implied by their observed proximities. Two measures commonly used for comparing a dendrogram with a proximity matrix or with a second dendrogram are the *cophenetic correlation* and *Goodman and Kruskal's* γ.

The starting point for either of these is the so-called *cophenetic matrix*. The elements of this matrix are the heights, h_{ij}, where two objects become members of the same cluster in the dendrogram. It is unaffected by the indeterminacy of the appearance of the dendrogram. The cophenetic matrix $\mathbf{H}$ for the single linkage example in Section 4.2.1 is

$$\mathbf{H} = \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 5.0 & 5.0 & 0.0 & & \\ 5.0 & 5.0 & 4.0 & 0.0 & \\ 5.0 & 5.0 & 4.0 & 3.0 & 0.0 \end{pmatrix}.$$

The *cophenetic correlation* is the product moment correlation between the $n(n-1)/2$ entries (h_{ij}) in the appropriate cophenetic matrices (excluding those on the diagonals). The matrix is most conveniently arranged in vector form. For example, the off-diagonal elements of $\mathbf{H}$ and the original proximity matrix $\mathbf{D}_1$ in Section 4.2.1 are as follows:

$$\begin{array}{lcccccccccc} \mathbf{H}: & 2, & 5, & 5, & 5, & 5, & 5, & 5, & 4, & 4, & 3 \\ \mathbf{D}_1: & 2, & 6, & 10, & 9, & 5, & 9, & 8, & 4, & 5, & 3. \end{array}$$

The cophenetic correlation between the distance matrix $\mathbf{D}_1$ and $\mathbf{H}$ is 0.82.

Another, nonparametric measure of association is Goodman and Kruskal's γ, defined as $(S_+ - S_-)/(S_+ + S_-)$, where S_+ and S_- are the number of concordances and discordances, respectively. A concordance or discordance in the context of matrix comparison is defined by comparing each pair of pairs. For example, the pairs h_{12} and h_{14} in $\mathbf{H}$ and d_{12} and d_{14} in $\mathbf{D}_1$ are discordant because $2 < 5$ in $\mathbf{H}$ and $2 < 10$ in D_1. For these data, γ is 1.0.

Further information on dendrogram comparison and some applications are given in Chapter 9.

4.4.3 Mathematical properties of hierarchical methods

A number of mathematical properties can be defined for clustering methods. One of these, the *ultrametric property,* was first introduced by Hartigan (1967), Jardine *et al.* (1967) and Johnson (1967), and has since been shown to be related to various features of clustering techniques, in particular the ability to represent the hierarchy by a dendrogram. The *ultrametric property* states that

$$h_{ij} \leq \max\left(h_{ij}, h_{jk}\right) \text{ for all } i, j \text{ and } k, \tag{4.12}$$

where h_{ij} is the distance between clusters i and j. An alternative way of describing this property is that for any three objects, the two largest distances between objects are equal. The property does not necessarily (or even usually) hold for the elements of proximity matrices. However, it does hold for the heights h_{ij} at which two objects become members of the same cluster in many hierarchical clustering techniques.

A consequence of failing to obey the ultrametric property is that *inversions* or *reversals* can occur in the dendrogram. This happens when the fusion levels do not form a monotonic sequence, so that a later fusion takes place at a lower level of dissimilarity than an earlier one. Morgan and Ray (1995) describe some empirical studies of inversions for a number of methods. Inversions are not necessarily a problem if the interest is in one particular partition rather than the complete hierarchical structure. They may also be useful in indicating areas where there is no clear structure (Gower, 1990). However, as Murtagh (1985) points out, reversals can make interpretation of the hierarchy very difficult, both in theoretical studies of cluster properties and also in applications where a hierarchical structure is an intrinsic part of the model. This is because the nested structure is not maintained, as shown in Figure 4.8. Both centroid and median clustering can produce reversals.

A related feature of clustering methods is their tendency to 'distort' space. The 'chaining' effect of single linkage, in which dissimilar objects are drawn into the same cluster, is an example of such distortion, *space contraction* in this case. The opposite type of distortion, where the process of fusing clusters tends to draw clusters together, is *space dilation,* as found in complete linkage. *Space-conserving* methods, such as group average linkage, obey the following inequality:

$$d_{iuv} \leq d_{i(uv)} \leq D_{iuv}, \tag{4.13}$$

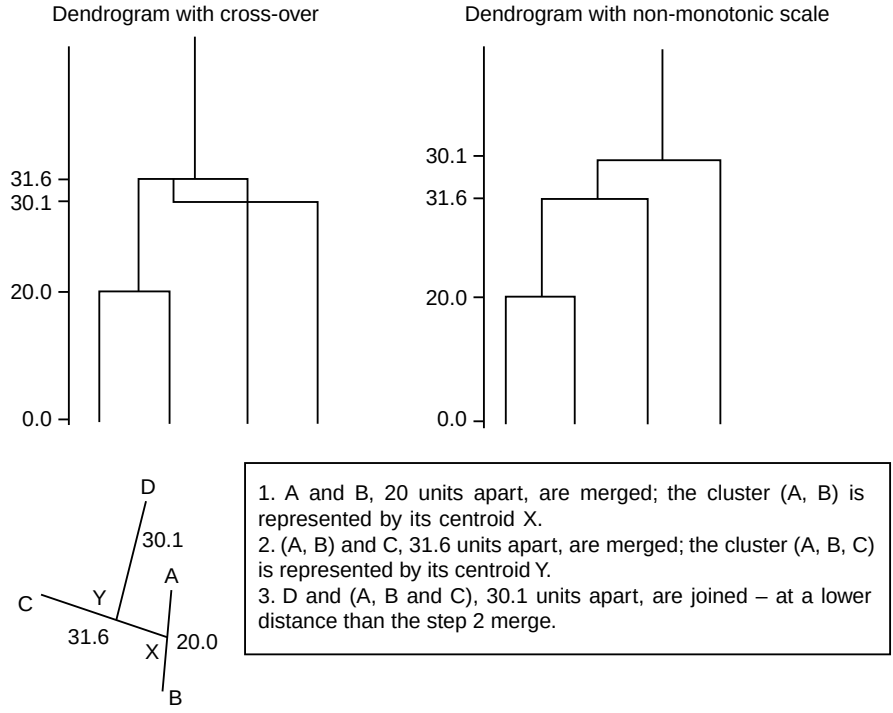


Figure 4.8 *Example of an inversion or reversal in a dendrogram.* (Adapted with permission of the publisher, Blackwell, from Morgan and Ray, 1995.)

where d_{iuv} and D_{iuv} are the minimum and maximum distances between object i and clusters u and v, respectively, and $d_{i(uv)}$ is the distance between object i and the fusion of clusters u and v. In other words, distances to merged clusters are intermediate between distances to the constituent clusters. Space-conserving methods can be thought of as 'averaging' the distances to clusters merged, while space-dilating (-contracting) methods move the merged clusters further from (closer to) each other.

A number of admissibility properties were introduced by Fisher and Van Ness (1971). Such properties would be desirable qualities, other things being equal, and as such they can aid in the choice of an appropriate clustering method. One of these, (*k-group*) *well-structured admissibility,* has been related to the Lance and Williams parameters by Mirkin (1996), who terms it *clump admissibility.* (There are a number of other subtypes of well-structured admissibility, but this one relates directly to space conservation and the ultrametric condition.) Mirkin defines this property as follows:

- *Clump admissibility*: there exists a clustering such that all within-cluster distances are smaller than all between-cluster distances.

Mirkin shows that clump admissibility and space conservation is equivalent to the following conditions, for any x and y such that $0 < x < 1$ and $y > 0$:

$$\begin{aligned} &\alpha(x,y) + \alpha(1-x,y) = 1; \\ &\beta(x,1-x,y) = 0; \\ &|\gamma(y)| \leq \alpha(x,y), \end{aligned} \tag{4.14}$$

where α, β and γ are the parameters in the Lance–Williams recurrence formula, expressed as functions of the cluster sizes, with $x = n_k/n_+$, $y = n_i/n_+$ and $z = n_j/n_+$, where $n_+ = n_i + n_j + n_k$ (see Equation (4.2) and Table 4.2).

Ohsumi and Nakamara (1989) and Chen and Van Ness (1996) also relate the Lance and Williams parameters to the space-conserving property. In addition to the well-structured admissibility property, a few of the more specialized, but nonetheless useful, properties are now summarized:

- *Convex admissibility*: if the objects can be represented in Euclidean space, the convex hulls of partitions never intersect.
- *Point proportional admissibility*: replication of points does not alter the boundaries of partitions.
- *Monotone admissibility*: monotonic transformation of the elements of the proximity matrix does not alter the clustering.

Table 4.2 summarizes the mathematical properties of the well-established hierarchical methods already introduced in Table 4.1, and also gives the Lance and Williams recurrence formula (see Section 4.2.3). The ultrametric property is denoted U, and the convex, point proportion and monotone admissibility properties are denoted C, P and M, respectively.

The convex admissibility property avoids one cluster 'cutting through' another and is only appropriate when the clusters are defined in Euclidean space. In the absence of further information about cluster structure, one would prefer to avoid such cuts, although many standard procedures (for example single and complete linkage) do not in fact have this property.

Point proportionality would be relevant in situations where samples might contain replicated observations (including cases in which sets of different objects had identical characteristics). Indeed, some software packages for clustering allow *case weights* to reflect replication. An example of replicated observations would be in systems for automatic monitoring of keywords contained in web pages; these are often derived from a number of different sources (see Kirriemuir and Willett (1995), for example). Point proportionality would also be helpful in achieving robustness if there were likely to be sets of observations differing only by a small amount.

The monotone property would be appropriate where only the rank-order information was reliable; for example, where the proximity matrix contained elements which had been obtained directly from subjective ratings, such as preference or brand switching matrices in market research (see the cola example in Chapter 1).

4.4.4 Choice of partition – the problem of the number of groups

It is often the case that an investigator is not interested in the complete hierarchy but only in one or two partitions obtained from it, and this involves deciding on the number of groups present. There are a variety of formal methods that apply equally well to hierarchical clustering and optimization methods. Some of these will be discussed in Section 5.5. In this chapter we concentrate on the methods that are specific to hierarchical techniques.

In standard agglomerative or polythetic divisive clustering, partitions are achieved by selecting one of the solutions in the nested sequence of clusterings that comprise the hierarchy, equivalent to cutting a dendrogram at a particular height (sometimes termed the *best cut*). This defines a partition such that clusters below that height are distant from each other by at least that amount, and the appearance of the dendrogram can thus informally suggest the number of clusters. Large changes in fusion levels are taken to indicate the best cut. A more flexible development of this idea is 'dynamic tree cutting' (Langfelder *et al.*, 2008). This allows for different branches of the tree to be cut at different levels. The process iterates until the number of clusters is stable by combining and decomposing clusters, making successive cuts of the sub-dendrograms within clusters based on their shape. This has been implemented as a package in `R` (`dynamicTreeCut`), and is particularly appropriate where there are sets of nested clusters. As with the more inflexible fixed-height cut methods, parameters for the cut heights and the minimum cluster sizes must be chosen, so there is the possibility of influence from *a priori* expectations.

More formal approaches to the problem of determining the number of clusters have been reviewed by Milligan and Cooper (1985). They identified five best-performing rules which were further investigated by Gordon (1998), who found that the two developed by Duda and Hart (1973) and Beale (1969a) are appropriate for the nested structure inherent in hierarchical methods, since they test whether a cluster should be divided. Both are based on the ratio of between-cluster to within-cluster sums of squares, when the cluster is optimally divided into two (see Section 5.5).

Other 'number of groups' procedures which are particularly suitable for hierarchical methods have been suggested by Mojena (1977). The first is based on the relative sizes of the different fusion levels in the dendrogram and is sometimes known as the *upper tail rule.* In detail, the proposal is to select the number of groups corresponding to the first stage in the dendrogram satisfying

$$\alpha_{j+1} > \bar{\alpha} + ks_{\alpha}, \tag{4.15}$$

where $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_{n-1}$ are the fusion levels corresponding to stages with n, $n-1, \ldots, 1$ clusters. The terms $\bar{\alpha}$ and s_{α} are respectively the mean and unbiased standard deviation of the j previous fusion levels, and k is a constant. Mojena suggests that values of k in the range 2.75–3.50 give the best results, although

Milligan and Cooper suggest 1.25. Alternatively, one can use a t-distribution (although this assumes an underlying normal distribution which is clearly not applicable to fusion levels). A visual approach is to identify breaks in the plot of the values $(\alpha_{j+1}-\bar{\alpha})/s_\alpha$ against the number of clusters j.

The second method proposed by Mojena is based on a moving average approach, familiar from ideas in quality control. Here the rule is to use the partition corresponding to the first stage j, in a partial cluster sequence from $j=r$ to $j=n-2$ clusters, satisfying

$$\alpha_{j+1} > \bar{\alpha} + L_j + b_j + ks_j, \tag{4.16}$$

where $\bar{\alpha}$ and s_j are the mean and standard deviation of the fusion values, now based on the previous t-values; L_j and b_j are corrections to the mean for the upward trend in fusion values (L_j is the 'trend lag', in quality control jargon, equal under certain simplifying assumptions to $(r-1)b_j/2$, where b_j is the moving least-squares slope of the fusion levels).

According to Wishart (1987), this second rule has the advantage that the fusion level being considered does not enter into the sample statistics, but the disadvantage is that the value of r has to be chosen by the investigator. In both cases it is usual to order the criterion values and then to choose the lowest number of clusters where the rule is satisfied. Results using the upper tail rule will be illustrated in Section 4.5.5.

Given the lack of consensus about which rule to apply (and the varying results on the same data), the following comment is appropriate (Baxter, 1994): 'informal and subjective criteria, based on subject expertise, are likely to remain the most common approach. In published studies practice could be improved by making such criteria more explicit than is sometimes the case'. A discussion of whether a data set shows *any* evidence of clustering is left to Chapter 9.

4.4.5 Hierarchical algorithms

It is worth distinguishing between hierarchical *methods* and hierarchical *algorithms* for computing the clustering. For any given hierarchical method, several different computational algorithms may be used to achieve the same result. Day (1996) discusses efficient algorithms for a wide variety of clustering methods, including hierarchical techniques, and gives a comprehensive bibliography. A more recent review is by Gascuel and McKenzie (2004). Many algorithms produce a nested structure by optimizing some criterion in a stepwise manner, whereas others operate globally, for example by minimizing the distortion, as discussed in Section 4.4.2, that results from representing the proximity matrix by a hierarchical structure. These global algorithms are known as *direct optimizing algorithms*. An early example was given by De Soete (1984a); Gordon (1998) and De Soete and Carroll (1996) give more examples and further references. A method for finding *parsimonious trees* (those with a minimum number of levels

in the hierarchy) was proposed by Sriram and Lewis (1993). Direct optimizing algorithms can be useful when elements of the proximity matrix are missing (De Soete, 1984b).

Zahn (1971) gives a number of graph-theoretical clustering algorithms based on the *minimum spanning tree*. A *graph* is a set of nodes and of relations between pairs of nodes indicated by joining the nodes by *edges*. A set of observations and their dissimilarities can be represented in a graph as nodes and edges, respectively. A spanning tree of a graph is a set of edges which provides a unique path between every pair of *nodes*, and a minimum spanning tree is the shortest of all such spanning trees. Minimum spanning trees are related to single linkage algorithms (see Gower and Ross, 1969).

4.4.6 Methods for large data sets

For very large data sets, where standard methods may be unable to cope, specialized methods have been developed, for example by Zupan (1982). The use of parallel computer hardware has become a possibility (see, for example, Tsai *et al.*, 1997; Rasmussen and Willett, 1989). Some methods for large data sets combine a hierarchical method with a preclustering or sampling phase. Three that have become relatively widely used are BIRCH, CURE and `SPSS TwoStep`. BIRCH (Zhang *et al.*, 1996) employs a preclustering phase where dense regions are summarized, the summaries being then clustered using a hierarchical method based on centroids. CURE (Guha *et al.*, 1998) starts with a random sample of points, and represents clusters by a smaller number of points that capture the shape of the cluster, which are then shrunk towards the centroid so as to dampen the effects of outliers; hierarchical clustering then operates on the representative points. CURE has been shown to be able to cope with arbitrary-shaped clusters, and in that respect may be superior to BIRCH, although it does require a judgement as to the number of clusters and also a parameter which favours more or less compact clusters. `Two-Step`'s first step is similar to BIRCH in that it forms 'preclusters' by detecting dense regions; at this step, outliers (clusters with few cases, e.g. $<5\%$,) can be rejected before the next stage. The second step has some overlap with the model-based methods described in Chapter 6, in that one of the possible distance measures used is a combination of the likelihoods calculated for the continuous (assuming multivariate mixtures of normal distributions) and for the categorical variables (assuming multinomial distributions). The original paper is that by Chiu *et al.* (2001); some further details were obtained by Bacher *et al.* (2004) by personal communication from these authors (and may be subject to change as the method is developed). Bacher *et al.* described an evaluation by simulation, and found that while the method worked well for continuous variables, it was outperformed by latent class models (see Chapter 6). They also point out an important aspect of the method, to do with the analysis of data with different measurement types, which is one of its features. This is the problem of commensurability – the need to standardize variables to a common scale of measurement – as discussed in general in Chapter 3. In the `TwoStep` method

using the log-likelihood distance, a difference of 1 in a categorical variable is equal to a difference of 2.45 scale units in a z-scored variable (and other standardizations give different relative weights to continuous and categorical variables). An interesting application of the method from meteorology is given by Michailidou *et al.* (2009).

4.5 Applications of hierarchical methods

In this section we describe applications of a number of the clustering techniques discussed earlier, with particular reference to some of the points raised in Section 4.4.

4.5.1 Dolphin whistles – agglomerative clustering

In a study of dolphin behaviour, Janik (1999) used two hierarchical methods of clustering to categorize dolphin whistles. The results were compared to human classification and to a previously published *k*-means approach (see Chapter 5 for a description of *k*-means). Bottle-nosed dolphins produce a variety of whistles, but each animal usually has a stereotypical 'signature whistle', a characteristic sound which is used exclusively and frequently by that animal. The data for this study were a sample of 104 line spectrograms of the whistles of four dolphins, both signature and nonsignature. A selection of signature spectrograms is shown in Figure 4.9, which also shows the human classification (D was subdivided into D_1 and D_2).

Proximity matrices were calculated using two methods: cross-correlation between the spectrograms, after first aligning them to have maximum correlation, thus producing a similarity coefficient; and average absolute difference between the spectrograms, thus producing a dissimilarity coefficient (the city block distance; see Chapter 3). Two procedures, average linkage and complete linkage, were applied to the data. The first was chosen as a representative of a commonly used procedure in biological science, and the second selected as a procedure that would be expected to identify very stereotypical whistle types.

The results for the average linkage procedure applied to the city block (average distance) proximity matrix are shown in Figure 4.10. The dendrogram has been cut at a level where the human-identified signature whistles cluster into reasonably distinct (although not perfectly separated) groups. The other procedure gave very similar results.

The categorization by a group of five humans was taken to be 'correct' for the signature whistles, since they showed high internal consistency. Neither human nor automatic methods showed consistency in grouping the 'nonsignature' whistles, and the authors concluded that the difficulty in using automated methods was not in the methods used but in the definition of similarity. They discuss the effect of duration of the whistle in the comparison of the spectrograms, the weight to be applied to different parts of the spectrogram, and

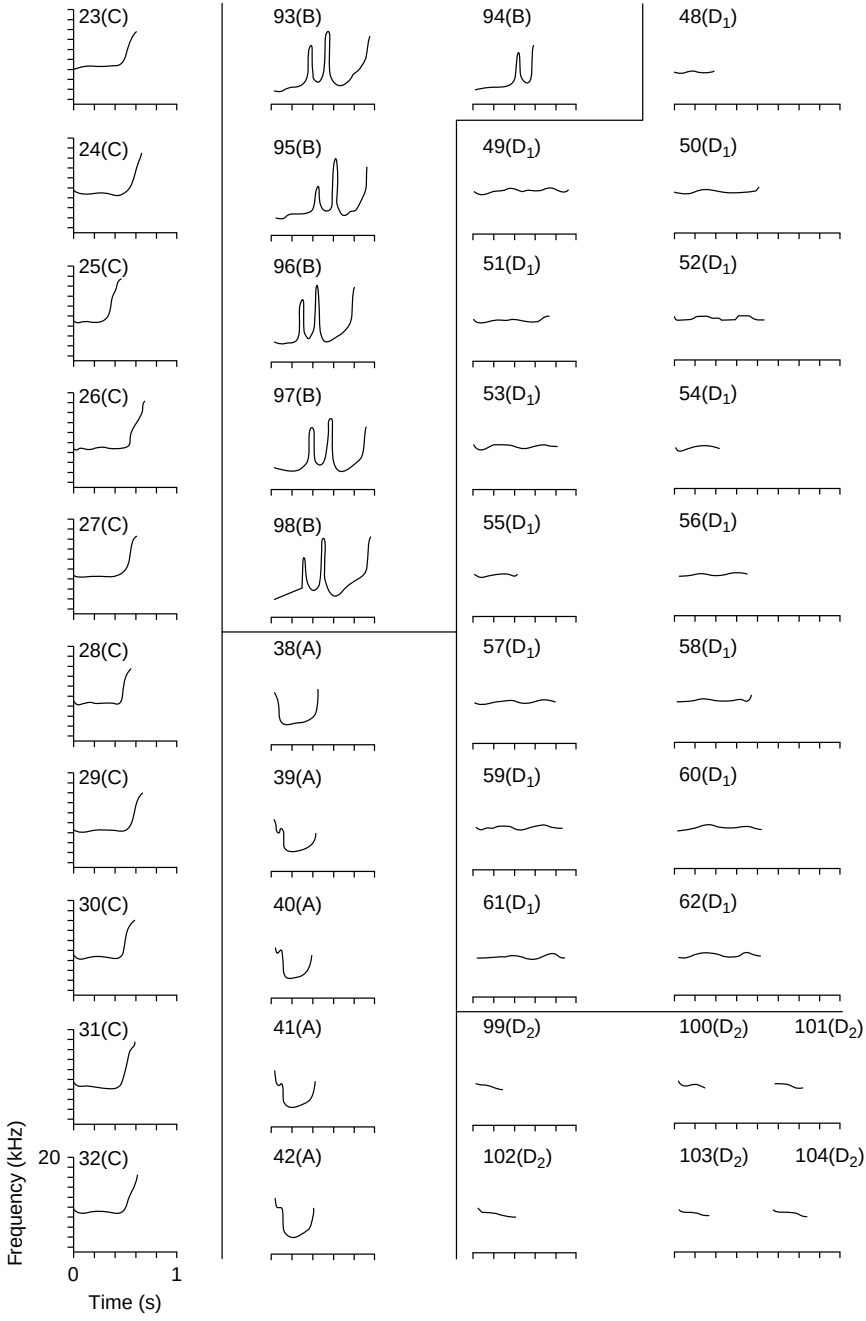


Figure 4.9 *Line spectrograms representing signature dolphin whistles. Letters indicate an expert's classification.* (Source: Janik, 1999.)

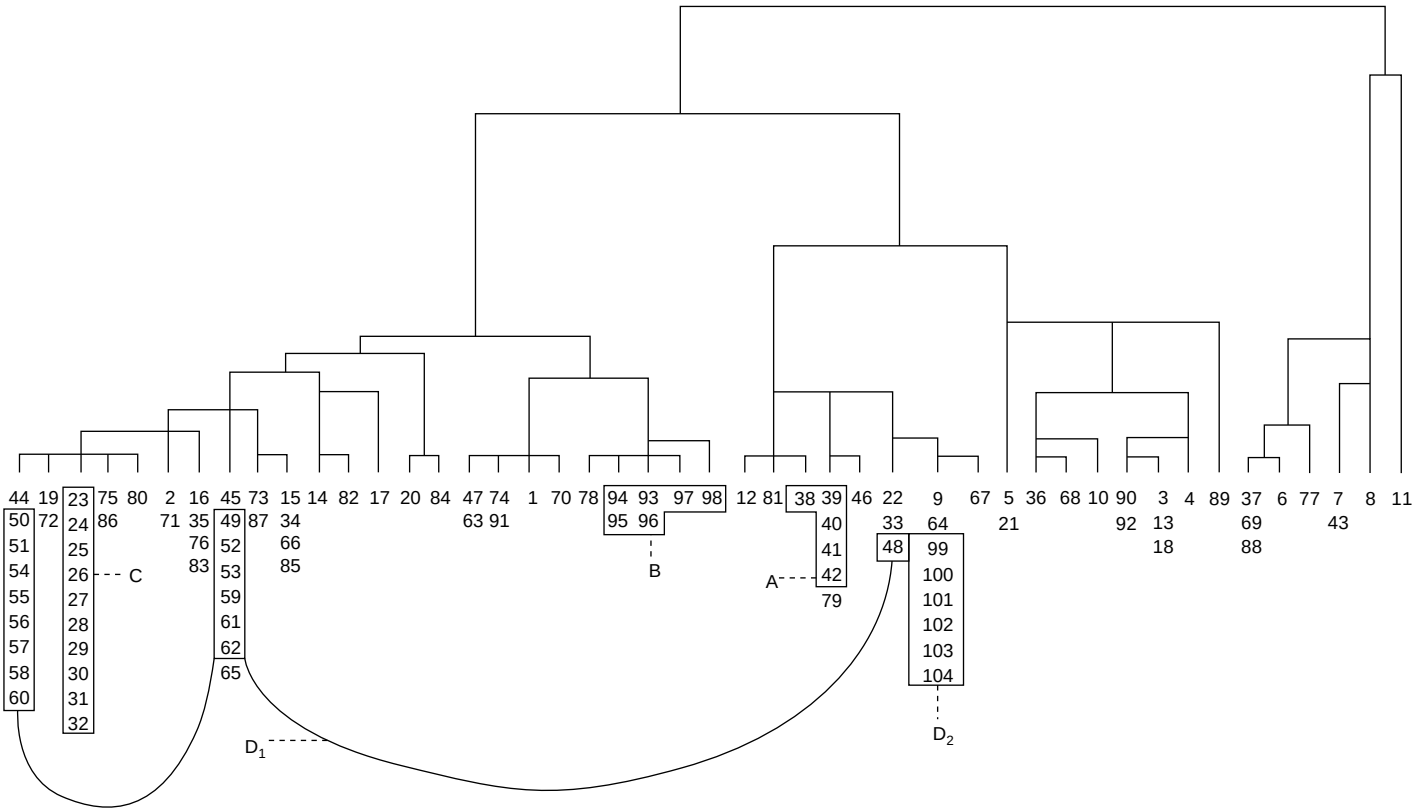


Figure 4.10 *Dendrogram of average linkage applied to dolphin whistle data, also showing human classification; see also Figure 4.9. Signature whistles are boxed.* (Source: Janik, 1999.)

whether dolphins are sensitive to the actual frequency or to the general shape of the signal. They conclude that proximity measures need to take account of the patterns that are relevant to the animal, and these need to be elicited by experimental methods.

4.5.2 Needs of psychiatric patients – monothetic divisive clustering

To illustrate the use of a monothetic divisive method, a data set from a European study of the needs, symptoms and mental health service use of schizophrenics in Amsterdam, Copenhagen, London, Verona and Santander (Becker *et al*., 1999) will be analysed. A subset of the data (male patients in London) is used. The binary variables are the presence or absence of a need in the following 'needs domains': accommodation, daytime activities, physical health, information about treatment, company. These are a subset of the 22 domains measured by the Camberwell Assessment of Need – European Version (McCrone *et al*., 2000). The data are shown in Table 4.3.

The criterion for splitting at each stage is the difference of cross-products (Equation (4.9)). At each stage the variable with the highest criterion value (summed over all the other variables) is chosen to make a split: by splitting on this variable one automatically carries other associated variables with it, so that the total number of splits is minimized. The results are illustrated in the banner plot shown in Figure 4.11.

In this case the 'best' splitting variable (that with the highest association with others) is 'information', which divides the data into 11 (with a need for information) and 23 (without an information need). Those with a need for information can then be divided into those with and without a need for daytime activities, and those without a need for information can be divided into those with and without a need for accommodation. At this separation stage there are four clusters. Looking at the right-hand side of the diagram, it can be seen that the final splits are based on the physical health needs, which are relatively unassociated with the other needs. There is one substantial group with completely homogeneous needs: five people have no need apart from physical health, a medical domain, rather than a social domain like the others. The identification of such subgroups could be used (in a larger study) to devise appropriate mental health services by grouping together people with similar needs.

4.5.3 Globalization of cities – polythetic divisive method

The following example of a polythetic divisive method involves the possible globalization of cities; that is, the tendency of cities across the globe to become (commercially) similar through the influence of multinational companies. The data describe the presence or absence of six multinational advertising agencies, and the status of the office of five multinational banks – head office, branch or agency, coded as 1, 2 or 3. The data are from Data Set 4 of the

Table 4.3 Needs of schizophrenic men in London according to the Camberwell Assessment of Need.

Patient	Accommodation	Daytime activities	Physical health	Information	Company
1	0	0	1	1	0
2	1	0	1	0	0
3	0	0	0	0	0
4	0	0	0	0	1
5	0	1	1	1	1
6	0	0	1	0	0
7	1	1	1	0	0
8	0	0	0	0	1
9	0	0	1	1	0
10	0	0	0	0	0
11	0	0	0	0	0
12	0	1	1	1	1
13	0	1	0	0	0
14	0	0	1	1	1
15	0	0	1	0	0
16	0	1	0	1	1
17	0	0	0	0	0
18	0	1	0	1	1
19	0	0	0	0	0
20	0	0	0	1	0
21	0	0	1	0	0
22	0	0	1	1	1
23	0	0	0	0	0
24	1	1	0	0	0
25	0	1	0	0	1
26	1	0	0	0	1
27	1	1	1	0	0
28	1	1	1	0	1
29	0	0	0	0	0
30	0	1	1	1	1
31	1	0	1	1	1
32	0	0	1	0	0
33	0	0	0	0	0
34	0	0	1	0	0

Globalization and World Cities Study Group and Network (GaWC) Research Group (Beaverstock *et al.*, 1999). Figure 4.12 shows the cities, which have been previously categorized as alpha, beta or gamma depending on the overall level of economic activity.

The proximity matrix was computed using Gower's mixed data coefficient (see Chapter 3), with the contributions from the binary variables (presence or absence of each advertising agency) entered into an asymmetric (Jaccard)

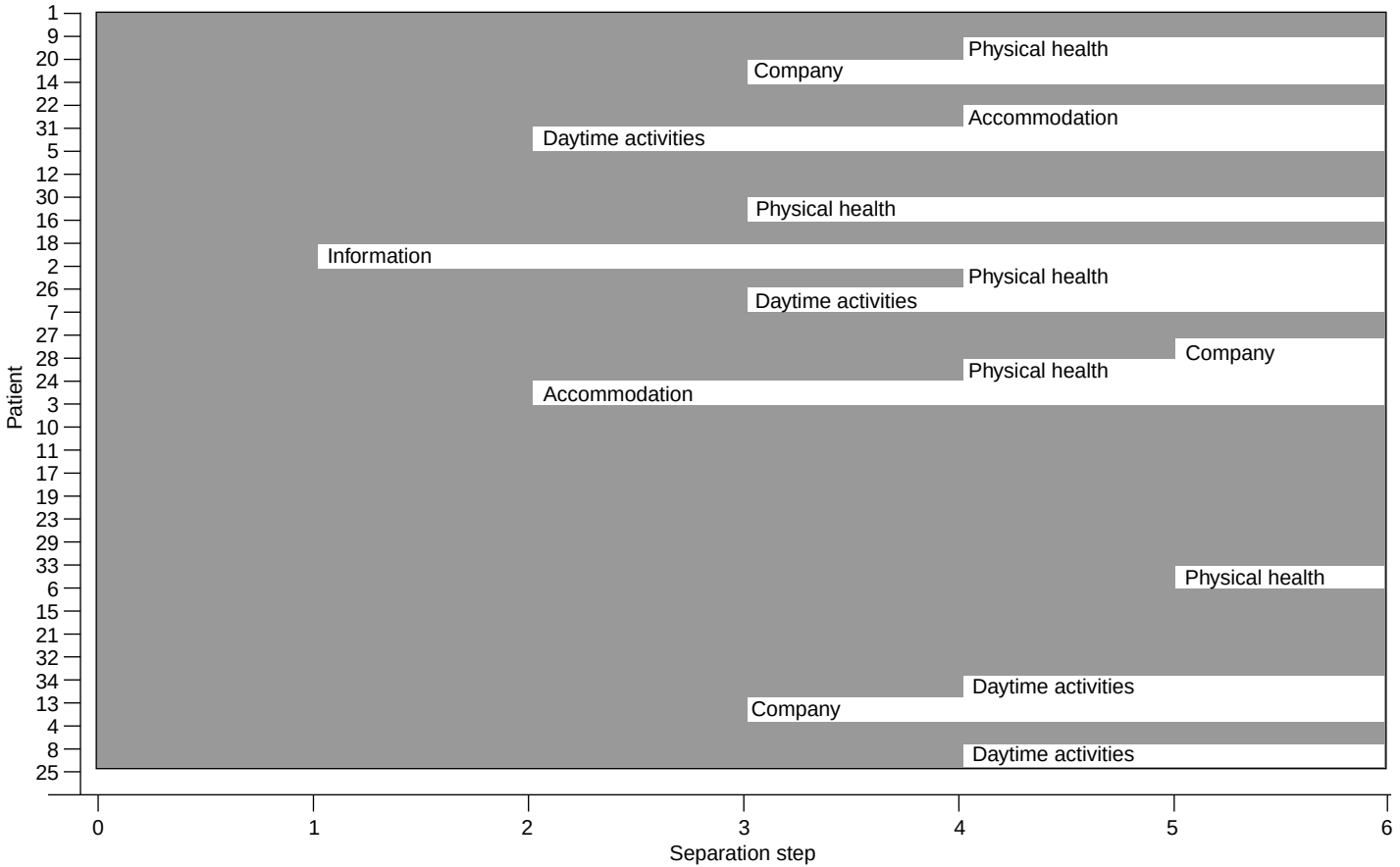
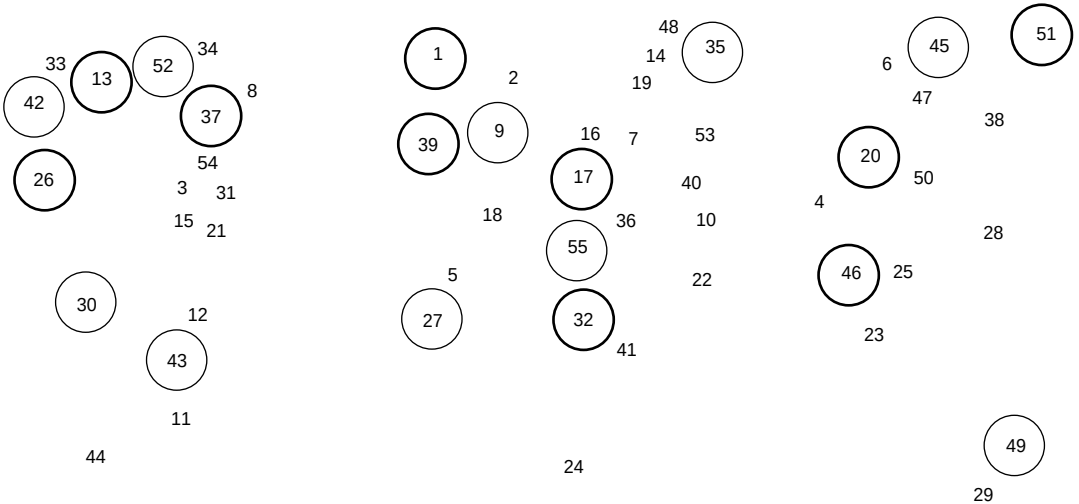


Figure 4.11 *Banner plot showing successive monothetic division of a set of male schizophrenics in London according to their needs* (see also Table 4.3)



Key Bold circles: alpha cities Other circles: beta cities Uncircled: gamma cities																					
		1	London	6	Beijing	11	Buenos Aires	16	Düsseldorf	21	Houston										
		2	Amsterdam	7	Berlin	12	Caracas	17	Frankfurt	22	Istanbul										
		3	Atlanta	8	Boston	13	Chicago	18	Geneva	23	Jakarta										
		4	Bangkok	9	Brussels	14	Copenhagen	19	Hamburg	24	Johannesburg										
		5	Barcelona	10	Budapest	15	Dallas	20	Hong Kong	25	Kuala Lumpur										
26	Los Angeles	31	Miami	36	Munich	41	Rome	46	Singapore	51	Tokyo										
27	Madrid	32	Milan	37	New York	42	San Francisco	47	Shanghai	52	Toronto										
28	Manila	33	Minneapolis	38	Osaka	43	São Paulo	48	Stockholm	53	Warsaw										
29	Melbourne	34	Montreal	39	Paris	44	Santiago	49	Sydney	54	Washington										
30	Mexico City	35	Moscow	40	Prague	45	Seoul	50	Taipei	55	Zurich										

Figure 4.12 *Diagrammatic map of global cities, classified according to overall economic activity as 'alpha', 'beta' or 'gamma'.* (Source: Beaverstock et al., 1999.)

coefficient matrix. The ordered variable (head office, subsidiary office or 'representative only', for banks) was included by treating the data as continuous and dividing by 3 to give a value between 0 and 1. An algorithm of Kaufman and Rousseeuw (1990) based on the method of MacNaughton-Smith *et al.* (1964) was applied, and this is illustrated with a dendrogram in Figure 4.13. Many of the so-called 'alpha' cities appear in the same cluster (far left); this is because the proximity measure is strongly influenced by overall activity, and there are many advertising agencies and head, rather than subsidiary, offices for banks in those areas.

Geographical clustering is also obvious and could be used to infer the common areas of influence of specific companies. The geographical clustering is not perfect: if this had been required, a *constrained* method could be used to allow clustering only of cities that were close either geographically as the crow flies or, more usefully, those with good transport links (see Chapter 8 for a more detailed discussion of constrained clustering).

4.5.4 Women's life histories – divisive clustering of sequence data

Piccarreta and Billari illustrate their CART-like monothetic divisive method described in Section 4.3.1 by analysing the employment and family trajectories of women using data from the British Household Panel Survey. The women are categorized in each month in terms of their work status (W), number of children (1, 2, $\geq$3) and whether in a co-resident partnership (U). The possible life states in each month from age 13–30 are then coded from these data, for example the state WU1 denotes work and in a partnership, with 1 child. As an illustration of the concepts described above, the sequence 0–0–W–W–WU–WU–W–W–W produces the auxiliary variable $\boldsymbol{W}_1$ (the month when a woman worked for the first time; in this example, 3), and $\boldsymbol{WU}_1$ (when she was in a partnership and work for the first time; in this example, 5). The state permanence sequence is 0^2–W^2–WU^2–W^3. Figure 4.14 shows the tree derived by the authors from the British Household Panel Survey using their new method.

A full discussion of the tree and its implications for life courses of women is given by the authors. As a starting point for orientation around the tree, we merely comment here that the first split is made on the basis of the time to reaching 'two children without work or further study'; the left-hand side of the tree thus contains women who wait to have two children later in life and who are interpreted as less family oriented. The characteristics of specific clusters are informative: for example, the medoid woman of cluster 3 (which contains 12.3% of the sample) has state permanence sequence 0^{105}–W^{31}–WU^{61}– $1WU^7$. She becomes employed just before 22 years, starts a union within 3 years after and becomes a mother about 5 years after that. This cluster is interpreted as a group of women who combine work and a (relatively delayed) family.

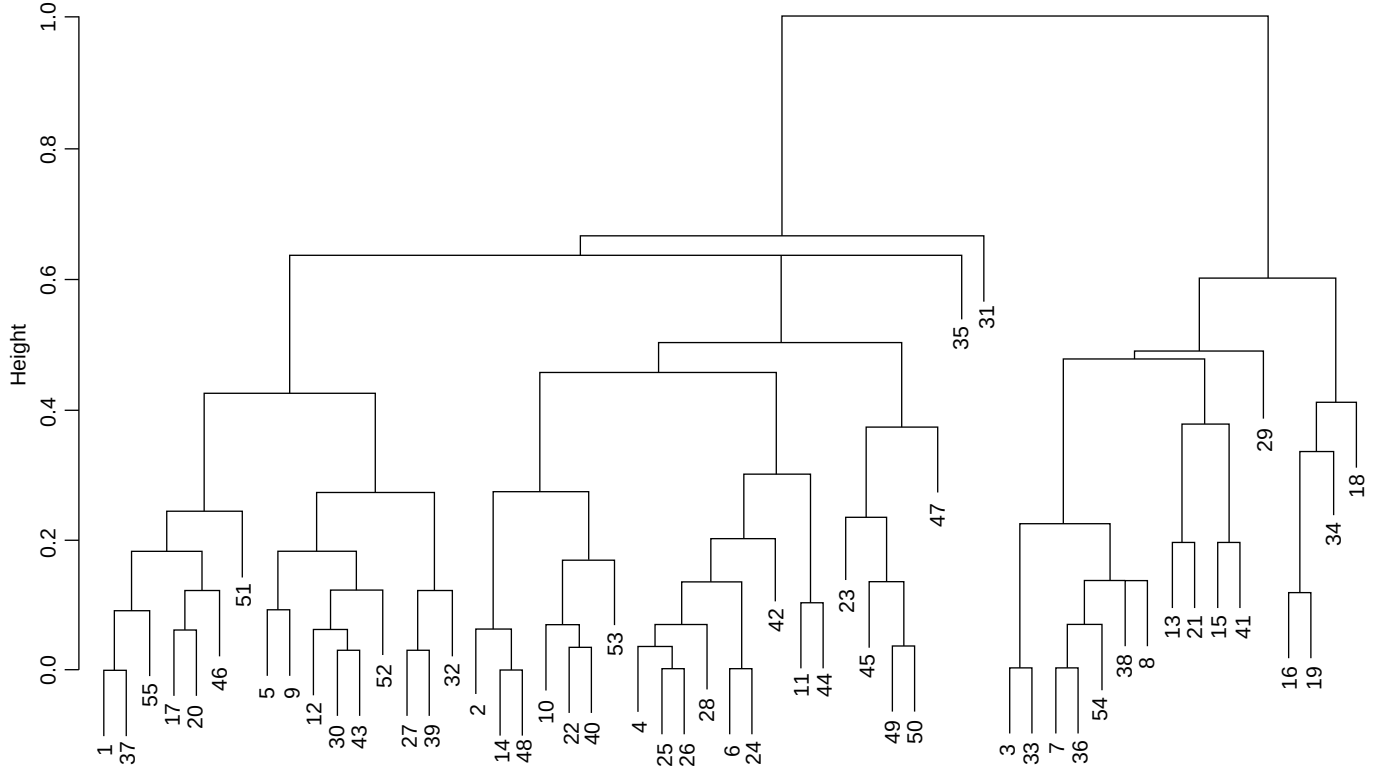


Figure 4.13 *Dendrogram showing polythetic divisive clustering of global cities, based on the strength of presence of advertising agencies and international banks (see also Figure 4.12).* Acknowledgement: the data used is from Data Set 4 from the GaWC Research Group and Network (www.lboro.ac.uk/gawc/datasets/da4.html). It was collected by J.V. Beaverstock, R.G. Smith and P.J. Taylor as part of their ESRC project 'The Geographical Scope of London as a World City' (R000222050)

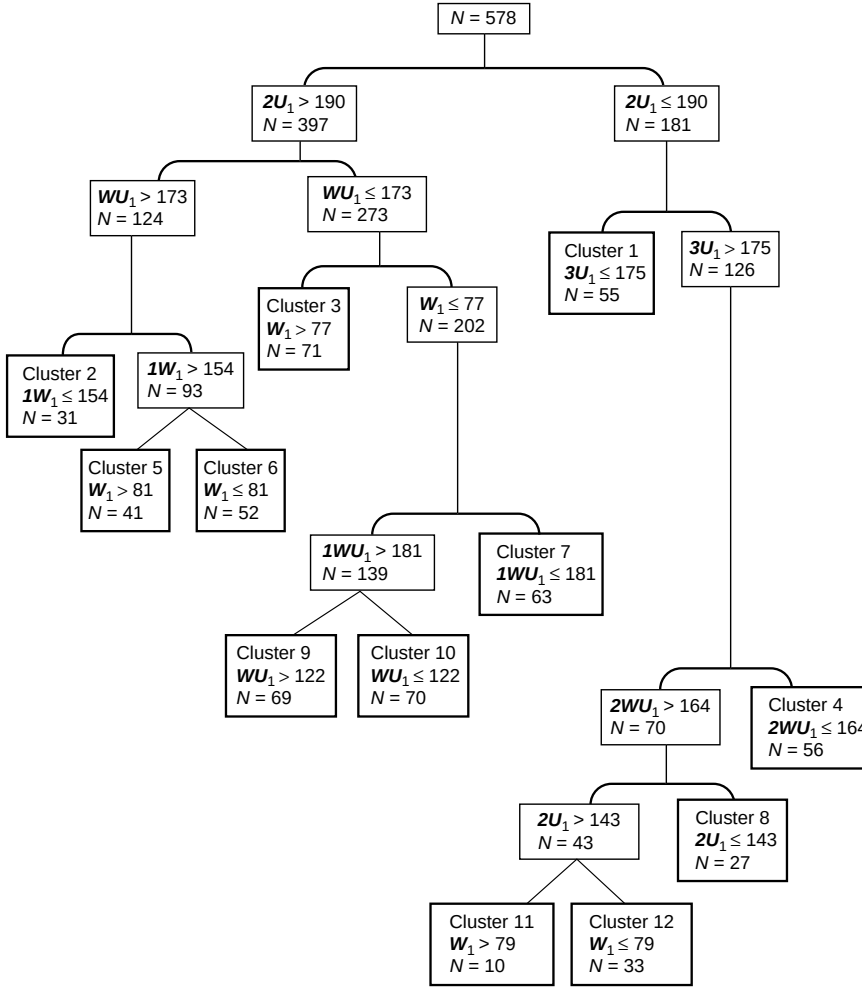


Figure 4.14 *Tree obtained from divisive cluster analysis after pruning. Auxiliary variables used are U_1 (time to being in a union); $1U_1$, $2U_1$, $3U_1$ (time to being in a union with 1, 2 or 3 or more children respectively); W_1 (time to first job); WU_1 (time to first job, being in a union without children); $1WU_1$, $2WU_1$ (time to first job, being in a union and with 1 or 2 children, respectively).*

4.5.5 Composition of mammals' milk – exemplars, dendrogram seriation and choice of partition

The compositions of various mammals' milk are given in Table 4.4 (Hartigan, 1975). Figure 4.15 shows an average linkage clustering based on Euclidean distances, in which the second dendrogram has been seriated so that the order of the labels is optimal. In this second dendrogram, unlike the first, all the equines are contiguous, as are the pairs of primates, camelids and bovines. The number of

Table 4.4 Composition of mammals' milk (percentage) standardized to have zero mean and unit standard deviation.

Mammal	Water	Protein	Fat	Lactose	Ash
Bison	0.681	−0.387	−0.818	0.856	0.073
Buffalo	0.307	−0.085	−0.229	0.310	−0.165
Camel	0.743	−0.742	−0.657	0.365	−0.303
Cat	0.268	1.064	−0.381	0.146	−0.224
Deer	−0.955	1.147	0.893	−0.836	1.063
Dog	−0.145	0.845	−0.077	−0.618	0.667
Dolphin	−2.592	1.201	2.338	−1.764	−0.660
Donkey	0.946	−1.235	−0.847	1.129	−0.918
Elephant	0.628	−0.715	0.693	0.801	−0.462
Fox	0.268	0.106	−0.419	0.419	0.132
Guinea pig	0.291	0.325	−0.295	−0.782	−0.026
Hippo	0.954	−1.536	−0.552	0.146	−1.512
Horse	0.930	−0.989	−0.885	1.511	−1.017
Llama	0.650	−0.633	−0.676	0.801	−0.125
Monkey	0.798	−1.098	−0.723	1.238	−1.353
Mule	0.923	−1.153	−0.809	0.747	−0.779
Orangutan	0.806	−1.317	−0.647	1.020	−1.234
Pig	0.362	0.243	−0.495	−0.236	0.469
Rabbit	−0.535	1.667	0.265	−1.218	2.846
Rat	−0.441	0.818	0.218	−0.454	1.063
Reindeer	−1.041	1.229	0.950	−0.891	1.063
Seal	−2.475	0.955	3.013	−2.256	−0.026
Sheep	0.299	−0.168	−0.372	0.310	0.093
Whale	−1.041	1.338	1.036	−1.382	1.658
Zebra	0.627	−0.879	−0.524	0.638	−0.323

(Taken with permission from Hartigan, 1975.)

clusters was assessed using the upper tail rule, as described in Equation (4.15); the criterion values are as follows (t-values in brackets, found by multiplying by the square root of $n-1$, where n is the number of objects):

- 2 clusters, 4.16 (20.4)
- 3 clusters, 1.59 (7.81)
- 4 clusters, 0.56 (2.73).

In this case, the four-cluster solution seems to be interpretable and is significant at $p = 0.05$, and exemplars for the four-cluster solution are underlined, with average compositions shown in Table 4.5. As might be expected, fat composition is the main identifying variable for cluster 4, with exemplar 'seal'. Cluster 3, with exemplar 'whale', is similar but with less fat and more ash. Cluster 1, with exemplar 'zebra', has the highest average lactose. Cluster 2, with exemplar 'dog', does not have any strong identifying features.

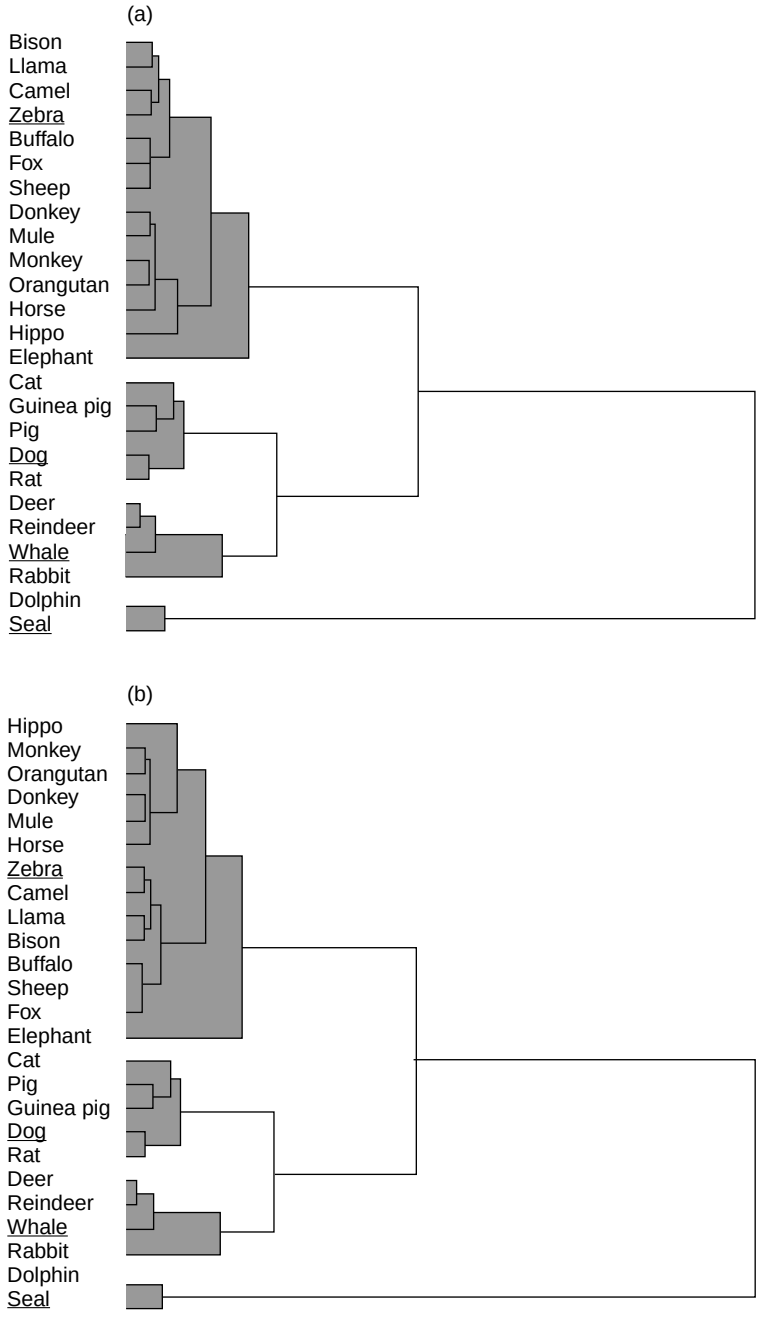


Figure 4.15 *Dendrograms for average linkage clustering of mammals' milk composition, showing 'best cut' (four-cluster solution) and the cluster exemplars; the lower dendrogram (b) was produced by seriating the objects (see also Tables 4.4 and 4.5)*

Table 4.5 Mean percentage composition of mammals' milk clusters.

Cluster (exemplar)	Water	Protein	Fat	Lactose	Ash
1 (Zebra)	85.76	3.39	4.70	5.48	0.58
2 (Dog)	79.02	8.62	8.14	3.42	1.06
3 (Whale)	66.71	11.12	18.58	2.15	1.70
4 (Seal)	45.68	10.15	38.47	0.45	0.69

4.6 Summary

Hierarchical methods form the backbone of cluster analysis in practice. They are widely available in software packages and they are easy to use, although clustering large data sets is time-consuming (methods to get round this can involve hybrid techniques which have preclustering or sampling phases and are usually available only in specialized packages). Choices that the investigator needs to make are the measure of proximity, the clustering method and, often, the number of clusters. The main problem in practice is that no particular clustering method can be recommended, since methods with favourable mathematical properties (such as single linkage) often do not seem to produce interpretable results empirically. Furthermore, to use the results involves choosing the partition, and the best way of doing this is unclear. When a particular partition is required and there is no underlying hierarchy, the methods of Chapter 5 may be more appropriate. Some of the problems of traditional hierarchical methods can be overcome by the use of model-based techniques, as will be discussed in Chapter 6.

5

Optimization clustering techniques

5.1 Introduction

In this chapter we consider a class of clustering techniques which produce a partition of the individuals into a specified number of groups, by either minimizing or maximizing some numerical criterion. Such *optimization methods* differ from the methods described in the previous chapter in not necessarily forming hierarchical classifications of the data. Differences between the methods in this class arise both because of the variety of clustering criteria that might be optimized and the various optimization algorithms that might be used. In the initial discussion of these methods it is assumed that the number of groups has been fixed by the investigator. Techniques potentially useful for suggesting the 'correct' number of clusters are described in Section 5.5.

The basic idea behind the methods to be described in this chapter is that associated with each partition of the n individuals into the required number of groups, g, is an index $c(n, g)$, the value of which measures some aspect of the 'quality' of this particular partition. For some indices high values are associated with a desirable cluster solution, whereas for others a low value is sought (see later). Associating an index with each partition allows them to be compared. A variety of such clustering criteria have been suggested. Some operate on the basis of the inter-individual dissimilarities; others employ the original data matrix.

We start by introducing cluster criteria derived from a dissimilarity matrix in Section 5.2, followed by criteria derived directly from continuous variables in Section 5.3, and then discuss algorithms that can be used to optimize these criteria in Section 5.4. Several examples of applications of cluster optimization methods are presented in Section 5.6.

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

5.2 Clustering criteria derived from the dissimilarity matrix

The concepts of *homogeneity* and *separation* can be employed to develop clustering indices. An informative partition of the objects should produce groups such that the objects within a group have a cohesive structure and with groups that are well isolated from each other (see Section 1.4). This approach is particularly useful for defining cluster criteria operating on the basis of the one-mode dissimilarity matrix Δ, with elements δ_{ij} measuring the dissimilarity between the ith and jth object. A variety of cluster criteria have been suggested, based on the δ_{ij}, that minimize the lack of homogeneity or maximize the separation of the groups. A number are listed in Table 5.1; further measures can be found in Rubin (1967), Belbin (1987) and Hansen and Jaumard (1997).

The first three measures, $h_1(m)$, $h_2(m)$ and $h_3(m)$, all quantify the lack of homogeneity, or *heterogeneity*, $h(m)$, of the mth group. The first index, $h_1(m)$, is the sum over all the (squared) dissimilarities between two objects from group m; the second, $h_2(m)$, is simply the maximum of the latter. When $r = 1$ and the dissimilarities are metric, $h_2(m)$ can be thought of as the *diameter* of the cluster. Index $h_3(m)$ measures the minimum sum over the (squared) dissimilarities between all the objects in group m and a single group member. For $r = 1$ and metric dissimilarities, the index is also referred to as the *star index*, the name arising from the graph formed by connecting all the objects of the group with the reference object. The smallest sum of distances is achieved when the reference object is located at the centre of the 'star'. In this connection, the star

Table 5.1 Measures of the (in)adequacy of the mth cluster containing n_m objects derived from a dissimilarity matrix Δ, with elements $\delta_{ql,kv}$ measuring the dissimilarity between the lth object in the qth group and the vth object in the kth group.

Measure	Index ($r \in \{1, 2\}$)
Lack of homogeneity	$h_1(m) = \sum_{l=1}^{n_m} \sum_{v=1, v\neq l}^{n_m} \left(\delta_{ml,mv}\right)^r$
Lack of homogeneity	$h_2(m) = \max_{\substack{l=1,\ldots,n_m \\ v=1,\ldots,n_m \\ v\neq l}} \left[\left(\delta_{ml,mv}\right)^r\right]$
Lack of homogeneity	$h_3(m) = \min_{v=1,\ldots,n_m} \left[\sum_{l=1}^{n_m} \left(\delta_{ml,mv}\right)^r\right]$
Separation	$i_1(m) = \sum_{l=1}^{n_m} \sum_{k\neq m} \sum_{v=1}^{n_k} \left(\delta_{ml,kv}\right)^r$
Separation	$i_2(m) = \min_{\substack{l=1,\ldots,n_m \\ k\neq m \\ v=1,\ldots,n_k}} \left[\left(\delta_{ml,kv}\right)^r\right]$

centre can be interpreted as a representative object or *examplar* of the group; Kaufman and Rousseeuw (1990) refer to it as the *medoid* (analogous to calling the group mean the *centroid*). Measures $i_1(m)$ and $i_2(m)$ quantify the separation, $i(m)$, of the mth group. Similarly to the first two heterogeneity criteria, $i_1(m)$ measures the sum over the (squared) dissimilarities between an object in the group and an object outside the group, and $i_2(m)$ is simply the minimum of the latter.

Having chosen an index measuring a group's lack of homogeneity or separation, cluster criteria can be defined by suitable aggregation over groups, for example

$$c_1(n, g) = \sum_{m=1}^{g} h(m), \tag{5.1}$$

$$c_2(n, g) = \max_{m=1,\ldots,g} [h(m)] \tag{5.2}$$

or

$$c_3(n, g) = \min_{m=1,\ldots,g} [h(m)]. \tag{5.3}$$

(Similarly, the same summary functions can be used for the separation indices.) Criterion (5.1) reflects the average lack of homogeneity, whereas criteria (5.2) and (5.3) measure the lack of homogeneity of the worst and best group, respectively. When dealing with lack of homogeneity criteria, a cluster solution is sought that minimizes the cluster criterion $c(n, g)$; when considering separation indices, the aim is to maximize $c(n, g)$. Note, however that the cluster criterion $\sum_{m=1}^{g} h_1(m)$ has the serious drawback that the number of dissimilarities contributing to it depends on the group sizes $n_m\left(\sum_{m=1}^{g} n_m = n\right)$; thus the sum might become large simply because the particular grouping into m groups produces many dissimilarities to be considered. This has led to the suggestion, that for index $h_1(m)$,

$$c_1^*(n, g) = \sum_{m=1}^{g} \frac{h_1(m)}{n_m} \tag{5.4}$$

would be a more appropriate summary function. (There are, of course, links between the cluster criteria: for example, minimizing $\sum_{m=1}^{g} h_1(m)$ is equivalent to maximizing $\sum_{m=1}^{g} i_1(m)$.) Cluster criteria might also be defined as a combination of homogeneity and separation measures.

5.3 Clustering criteria derived from continuous data

The most commonly used clustering criteria derived from a (two-mode) $n \times p$ matrix, $\mathbf{X}$, of continuous data make use of a decomposition of the $p \times p$ dispersion matrix, $\mathbf{T}$, given by

$$\mathbf{T} = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}})(\mathbf{x}_{ml} - \bar{\mathbf{x}})', \tag{5.5}$$

where $\mathbf{x}_{ml}$ is the p-dimensional vector of observations of the lth object in group m, and $\bar{\mathbf{x}}$ the p-dimensional vector of overall sample means for each variable. This total dispersion matrix can be partitioned into the within-group dispersion matrix

$$\mathbf{W} = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)', \tag{5.6}$$

where $\bar{\mathbf{x}}_m$ is the p-dimensional vector of sample means within group m, and the between-group dispersion matrix

$$\mathbf{B} = \sum_{m=1}^{g} n_m (\bar{\mathbf{x}}_m - \bar{\mathbf{x}})(\bar{\mathbf{x}}_m - \bar{\mathbf{x}})', \tag{5.7}$$

so that

$$\mathbf{T} = \mathbf{W} + \mathbf{B}. \tag{5.8}$$

For univariate data ($p = 1$), Equation (5.8) represents the division of the total sum-of-squares of a variable into the within- and between-groups sum of squares, familiar from a one-way analysis of variance. In the univariate case, a natural criterion for grouping would be to choose the partition corresponding to the minimum value of the within-group sum of squares or, equivalently, the maximum value of the between-group sum of squares.

5.3.1 Minimization of trace(W)

In the multivariate case ($p > 1$), the derivation of a clustering criterion from Equation (5.8) is not so clear cut as when $p = 1$, and several alternatives have been suggested. An obvious extension, to the multivariate case, of the minimization of the within-group sum-of-squares criterion in the univariate case is to minimize the sum of the within-group sums of squares, over all the variables; that is, to minimize trace(**W**) (which is, of course, equivalent to maximizing trace(**B**)). This can be shown to be equivalent to minimizing the sum of the squared Euclidean distances between individuals and their group mean, that is,

$$E = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)'(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m) = \sum_{m=1}^{g} \sum_{l=1}^{n_m} d_{ml,m}^2, \tag{5.9}$$

where $d_{ml,m}$ is the Euclidean distance between the lth individual in the mth group and the mean of the mth group. The criterion can also be derived on the basis of

the distance matrix

$$E = \sum_{m=1}^{g} \frac{1}{2n_m} \sum_{l=1}^{n_m} \sum_{v=1}^{n_m} d^2_{ml,mv}, \tag{5.10}$$

where $d_{ml,mv}$ is the Euclidean distance between the lth and vth individual in the mth group; thus the minimization of trace($\mathbf{W}$) is equivalent to the minimization of the lack of homogeneity criterion (5.4) for Euclidean distances and $r = 2$ in the definition of $h_1(m)$ in Table 5.1. Ward (1963) first used this criterion in the context of hierarchical clustering where only the hierarchical building process is optimized (see Section 4.2). The clustering criterion was explicitly suggested by Singleton and Kautz (1965), and is implicit in the clustering methods described by Forgy (1965), Edwards and Cavalli-Sforza (1965), Jancey (1966), MacQueen (1967) and Ball and Hall (1967).

5.3.2 Minimization of det(W)

In multivariate analysis of variance, one of the tests for differences in group mean vectors is based on the ratio of the determinants of the total and within-group dispersion matrices (see Krzanowski, 1988). Large values of $\det(\mathbf{T})/\det(\mathbf{W})$ indicate that the group mean vectors differ. Such considerations led Friedman and Rubin (1967) to suggest, as a clustering criterion, the maximization of this ratio. Since for all partitions of n individuals into g groups, $\mathbf{T}$ remains the same, maximization of $\det(\mathbf{T})/\det(\mathbf{W})$ is equivalent to the minimization of $\det(\mathbf{W})$. This criterion has been studied in detail by Marriott (1971, 1982).

5.3.3 Maximization of trace ($\mathbf{BW^{-1}}$)

A further criterion suggested by Friedman and Rubin (1967) is the maximization of the trace of the matrix obtained from the product of the between-groups dispersion matrix and the inverse of the within-groups dispersion matrix. This function is a further test criterion used in the context of multivariate analysis of variance, with large values of trace$\left(\mathbf{BW^{-1}}\right)$ indicating that the group mean vectors differ.

5.3.4 Properties of the clustering criteria

Of the three clustering criteria mentioned above, the minimization of trace($\mathbf{W}$) is perhaps the one most commonly used. It is, however, well known to suffer from some serious problems. Firstly, the method is scale dependent. Different solutions may be obtained from the raw data and from the data standardized in some way. That this is so can be seen from the criterion's equivalent definition in terms of Euclidean distances in Equation (5.9), and the effect of weighting on the latter (see Section 3.7). Clearly this is of considerable practical importance because of the need for standardization in many applications. A further problem

with the use of this criterion is that it may impose a 'spherical' structure on the observed clusters even when the 'natural' clusters in the data are of other shapes. To illustrate this problem, a two-group cluster solution was produced for two well-separated elliptical clusters generated from bivariate normal distributions used previously (see Section 4.2.4), by minimizing trace(**W**). The cluster solution is shown in Figure 5.1(a). Clearly the 'wrong' structure has been forced upon the data.

The scale dependency of the trace(**W**) method was the motivation of Friedman and Rubin's (1967) search for alternative criteria not affected by scaling. The scale independence of the criteria based on maximizing $\det(\mathbf{T})/\det(\mathbf{W})$ or $\operatorname{trace}(\mathbf{B}\mathbf{W}^{-1})$ can be seen by formulating these functions in terms of the eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\mathbf{B}\mathbf{W}^{-1}$; that is

$$\operatorname{trace}(\mathbf{B}\mathbf{W}^{-1}) = \sum_{k=1}^{p} \lambda_k \tag{5.11}$$

and

$$\frac{\det(\mathbf{T})}{\det(\mathbf{W})} = \prod_{k=1}^{p} (1 + \lambda_k). \tag{5.12}$$

Because the eigenvalues of the matrix $\mathbf{B}\mathbf{W}^{-1}$ are the same irrespective of whether this matrix is obtained from the original data matrix **X** or the weighted matrix $\mathbf{X}\,\operatorname{diag}(w_1, \ldots, w_p)$, the optimization criteria are not affected by scaling. Of course, the implication of this is that these criteria are not suitable for cluster applications where the investigator wishes to employ the variables on their original scale or wants to introduce weights based on subjective judgements.

Out of Friedman and Rubin's (1967) suggestions, the criterion of minimizing det(**W**) has been most widely used. This criterion also does not restrict clusters to be spherical, as is clearly seen in its two-group solution for the data presented in Figure 5.1. In contrast to the trace(**W**) criterion, the det(**W**) criterion is able to identify the elliptical clusters (Figure 5.1(b)). However, both the trace(**W**) criterion and the det(**W**) criterion have been found to produce groups that contain roughly equal numbers of objects, and the det(**W**) criterion, while allowing for elliptical clusters, assumes that the clusters have the same shape (i.e. the same orientation and the same elliptical degree). Again this can cause problems when the data do not conform to these requirements, when other cluster criteria may be needed.

5.3.5 Alternative criteria for clusters of different shapes and sizes

In an attempt to overcome the 'similar shape' problem of the det(**W**) criterion, Scott and Symons (1971) suggested a clustering method based on the

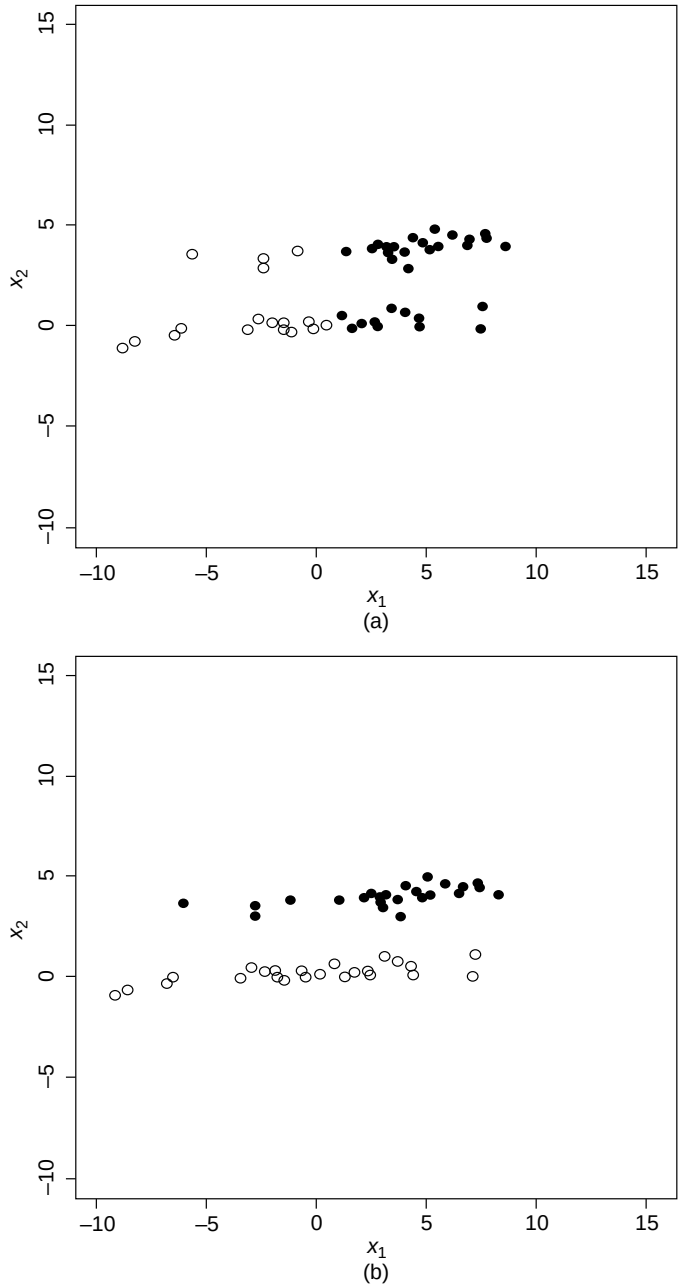


Figure 5.1 *Two well-separated clusters and cluster solutions by different optimization criteria: two-group solution obtained by minimizing (a) trace(**W**); (b) det(**W**).*

minimization of

$$\prod_{m=1}^{g} [\det(\mathbf{W}_m)]^{n_m}, \tag{5.13}$$

where $\mathbf{W}_m$ is the dispersion matrix within the mth group:

$$\mathbf{W}_m = \sum_{l=1}^{n_m} (\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)(\mathbf{x}_{ml} - \bar{\mathbf{x}}_m)', \tag{5.14}$$

and n_m is the number of individuals in the mth group. (This method is restricted to cluster solutions where each cluster contains at least $p + 1$ individuals. The restriction is necessary to avoid singular dispersion matrices, the determinants of which would be zero.) An alternative criterion described by Maronna and Jacovkis (1974) is the minimization of

$$\sum_{m=1}^{g} (n_m - 1)[\det(\mathbf{W}_m)]^{1/p}. \tag{5.15}$$

As an illustration of how such criteria may perform more successfully than the simpler minimization of det($\mathbf{W}$), 200 bivariate observations were generated by sampling 100 from each of two bivariate normal populations with mean vectors $\boldsymbol{\mu}'_1 = (0, 0)$ and $\boldsymbol{\mu}'_2 = (4, 4)$ and covariance matrices

$$\sum_1 = \begin{pmatrix} 0.94 & -0.3 \\ -0.3 & 2.96 \end{pmatrix} \text{ and } \sum_2 = \begin{pmatrix} 0.8 & 1.27 \\ 1.27 & 2.27 \end{pmatrix}.$$

The resulting two elliptical clusters are of the same size but of different shapes, as shown in Figure 5.2(a). The two-group solution given by minimizing det($\mathbf{W}$) is shown in Figure 5.2(b). Six observations belonging to the less elliptical group are misplaced because of the tendency to find clusters of similar shape. The method suggested by Scott and Symons was also applied to these data and resulted in only two misclassified objects (Figure 5.2(c)).

The tendency for criteria such as trace($\mathbf{W}$) and det($\mathbf{W}$) to give groups of equal size has been commented on by a number of authors, for example Scott and Symons (1971). An illustration of the 'similar size' problem is given in Figure 5.3.

In an attempt to overcome this problem, Symons (1981) suggested two further criteria for minimization:

$$\prod_{m=1}^{g} \left[\det\left(W/n_m^2\right)\right]^{n_m} \tag{5.16}$$

(a modification of the determinant criteria) and

$$\prod_{m=1}^{g} \left[\det\left(W_m/n_m^2\right)\right]^{n_m} \tag{5.17}$$

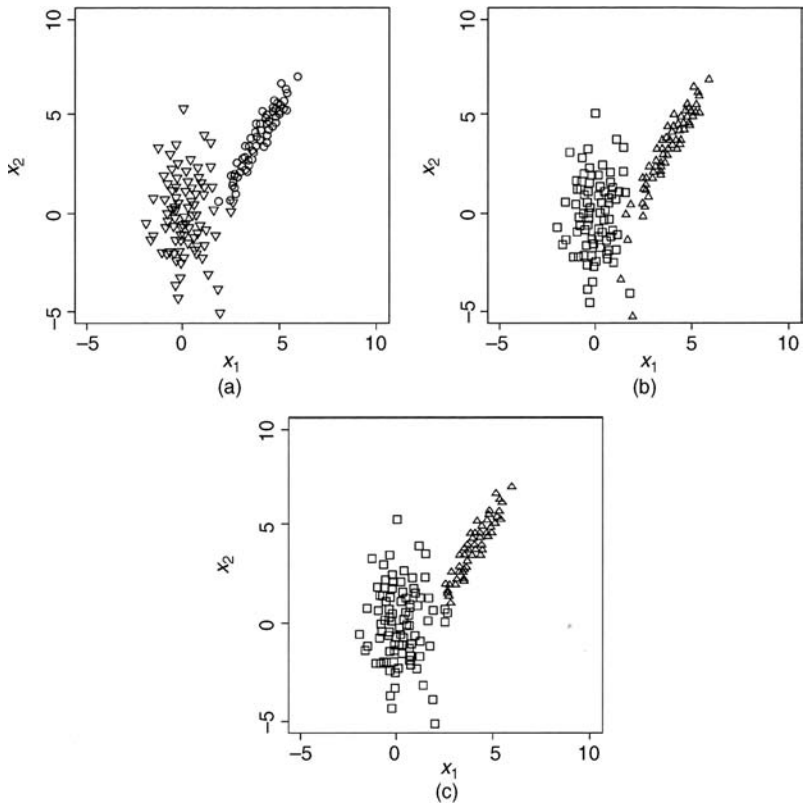


Figure 5.2 *Two elliptical clusters of same size but different shapes, and cluster solutions by different optimization criteria: (a) two simulated clusters; (b) two-group solution obtained by minimizing det(**W**); (c) two-group solution obtained by minimizing* $\prod_{m=1}^{g} [\det(\mathbf{W}_m)]^{n_m}$.

(a modification of criterion (5.13)). Marriott (1982) considered these criteria and the earlier suggestions by Scott and Symons (1971) and Maronna and Jacovkis (1974), in terms of change produced by adding an individual to a cluster. He concluded that the criteria of Symons (1981) may have several desirable properties.

The majority of the cluster criteria introduced above are heuristic. This does not mean, however, that there are no assumptions about the class structure involved. In fact, some of the criteria can be shown to be equivalent to more formal statistical criteria, in which optimizing some criterion is equivalent to maximizing the likelihood of a specific underlying probability model (Scott and Symons, 1971; Symons, 1981). Such statistical probability models will be considered in detail in the next chapter. They can help us to understand when existing cluster criteria are likely to be successful, and can be used to suggest

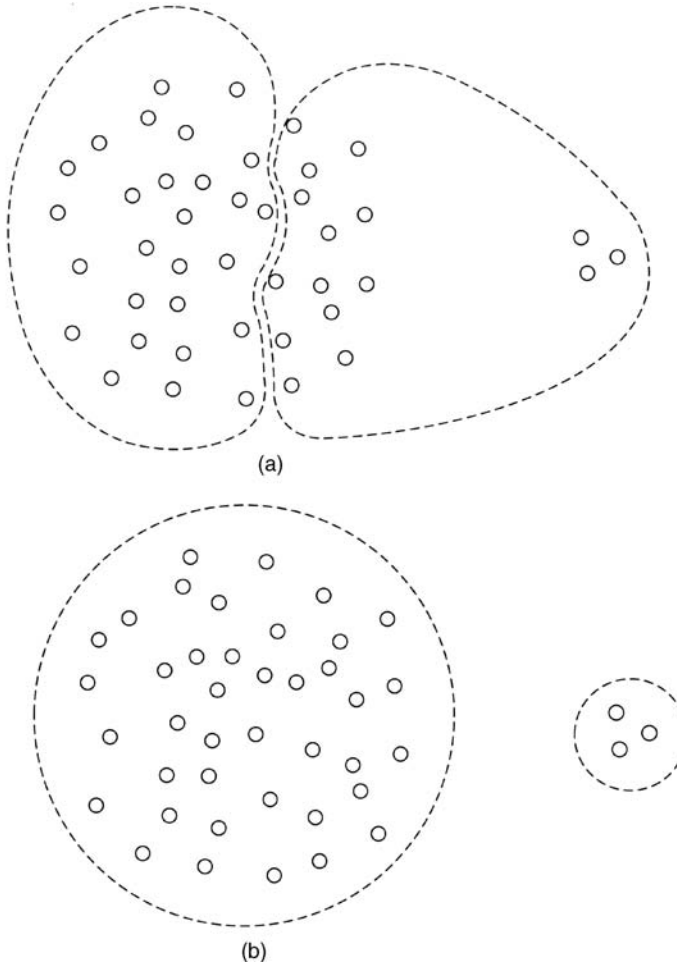


Figure 5.3 *Illustration of the 'equal-sized groups' problem of minimization of trace(**W**) clustering.*

further criteria for new situations. For example, criteria have been proposed for clusters known to be elliptical with roughly the same size and shape, but oriented in different directions (Murtagh and Raftery, 1984; Banfield and Raftery, 1993; Celeux and Govaert, 1995).

All the criteria mentioned in this section are essentially most suitable for data where all the variables are measured on a continuous scale. When the variables are not continuous, a suitable dissimilarity matrix can be generated using the measures introduced in Chapter 3, and a cluster criterion operating on the basis of the dissimilarity matrix employed. Alternatively, the dissimilarity matrix can

be transformed into a Euclidean distance matrix (see Section 3.3), and the clustering based on the representation of the objects in Euclidean space. In addition, Gower (1974) and Späth (1985) describe criteria suitable for binary and ordinal data.

5.4 Optimization algorithms

Having decided on a suitable clustering criterion, consideration needs to be given to how to find a partition into g groups that optimizes the criterion. (There might, of course, exist more than one partition that optimizes the clustering criterion.) In theory one would calculate the value of the criterion for each possible partition and choose a partition that gives an optimal value for the criterion. In practice, however, the task is not so straightforward. The number of different partitions of n objects into g groups is given by Liu (1968) as

$$N(n, g) = \frac{1}{g!}\sum_{m=1}^{g}(-1)^{g-m}\binom{g}{m}m^n. \tag{5.18}$$

There are a large number of possible partitions even for moderate n and g; for example,

$$\begin{aligned} N(2,\ 5) &= 15 \\ N(10,\ 3) &= 9330 \\ N(50,\ 4) &\approx 5.3 \times 10^{28} \\ N(100,\ 5) &\approx 6.6 \times 10^{67}. \end{aligned}$$

Thus, even with today's computers, the number of computations is so vast that complete enumeration of every possible partition is simply not possible. For some criteria an optimal partition can be identified without enumerating all partitions (Gordon, 1999, gives an example of such an exception). For other cluster criteria, explicit optimization techniques such as *dynamic programming* (Jensen, 1969; van Os and Meulman, 2004) or *branch and bound algorithms* (Koontz *et al*., 1975; Diehr, 1985; Brusco, 2006; Brusco and Kohn, 2008) can be utilized to reduce the necessary enumerations, but even with such savings global searches remain impracticable.

This problem has led to the development of algorithms designed to search for the optimum value of a clustering criterion by rearranging existing partitions and keeping the new one only if it provides improvement; these are so-called *hill-climbing algorithms* (Friedman and Rubin, 1967), although in the case of criteria which require minimization they should perhaps be termed *hill descending*. The essential steps in these algorithms are:

- Find some initial partition of the n objects into g groups.
- Calculate the change in clustering criterion produced by moving each object from its own to another group.

- Make the change which leads to the greatest improvement in the value of the clustering criterion.
- Repeat the previous two steps until no move of a single object causes the cluster criterion to improve.

An initial partition can be obtained in a number of ways. It might, for example be specified on the basis of prior knowledge, or it might be the result of a previous application of another cluster method, perhaps one of the hierarchical techniques described in the last chapter. Alternatively, an initial partition might be chosen at random or, when the objects can be represented as points in Euclidean space, g points might be selected in some way to act as cluster centres. All these methods and others have been used at some time or another – for details, see MacQueen (1967), Beale (1969a,b), Thorndike (1953), McRae (1971), Friedman and Rubin (1967) and Blashfield (1976). The last two authors find, not surprisingly, that the results from an optimization method can be radically affected by the choice of the starting partition. Different initial partitions may lead to different *local* optima of the clustering criterion, although with well-structured data it is reasonable to expect convergence to the same, hopefully *global*, optimum from most starting configurations (Hartigan (1975) considers empirical and analytical results connecting local and global optima). Marriott (1982) suggests that slow convergence and widely different groupings given by different initial partitions usually indicate that g is wrongly chosen, in particular that there is no clear evidence of clustering. It is therefore advisable to run an optimization algorithm several times with varying initial partitions.

One of the earliest hill-climbing algorithms consisted of iteratively updating a partition by simultaneously relocating each object to the group to whose mean it was closest and then recalculating the group means (Ball and Hall, 1967). Although not explicitly stated, it can be shown that, under some regularity conditions, this is equivalent to minimizing trace($\mathbf{W}$) when Euclidean distances are used to define 'closeness'. Such algorithms, involving the calculation of the mean (centroid) of each cluster, are often referred to as *k-means algorithms*. Algorithms that relocate an object into the group to whose exemplar it is nearest in terms of some dissimilarity measure, and then re-evaluate the group exemplars, have received attention in more recent years. In contrast to the group means (centroids), the group exemplars (medoids) correspond to actual objects in the data set. Minimizing dissimilarities to the exemplars is equivalent to minimizing cluster criterion $c_1(n, g) = \sum_{m=1}^{g} h_3(m)$ for $r = 1$ for a given dissimilarity matrix – also sometimes called the 'sums of the stars criterion' (Hansen and Jaumard, 1997). Depending on their originators, such algorithms are referred to as *partitioning around medoids* or PAM (Kaufman and Rousseeuw, 1990) or as *k-median algorithms* (Brusco and Kohn, 2009; Kohn *et al*., 2010). Although the four steps outlined above capture the essence of a hill-climbing algorithm, there are differences in their detailed implementations. For example, implementations of the k-means algorithm for trace($\mathbf{W}$) minimization differ in whether the objects are relocated simultaneously or singly. Single objects can be considered for relocation in a variety of orders (Friedman and Rubin, 1967; MacQueen, 1967;

Hartigan and Wong, 1979; Ismail and Kamel, 1989); for example in a random or systematic order. Objects can be relocated to the nearest group or the one that results in the largest improvement in the clustering criterion. Group means can be updated after every single relocation or after a fixed number of object relocations. Finally, a variation of the *k*-means algorithm considered by Banfield and Bassill (1977) is the pairwise exchange of group membership of two objects.

In recent years a number of new hill-climbing algorithms have been put forward which, while not using global searches, hope to escape local optima. *Simulated annealing* algorithms (Kirkpatrick *et al*., 1983; Klein and Dubes, 1989), *tabu search* algorithms (Pacheco and Valencia, 2003), *genetic algorithms* (Maulik and Bandyopadhyay, 2000) and *variable neighbourhood search* procedures (Hansen and Mladenovic, 2001) have all been suggested for their potential to find global optima. Brusco and Steinley (2007) compared the performance of nine algorithms for trace(**W**) minimization in a simulation study, and found a genetic algorithm and a variable neighbourhood search procedure to be the most effective, with a simpler procedure that consisted of applying the basic *k*-means algorithms of MacQueen (1967) using an initial partition derived from Forgy's 1965 algorithm also yielding good performance. It is interesting to note that an algorithm first suggested over 40 years ago remains competitive with much more recently suggested approaches.

Several authors have suggested modifications of the basic hill-climbing algorithm. For example, the number of groups can be modified during the process of the iterative relocation (Ball and Hall, 1967; MacQueen, 1967), or a limited number of 'atypical' objects can be allocated to an outlier group which is disregarded when the cluster criterion is evaluated (Wishart, 1987). In simulated annealing algorithms, a relocation of an object which reduces the quality of the partitions is not ruled out; instead it is assigned a small probability in an attempt to prevent the algorithm from becoming trapped in an inferior local optimum (Klein and Dubes, 1989; Selim and Asultan, 1991; Sun *et al*., 1994). In all these algorithms the cluster solution is affected by the choice of internal parameters, for example the value of the probability for carrying out a nonbeneficial relocation or the maximum size of the outlier group. This, perhaps, explains why such algorithms have not yet been widely used in cluster applications and *k*-means-type algorithms have remained popular.

We have previously discussed the problem of appropriate standardization when optimizing a criterion that is not invariant under scaling, such as trace(**W**) (see Section 3.8). So-called *adaptive procedures* can be used to resolve the circular problem of (i) needing to know the group membership to standardize variables to unit within-group variance, while (ii) needing to know the standardized variables to estimate group membership. For example, *k*-means clustering can be modified to become such an adaptive procedure by recalculating distances at each step of the procedure using Mahalanobis's generalized distance measure based on the current grouping (Diday and Govaert, 1977). Similar ideas underlie the adaptive clustering algorithms suggested by Lefkovitch (1978, 1980), De Sarbo *et al*. (1984), Fowlkes *et al*. (1988), Gnanadesikan *et al*., (1995) and Steinley and Brusco (2008a).

Finally, the use of mathematical programming algorithms for finding optima of clustering criteria is discussed by Hansen *et al.* (1994); Hansen and Jaumard (1997) and Gordon (1999). Further algorithms for obtaining optimal partitions, for example the use of minimum spanning trees, are reviewed in Gordon (1999).

5.4.1 Numerical example

As a simple illustration of a *k*-means algorithm we will consider the following data set consisting of the scores of two variables on each of seven individuals.

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

This data set is to be clustered into two groups using the minimization of trace(**W**) cluster criterion. As a first step in finding a sensible initial partition, let the variable values of the two individuals furthest apart (using the Euclidean distance measure) define the initial cluster means, giving

	Individual	Mean vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

The remaining individuals are now examined in sequence and allocated to the group to which they are closest, in terms of Euclidean distance to the group mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps:

	Group 1		Group 2	
	Individual	Mean vector	Individual	Mean vector
Step 1	1	(1.0, 1.0)	4	(5, 7)
Step 2	1, 2	(1.2, 1.5)	4	(5, 7)
Step 3	1, 2, 3	(1.8, 2.3)	4	(5, 7)
Step 4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
Step 5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
Step 6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

This gives the initial partition, the two groups at this stage having the following characteristics:

	Individual	Mean vector	trace($\mathbf{W}_i$)
Group 1	1, 2, 3	(1.8, 2.3)	6.84
Group 2	4, 5, 6, 7	(4.1, 5.4)	5.38

Thus at this point we have trace($\mathbf{W}$) = 6.84 + 5.38 = 12.22. The first relocation step of the k-means algorithm now compares each individual's distance to its own group mean against its distance to the opposite group mean. We find:

Individual	Distance to mean of group 1	Distance to mean of group 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Only individual 3 is nearer to the mean of the opposite group (group 2) than its own (group 1). Thus individual 3 is relocated to group 2, resulting in the new partition:

	Individual	Mean vector	trace($\mathbf{W}_i$)
Group 1	1, 2	(1.3, 1.5)	0.63
Group 2	3, 4, 5, 6, 7	(3.9, 5.1)	7.9

The move causes a decrease in clustering criterion to trace($\mathbf{W}$) = 0.63 + 7.9 = 8.53. The iterative relocation would now continue from this new partition; however, in this example each individual is now nearer its own group mean than that of the other group and the iteration stops, choosing the latest partitioning as the final cluster solution.

5.4.2 More on *k*-means

The k-means algorithm described earlier in the chapter continues to attract attention despite having been around for more than 40 years. In the context of choosing starting values, for example, Steinley (2003) pointed out that the number of local optima for data sets of moderate size can run into the thousands, so that the results from studies using a small number of random starts may remain misleading. Elaborate approaches for determining starting values have been suggested; for

example, a bootstrap-like approach to 'refine' initial seeds (Bradley and Fayyad, 1998) and methods that restrict initial centroids to be chosen from areas of high density (e.g. Cerioli and Zani, 2001). Steinley (2003) suggests the following procedure: repeatedly (>5000 times) randomly partition the objects into k groups, calculate initial centroids from these partitions, run a k-means algorithm for each of these starting values and choose as the final solution the partition that minimizes trace($\mathbf{W}$) over the repeated runs. The author shows that his approach outperforms several other methods used in commercial software packages, and in later simulations shows that it also outperforms other competitor approaches (Steinley and Brusco, 2007). Further use can be made of the set of locally optimal k-means solutions, by combining information regarding the cluster problem (number of clusters, number of variables, sample size, etc.) with the distribution of the local optima created from the multiple initializations, to determine the quality of a cluster solution (Steinley, 2006a). Finally, the set of k-means solutions can provide the basis of a so-called 'stability analysis' which can identify the most stable solution, suggest the number of clusters and empirically derive cluster membership probabilities (Steinley, 2008). For more details on developments with the k-means approach see Steinley (2006b).

5.4.3 Software implementations of optimization clustering

Nowadays most general-purpose statistical packages such as `SPSS`, `Stata`, `R` and `SAS` have implemented algorithms for optimization clustering. Virtually all packages provide a k-means algorithm for trace($\mathbf{W}$) minimization. The k-median algorithm, which is often suggested as an alternative to k-means due to its appealing use of exemplars instead of means, robustness towards outliers and its ability to accommodate a range of dissimilarity measures, is available in `Stata` and `R` (`pam` from package `cluster`, or `kcca` from package `flexclust`). The `R` toolbox implemented in `R` package `flexclust` further provides a framework for dealing with more unusual dissimilarity measures such as the jacknife correlation, which has received some attention in clustering gene expression profiles; for more details see Leisch (2006).

5.5 Choosing the number of clusters

In most applications of optimization methods of cluster analysis, the investigator will have to 'estimate' the number of clusters in the data. A variety of methods have been suggested which may be helpful in particular situations. Most are relatively informal and involve, essentially, plotting the value of the clustering criterion against the number of groups. Large changes of levels in the plot are usually taken as suggestive of a particular number of groups. Like similar procedures for judging dendrograms (see Section 4.4.4), this approach may be very subjective, with 'large' often being a function of the user's prior expectations.

A number of more formal techniques have been suggested which try to overcome the problem of subjectivity (Milligan and Cooper (1985) list 30 such

methods). While many different methods have been proposed, there have been only limited investigations into their properties. The most detailed comparative study of the performance of techniques for determining the number of groups has been carried out by Milligan and Cooper (1985), and more recently 15 indices for high-dimensional binary data were assessed by Dimitriadou *et al.* (2002). Both studies assess the ability of formal/automated methods to detect the correct number of clusters in a series of simulated data sets. As with all simulation studies, their findings cannot be generalized, since the performance of a method might depend on the (unknown) cluster structure as well as the cluster algorithm employed to determine group membership. However, the studies are valuable in identifying methods which perform poorly even under well-defined cluster structures.

The two top performers in Milligan and Cooper's study were techniques introduced by Calinski and Harabasz (1974) and Duda and Hart (1973) for use with continuous data. Calinski and Harabasz (1974) suggest taking the value of g, the number of groups, which corresponds to the maximum value of $C(g)$, where $C(g)$ is given by

$$C(g) = \frac{\text{trace}(\mathbf{B})}{(g-1)} \Big/ \frac{\text{trace}(\mathbf{W})}{(n-g)}. \tag{5.19}$$

As with all techniques for determining the number of groups, the evaluation of this criterion at a given number of groups, g, requires knowledge of the group membership to determine the matrices $\mathbf{B}$ and $\mathbf{W}$. In general, the number of groups chosen depends on the cluster method (and implementation) used.

Duda and Hart (1973) offer a criterion for dividing the mth cluster into two subclusters. They compare the within-cluster sum of squared distances between the objects and the centroid, $J_1^2(m)$, with the sum of within-cluster sum of squared distances when the cluster is optimally divided into two, $J_2^2(m)$. The null hypothesis that the cluster is homogeneous should be rejected (and the cluster subdivided) if

$$L(m) = \left(1 - \frac{J_2^2}{J_1^2} - \frac{2}{\pi p}\right) \left\{ \frac{n_m p}{2[1-8/(\pi^2 p)]} \right\}^{1/2} \tag{5.20}$$

exceeds the critical value from a standard normal distribution (here p is the number of variables and n_m the number of objects in cluster m). In contrast to (5.19), Duda and Hart's suggestion represents a *local* criterion. This can be converted into a *global* criterion for deciding whether an additional group is present by considering the set of test statistics, $\{L(m) : m = 1, \ldots, g\}$, for all groups. The null hypothesis of homogeneous groups is rejected in favour of a further group when at least one test statistic exceeds its critical value. (However, when proceeding in this way the significance levels should not be interpreted strictly because of the multiple testing.)

A further rule operating on the sum of the squared distances, which was also one of the better performers in Milligan and Cooper's study, is an 'F-test' proposed by Beale (1969a). Let S_g^2 denote the sum of the square deviations from cluster

centroids in the sample. Then a division of the n objects into g_2 clusters is significantly better than a division into g_1 clusters ($g_2 > g_1$) if the test statistic

$$F(g_1, g_2) = \frac{\left(S^2_{g_1} - S^2_{g_2}\right)/S^2_{g_2}}{[(n-g_1)/(n-g_2)](g_2/g_1)^{2/p} - 1}. \tag{5.21}$$

exceeds the critical value from an F-distribution with $p(g_2 - g_1)$ and $p(n - g_2)$ degrees of freedom.

Marriott (1971) proposes a possible procedure for assessing the number of groups when using minimization of det($\mathbf{W}$) as the chosen clustering criterion. He suggests taking the value of g for which $g^2\det(\mathbf{W})$ is a minimum. For unimodal distributions, Marriott shows that this is likely to lead to accepting a single group ($g = 1$), and for strongly grouped data it will lead to the appropriate value of g. In addition, for given g, the associated statistic, $g^2\det(\mathbf{W})/\det(\mathbf{T})$, whose value decreases with increasing degree of clustering, can be used to test for evidence of cluster structure. In particular, if the test statistic has a value greater than 1 for all possible subdivisions, the objects should be regarded as forming a single group. The sampling properties of the test statistic under the uniformity hypothesis (a special case of absence of clustering) can be investigated by Monte Carlo methods (more details are given in Chapter 9). The rule was found by Milligan and Cooper (1985) to have a tendency to specify a constant number of clusters.

The methods for choosing the number of groups described so far all assume that the variables are measured on a continuous scale. An example of a procedure which can also be used for categorical data is an adaptation of Goodman and Kruskal's gamma statistic for use in classification studies (Baker and Hubert, 1975). This procedure operates on a dissimilarity matrix, with each within-group dissimilarity being compared with each between-group dissimilarity. In this context a pair of dissimilarities is deemed *concordant* (*discordant*) if the within-cluster dissimilarity is strictly less (strictly greater) than the between-cluster dissimilarity. The concordance index, $I(g)$, is defined as

$$I(g) = \frac{S_+ - S_-}{S_+ + S_-} \in [-1, 1], \tag{5.22}$$

where S_+ and S_- are the number of concordant and discordant pairs, respectively. The number of groups, g, is chosen so that $I(g)$ is a maximum. This rule was found to perform well by Milligan and Cooper (1985).

A further diagnostic that is helpful for determining the number of groups which also operates on the basis of the dissimilarity matrix is the *silhouette plot* suggested by Kaufman and Rousseeuw (1990) and implemented in the `R` package `cluster`. For each object i they define an index $s(i) \in [-1, 1]$, which compares object i's separation from its cluster against the heterogeneity of the cluster (for the exact definition of this index on the basis of the dissimilarities, see Kaufman and Rousseeuw, 1990). When $s(i)$ has a value close to 1, the

heterogeneity of object i's cluster is much smaller than its separation and object i is taken as 'well classified'. Similarly, when $s(i)$ is close to -1 the opposite relationship applies and object i is taken to be 'misclassified'. When the index is near zero it is not clear whether the object should have been assigned to its current cluster or a neighbouring cluster. In the silhouette plot the $s(i)$ are displayed as horizontal bars, ranked in decreasing order for each cluster (an example will be shown in the next section, see Figure 5.5). The silhouette plot is a means of assessing the quality of a cluster solution, enabling the investigator to identify 'poorly' classified objects and so distinguishing clear-cut clusters from weak ones. Silhouette plots for cluster solutions obtained from different choices for the number of groups can be compared, and the number of groups chosen so that the quality of the cluster solution is maximized. In this respect the *average silhouette width* – the average of the $s(i)$ over the entire data set – can be maximized to provide a more formal criterion for selecting the number of groups. Kaufman and Rousseeuw (1990) also give some guidance as to the desirable size of the silhouette width; they consider a reasonable classification to be characterized by a silhouette width above 0.5 and point out that a small silhouette width, say an average width below 0.2, should be interpreted as a lack of substantial cluster structure.

More recently, the so-called GAP-statistic has been proposed in the statistical literature as a measure for estimating the number of clusters (Tibshirani *et al.*, 2001). Tibshirani and colleagues develop an approach that formalizes the idea of finding an 'elbow' in the plot of the optimized cluster criterion against the number of clusters, g. Their idea is to standardize the graph of $\log[C(n, g)]$ against the number of clusters, where $C(n, g)$ is a cluster criterion that has been minimized for g clusters, by comparing it with its expectation under an appropriate null reference distribution. For this purpose, letting E_n^* denote the expectation under a sample size of n from the reference distribution, they suggest that the optimal value for the number of clusters is the value g for which the 'gap'

$$\mathrm{GAP}_n(g) = \mathrm{E}_n^*\{\log[C(n, g)]\} - \log[C(n, g)] \tag{5.23}$$

is largest. They show that Monte Carlo simulation from a uniform distribution over a box aligned with the principal components of the data can serve to generate a suitable null reference distribution, and demonstrate a promising performance of the GAP-statistic in a simulation study (Tibshirani *et al.*, 2001). Their procedure transforms the optimized cluster criterion onto the log-scale, so that maximizing the absolute discrepancy with expected values amounts to maximizing the relative (factor) discrepancy on the original scale. Importantly their procedure allows evaluation of the quality of the single cluster solution, and this enables an investigator to address the question of whether there is any evidence for the existence of distinct clusters in the data or whether they are best regarded as a single homogeneous group (for applications see, e.g., Nazareth *et al.*, 2006 or King *et al.*, 2007). We return to this important issue in Chapter 9.

General-purpose statistical packages tend to contain one or other of the better-known stopping rules. For example, `Stata` provides the Calinski–Harabasz and Duda–Hart procedures for determining the number of clusters. The `R` package `clusterSim` also contains some of the more recently introduced criteria (including Silhouette index and GAP-statistics).

In conclusion, it is advisable not to depend on a single rule for selecting the number of groups but to synthesize the results of several techniques. Also, like the cluster criteria themselves, some rules for choosing the number of clusters make assumptions about the cluster structure and will only perform well when these assumptions are met. For example, experience with Beale's rule suggests that it will only be successful when the clusters are fairly well separated and approximately spherical in shape. Some examples of the use of techniques for selecting the number of groups will be given in the next section.

5.6 Applications of optimization methods

In this section several applications of optimization-type clustering are discussed, beginning with two data sets introduced in previous chapters.

5.6.1 Survey of student attitudes towards video games

To illustrate the application of optimization methods operating on the dissimilarity matrix, we return to the survey of university students' preferences and attitudes towards video games presented in Table 3.5. The table lists seven variables assessing students' attitudes towards video gaming. We would like to discover whether there are groups of university students that are similar in their attitudes and different from students in other groups. In Chapter 3 we used Gower's general similarity measure to construct a dissimilarity matrix (displayed in Table 3.6) since the variables were of mixed type. We obtained a cluster solution by minimizing the sum of the dissimilarities between the students and their cluster exemplar students (medoids). A partitioning around medoids algorithm (`pam` in `R` – for details see Kaufman and Rousseeuw (1990)) was run five times to produce partitions into two to six groups, and silhouette plots were employed to choose the number of groups. We obtained the following average silhouette widths:

No. of groups (g)	Average silhouette width
2	0.32
3	0.32
4	0.26
5	0.30
6	0.32

In this example the highest average widths were achieved by a two-, three- and six-cluster solution. The six-cluster solution created clusters with less

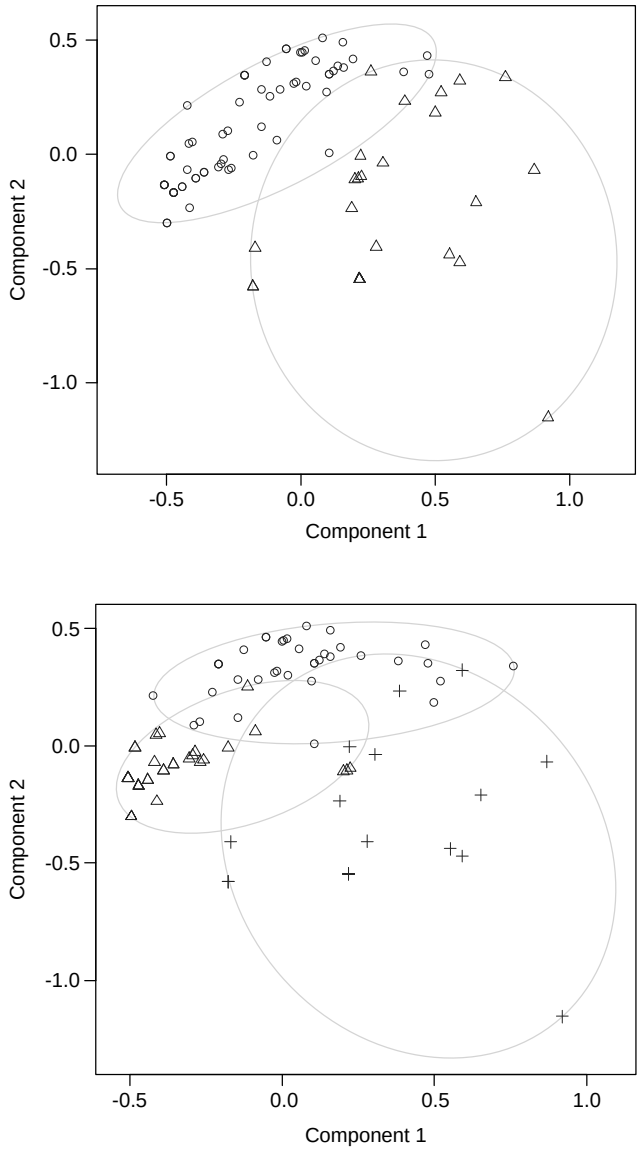
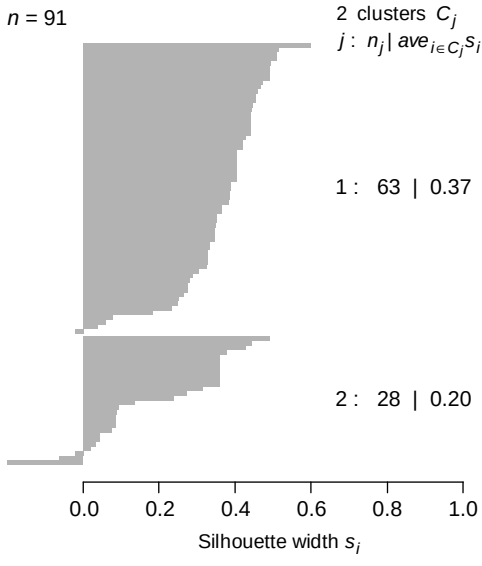


Figure 5.4 *Cluster solutions for video survey data obtained by minimization of the sum of the Gower dissimilarities between the students and their cluster medoids, displayed in the space of the first two principal components. (a) Two-group solution; (b) three-group solution.*

than 10 students and was not considered further. Figure 5.4 displays the two remaining competing partitions in the space of the first two principal components, and Figure 5.5 shows the full silhouette plots for the cluster solutions. The figures show that the three-cluster solution is more or less created by splitting



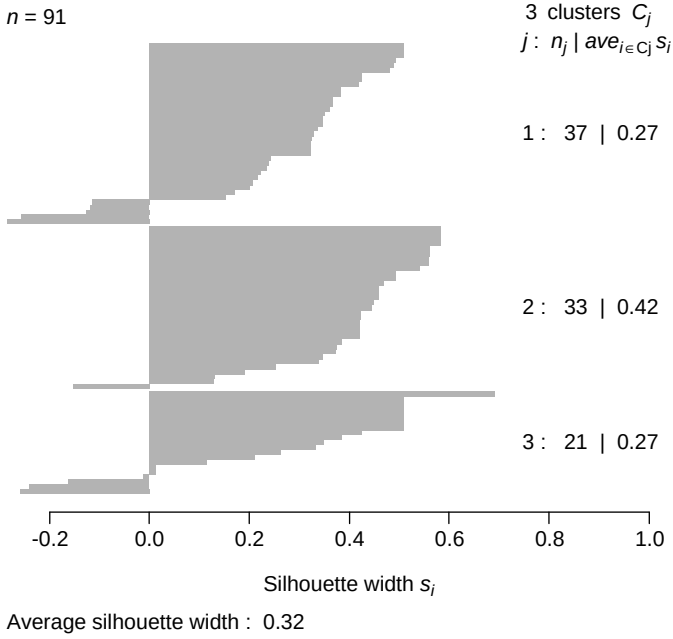


Figure 5.5 *Silhouette plots for cluster solutions for video survey data obtained by minimization of the sum of the dissimilarities between the students and their cluster medoids. (a) Two-group solution; (b) three-group solution.*

cluster 1 from the two-cluster solution into two subgroups. Although none of the partitions would be called a 'reasonable classification' by Kaufman and Rousseeuw (1990), here, for illustrative purposes, we do consider the more parsimonious two-cluster solution further.

The cluster medoids can be used to represent the clusters. The two groups of students can be characterized as follows:

- Cluster 1 ($n = 63$): students who have played video games before and did not spend any time playing video games in the week prior to the survey, who generally like it 'somewhat', and play on home computers with a semesterly frequency but do not play when they are busy.
- Cluster 2 ($n = 28$): students who have played video games before and did not spend any time playing video games in the week prior to the survey, but did not provide any information on detailed behaviour.

Here the medoid for cluster 2 is of limited descriptive use, as the exemplar student is one who did not answer most of the questions. Further assessment of the 28 students who were allocated to cluster 2 shows that when they answered respective questions they included students who spend many hours (>10 hours in the week prior to the survey) playing video games, were more likely to play in the home at a computer and a system, and to play both at home and in arcades, tended to play more frequently (more daily) and were more likely to play during busy periods.

5.6.2 Air pollution indicators for US cities

We now return to the data shown in Table 2.3. These data relate to air pollution in 41 US cities. Air pollution is measured by the annual mean concentration of sulphur dioxide, in micrograms per cubic metre. The other six variables are indicators of climatic conditions and human ecology. Here we employ an optimization clustering technique to investigate whether the US cities could be grouped with regard to climatic and ecological indicators and then such a grouping used to predict air pollution. We used a k-means clustering algorithm which minimized the sum of the squared Euclidean distances between the cities and their cluster means. Since the indicators were measured on different scales (for example average annual temperature was measured in degrees Fahrenheit while average wind speed was measured in miles per hour) and the k-means algorithm is not invariant under scaling, each of the indicators was standardized to unit variance prior to cluster analysis.

The algorithm was run to obtain two- to six-group cluster solutions, and Calinski and Harabasz's (1974) rule used to choose the number of clusters. We obtained the following value for criterion $C(g)$, defined in Equation (5.19):

No. of groups (g)	$C(g)$
2	11.6
3	13.8
4	17.0
5	14.3
6	15.8

Since the criterion had its maximum value for $g = 4$, we chose the four-group cluster solution. (In fact other rules, for example average silhouette widths, would also have led to the choice of four groups). Figure 5.6 displays this solution on a map of the USA with each city's state being labelled with the appropriate cluster number. Clusters appear to correspond to geographical location, with cities in group 3 being located in the west, cities in group 2 mainly in the north east, and cities in group 1 in the south east of the USA. The cities' group memberships together with group means are shown in Table 5.2.

The group means (Table 5.2) can be used to interpret the four-group cluster solution. However, when the variables are continuous the interpretation can be aided by displaying the whole set of objects in the space of the original variables indicating their cluster membership; here we achieve this by means of a scatterplot matrix (Figure 5.7). On the basis of these diagnostics we labelled the clusters as follows:

1. *Humid climate*: a group of cities showing large amounts of precipitation in combination with high temperatures.
2. *Wet, cold and windy climate*: a group of cities characterized by a large number of days with precipitation per year in combination with low temperatures and high wind speeds.
3. *Dry climate*: a group of cities characterized by little precipitation (less than 21 inches of average annual precipitation and less than 90 days with precipitation per year).
4. *Dense population*: a group of cities characterized by high values for human ecology indicators. The two group members Chicago and Philadelphia have both the largest numbers of manufacturing enterprises employing 20 or more workers and the largest population sizes.

It is also interesting to note that the groups identified by the formal application of a cluster algorithm are consistent with those indicated by the explorative kernel estimators in Section 2.2.4: the outlying observations in Figure 2.10 now constitute group 4, and the small centre in the plot of amount of precipitation versus wind speed or days with precipitation has become group 3.

Finally, to answer the question of whether the human ecology and climatic indicators could be used to predict air pollution, we compared sulphur dioxide levels between the four groups of cities identified by the cluster analysis. This

Figure 5.6 *Cluster solution of US cities displayed in geographical space. The US states are labelled according to the group membership of their cities ('1' = group 1, '2' = group 2, '3' = group 3, '4' = group 4, 'NaN' = not known).*

Table 5.2 Clustering of US cities: results from k-means clustering.

Group 1: $n = 16$
Little Rock, Wilmington, Washington, Jacksonville, Miami, Atlanta, Louisville, New Orleans, Cincinnati, Memphis, Nashville, Dallas, Houston, Norfolk, Richmond, Charleston

Group 2: $n = 18$
Hartford, Indianapolis, Des Moines, Wichita, Baltimore, Detroit, Minneapolis, Kansas City, St Louis, Omaha, Albany, Buffalo, Cleveland, Columbus, Pittsburgh, Providence, Seattle, Milwaukee

Group 3: $n = 5$
Phoenix, San Francisco, Denver, Albuquerque, Salt Lake City

Group 4: $n = 2$
Chicago, Philadelphia

Group means (for SO_2 content, standard error of mean in brackets)

	1	2	3	4	5	6	7
Group 1	20.19 (2.39)	61.50	285.9	472.7	8.79	46.03	114.8
Group 2	36.22 (5.50)	50.58	448.5	546.5	10.29	35.17	125.6
Group 3	15.60 (3.33)	57.34	260.6	446.6	8.26	12.72	67.2
Group 4	89.50 (20.50)	52.60	2518.0	2659.5	10.00	37.19	118.5

Variables
1. SO_2 content of air in micrograms per cubic metre, not used in clustering
2. Average annual temperature in °F
3. Number of manufacturing enterprises employing 20 or more workers
4. Population size in thousands
5. Average wind speed in miles per hour
6. Average annual precipitation in inches
7. Average number of days with precipitation per year.

showed that pollution levels differed significantly between the groups (Kruskal–Wallis test: $p = 0.01$), with the highest sulphur dioxide concentrations found for the dense population group, followed by the two wet-climate groups, and the least pollution found for the dry climate group (the mean sulphur dioxide concentrations per group are displayed in Table 5.2).

5.6.3 Aesthetic judgement of painters

Davenport and Studdert-Kennedy (1972) present a data set where a seventeenth-century critic, Roger de Piles, expressed in quantitative terms a series of aesthetic judgements on 56 painters, using four standard but complex conceptual judgements. De Piles set out to divide 'the chief parts of the art into four columns to wit *Composition*, *Design*, *Colouring* and *Expression*', and in each dimension scored his 56 painters on a scale between 0 and 20, with the latter score reserved for 'sovereign perfection, which no man has fully arrived at'. The data are given in Table 5.3. (Two missing values present in Davenport and

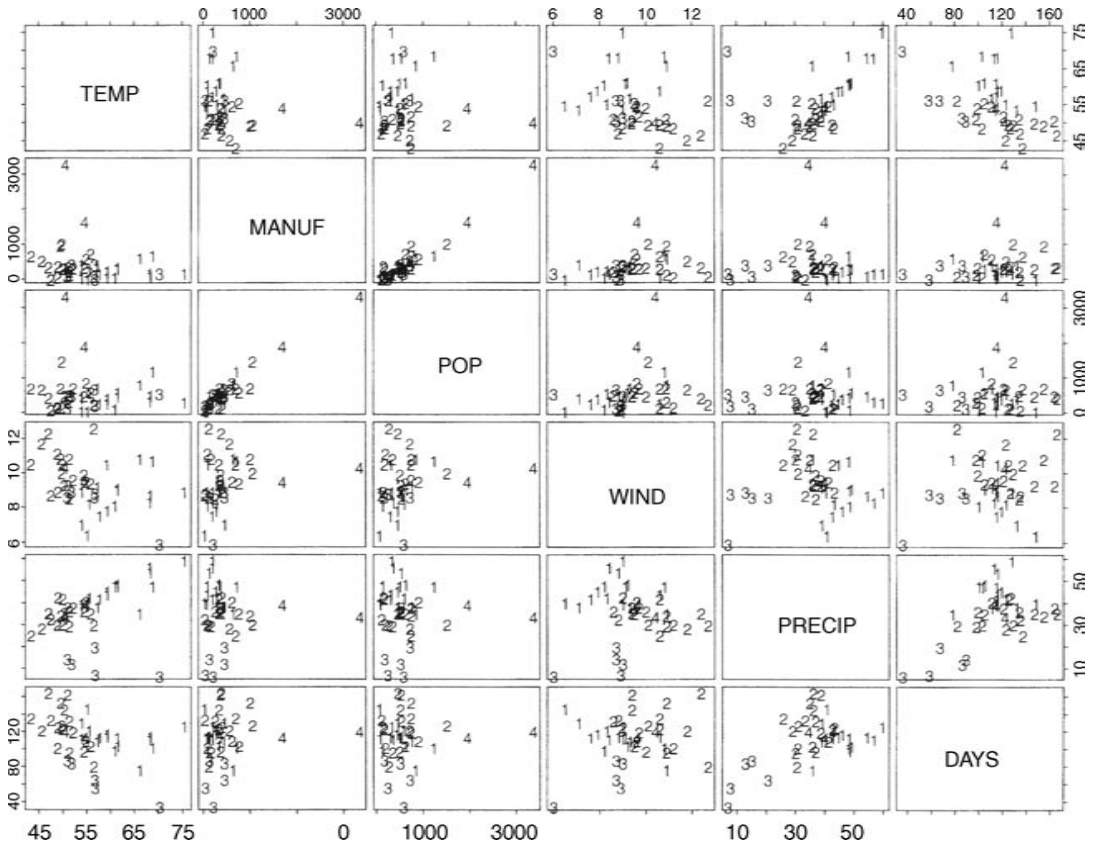


Figure 5.7 *Scatter plot matrix of pollution indicators for 41 US cities ('1' = group 1, '2' = group 2, '3' = group 3, '4' = group 4).*

Table 5.3 Artist data (reproduced with permission from Davenport and Studdert-Kennedy, 1972).

Painter		Composition	Drawing	Colouring	Expression	School
1	Albani	14	14	10	6	E
2	Durer	8	10	10	8	F
3	Del Sarto	12	16	9	8	A
4	Barocci	14	15	6	10	C
5	Bassano	6	8	17	0	D
6	Del Piombo	8	13	16	7	A
7	Bellini	4	6	14	0	D
8	Bourdon	10	8	8	4	H
9	Le Brun	16	16	8	16	H
10	Veronese	15	10	16	3	D
11	The Carracci	15	17	13	13	E
12	Corregio	13	13	15	12	E
13	Volterra	12	15	5	8	B
14	Dipenbeck	11	10	14	6	G
15	Domenichino	15	17	9	17	E
16	Giogione	8	9	18	4	D
17	Guercino	18	10	10	4	E
18	Guido Reni	13.53[a]	13	9	12	E
19	Holbein	9	10	16	13	F
20	Da Udine	10	8	16	3	A
21	J. Jordaens	10	8	16	6	G
22	L. Jordaens	13	12	9	6	C
23	Josepin	10	10	6	2	C
24	Romano	15	16	4	14	A
25	Lanfranco	14	13	10	5	E
26	Da Vinci	15	16	4	14	A
27	Van Leyden	8	6	6	4	F
28	Michelangelo	8	17	4	8	A
29	Caravaggio	6	6	16	0	E
30	Murillo	6	8	15	4	D
31	Venius	13	14	10	10	G
32	Vecchio	5	6	16	0	D
33	Giovane	12	9	14	6	D
34	Parmigiano	10	15	6	6	B
35	Penni	0	15	8	0	A
36	Perino del Vaga	15	16	7	6	A
37	Corton	16	14	12	6	C
38	Perugino	4	12	10	4	A
39	Polidore da Cara	10	17	7.58[a]	15	A
40	Pordenone	8	14	17	5	D
41	Pourbus	4	15	6	6	F
42	Poussin	15	17	6	15	H
43	Primaticcio	15	14	7	10	B
44	Raphael	17	18	12	18	A
45	Rembrandt	15	6	17	12	G

Table 5.3 (*Continued*)

Painter		Composition	Drawing	Colouring	Expression	School
46	Rubens	18	13	17	17	G
47	Salvata	13	15	8	8	B
48	Le Sueur	15	15	4	15	H
49	Teniers	15	12	13	6	G
50	Testa	11	15	0	6	C
51	Tintoretto	15	14	16	4	D
52	Titian	12	15	18	6	D
53	Van Dyck	15	10	17	13	G
54	Vanius	15	15	12	13	C
55	T. Zuccaro	13	14	10	9	B
56	F. Zuccaro	10	13	8	8	B

[a]This variable value is missing in the original data. These values were estimated using the EM algorithm described in Little and Rubin (1987).
A = Renaissance, B = Mannerist, C = Seicento, D = Venetian, E = Lombard, F = Sixteenth century, G = Seventeenth century, H = French.

Studdert-Kennedy's listing of these data have been replaced by estimated values from the application of the EM algorithm described in Little and Rubin, 1987.)

In an attempt to organize the data in a manner which might shed further light on the nature of de Piles and, by extension, early eighteenth-century artistic judgement, the painters were clustered using the minimization of det($\mathbf{W}$) method. Solutions from two to four groups were found, with, in each case, four random starting configurations being considered. The composition of the groups and the group mean vector for each solution are given in Table 5.4. Values of $g^2\det(\mathbf{W})/\det(\mathbf{T})$ were calculated to assess whether a particular number of groups was clearly indicated for these data. The results obtained were:

No. of groups (g)	$g^2\det(\mathbf{W})/\det(\mathbf{T})$
1	1.000
2	0.785
3	0.643
4	0.611

For these data this index for the selection of the number of groups does not appear to be particularly helpful, although the decreasing values of the index do seem to suggest that the data have *some* structure. Figure 5.8 displays the four-group solution in the space of the first two principal components.

It is difficult to speculate on these results without being an informed art historian. One thing which is apparent, however, in the solutions shown in Table 5.4 and in solutions for larger numbers of clusters not given here, is that correspondence between clusters and the school of an artist is relatively small.

Table 5.4 Clustering of painters: results from minimization of det(**W**).

Two groups

Group 1: $n = 35$
1, 3, 4, 9, 10, 11, 12, 13, 15, 17, 18, 22, 23, 24, 25, 26, 28, 31, 34, 36, 37, 39, 42, 43, 44, 46, 47, 48, 49, 50, 51, 52, 54, 55, 56

Group 2: $n = 21$
2, 5, 6, 7, 8, 14, 16, 19, 20, 21, 27, 29, 30, 32, 33, 35, 38, 40, 41, 45, 53

	Composition	Drawing	Colouring	Expression
Group 1	13.7	14.4	9.2	9.6
Group 2	7.9	9.4	13.7	5.0

Three groups

Group 1: $n = 13$
5, 6, 7, 16, 20, 29, 30, 32, 35, 38, 40, 41, 52

Group 2: $n = 27$
1, 2, 3, 4, 8, 10, 13, 14, 17, 21, 22, 23, 25, 27, 28, 31, 33, 34, 36, 37, 43, 47, 49, 50, 51, 55, 56

Group 3: $n = 16$
9, 11, 12, 15, 18, 19, 24, 26, 39, 42, 44, 45, 46, 48, 53, 54

	Composition	Drawing	Colouring	Expression
Group 1	6.2	10.4	14.4	3.0
Group 2	12.4	12.5	9.2	6.4
Group 3	14.5	14.3	10.7	14.3

Four Groups

Group 1: $n = 16$
2, 5, 7, 8, 14, 16, 19, 20, 21, 27, 29, 30, 32, 33, 45, 53

Group 2: $n = 15$
9, 11, 12, 15, 18, 24, 26, 31, 39, 42, 44, 46, 48, 54, 56

Group 3: $n = 18$
1, 3, 4, 10, 13, 17, 22, 23, 25, 34, 36, 37, 43, 47, 49, 50, 51, 55

Group 4: $n = 7$
6, 28, 35, 38, 40, 41, 42

	Composition	Drawing	Colouring	Expression
Group 1	8.9	8.0	14.4	5.2
Group 2	14.4	15.3	9.3	13.9
Group 3	13.6	13.5	8.9	6.3
Group 4	6.3	14.4	11.3	5.1

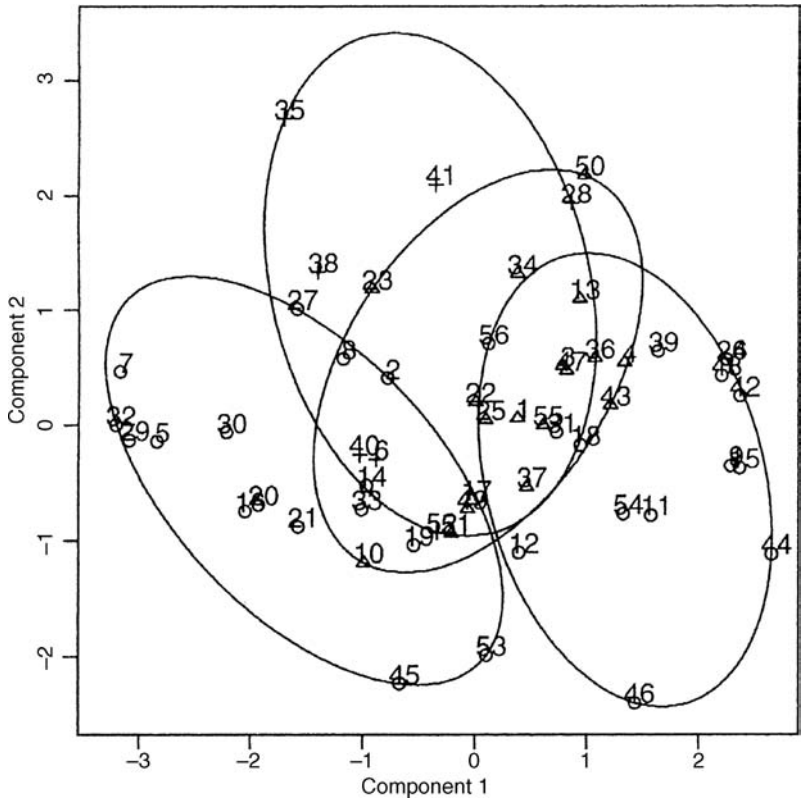


Figure 5.8 *Cluster solutions for painters given by minimization of det(**W**) on data in Table 5.3, displayed in the space of the first two principal components.*

A further small point is that several of those masters, whom de Piles and his contemporaries would have considered as embodying and continuing the 'grand tradition', are always grouped together – see, for example, Le Brun (9), Da Vinci (26), Poussin (42) and Le Sueur (48).

For these data the solutions given by minimizing trace(**W**) are very similar to those detailed in Table 5.4; see Davenport and Studdert-Kennedy (1972) for details.

5.6.4 Classification of 'nonspecific' back pain

The ambiguity of available diagnostics for low back pain was, according to Heinrich *et al.* (1985), 'detrimental to the patient's morale and impedes research for optimal treatment and prevention'. Consequently these authors applied several methods of multivariate analysis to a set of 132 signs and symptoms collected on 301 patients suffering from nonspecific low back pain, in the search for a useful classification. Amongst these techniques were minimization of

trace(**W**) and minimization of det(**W**) clustering. Although the results from the different methods were not entirely consistent, five strands of stable group description could be identified:

(i) A group of patients showing high scores on the general pain indices.
(ii) A group of patients characterized by high scores on the bilateral pain indices.
(iii) A group of patients showing most frequently their pain switching sides.
(iv) A group of patients labelled by the absence of signs and symptoms.
(v) A group of patients predominantly showing the presence of anterior disc changes, the absence of reflexes, the presence of sciatica and ipsilateral pain in connection with an acute condition.

5.7 Summary

Several optimization clustering techniques have been suggested, but two remain most popular – minimization of trace(**W**) and minimization of det(**W**). Presumably due to the availability of k-means algorithms in many software packages, the former is most commonly used (for example, the Office for National Statistics classification of local and health authorities of Great Britain resulted from a k-means algorithm; see Wallace and Denham, 1996). The latter has considerable advantages over the former, particularly in being invariant to scale changes in the observed variables; additionally it does not make such restrictive assumptions about the shape of the clusters. It appears that optimization clustering methods have not found as wide a degree of acceptance as the hierarchical procedures described in the last chapter. Certainly the number of applications found in the literature is far fewer. In some respects this is a pity, because of the link between many of the clustering criteria and particular types of formal probability models for cluster analysis to be described in the next Chapter.

6

Finite mixture densities as models for cluster analysis

6.1 Introduction

The majority of cluster criteria discussed in Chapters 4 and 5 were heuristic in the sense that assumptions about the class structure were not explicitly stated. But this does not necessarily mean that such assumptions are not made, so that the different methods for producing the clusters and for determining the number of clusters may often give conflicting solutions. Procedures for deciding upon a final cluster solution when using these methods were, in general, informal and subjective. In Chapter 5 some rather more formalized methods based on the optimization of numerical criteria were described, but these methods still relied on rather ad hoc approaches when, for example, deciding on the number of clusters.

In this chapter we introduce an alternative approach to clustering which postulates a formal statistical model for the population from which the data are sampled, a model that assumes that this population consists of a number of subpopulations (the 'clusters') in each of which the variables have a different multivariate probability density function, resulting in what is known as a *finite mixture density* for the population as a whole. By using finite mixture densities as models for cluster analysis, the clustering problem becomes that of estimating the parameters of the assumed mixture and then using the estimated parameters to calculate the posterior probabilities of cluster membership. And determining the number of clusters reduces to a model selection problem for which objective procedures exist.

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

Finite mixture densities often provide a sensible statistical *model* for the clustering process, and cluster analyses based on finite mixture models are also known as *model-based* clustering methods (Banfield and Raftery, 1993). Finite mixture models are being increasingly used in recent years to cluster data in a variety of disciplines, including behavioural, medical, genetic, computer and environmental sciences, robotics and engineering; see, for example, Everitt and Bullmore (1999); Benitez and Nenadic (2008); Bouguila and Amayri (2009); Branchaud *et al.* (2010); Dai *et al.* (2009); Dunson (2009); Ganesalingam *et al.* (2009); Marin *et al.* (2005); Meghani *et al.* (2009); Pledger and Phillpot (2008) and van Hattum and Hoijtink (2009).

Finite mixture modelling can be seen as a form of *latent variable analysis* (see, for example, Skrondal and Rabe-Hesketh, 2004) with 'subpopulation' being a latent categorical variable and the latent classes being described by the different components of the mixture density; consequently cluster analysis based on such models is also often referred to as *latent class cluster analysis*.

This chapter introduces the concept of model-based cluster analysis using finite mixture models, and gives details of estimation, model selection and the use of a Bayesian approach, etc. Finite mixture models can also be very useful in the clustering of structured data, and this will be the subject of Chapter 7.

6.2 Finite mixture densities

Finite mixture densities are described in detail in Everitt and Hand (1981), Titterington *et al.* (1985), McLachlan and Basford (1988), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006); they are a family of probability density functions of the form:

$$f(\mathbf{x}; \mathbf{p}, \boldsymbol{\theta}) = \sum_{j=1}^{c} p_j g_j(\mathbf{x}; \boldsymbol{\theta}_j), \tag{6.1}$$

where $\mathbf{x}$ is a p-dimensional random variable, $\mathbf{p}' = (p_1, p_2, \ldots, p_{c-1})$ and $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \boldsymbol{\theta}'_2, \ldots, \boldsymbol{\theta}'_c)$, with the p_j being known as mixing proportions and the g_j, $j = 1, 2, \ldots, c$, being the component densities, with density g_j being parameterized by $\boldsymbol{\theta}_j$. The mixing proportions are nonnegative and are such that $\sum_{j=1}^{c} p_j = 1$. The number of components forming the mixture, that is, the postulated number of clusters, is c.

Finite mixtures provide suitable models for cluster analysis if we assume that each group of observations in a data set suspected to contain clusters comes from a population with a different probability distribution. The latter may belong to the same family, but differ in the values they have for the parameters of the distribution. An example of this type that we shall discuss in detail later is a mixture in which the component densities are multivariate normal with different mean vectors and possibly different covariance matrices. In other cases, mixtures may comprise sums of *different* component densities – an example will be described in Section 6.9.4.

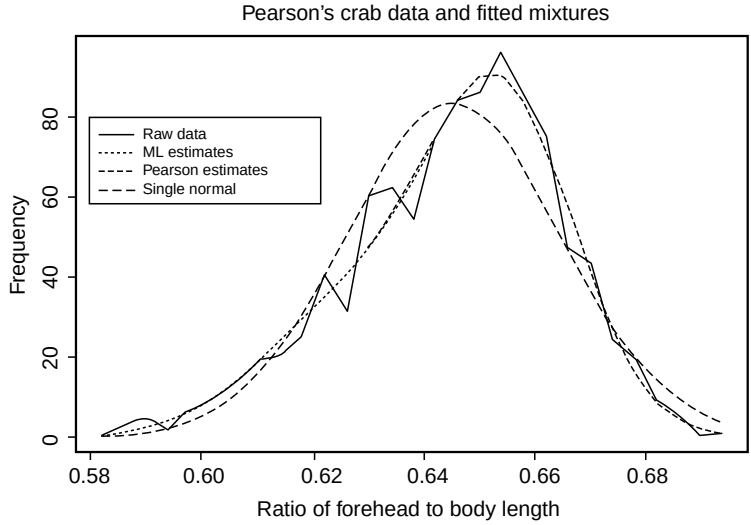


Figure 6.1 *Frequency polygon of ratio of forehead to body length in 1000 crabs and two-component normal mixtures fitted by method of moments (Pearson) and maximum likelihood (ML).*

Having estimated the parameters of the assumed mixture distribution, observations can be associated with particular clusters on the basis of the maximum value of the following estimated posterior probability:

$$\Pr(\text{cluster } j|\mathbf{x}_i) = \frac{\hat{p}_j g_j(\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{f(\mathbf{x}_i; \hat{\mathbf{p}}, \hat{\boldsymbol{\theta}})} \qquad j = 1, 2, \ldots, c. \tag{6.2}$$

One of the earliest applications of a finite mixture density was made by Karl Pearson (Pearson, 1894), who applied a two-component univariate Gaussian mixture to a set of measurements on the ratio of forehead to body length of 1000 Naples crabs from a mixture of two species. Figure 6.1 shows a frequency polygon of the original data together with the two-component mixture as estimated by Pearson using the method of moments (see Everitt and Hand, 1981, for details). In addition, Figure 6.1 shows the two-component mixture as estimated by maximum likelihood (see the next section) and a single normal density having the sample mean and variance of the 1000 observations. Here the method of moments and maximum likelihood solutions are very close, although it is known that, in general, maximum likelihood is far more efficient (see Tan and Chang, 1972).

6.2.1 Maximum likelihood estimation

Given a sample of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from the mixture density given in (6.1), the log-likelihood function, ℓ, is

$$\ell(\mathbf{p}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i; \mathbf{p}, \boldsymbol{\theta}). \tag{6.3}$$

Estimates of the parameters in the density would usually be obtained as a solution of the likelihood equations

$$\frac{\partial \ell(\boldsymbol{\phi})}{\partial(\boldsymbol{\phi})} = 0, \tag{6.4}$$

where $\boldsymbol{\phi}' = (\mathbf{p}', \boldsymbol{\theta}')$. In the case of finite mixture densities, the likelihood function is too complicated to employ the usual methods for its maximization, for example an iterative Newton–Raphson method which approximates the gradient vector of the log-likelihood function $\ell(\boldsymbol{\phi})$ by a linear Taylor series expansion (Everitt, 1987).

Because of this complexity, the required maximum likelihood estimates of the parameters in a finite mixture model have to be computed in some other way. Perhaps the most widely used approach is that of the iterative *expectation maximization* (EM) algorithm described in Dempster *et al.* (1977), which will be explained in detail in the next section for a model in which the component densities are multivariate normal. As an alternative to the EM algorithm, Bayesian estimation methods using the Gibbs sampler or other Monte Carlo Markov chain (MCMC) methods are becoming increasingly popular – see, for example, Marin *et al.,* (2005) and McLachlan and Peel (2000). A relatively brief account of Bayesian estimation is given in this chapter – see Section 6.4.

6.2.2 Maximum likelihood estimation of mixtures of multivariate normal densities

In the case of a mixture in which the jth component density is multivariate normal with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, the application of maximum likelihood results in the following series of equations (see Everitt and Hand, 1981, for details).

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} \Pr(j|\mathbf{x}_i) \tag{6.5}$$

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{n\hat{p}_j} \sum_{i=1}^{n} \mathbf{x}_i \Pr(j|\mathbf{x}_i) \tag{6.6}$$

$$\hat{\boldsymbol{\Sigma}}_j = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)' \Pr(j|\mathbf{x}_i), \tag{6.7}$$

where the $\Pr(j|\mathbf{x}_i)$ are the estimated posterior probabilities given in Equation (6.2).

Hasselblad (1966, 1969), Wolfe (1970) and Day (1969) all suggest an iterative scheme for solving the likelihood equations given above, which involves finding initial estimates of the posterior probabilities from given initial estimates of the parameters of the mixture, and then evaluating the right-hand sides of Equations (6.5) to (6.7) to give revised values for the parameters. From these, new estimates of the posterior probabilities are derived and the procedure is repeated until some suitable convergence criterion is satisfied. There are potential problems with this process unless the component covariance matrices are constrained in some way; for example, they are all assumed to be the same – again see Everitt and Hand (1981) for details.

This procedure is a particular example of the iterative expectation maximization (EM) algorithm described by Dempster *et al.* (1977) in the context of likelihood estimation for incomplete data problems. In estimating parameters in a mixture, it is the 'labels' of the component density from which an observation arises that are missing. Further comments about properties and problems of the EM algorithm, such as its convergence rate, local minima, etc., will be left until Section 6.2.3.

Scott and Symons (1971), Celeux and Govaert (1992) and Banfield and Raftery (1993) also consider the use of mixture models for cluster analysis, but in a somewhat different way to that described above. They begin with the same probability model as that described above, namely one that assumes that the population of interest consists of c subpopulations and that the density of a p-dimensional observation from the jth subpopulation is $g_j(\mathbf{x};\boldsymbol{\theta}_j)$ for some unknown vector of parameters, $\boldsymbol{\theta}_j$. But then they introduce the *classification maximum likelihood* procedure which involves choosing $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ to maximize the likelihood given by:

$$L(\boldsymbol{\theta},\boldsymbol{\gamma}) = \prod_{i=1}^{n} g_{\gamma_i}(\mathbf{x}_i;\boldsymbol{\theta}_{\gamma_i}), \tag{6.8}$$

where $\boldsymbol{\gamma}' = (\gamma_1,\ldots,\gamma_n)$ and $\gamma_i = j$ if $\mathbf{x}_i$ is from the jth subpopulation; the γ_i label the subpopulation of each observation, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. If $g_j(\mathbf{x};\boldsymbol{\theta}_j)$ is a multivariate normal density with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$, this likelihood has the form

$$L(\boldsymbol{\theta},\boldsymbol{\gamma}) = \text{const}\prod_{j=1}^{c}\prod_{i\in E_j}|\boldsymbol{\Sigma}_j|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}+(\mathbf{x}_i-\boldsymbol{\mu}_j)\right\}, \tag{6.9}$$

where $E_j = \{\mathrm{i} : \gamma_i = j\}$. The maximum likelihood estimator of $\boldsymbol{\mu}_j$ is $\bar{\mathbf{x}}_j = n_j^{-1}\boldsymbol{\Sigma}_{i\in E_j}\mathbf{x}_i$, where n_j is the number of elements in E_j. Replacing $\boldsymbol{\mu}_j$ in (6.9) with the maximum likelihood estimator $\bar{\mathbf{x}}_j$ yields the following log-likelihood:

$$\ell(\boldsymbol{\theta},\boldsymbol{\gamma}) = \text{const}-\frac{1}{2}\sum_{j=1}^{c}\left\{\ tr(\mathbf{W}_j\boldsymbol{\Sigma}_j^{-1}) + n_j\log|\boldsymbol{\Sigma}_j|\right\}, \tag{6.10}$$

where $\mathbf{W}_j$ is the sample cross-product matrix for the jth subpopulation introduced in the previous chapter (see Section 5.3.5).

Banfield and Raftery (1993) demonstrate the following:

- If $\boldsymbol{\Sigma}_j = \sigma^2\mathbf{I}(j = 1, \ldots, c)$ then the log-likelihood in (6.10) is maximized by choosing $\boldsymbol{\gamma}$ to minimize tr($\mathbf{W}$), where $\mathbf{W} = \boldsymbol{\Sigma}_{j=1}^{c}\mathbf{W}_j$. This is one of the criteria introduced in Chapter 5 and is also the sum-of-squares criterion which underlies Ward's agglomerative hierarchical clustering method described in Chapter 4.
- If $\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}$ $(j = 1, \ldots, c)$ then the log-likelihood in (6.10) is maximized by choosing $\boldsymbol{\gamma}$ to minimize det($\mathbf{W}$), a further criterion discussed in Chapter 5 (see Section 5.3.2).
- If $\boldsymbol{\Sigma}_j$ is not constrained, the log-likelihood in (6.10) is maximized by choosing $\boldsymbol{\gamma}$ to minimize $\prod_{j=1}^{c}[\det(\mathbf{W}_j/n_j)]^{n_j}$, a criterion similar to that suggested by Scott and Symons (1971) (see Section 5.3.5).

The classification maximum likelihood approach nicely links the approach of this chapter and the more intuitive approach used in Chapter 5. The difference between the classification likelihood procedure and the finite mixture approach described earlier is that in the latter it is assumed that the $\mathbf{x}_i$ arises from mixture distributions, the parameters of which are estimated and then cluster membership determined by the maximum values of the estimated posterior probabilities. In contrast, in the classification likelihood approach, it is assumed that $\mathbf{x}_i$ arises from the single distribution $g_{\gamma_i}(\mathbf{x}_i, \boldsymbol{\theta}_{\gamma_i})$ which is determined by the unknown parameter label γ_i. Cluster membership is then directly estimated by finding labels that maximize the classification likelihood. Banfield and Raftery (1993) implement the classification likelihood approach as a hierarchical method in which, at each step, the two clusters that give the greatest increase in the classification likelihood are merged.

Banfield and Raftery (1993) then develop their ideas further to lead to new criteria for clustering which are more general than minimizing det($\mathbf{W}$) but more parsimonious than the completely unconstrained model. Specifically, they suggest criteria which allow some, but not all, of the features of cluster distributions (orientation, size and shape) to vary between clusters, while constraining others to be the same. These new criteria arise from considering the reparameterization of the covariance matrix $\boldsymbol{\Sigma}_j$ in terms of its eigenvalue description

$$\boldsymbol{\Sigma}_j = \mathbf{D}_j\boldsymbol{\Lambda}_j\mathbf{D}'_j, \tag{6.11}$$

where $\mathbf{D}_j$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}_j$ is a diagonal matrix with the eigenvalues of $\boldsymbol{\Sigma}_j$ on the diagonal (this is simply the usual principal components transformation – see Chapter 2). The orientation of the principal components of $\boldsymbol{\Sigma}_j$ is determined by $\mathbf{D}_j$, whilst $\boldsymbol{\Lambda}_j$ specifies the size and shape of the density contours. Specifically, we can write $\boldsymbol{\Lambda}_j = \lambda_j\mathbf{A}_j$, where λ_j is the largest eigenvalue of $\boldsymbol{\Sigma}_j$ and $\mathbf{A}_j = \text{diag}(1, \alpha_2, \ldots, \alpha_p)$ contains the eigenvalue ratios after division by λ_j. Hence λ_j controls the size of the jth cluster and $\mathbf{A}_j$ its shape. (Note that the term 'size' here refers to the volume occupied in space, not the number of objects in the cluster.) In

two dimensions the parameters would reflect, for each cluster, the correlation between the two variables, and the magnitudes of their standard deviations. More details are given in Banfield and Raftery (1993) and Celeux and Govaert (1995), but Table 6.1 gives the constraints imposed on clusters by the various criteria proposed. The resulting models make up what Fraley and Raftery (1998, 2003, 2006) term the MCLUST family of mixture models.

For small sample sizes and where the data consist of well-separated and approximately equal-sized clusters, the classification likelihood approach may recover the underlying clustering structure somewhat better than the mixture likelihood approach (Ganesalingam, 1989; Govaert and Nadif, 1996; Celeux and Govaert 1993). But it is known to yield inconsistent component parameter estimates $\boldsymbol{\theta}$ and mixing proportions $\mathbf{p}$ (Bryant and Williamson 1978, 1986), which is not the case when maximizing the mixture likelihood function. For this

Table 6.1 Constraints imposed on clusters by different criteria.

Model	Distribution	Volume λ_j	Shape $\mathbf{A}_j$	Orientation $\mathbf{D}_j$	Criterion	Origin
EII	Spherical	Same	Same	None	trace($\mathbf{W}$)	Ward (1963)
VII	Spherical	Different	Same	None	$\prod_{j=1}^{c}\left[\text{trace}\left(\mathbf{W}_j/n_j\right)\right]^{n_j}$	Banfield and Raftery (1993)
EEI	Diagonal	Same	Same	Coordinate axis		
VEI	Diagonal	Different	Same	Coordinate axis		
EVI	Diagonal	Same	Different	Coordinate axis		
VVI	Diagonal	Different	Different	Coordinate axis		
EEE	Ellipsoidal	Same	Same	Same	det($\mathbf{W}$)	Friedman and Rubin (1967)
VEE	Ellipsoidal	Different	Same	Same		
EEV	Ellipsoidal	Same	Same	Different	S	Murtagh and Raftery (1984)
VEV	Ellipsoidal	Different	Same	Different	S^*	Banfield and Raftery (1993)
VVE	Ellipsoidal	Different	Different	Same	$\left\|\boldsymbol{\Sigma}_j\right\|\left\|\mathbf{W}_j\right\|^{1/d}$	(Celeux and Govaert, 1995)
VVV	Ellipsoidal	Different	Different	Different	$\prod_{j=1}^{c}\left[\det\left(\mathbf{W}_j/n_j\right)\right]$	Scott and Symons (1971)

Parameterizations of the multivariate normal mixture model using eigenvalue decomposition based on Banfield and Raftery (1993) and Fraley and Raftery (1998, 2002). The orientation 'Coordinate axis' refers to models with components parallel to the coordinate axis.

Model names describe model restrictions of volume λ_j, shape $\mathbf{A}_j$ and orientation $\mathbf{D}_j$, respectively: V = variable, parameter unconstrained; E = equal, parameter constraint; I = matrix constrained to identity matrix.

Criteria: eigenvalue decomposition corresponds to minimizing of the criteria if classification maximum likelihood estimation is used.

$S = \sum_{j=1}^{c} S_k$ with $S_k = tr(\mathbf{A}^{-1}\,\boldsymbol{\Omega}_j)$ and $\boldsymbol{\Omega}_j = diag\left(\omega_{1j}, \ldots, \omega_{pj}\right)$, where ω_{pj} is the pth eigenvalue of $\mathbf{W}_j$;

$S^* = \sum_{j=1}^{c} n_c \log(S_k/n_k)$ (for further details see Banfield and Raftery, 1993).

reason, the mixture likelihood approach based on the EM algorithm for parameter estimation is generally preferred over classification likelihood, and it is the former that is implemented in the `mclust` function in `R`, which fits the models in the MCLUST family described in Table 6.1.

One final feature of the Banfield and Raftery approach to clustering is that they allow for observations that do not follow the assumed distribution, for example outliers and other 'noise' points. Such points are catered for by assuming that they arise from a Poisson process, and the likelihood function is modified accordingly (see Banfield and Raftery, 1993, for details).

6.2.3 Problems with maximum likelihood estimation of finite mixture models using the EM algorithm

A comprehensive account of the EM algorithm is available in McLachlan and Krishnan (1997, 2008) and Watanabe and Yamaguchi (2004); here we give a fairly brief account of some of its properties and some of its potential problems when used to estimate the parameters in a finite mixture model.

Convergence problems with EM estimations

Multiple maxima are a common feature of finite mixture likelihoods, and convergence to a local maximum can produce suboptimal maximum likelihood estimates – not 'true' maximum likelihood estimates. As a consequence, the EM algorithm ideally needs to be run repeatedly using different sets of starting values. Observing the same log-likelihood values from multiple starting points increases the confidence that the solution is a global maximum. Initial starting values for application of the EM algorithm can be based on solutions from standard heuristic hierarchical clustering methods, k-means clustering or indeed almost any other reasonable method of cluster analysis.

Celeux and Govaert (1992) suggested the use of *stochastic* EM algorithms to obtain initial starting values. The stochastic EM algorithm restores the unknown component labels $\mathbf{z}_i$, $i = 1, \ldots, n$ at each iteration between the E and M steps, by drawing them at random from their current conditional distribution. The stochastic EM algorithm prevents convergence to the first local maximum of the likelihood and thereby avoids approaching a local maximum of the log-likelihood function. In a second step, parameter estimates obtained from the stochastic EM are used as initial values for the EM (Biernacki *et al.*, 2003, 2006). Berchtold (2004) demonstrated that such a two-stage procedure, which combines an explorative phase and an optimization phase, yields the best results.

Singularities in likelihood function and degenerate distributions

A further problem with using maximum likelihood to estimate the parameters in a finite mixture model is that of *singularities* in likelihood function; that is, points where the likelihood becomes infinite, giving rise to degenerate distributions.

Singularities in the likelihood arise if the number of parameters to be estimated is large in relation to sample size, for example in models with unrestricted covariances and large numbers of components. For such mixtures the likelihood is not bounded and the EM algorithm may fail to converge, instead diverging to a point of infinite likelihood (Titterington *et al*., 1985; Fraley and Raftery, 2007).

Constraining the covariances will often overcome the problems of singularities, as will other approaches described in Chapter 7. Fraley and Raftery (2007) propose the use of a Bayesian approach in which the maximum likelihood estimate is replaced by the mode of the posterior (maximum *a posteriori* estimate) also found by the EM algorithm. More will be said about Bayesian mixture models in Section 6.4.

6.3 Other finite mixture densities

Although mixture densities with normal components are those used most often in practice, mixture densities with other types of components are also of importance in particular applications, and a number of possibilities will now be considered.

6.3.1 Mixtures of multivariate *t*-distributions

Finite mixtures with multivariate normal components have been widely used to model continuous multivariate data suspected of containing subgroups. One advantage of such mixtures is their computational convenience. They can be easily fitted iteratively by maximum likelihood either directly or via the EM algorithm of Dempster *et al.* (1977). And although finite mixture models with normal component distributions allow an arbitrarily close modelling of *any* distribution simply by increasing the number of components, such a mixture can be misleading if the true component distributions are skewed, because several normal distributions are perhaps being used to capture the skewness of a single cluster and may not reflect the subgroups of individuals in the sample. As a result, the number of clusters in the data will be overestimated; in such cases it is clear that the use of normal component mixtures will not be appropriate (Bauer and Curran, 2003; Jasra *et al*., 2006; Fraley and Raftery, 2007).

For such reasons, Peel and McLachlan (1999) consider fitting mixtures of *multivariate t-distributions*, where the component densities themselves capture skewness and excess kurtosis. The *t*-distribution provides a longer-tailed alternative to the normal distribution and thus provides a more robust approach to the fitting of mixture models, as observations that are atypical of a component are given reduced weight in the calculation of its parameters. Also, the use of *t* components gives less extreme estimates of the posterior probabilities of component membership of the mixture model, as demonstrated by Peel and McLachlan (2000). Henning (2004) demonstrated that the use of mixtures of *t* components instead of normals in the presence of gross outliers adds stability to the clustering process, although the number of outliers needed for breakdown (that

is the situation where the estimator $\boldsymbol{\theta}_j$ for a fixed number of components takes on values that are arbitrarily large and meaningless) is almost the same as with normal mixture models.

Details of the fitting and estimation process for mixtures with multivariate t components are given in Peel and McLachlan (2000), who also report an example involving part of the crab data set of Campbell and Mahon (1974) met earlier in Chapter 2. Peel and McLachlan analyse the data from 100 blue crabs, 50 of which are male and 50 female. A mixture of two t components results in one cluster containing 39 male crabs, and the other containing the 50 female crabs and the remaining 11 males. Using the normal mixture model results in one additional male crab being assigned to the second cluster. Wang *et al.* (2004) extended Peel and McLachlan's modelling approach and presented a framework for fitting mixtures of multivariate t-distributions when data are missing at random.

McLachlan *et al*. (2006) describe the fitting of multivariate skewed t-distribution mixtures to highly asymmetric multivariate data using a t-factor analyser model (see Chapter 7 and Wang *et al.,* 2009). Frühwirth-Schnatter and Pyne (2010) introduce a Bayesian approach to model skew-normal and skew-t-mixture distributions (generalizations of distributions that allow for nonzero skewness), using data augmentation and Markov chain Monte Carlo estimation. These approaches are particularly useful for the analysis of high-dimensional data with skewed distributions, such as cell protein or gene-expression data.

6.3.2 Mixtures for categorical data – latent class analysis

The mixture models considered in Section 6.2 based on Gaussian components will, of course, not be suitable for data sets when the variables recorded are categorical, since they assume that within each group the variables have a multivariate normal distribution. Clearly, to provide suitable models for categorical data, the mixture will need to be based on more-suitable component densities. Widely used are *multivariate Bernoulli densities* which arise from assuming that, within each group, the categorical variables are independent of one another, the so-called *conditional independence assumption*. It is this approach which is the basis of *latent class analysis* (see Lazarsfeld and Henry, 1968; Goodman, 1974; Clogg, 1996; Hagenaars and McCutcheon, 2002). Here the appropriate model for binary variables is described.

Suppose that there are g groups in the data and that in group i the vector $\boldsymbol{\theta}_i$ gives the probabilities of a '1' response on each variable; that is,

$$\Pr(x_{ij} = 1|\text{group } i) = \theta_{ij}, \tag{6.12}$$

where x_{ij} is the value taken by the jth variable in group i. From the conditional independence assumption it follows that the probability of an observed vector of scores $\mathbf{x}$ in group i is given by

$$\Pr(\mathbf{x}|\text{group } i) = \prod_{j=1}^{p} \theta_{ij}^{x_{ij}} (1-\theta_{ij})^{1-x_{ij}}. \tag{6.13}$$

If the proportions of each group in the population are $p_1, p_2, \ldots, p_g$, then the unconditional probability of the observation $\mathbf{x}$ is given by the mixture density

$$\Pr(\mathbf{x}) = \sum_{i=1}^{g} p_i \prod_{j=1}^{p} \theta_{ij}^{x_{ij}} (1-\theta_{ij})^{1-x_{ij}}. \tag{6.14}$$

Estimation of the parameters is again via maximum likelihood, and formation of groups by considering the estimated posterior probabilities – see McLachlan and Basford (1988) for details.

Often the observed association between variables cannot entirely be explained by their association with the components. A common procedure is to increase the number of clusters until the assumption of local independence holds. This can lead to spurious components and models that are a poor fit. It is better to relax the assumption of local independence by allowing for the residual association between observed variables that are responsible for local dependence (e.g. related questions in a psychometric questionnaire), by using the appropriate multivariate distribution for a set of locally dependent variables. In this case the index j of Equations (6.13) and (6.14) would not be a single, but a set of indicators. Another possibility is the use of factor mixture models, which are useful if groups of variables account for local dependency (e.g. sub-constructs in a psychometric measurement scale). See Hagenaars (1988) for a detailed description of how to define latent class models with local dependencies, and Magidson and Vermunt (2001) for latent class factor mixture cluster models. Finite mixtures of factor models will be discussed in Section 7.3 in the next chapter. Reboussin *et al.* (2008) propose locally dependent latent class models with covariates, which they used to identify subtypes of underage drinkers on the basis of their observed patterns of drinking.

6.3.3 Mixture models for mixed-mode data

Multivariate normal mixtures are applicable to data containing continuous variables, latent class models to data with categorical variables. In practice, of course, many data sets will contain variables of both types, and Everitt (1988) suggests how the mixture approach can be applied to such *mixed-mode* data.

Suppose the data consist of p continuous variables $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$ and q categorical variables, $\mathbf{z}' = (z_1, z_2, \ldots, z_q)$, with z_i having c_i categories. It is assumed that each categorical variable z_i arises from an unobservable continuous random variable, y_i, by applying a series of thresholds; that is,

$$z_j = k \quad \text{if} \quad \alpha_{j,k-1} \leq y_j < \alpha_{j,k} \tag{6.15}$$

for $k = 1, 2, \ldots, c_j$ and $\alpha_{j,0} = -\infty$, $\alpha_{j,c_j} = \infty$, $j = 1, 2, \ldots, q$. The threshold values $\alpha_{j,k}$, for $j = 1, 2, \ldots, q$, $k = 1, 2, \ldots, c_j - 1$ are considered as unknown parameters to be estimated from the data.

The density function of the observed continuous variables, **x,** and the continuous latent variables, $\mathbf{y}' = (y_1, \ldots, y_q)$, is assumed to be the multivariate normal mixture, with assumed common covariance matrix for the component densities. The density function of the observed variables **x**, **z** is then given by

$$h(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{g} p_i \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_q}^{b_q} \text{MVN}(\mathbf{x}, \mathbf{y}; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}) dy_1 dy_2 \cdots dy_q, \qquad (6.16)$$

where the limits of integration correspond to the threshold values appropriate to the particular values in the vector **z**.

Maximum likelihood can be used to estimate the parameters of the model, although the presence of the multidimensional integral in (6.16) makes the procedures computationally impracticable if the number of categorical variables is greater than four. Examples of how this method performs in comparison with other more routine approaches to the clustering of mixed-mode data are given in Everitt and Merette (1989). Lawrence and Krzanowski (1996), Willse and Boik (1999) and Muthén and Shedden (1999) introduced models which are computationally more feasible for a larger number of variables. Hunt and Jorgensen (2003) extended Muthén and Shedden's model to mixed-mode data where some of the data are missing at random. Further developments are described in Ganesalingam (1989), Govaert and Nadif (1996), Jorgensen and Hunt (1999), McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

6.4 Bayesian analysis of mixtures

With the development of the *Markov chain Monte Carlo* (MCMC) sampling method for estimating the parameters of Bayesian models (see, for example, Geman and Geman, 1984; Tanner and Wong, 1988; Gelfand and Smith, 1990; Green, 1995; Gilks *et al*., 1998; for a review see Chen *et al*., 2000 or Gammermann and Lopez, 2006), and the growth of computing power in the 1990s, *Bayesian statistics* (see, for example, Gelman *et al.,* 2003) has been revolutionized and has attracted increasing interest amongst statisticians, not least in the area of using finite mixture models for cluster analysis. Richardson and Green (1997), for example, consider a Bayesian approach to the analysis of univariate normal mixtures in which the mixing proportions, component means, component variances and the number of components are regarded as drawn from appropriate prior distributions. Because Bayesian inference usually cannot be performed analytically, MCMC algorithms are used for estimation (see Gilks *et al*., 1996).

There are two main reasons why a Bayesian approach to fitting finite mixture models is worth considering. First, Bayesian modelling allows parameter estimation for models where the likelihood method fails because of singularities in the likelihood surface; here Bayesian modelling is employed primarily for pragmatic reasons.

The second reason for the increasing interest in Bayesian mixture modelling is philosophical, because the Bayesian approach allows probabilistic statements to be made directly about the unknown parameters, and findings from previous research or from expert opinion can be incorporated with the prior distribution. Richardson and Green (1997) argue that the Bayesian paradigm is the only sensible approach to model-based clustering if the number of components is unknown.

An introduction to Bayesian modelling of finite mixtures is given by Marin *et al.* (2005). A detailed account of Bayesian methods for finite mixtures is given in the book by Frühwirth-Schnatter (2006). A Bayesian framework for estimating finite mixtures of the latent variable analysis or structural equation modelling approach is described by Zhu and Lee (2001).

Bayesian finite mixture modelling is, however, not without its problems, of which the two most important are (i) the choice of a prior distribution and (ii) the component label-switching problem during MCMC sampling. Both problems are described in the following two sections.

6.4.1 Choosing a prior distribution

A key feature of Bayesian inference about the parameter $\boldsymbol{\theta}$ is the requirement for the prior distribution $\pi(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, which describes the information about the parameter before the data are seen. After observing the data the prior distribution will be updated to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{x})$using the likelihood of the data given $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto L(\mathbf{x}\,|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{6.17}$$

The posterior distribution provides the basis for statistical inferences.

The choice of the prior, which should reflect the available knowledge before the data are seen, is important since it influences the posterior inference. In cluster analysis the number of clusters and the parameters of the cluster model are usually unknown. In this case the prior should have little influence on inference, and the data should mainly determine the posterior distribution through the likelihood. For finite mixtures, conjugate priors – where priors and likelihood are of such a form that the posterior distribution follows the same distribution – are commonly used because they are computationally traceable (Roeder and Wasserman, 1997; Rufo *et al.,* 2006).

For normal finite mixture models with a fixed number of components, Fraley and Raftery (2007) used a normal conjugate prior for the means of the components. Conjugate priors for restricted covariance matrices through eigenvalue decomposition are inverse Wishart priors for the ellipsoidal models of Table 6.1 and inverse gamma priors for diagonal and spherical models of the MCLUST family of models. Fraley and Raftery used a uniform prior distribution on the simplex for the mixing proportions, giving the same weight for each component. This prior is a special case of the Dirichlet distribution, a commonly used conjugate prior for a multinomial distribution. The posterior expectation of

Fraley and Raftery's model can be written in closed form, and mixture estimation can be performed using the EM algorithm and does not require MCMC simulations. Fraley and Raftery used highly dispersed priors, and their method avoided the degeneracies and singularities of standard EM estimation but provided similar results to standard estimation methods if no singularity and degeneracy problems are present.

Conjugate prior distributions for the parameters $\boldsymbol{\theta}$ of other members of the exponential family are described in Morris (1983), and Rufo *et al.* (2006) discuss the use of conjugate priors for finite mixtures, and provide an algorithm to calculate the parameters in the prior distribution from the exponential families that obtain the least information on the mixture parameters. Non-conjugate priors (such as partially proper priors and improper priors) were used by Mengersen and Robert (1996), Roeder and Wasserman (1997) and Robert and Titterington (1998), and their use is discussed in Marin *et al.* (2005).

6.4.2 Label switching

A second possible problem with a Bayesian approach to cluster analysis via finite mixture distributions is that of label switching during MCMC sampling, which arises because of the symmetry in the likelihood of the model parameters. For Bayesian mixtures the likelihood function and the resulting posterior distribution is invariant under permutations with respect to component labels. The labels of the components during one run of an MCMC sampler may be switched on different iterations, and the lack of identifiability gives rise to problems when making inferences about the individual components. To obtain a meaningful interpretation of the components it is necessary to account for label switching so that components are in the same order at each iteration. One way to deal with the label-switching problem and to 'relabel' the components is to impose identifiable constraints on a particular set of model parameters, such as $\mu_1 < \mu_2 < \cdots < \mu_c$ in mixtures with component distributions that are univariate normals, or alternatively $p_1 < p_2 < \cdots < p_c$ on the mixing distribution (Richardson and Green, 1997; Roeder and Wasserman, 1997). However, Celeux *et al.* (2000) demonstrated that this 'identifiability constraint' does not always work and can seriously distort the results. Stephens (2000) suggests an alternative approach to overcome the label-switching problem by inspecting the posterior probabilities p_{ij}. Two sampling points are labelled as the same if divergence from one scaled distribution to another is small (Stephens, 2000; Rufo *et al.*, 2006). A second alternative imposes a reordering constraint after the simulations have been done, by selecting one of the $c!$ modal regions of the posterior distribution and performing relabeling in terms of proximity to this region (Marin *et al.*, 2005).

Jasra *et al.* (2005) reviewed the solutions to the label-switching problem and introduced a further probabilistic relabeling approach. In contrast to other approaches which assume a correct relabeling, it incorporates the uncertainty of the relabeling at each iteration.

6.4.3 Markov chain Monte Carlo samplers

Having chosen an appropriate prior distribution and decided how to deal with the label-switching problem, Bayesian model estimation is accomplished by simulating from a mixture posterior distribution using data augmentation – that is, by simulating the unobserved $\mathbf{z}$, where $\mathbf{z}$ encodes the missing component information – and MCMC samplers. Commonly used MCMC samplers for finite mixture models are the Gibbs, and modifications of the Metropolis–Hastings sampler. For details about data augmentation and Gibbs sampling, see Dempster *et al.* (1977) and Diebolt and Robert (1994). If the number of components is fixed, the tempering MCMC (Neal, 1996; Celeux *et al.*, 2000), or the more advanced evolutionary MC (Liang and Wong, 2001) algorithms are commonly used. Further MCMC samplers which allow sampling from the joint posterior distribution of all mixture parameters, including c, the number of mixture components, are presented in the Section 6.5.4. Jasra *et al.* (2005) present a review of various MCMC sampling schemes that have been suggested for mixture models.

6.5 Inference for mixture models with unknown number of components and model structure

Determining the number of clusters in a data set is a problem that we have met when discussing a wide range of heuristic clustering methods in previous chapters. When using finite mixture models as the basis of clustering, the 'number of clusters' problem remains, but now there are a number of possible model selection procedures that might provide a more convincing solution to the problem than the generally ad hoc approaches used in earlier chapters. And a further possible advantage for finite mixture modelling in this context is the possibility that these model selection methods might possibly be applied to *simultaneously* determine both the number of clusters *and* the most appropriate form of the component distributions in the mixture (in terms of the decompositions given in Table 6.1 and others to be introduced in the next chapter).

We will consider four possible model selection procedures for making inferences about the number of clusters and/or cluster models; these are as follows:

1. Log-likelihood ratio test statistics
2. Information theoretic approaches
3. Bayes factors
4. Markov chain Monte Carlo methods using reversible jump MCMC or birth and death process methodology.

6.5.1 Log-likelihood ratio test statistics

A natural candidate for testing the hypothesis $c = c_0$ against $c = c_1 (c_1 > c_0)$, where c is the number of components in a mixture density, is the likelihood ratio statistic, λ, given by

$$\lambda = \frac{\text{likelihood with } c = c_0}{\text{likelihood with } c = c_1}. \tag{6.18}$$

Unfortunately this does not lead to a suitable significance test, since for mixture densities regularity conditions do not hold for $-2\ln\lambda$ to have its usual asymptotic null distribution, that is, a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters in the two cases. The problem is that the null distribution is on the edge of the parameter space, in the sense that when two components coincide, their mixing proportions become unidentifiable. As a consequence, the log-likelihood ratio test (LRT) tends to overestimate the number of clusters.

The problem has been considered by a number of authors for a variety of circumstances (which are summarized by McLachlan and Peel, 2000; Schlattmann, 2009). Wolfe (1971), for example, on the basis of a limited simulation study, suggests that the null distribution of the modified likelihood ratio statistic for a finite mixture of a p-variate normal, $\{-2[n-1-p-0.5(c+1)]/n\}$, for testing $c+1$ against c components, is chi-squared with $2v-2$ degrees of freedom, where v is the number of extra parameters in the $c+1$ component mixture. Later authors, such as Hernández-Avilia (1979), Everitt (1981) and Thode *et al.* (1989) have shown that Wolfe's suggestion behaves reasonably in particular circumstances but may lack power. Lo *et al.* (2001) propose using an approximation of the distribution of the difference of the two log-likelihoods instead of using the traditional chi-squared distribution to compare models with c and $c+1$ components. However, Jeffries (2003) identified a flaw in the mathematical proof of the Lo–Mendell–Rubin LRT, and showed that the problem is also present in its empirical evaluation. Simulation studies by Lo *et al.* (2001) and Tofighi and Enders (2007) suggest that the test is still useful as an empirical tool for determining the number of clusters.

McLachlan (1987) proposed the use of the parametric bootstrap approach for assessing the number of components in normal finite mixtures using the log-likelihood ratio test statistic (LRTS). Bootstrap samples are used to empirically estimate the distribution of the LRTS. To test the null hypothesis of a mixture with c components against the alternative of a $(c+1)$-component mixture using the bootstrap LRTS, n samples are repeatedly simulated from a c-component mixture where the parameters are replaced by their likelihood estimates obtained from fitting the c-component model to the original sample. The LRTS is calculated for this bootstrap sample after fitting mixture models with c and $c+1$ components to the simulated data. This sampling and model fitting procedure is repeated independently B times, and the replicated LRTSs provide an approximation of the true distribution under the null hypothesis. This distribution allows an estimate of the actual level of significance, P, for the corresponding LRTS obtained from the original sample. Specifically, the probability P under the null hypothesis of obtaining an LRTS as large as or larger than the observed one is

$$P = (m+1)/(B+1), \tag{6.19}$$

where m is the total number of bootstrap samples with an LRTS larger than the original one.

As in the Bayesian MCMC sampling, label switching may occur across replications, resulting in distortion of the mixing distribution. Again, one way to deal with the label-switching problem and to 'relabel' the components is to impose identifiable constraints on a particular set of model parameters, such as $\mu_1 < \mu_2 < \cdots < \mu_c$ in mixtures with component distributions that are univariate normals, or alternatively $p_1 < p_2 < \cdots < p_c$ on the mixing distribution. McLachlan and Peel (2000) avoided label switching by using the parameter estimates of the original sample as starting values for the EM model estimation of each bootstrap sample, while Dias and Vermunt (2008) suggested using the method proposed by Stephens (2000, see Section 6.4.2) to determine the right order of the components during the bootstrapping process. The use of the bootstrap likelihood ratio test is still not completely satisfactory because the parameter estimate $\hat{\boldsymbol{\Psi}}$ in place of $\boldsymbol{\Psi}$ under the null hypothesis affects the accuracy of the bootstrap (McLachlan and Peel, 2000; Seidel *et al.,* 2000).

An empirical simulation study by Nylund *et al.* (2007) compared the performance of the LRT, the Lo–Mendell–Rubin LRT and the bootstrap LRT across different settings, and concluded that the bootstrap LRT outperformed the other two tests in correctly identifying the correct number of components. The bootstrap LRT consistently showed type I error rates close to the expected values of 0.05 across all settings. Except for simulations with categorical data, small sample size and unequal cluster sizes, the power to select the right number of clusters was close to 100%. The Lo–Mendell–Rubin test performed well in settings where continuous outcomes were used, while the LRT in general performed as expected – poorly. The bootstrap LRT seems to be the preferred method to select the number of clusters based on a statistical test. However, little is known about the performance of the test if the distributional and model assumptions are violated, and the Lo–Mendell–Rubin test may perform better in these contexts (Nylund *et al*., 2007).

An interesting alternative to the bootstrap likelihood ratio test to determine the number of clusters was proposed by Böhning (2000) and Schlattmann (2003, 2005 and 2009). In their approach, N data are resampled with replacement B times and the number of components c is determined from each bootstrap sample using a hybrid mixture algorithm. This estimation method combines a mixture estimator (VEM: vertex exchange method) for flexible support size to determine the number of parameters based on topology and geometry of the likelihood (see Böhning (2000) and Schlattmann (2009)) and the standard EM estimation method. The final number of components is obtained as the mode of the bootstrap distribution of c.

The likelihood ratio test and its modifications allow one to assess only nested models, and acceptance and rejection of a model are influenced by sample size. The use of the traditional approach of null hypothesis testing has been criticised for a variety of reasons (Cohen, 1994; Nickerson, 2000), and alternative model selection methods using information theoretic or Bayesian approaches such as Bayes factors have been advocated in recent years; see, for example,

Buckland *et al.* (1997) for information theoretic methods and Kass and Raftery (1995) and Marden (2000) for Bayes factors.

6.5.2 Information criteria

The use of information theoretic methods has been increasingly popular in model-based cluster analysis.

Model selection procedures based on information theoretical methods use a measure of the information lost when a particular model is used to approximate the (unknown) true model. A set of competing models, such as different numbers of clusters, are ranked according to their relative information loss. The model with the lowest relative information loss (or lowest information criterion) is then preferred. Unlike the LRT-based tests, information criteria allow one to quantify the differences between a candidate set of models and there may not be a single best model. The two most popular information criteria used for model selection are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC).

Akaike (1973) showed that an approximately unbiased estimator of the relative distance between a fitted model and the true unknown mechanism is

$$AIC = -2LL(\hat{\boldsymbol{\Psi}}|\text{data}) + 2p, \tag{6.20}$$

where $-2LL$ is twice the negative maximized log-likelihood (or deviance) and p is the number of parameters.

The first term measures the lack of fit of the data, while the second term is a penalty for model complexity. The AIC is therefore, like other information criteria, a penalized log-likelihood. The preferred model is one with the fewest parameters that still provides an adequate fit to the data.

The AIC identifies the model, from the set of candidate models, that produces a predictive density which is on average the closest to the true density. Use of the AIC as a general model selection criterion including finite mixture models is advocated by Burnham and Anderson (2002). However, the AIC is inconsistent, and studies have shown that it tends to overfit models and therefore overestimates the number of components (Koehler and Murphree, 1988; Celeux and Soromenho, 1996).

Another information criterion which is predominantly used in finite mixture model selection is the Bayesian information criterion (BIC; Schwarz, 1978), which is derived from a Bayesian framework but can also be applied in a frequentist way. The BIC is calculated as

$$BIC = -2LL(\hat{\boldsymbol{\Psi}}|\text{data}) + p\ln(n), \tag{6.21}$$

where p is the number of parameters and n is the sample size.

A sample-size adjusted version of the BIC replaces $\ln(n)$ with $\ln[(n+2)/24]$ (originally derived by Rissanen (1978), and suggested among others for finite mixture model selection by Sclove (1987)).

The BIC and adjusted BIC increase the penalty for complexity as sample size increases and the BIC especially places a high value on parsimony. The BIC evaluates the posterior probability of the competing models with specified priors

and thus provides an approximation to Bayes factors for regular models (Yang, 2006). Kass and Wasserman (1995) showed that for regular models the BIC provides an approximation to Bayes factors when prior odds are equal to 1 (unit information prior).

The approximations used to derive AIC and BIC values, and thus the validity of decisions made on number of clusters, depend strongly on regularity conditions. The regularity conditions do not necessarily hold for mixture models, especially if high-dimensional data are analysed (Aitkin and Rubin, 1985; Ray and Lindsay, 2008), but there is considerable theoretical and empirical support for the use of the BIC in model selection. Roeder and Wasserman (1997) and Keribin (2000) showed that the BIC leads to a consistent estimator of the correct number of components in mixtures. Campbell *et al.* (1997) and Dasgupta and Raftery (1998) demonstrated that BIC model selection performs well and usually outperforms the AIC in both simulations and real data sets with known component membership. Other authors confirmed that the AIC tends to overestimate the number of components in mixture models (Soromenho, 1993; Celeux and Soromenho, 1996; Andrews and Currim, 2003). However, a study by Brame *et al.* (2006) showed that the AIC was superior in recovering components when the components were less separated. A simulation study by Yang and Yang (2007) compared a variety of information criteria, including AIC, BIC and sample-size adjusted BIC, and concluded that the sample-size adjusted BIC performed well in most of their simulations unless sample size is limited and components are not well separated. The adjusted BIC also performed better in a simulation study to select the number of components in finite mixture regressions (Sarstedt and Schwaiger, 2008). The simulation studies by Nylund *et al.* (2007) showed that the bootstrap likelihood ratio test performed better in detecting the components in simulations than AIC, BIC and sample-size adjusted BIC.

All simulation results depend heavily on assumptions about the population (scale of observed variables, separation of latent classes, and number of classes) and the estimation model. Furthermore, none of the studies assessed the robustness of selection criteria if data do not match the assumptions of the estimation model. It should be kept in mind that any given continuous distribution can be approximated by a mixture model, and mixture models with as few as two or three components can provide a very good approximation to non-normal homogeneous distributions such as the beta or gamma distribution (Tarpey *et al.,* 2008). Research on model selection criteria has not yet provided an unequivocal answer to the basic question of selecting the right number of components. Several authors therefore recommend the use of multiple statistics along with theoretical and practical considerations (Bauer and Curran, 2004; Nagin and Tremblay, 2005; Nagin, 2010).

6.5.3 Bayes factors

Bayes factors can be used to compare nested and non-nested models with different parameterization, differing numbers of components or both. The Bayes factor is the posterior odds of one model against another model if neither model is favoured *a priori.* An introduction to Bayes factors is presented by Kass and Raftery (1995).

Estimation of Bayes factors involves the integration of the marginal likelihood, which is the limitation of a widespread use of Bayes factors for model selection in finite mixtures modelling. Carlin and Chib (1995), Chib (1995), Lewis and Raftery (1997), Verdinelli and Wasserman (1995) and Han and Carlin (2001) discuss the computation of Bayes factors via MCMC methods. However, Raftery (1995) shows that an approximation to the integral can be derived from the Bayesian information criteria (BIC), which is relatively easily computed (see formula 6.21) and does not require evaluation of prior distributions. The difference in BIC between two models i and j is approximately equal to -2 times the logarithm of the Bayes factor for model i versus model j. Simulation studies conducted by Roeder and Wasserman (1997) showed that the BIC provides a very good approximation to the marginal likelihood as long as the priors are not strong. This relationship between BIC and Bayes factor explains the popular use of the BIC for model selection.

Qu and Qu (2000) describe the implementation of a Bayesian approach to probit mixture models using Bayes factors to select the number of components. They applied their model to a study of insecticide tolerance of sheep worms (*Ostertagia spp*.), with the aim of determining subpopulations of susceptible and resistant strains after a treatment with an insecticide. Qu and Qu fitted a series of models with increasing number of components using data augmentation and Gibbs sampling, and selected a three-component model with equal variances using approximate Bayes factors. The three components corresponded to susceptible, moderately resistant and severely resistant strains. They concluded that the resistance is inherited, with the extremely resistant group corresponding to a homozygote genotype *aa* with two resistance alleles, while the moderately resistant strain corresponds to a heterozygous genotype *Aa* with only one resistance allele.

6.5.4 Markov chain Monte Carlo methods

If the number of components is not known, the application of the reversible jump MCMC algorithm of Green (1995) has been proposed by Richardson and Green (1997) for finite mixtures. This algorithm allows sampling from the joint posterior distribution of all mixture parameters, including c, the number of mixture components. The Markov chain thus 'jumps' between parameter subspaces, that is different models with different numbers of components. Jumps are achieved by adding new or deleting old components, or by splitting and merging existing ones. The posterior of the number of components, c, can be estimated by the relative frequency with which each model is visited during each iteration. Stephens (2000) suggested the use of a birth and death MCMC and applied it to multivariate mixtures of normal and t-distributions with unknown number of components. His method is easier to implement and handles multivariate cases better. Recently, Nobile and Fearnside (2007) developed a Markov chain Monte Carlo method using an allocation sampler for the Bayesian analysis of finite mixture distributions with an unknown number of components that can be used for mixtures of components from any parametric family.

It should be noted that the methods described above are applicable for selecting models in general. However, the regularity problems for comparing models with different numbers of components using likelihood-related tests or information criteria are less of an issue if different parameterizations of the clusters are compared.

In the following section we will discuss another model selection problem, namely how to select a subset of variables from a high-dimensional data set.

6.6 Dimension reduction – variable selection in finite mixture modelling

The cluster structure of interest in a data set is often confined to a subset of the available variables. This is frequently the case in the analyses of DNA microarray studies where the expression of thousands of genes on a small number of subjects is quantified and only a small subset of the genes is assumed to be responsible for the cluster structure. The analysis of a gene-expression data set is a typical example of high-dimensional data where the number of variables (genes) is considerably larger than the number of cases (subjects). The inclusion of redundant variables complicates or even masks the clustering process. Removing redundant variables would decrease the number of parameters to estimate the model, which then can be estimated with more precision (Raftery and Dean, 2006). Furthermore, using an informative subset of variables will often be of help in interpreting the fitted finite mixture model. In Section 3.7 we discussed ways to weight a variable to give it greater or lesser importance than other variables when using these to determine the proximity between two objects. In this paragraph we will concentrate on variable selection methods.

Standard procedures in cluster analyses in DNA microarray studies separate the variable selection and clustering process. Common methods are filtering out variables such as genes that are not expressed or do not vary across samples, and/or applying dimension reducing techniques such as principal component or principal coordinate analysis prior to the cluster analysis (Chae and Warde, 2006; Ghosh and Chinnaiyan, 2002; Shannon *et al.* 2003). The first approach does not assess the joint effect of multiple variables and may remove potentially important variables, which only provide significant improvement in conjunction with others (Tadesse *et al*., 2005). The second approach, focusing on the leading components of a principal component analysis, is also often not advisable. Chang (1983) showed for a two-component mixture model that the components with the largest eigenvalues do not necessarily provide most information about the clustering structure. Further empirical studies using real gene-expression and simulated data showed that using principal components instead of the original data often worsens the recovery of the underlying cluster structure (Yeung and Ruzzo, 2001). In addition, using principal components results in linear combinations and does not allow evaluation of the original variable.

In recent years two techniques have been introduced that recast the variable selection problem into a Bayesian model selection problem within the settings of

finite mixtures. Both techniques jointly estimate the cluster pattern and select the set of variables which best describes the clustering process. The first technique was introduced by Raftery and Dean (2006) for Gaussian mixtures. Their method assesses a variable's importance by comparing two models given the clustering variables already selected. The first model includes the potential cluster variable conditioned on clustering variables already selected, while in the second model the variable is not included. Variables that are neither selected for the clustering process nor assessed as a possible cluster variable are assumed to be conditionally independent of the clustering given the other variables. Again the BIC, as an easy-to-compute approximation for the Bayes factor, is used to compare the two competing models.

These authors recommend a 'greedy' search algorithm to search the parameter space and to select the clustering variables simultaneously with the number of clusters and parameterizations. The algorithm is similar to forward stepwise regression model selection, and thus vulnerable to the same problems, such as selecting an unstable model as the final model or underestimating the standard errors of the parameters. However, Raftery and Dean showed the usefulness of their model selection procedure/algorithm using simulated data and examples. They reanalysed the crab data set of Campbell and Mahon (1974), which was introduced in Chapter 2 with its five variables. Their procedure selected four out of the five variables and the correct number of clusters. Only 7.5% of the data points were erroneously classified, while the same finite mixture analyses without variable selection overestimated the number of clusters (7), leading to a high error rate of 45.5%. Even if the number of groups was known, the error classification was 42%. Dean and Raftery (2010) extended the procedure to latent class (multinomial mixture) models.

Maugis *et al.* (2009a, 2009b) proposed further extensions of Raftery and Dean's model selection procedure. The authors claim that their more versatile method improves the clustering and its interpretation especially for high-dimensional data.

The second approach to Bayesian model selection techniques was proposed by Tadesse *et al.* (2005) for high-dimensional data. It is a fully Bayesian method using a stochastic search and reversible jump MCMC procedure to jointly estimate the clustering process and the best subset of variables. Swartz *et al.* (2008) extended the method of Tadesse *et al.* (2005) for data with a known substructure, such as the structure imposed by an experimental design. They analysed microarray gene-expression data from a colon carcinogenesis study in rats which were fed with two different diets (either high in corn oil or high in fish oil). Their method allows one to answer the question as to whether the treatment affects the subjects differently and which genes define those differences.

McLachlan *et al.* (2002) used a different kind of approach to analyse high-dimensional data by using a mixture of factor analysers to reduce the extremely high dimensionality of a gene-expression problem. This method will be discussed in Chapter 7.

6.7 Finite regression mixtures

An obvious extension of the finite mixture model is a mixture of generalized linear models (GLMs) by estimating a generalized linear model for each component (Aitkin, 1996; Jansen, 1993; Wedel, 2002; Wedel and Desarbo, 1994). Conventional regression models implicitly assume that the regression coefficients β_j are the same over all observations, and treat the sample as a homogeneous group. This assumption is often violated if important variables are not included in the model. The sample may contain several unknown subpopulations with different sets of regression parameters. The application of finite mixtures of generalized linear models does not require knowledge of the subpopulations but tries to identify the populations and allows regression coefficients to vary between the subpopulations to adequately capture their characteristics.

Typically each component is described by a generalized linear model belonging to the same exponential family (the models have the same error distribution and link function) but different linear predictors. In this case the semiparametric mixing distribution of $\mathbf{p}$ of (6.1) would be

$$\mathbf{p}^* = \begin{pmatrix} \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_{c-1} \\ \mathbf{p}_1, \ldots, \mathbf{p}_{c-1} \end{pmatrix},$$

where $\boldsymbol{\lambda}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{im})$, with β_i denoting the regression parameters of the ith component, m is the number of covariates in the generalized linear model and $\mathbf{p}_j$, $(j = 1, \ldots, c)$ denotes the mixing weights of the c components.

A mixture of generalized linear models was used by Aitkin (1999) to assess heterogeneity of treatment effects in a meta-analysis. He reanalysed a 22-centre trial studying the effect of beta-blockers on mortality after myocardial infarction, which was originally published by Yusuf *et al.* (1985). A finite mixture of binomial regression models with treatment as fixed and centre as a random effect was used to describe the data. In the finite mixture model the random effect was allowed to follow a mixture distribution instead of the standard normal distribution. A three-component model with the intercept allowed to follow a mixture distribution, and fixed treatment effect, produced a considerably better fit than a standard random effects meta-analysis. The treatment effect on the logit scale was the same within each component, but overall mortality rates differed between the three subpopulations.

There is a growing body of literature on finite mixture regressions and their application (see Wedel, 2001; Wedel and Desarbo, 2002; Grun and Leisch, 2007; Schlattmann, 2009 for details).

6.8 Software for finite mixture modelling

Several user-written packages for the open-source software `R` are available. One that has already been mentioned is `mclust`, a multipurpose package for normal

mixture modelling via the EM algorithm. Within `mclust`, the function `Mclust` allows the fitting of a variety of covariance structures using eigenvalue decomposition, and performs automatic model selection for the number of clusters and the best covariance structure based on BIC criteria. The package `clustvarsel` written by Dean and Raftery (2009) extends the functionality of `mclust` by allowing automatic search for an optimal subset of variables based on BIC (Raftery and Dean, 2006).

Finite mixture regression modelling can be performed with the `R` package `flexmix` written by Bettina Gruen (Vienna University of Economics and Business) and Friedrich Fleisch (Ludwig Maximilian University of Munich) or `CAMAN` written by Peter Schlattmann (Charité, Berlin). The accompanying book by Schlattmann (2009) provides a variety of worked examples using `CAMAN`. User-friendly and powerful software for finite mixture modelling is `Latent GOLD` (Statistical Innovations). Powerful, specialized software for structural equation modelling which allow the estimation of finite mixture models includes `Mplus` (Muthén and Muthén, 2010) and `gllamm` (Rabe-Hesketh *et al.* (2004)), a user-contributed program for `Stata`.

6.9 Some examples of the application of finite mixture densities

In this section we give a number of examples of how finite mixture densities are used in practice, beginning with those involving Gaussian components, the first univariate and the second multivariate.

6.9.1 Finite mixture densities with univariate Gaussian components

We begin with an example involving the age of onset of schizophrenia.

A sex difference in the age of onset of schizophrenia was noted by Kraepelin (1919). Subsequently, it has been one of the most consistent findings in the epidemiology of the disorder. Levine (1981) collated the results of 7 studies on the age of onset of the illness, and 13 studies on age at first admission, and showed that all these studies were consistent in reporting an earlier onset of schizophrenia in men than in women. Levine suggested two competing models to explain these data:

> The timing model states that schizophrenia is essentially the same disorder in the two sexes but has an early onset in men and a late onset in women . . . In contrast with the timing model, the subtype model posits two types of schizophrenia. One is characterized by early onset, typical symptoms, and poor premorbid competence, and the other by late onset, atypical symptoms, and good premorbid competence . . . the early onset typical schizophrenia is largely a disorder of men, and late onset, atypical schizophrenia is largely a disorder in women.

Table 6.2 Data set for age of onset of schizophrenia (years).

(i) Women

20	30	21	23	30	25	13	19	16	25	20	25	27	43	6	21	15	26	23	21	23	23
34	14	17	18	21	16	35	32	48	53	51	48	29	25	44	23	36	58	28	51	40	43
21	48	17	23	28	44	28	21	31	22	56	60	15	21	30	26	28	23	21	20	43	39
40	26	50	17	17	23	44	30	35	20	41	18	39	27	28	30	34	33	30	29	46	36
58	28	30	28	37	31	29	32	48	49	30											

(ii) Men

21	18	23	21	27	24	20	12	15	19	21	22	19	24	9	19	18	17	23	17	23	19
37	26	22	24	19	22	19	16	16	18	16	33	22	23	10	14	15	20	11	25	9	22
25	20	19	22	23	24	29	24	22	26	20	25	17	25	28	22	22	23	35	16	29	33
15	29	20	29	24	39	10	20	23	15	18	20	21	30	21	18	19	15	19	18	25	17
15	42	27	18	43	20	17	21	5	27	25	18	24	33	32	29	34	20	21	31	22	15
27	26	23	47	17	21	16	21	19	31	34	23	23	20	21	18	26	30	17	21	19	22
52	19	24	19	19	33	32	29	58	39	42	32	32	46	38	44	35	45	41	31		

The subtype model implies that the age of onset distribution for both male and female schizophrenics will be a mixture, with the mixing proportion for early onset schizophrenia being larger for men than for women. The data are shown in Table 6.2. To investigate this model, age of onset (determined as age on first admission) of 99 female and 152 male schizophrenics was fitted using finite mixture distributions with normal components.

The data were analysed using the statistical open source software `R` (R development core team, 2010) and the user-contributed package `mclust` (Fraley and Raftery, 2010). `mclust` applies the BIC to choose the number of components and their parameterization. In the univariate case only two models are possible, with equal and unequal variances. Figure 6.2 shows the BIC values as a function of the number of components and type of variance model. By convention, differences of 2 or more in BIC values between two models are considered as positive evidence for model differences, differences of between 6 and 10 as strong evidence, and differences of more than 10 as decisive evidence (Kass and Raftery, 1995).

For both sets of data, model selection using BIC provides decisive evidence that a two-component mixture model provides the best fit. However, BIC model selection suggests that the variances of the two mixtures of the female data set are equal, while they are different for males. The parameter estimates of the selected models for males and females are shown in Table 6.3. Confidence intervals were obtained by using the bootstrap (see Efron and Tibshirani, 1993) using the `R` package `boot` (Canty and Ripley, 2010); see Everitt and Hothorn (2009) for programming details. The bootstrap distributions for each parameter and data set are shown in Figures 6.3 and 6.4.

Histograms of the data showing both the fitted two-component mixture distribution and a single normal fit are shown in Figure 6.5. The fitted two-component

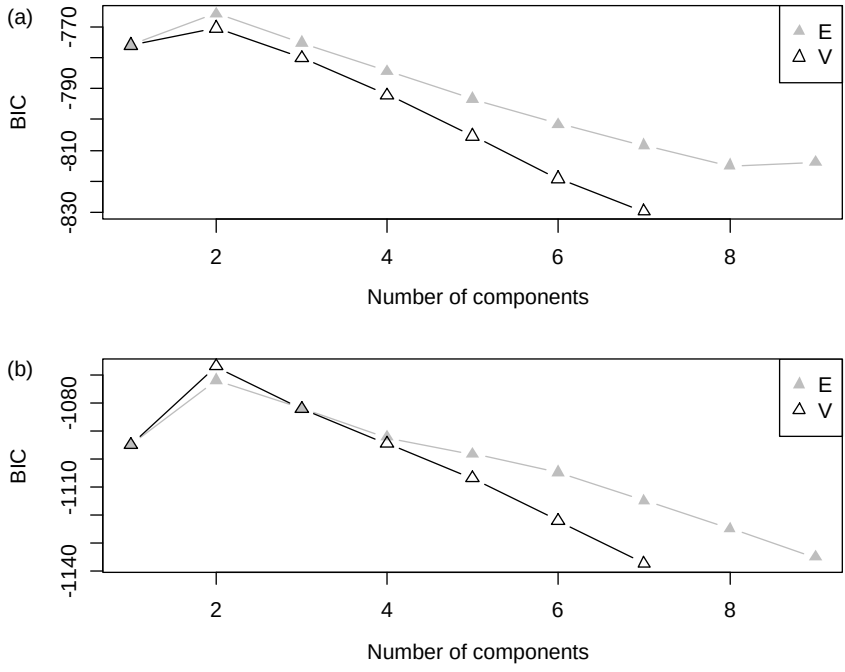


Figure 6.2 *BIC values for equal (E) and unequal (V) variance model parameterization and up to nine component models for (a) females and (b) males.*

mixture provides a reasonable fit to the data of both males and females, which provides support for Levine's *subtype model*. However, it is difficult to draw convincing conclusions about the proposed subtype model of schizophrenia because of the very wide confidence intervals for the parameters. Far larger sample sizes are required than those used here.

Table 6.3 Age of onset of schizophrenia: results of fitting finite two-mixture densities.

	Parameter	Final value	Bootstrap 95% CI[a]
(i) Women			
	p_1	0.746	(0.634, 0.845)
	μ_1	24.92	(23.24, 26.68)
	μ_2	46.82	(43.23, 50.57)
	σ^2	44.51	(33.70, 66.84)
(ii) Men			
	p_1	0.52	(0.340, 0.733)
	μ_1	20.28	(19.16, 21.47)
	σ_1^2	9.88	(5.64, 32.49)
	μ_2	27.92	(24.61, 33.45)
	σ_2^2	113.37	(81.10, 184.8)

[a] 95% bias-corrected and accelerated (BCa) bootstrap confidence interval based on 1000 bootstrap samples.

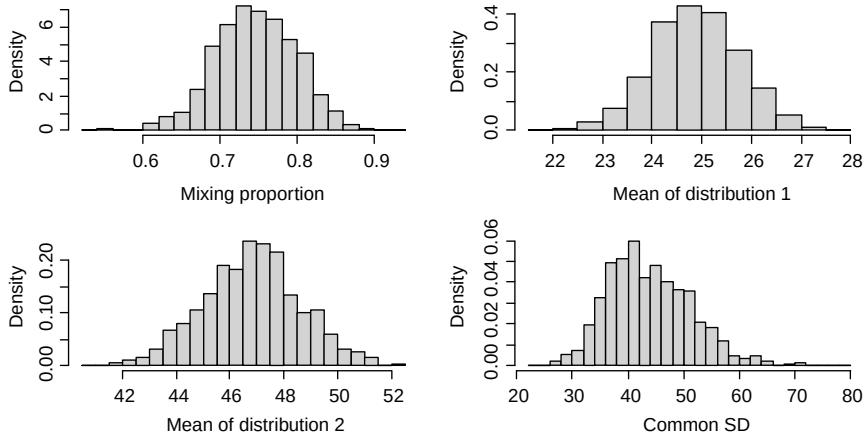


Figure 6.3 *Bootstrap distributions for four parameters of the two-component normal mixture fitted to age of onset data for females. The mixing proportion f, mean of first and second distribution and the common standard deviation (SD) of both distributions are shown.*

The second univariate example involves using gene-expression data obtained by microarray analyses. The advancement of microarray technology in this millennium allows the monitoring of the expression of thousands of genes simultaneously; for an overview of the technology, experimental design and analysis of microarray experiments see Russel *et al.* (2008). The aim of analyzing the expression data generated from microarray experiments is to detect genes with differential expression under two (or more) conditions. For example, in their seminal paper, Golub *et al.* (1999) used microarrays to identify genes in bone marrow tissue of patients with leukaemia whose expression differs between two diagnostic categories of leukaemia: acute myeloid leukaemia (AML) and acute lymphatic leukaemia (ALL). The authors assessed the gene expression of 7129 human genes of 45 AML and 27 ALL patients. One approach to analyse microarray experiments is to test for statistical differences in gene-expression level between the two conditions by performing a t-test for each of the N genes. The study of Golub revealed 2046 genes with a p-value smaller than 0.05. Avoiding a large proportion of false positives while keeping a reasonable power to detect genes with differential expression between the two groups is a statistical challenge. A standard procedure is to use false-detection-rate methods to adjust the p-values for multiple testing (e.g. Benjamini and Hochberg, 1995), to keep the proportion of false positives at a predefined level while maintaining a reasonable power. As an alternative, finite mixture models have been proposed to find differentially expressed genes.

A univariate finite mixture model is used to identify differently expressed genes using the data of Golub *et al.* (1999). The method suggested by Schlattmann (2009) was used. This approach assumes in the simplest case that a gene is either differentially expressed or not, and that we can rephrase the problem of finding differentially expressed genes into the framework of fitting a two-component

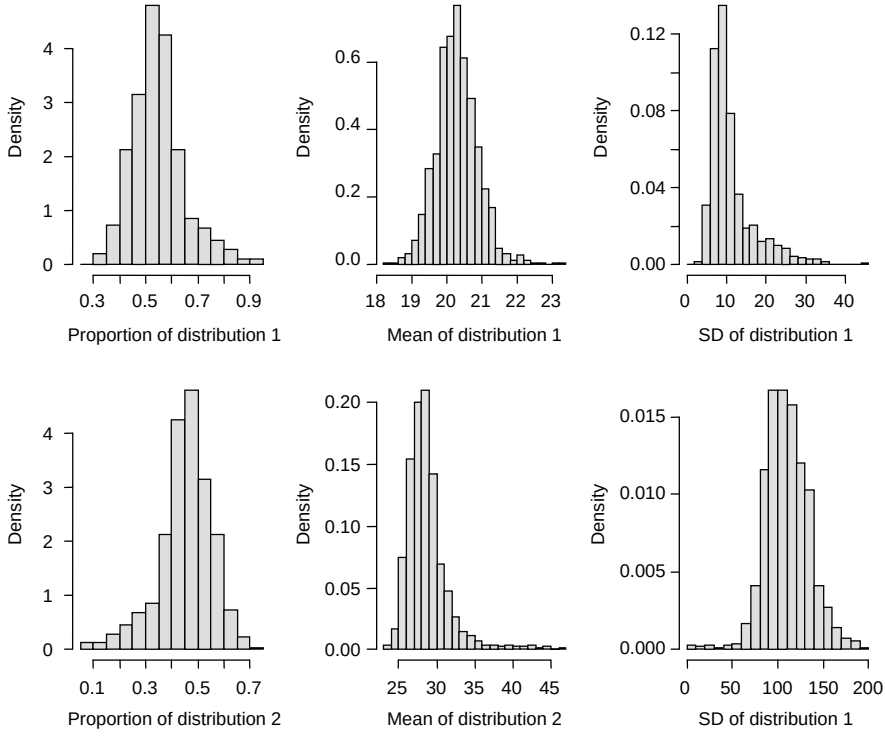


Figure 6.4 *Bootstrap distributions for six parameters of the two-component normal mixture fitted to age of onset data for males. The first row shows the proportion, mean and standard deviation of the first distribution, the second row shows the same parameters for the second distribution.*

mixture model of p-values of the statistical tests or other test statistics to model a null density for the non-differentially expressed genes and a density for the differentially expressed genes. Efron suggested the use of z-values instead of p-values P_i by

$$z_i = \Phi^{-1}(1-P_i) \quad i = 1, 2, \ldots, N, \tag{6.22}$$

where $\Phi^{-1}(\bullet)$is the inverse of the standard normal cumulative distribution function N(0,1).

Assuming independence of gene-expression levels across genes, under the null hypothesis the distribution of z_i will be standard normal N(0,1). In contrast, under the alternative hypothesis, the distribution of z-values will tend to cluster at a value larger than 0. Table 6.4 shows the first 50 p-values and their corresponding z-values for the leukaemia microarray study.

Finite mixture models of normals can be used to empirically estimate the null distribution of the z-scores and to identify a small percentage of potentially

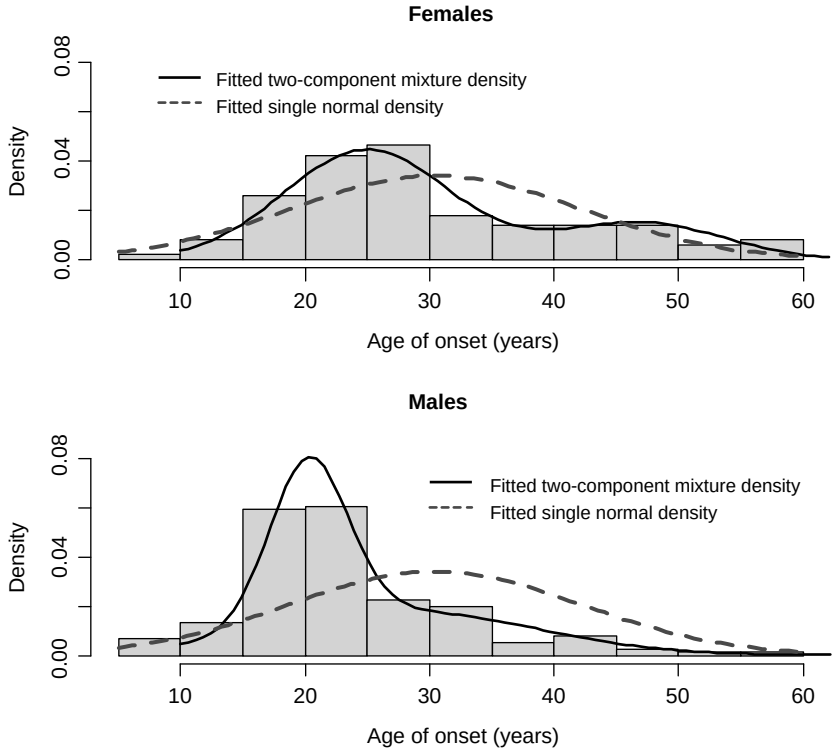


Figure 6.5 *Histograms and fitted mixture distributions for age of onset data for women and men.*

differentiated genes. The user-written `R` package `CAMAN` (Schlattmann and Höhne, 2009) was used to estimate the parameters of the mixture models. The package is described in detail in the book by Schlattmann (2009).

Table 6.5 shows the result of this analysis. The best model (with the smallest BIC) is a three-component model with equal variances. Figure 6.6 shows the histogram of the z-values, the fitted three-mixture distribution and the mixture density. The first component, with mean z-score of 0.15 and a variance of 1.08, models the z-distribution of the non-differentially expressed genes. The empirical null distribution differs slightly from the theoretical distribution. About 61% of the genes belong to this component. The second component includes 34.5% of the genes and shows some evidence of differential expression with a mean z-score of 1.8. The smallest component with almost 5% (or 335 genes) includes the differentially expressed genes with a mean z-score of almost 4. Further developments in using finite mixtures to analyse microarray data are described in Lee *et al.* (2000), Efron (2004), Allison *et al.* (2002), Jiao and Zhang (2008), Pan *et al.* (2003), McLachlan *et al.* (2006), and Khalili *et al.* (2009).

Table 6.4 *p*- and *z*-values for the statistical comparison of gene expressions between AML and ALL groups for the first 50 genes of the data of Golup *et al.* (1999); *p*-values are results of independent *t*-tests.

Gene	*p*	*z*
AFFX-BioB-5_at	0.520	0.528
AFFX-BioB-M_at	0.415	0.828
AFFX-BioB-3_at	0.863	−0.625
AFFX-BioC-5_at	0.110	1.980
AFFX-BioC-3_at	0.369	0.962
AFFX-BioDn-5_at	0.506	0.568
AFFX-BioDn-3_at	0.229	1.421
AFFX-CreX-5_at	0.112	1.966
AFFX-CreX-3_at	0.407	0.850
AFFX-BioB-5_st	0.817	−0.422
AFFX-BioB-M_st	0.418	0.816
AFFX-BioB-3_st	0.131	1.865
AFFX-BioC-5_st	0.407	0.848
AFFX-BioC-3_st	0.154	1.742
AFFX-BioDn-5_st	0.508	0.562
AFFX-BioDn-3_st	0.875	−0.700
AFFX-CreX-5_st	0.120	1.925
AFFX-CreX-3_st	0.349	1.025
hum_alu_at	0.392	0.893
AFFX-DapX-5_at	0.507	0.567
AFFX-DapX-M_at	0.673	0.079
AFFX-DapX-3_at	0.492	0.608
AFFX-LysX-5_at	0.807	−0.380
AFFX-LysX-M_at	0.875	−0.703
AFFX-LysX-3_at	0.215	1.471
AFFX-PheX-5_at	0.716	−0.055
AFFX-PheX-M_at	0.154	1.744
AFFX-PheX-3_at	0.186	1.594
AFFX-ThrX-5_at	0.386	0.913
AFFX-ThrX-M_at	0.488	0.616
AFFX-ThrX-3_at	0.369	0.963
AFFX-TrpnX-5_at	0.640	0.177
AFFX-TrpnX-M_at	0.653	0.136
AFFX-TrpnX-3_at	0.422	0.805
AFFX-HUMISGF3A/M97935_5_at	0.417	0.820
AFFX-HUMISGF3A/M97935_MA_at	0.172	1.656
AFFX-HUMISGF3A/M97935_MB_at	0.698	0.001
AFFX-HUMISGF3A/M97935_3_at	0.916	−0.986
AFFX-HUMRGE/M10098_5_at	0.114	1.956
AFFX-HUMRGE/M10098_M_at	0.057	2.364
AFFX-HUMRGE/M10098_3_at	0.095	2.074
AFFX-HUMGAPDH/M33197_5_at	0.336	1.066
AFFX-HUMGAPDH/M33197_M_at	0.454	0.716

Table 6.4 (*Continued*)

Gene	p	z
AFFX-HUMGAPDH/M33197_3_at	0.116	1.948
AFFX-HSAC07/X00351_5_at	0.518	0.532
AFFX-HSAC07/X00351_M_at	0.532	0.493
AFFX-HSAC07/X00351_3_at	0.509	0.560
AFFX-HUMTFRR/M11507_5_at	0.086	2.138
AFFX-HUMTFRR/M11507_M_at	0.086	2.134

Table 6.5 Finite mixture models for the data of Golub *et al.* (1999).

Number of components	LL	BIC	Delta BIC ($BIC_{model\ i} - BIC_{best\ model}$)	Component weight p_i	Component mean (variance)
1	−12874	25757	98	1	0.90 (2.13)
2	−12768	25564	5	C1: 0.860	0.57 (1.46)
				C2: 0.140	2.99 (1.46)
3	−12757	25559	0	C1: 0.609	0.15 (1.08)
				C2: 0.343	1.82 (1.08)
				C3: 0.047	3.96 (1.08)
4	−12753	25568	9	C1: 0.430	−0.18 (0.85)
				C2: 0.460	1.31(0.85)
				C3: 0.120	3.10 (0.85)
				C4: 0.010	5.00 (0.85)

Results are presented only for models with equal variances for the different components. LL is the log-likelihood value for a model. Delta BIC is the difference in BIC between model *i* and the best model. Differences in BIC between two models of 2 or more are considered as positive evidence for model differences, differences between 6 and 10 as strong evidence, and differences of more than 10 as decisive evidence (Kass and Raftery, 1995).

6.9.2 Finite mixture densities with multivariate Gaussian components

In this section we will look at an example involving multivariate data; the example is concerned with diabetes diagnosis (Reaven and Miller, 1979). The aim of the study was to assess the possibility of using the three indirect variables for a reliable classification of diabetes type status that is close to the clinical classification. The data set consists of 145 patients with measurements of the following three physiological variables:

- *glucose* intolerance: plasma glucose level response after a glucose drink;
- *insulin* response: plasma insulin changes as a response to oral glucose over a period of time;
- *SSPG* (steady-stage plasma glucose level): measure of insulin resistance.

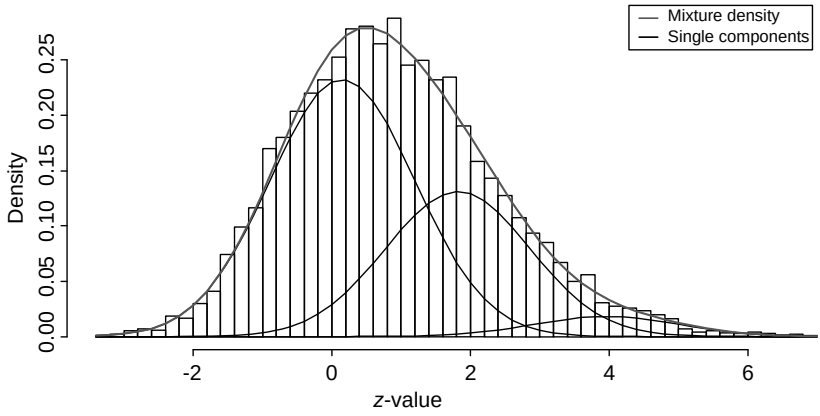


Figure 6.6 *Histograms of z-scores, mixture density and single components of the three-mixture model fitted to the data of Golub et al. (1999).*

The subjects were clinically classified as normal, chemical (mild) diabetic (pre-diabetes stage) and overt (clinical) diabetic, which allows a comparison with the classification obtained from the mixture model. The data set is shown in Table 6.6; Figure 6.7 shows a scatterplot matrix of the three variables with densities and rug plots in the margins of the diagonal cells. The bivariate scatterplots suggest some clustering in the data. One to nine components with different variance–covariance structures based on the eigenvalue decomposition (see Table 6.1) were fitted to the data, and BIC was used to select the best fitting model. Figure 6.8 shows that the best fitting model with the lowest BIC value was a three mixture with unrestricted variance–covariance structure ('VVV').

The scatterplot matrix of the three variables with different symbols for the classification based on the mixture model is shown in Figure 6.9. The three-dimensional shape of the data resembles a 'boomerang with two wings and a fat middle' (Reaven and Miller, 1979). The two wings are interpreted as representing patients with chemical diabetes and with overt diabetes, respectively, while the spherical middle corresponds to normal subjects. Table 6.7 presents the mixture probabilities and the means and standard deviations of each clinical variable for each cluster. A comparison of the three-component classification with the clinical assessment shows 88.3% agreement of the data. Details about the comparison of clinical assessment and component classification are shown in Table 6.8.

Figure 6.10 shows a scatterplot matrix of the three variables with filled symbols corresponding to misclassification. Misclassification occurs mainly in the 'fat middle'. This corresponds to areas of high uncertainty, the 1-probability of case i belonging to a cluster j, which are plotted in Figure 6.11. Larger dots mean greater uncertainty, and most misclassified cases show a high classification uncertainty, suggesting a suitable model for the data as clinically classified. Finally, the uncertainty plots also show the ellipses with axes corresponding to the variance of each component.

Table 6.6 Diabetes data set.

Clinical classification	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	
Glucose	98	88	100	89	91	90	94	85	96	111	107	114	101	108	112	105	103	99	102	110	102	96	95	112
Insulin	478	439	429	472	436	413	426	425	465	558	503	540	469	486	568	527	537	466	599	477	472	456	517	503
SSPG	151	208	201	162	148	344	213	143	237	748	320	188	607	297	232	480	622	287	266	124	297	326	564	408
Clinical classification	C	C	C	C	C	C	C	C	C	C	C	C	N	N	N	N	N	N	N	N	N	N	N	N
Glucose	110	92	104	75	92	92	92	93	112	88	114	103	80	97	105	90	90	86	100	85	97	97	91	87
Insulin	522	476	472	45	442	541	580	472	562	423	643	533	356	289	319	356	323	381	350	301	379	296	353	306
SSPG	325	433	180	392	109	313	132	285	139	212	155	120	124	117	143	199	240	157	221	186	142	131	221	178
Clinical classification	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Glucose	78	90	86	80	90	99	85	90	90	88	95	90	92	74	98	100	86	98	70	99	75	90	85	99
Insulin	290	371	312	393	364	359	296	345	378	304	347	327	386	365	365	352	325	321	360	336	352	353	373	376
SSPG	136	200	208	202	152	185	116	123	136	134	184	192	279	228	145	172	179	222	134	143	169	263	174	134
Clinical classification	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
Glucose	100	78	106	98	102	90	94	80	93	86	85	96	88	87	94	93	86	86	96	86	89	83	100	110
Insulin	367	335	396	277	378	360	291	269	318	328	334	356	291	360	313	306	319	349	332	323	323	351	398	426
SSPG	182	241	128	222	165	282	94	121	73	106	118	112	157	292	200	220	144	109	151	158	73	81	122	117
Clinical classification	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	O	O	O	O	O	O	O	O
Glucose	80	96	95	82	84	100	86	93	107	112	93	93	90	99	93	89	300	303	125	280	216	190	151	303
Insulin	333	418	391	390	416	385	393	376	403	414	364	391	356	398	393	318	1468	1487	714	1470	1113	972	854	1364
SSPG	131	130	137	375	146	192	115	195	267	281	156	221	199	76	490	73	28	23	232	54	81	87	76	42
Clinical classification	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O	O
Glucose	173	203	195	140	151	275	260	149	233	146	124	213	330	123	130	120	138	188	339	265	353	180	213	328	346
Insulin	832	967	920	613	857	1373	1133	849	1183	847	538	1001	1520	557	670	636	741	958	1354	1263	1428	923	1025	1246	1568
SSPG	102	138	160	131	145	45	118	159	73	103	460	42	13	130	44	314	219	100	10	83	41	77	29	124	15

Diagnostic groups: N = Normal, C = chemical diabetes, O = overt diabetes.

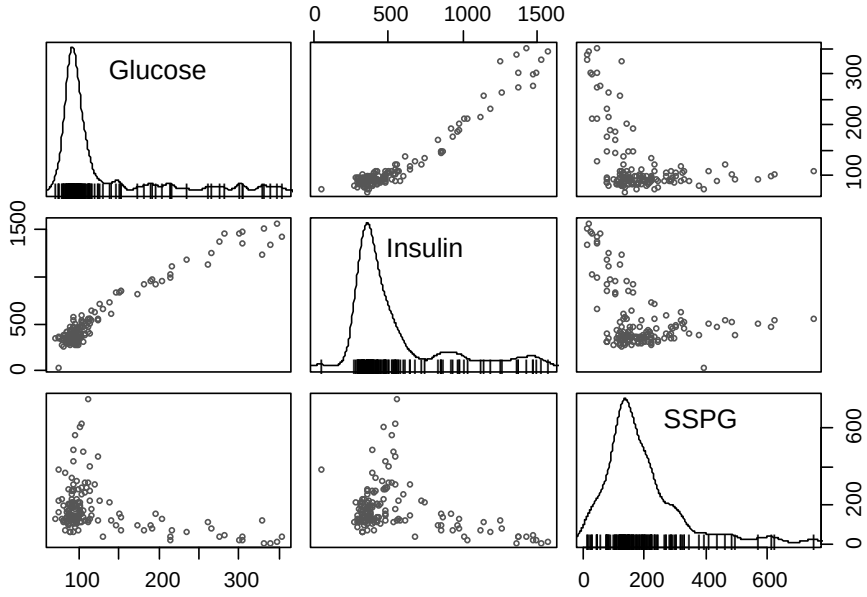


Figure 6.7 *Scatterplot matrix of the three variables in the diabetes data set with densities and rug plots in the margins of the diagonal cells.*

It may be relevant to assess if the structure of interest can be similarly represented by a smaller number of variables. In the diabetes data set a suitable subset of two variables may reduce time and costs for a screening procedure. The `R` package `clustvarsel` (Dean and Raftery, 2009) extends the functions of

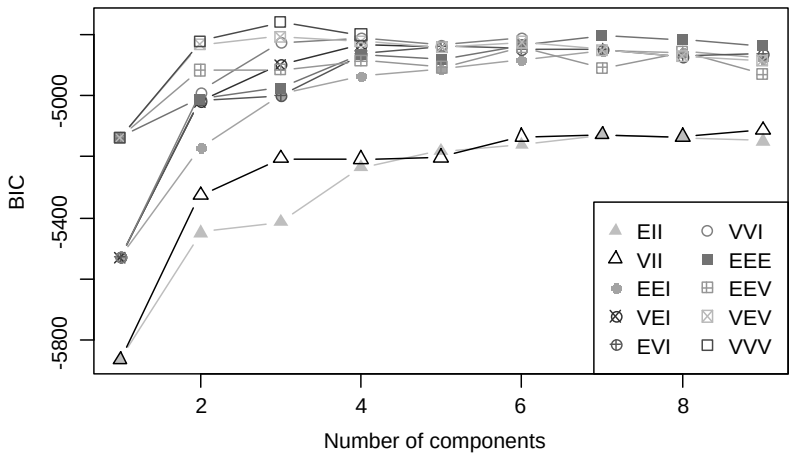


Figure 6.8 *BIC values for 10 different model parameterizations available in* `mclust` *(see Table 6.1 for details) and up to 9 component models for diabetes data.*

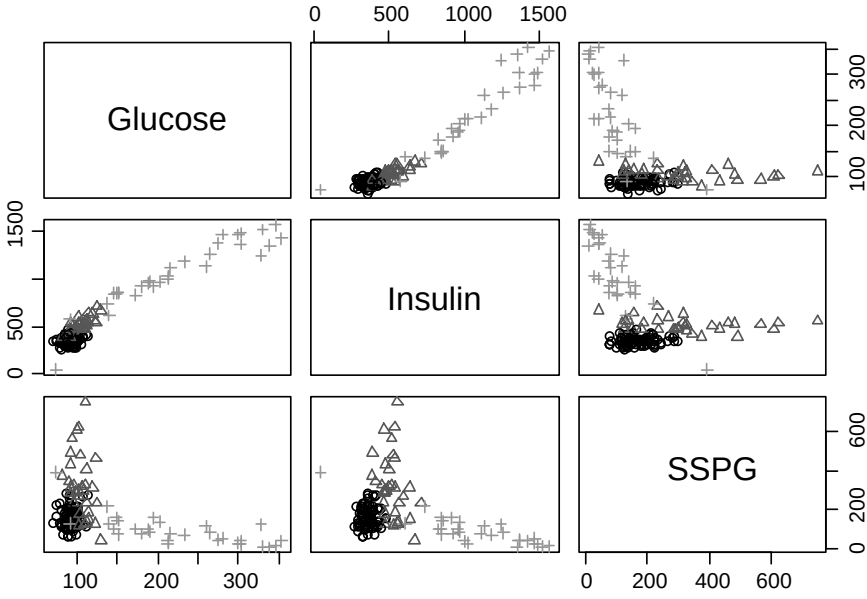


Figure 6.9 *Bivariate scatterplot matrix of glucose, insulin and SSPG variables with different symbols indicating the classification corresponding to the three-component model projection with unrestricted variance–covariance matrix.*

`mclust` by implementing the subset variable selection method proposed by Raftery and Dean (2006) and described in Section 6.6. Applying `clustvarsel` to the diabetes data revealed that all three variables are necessary for an adequate mixture model fit of the diabetes data.

6.9.3 Applications of latent class analysis

As part of their detailed statistical modelling of an extensive body of educational research data on teaching, Aitkin *et al.* (1981) fitted latent class models to observations on 38 binary variables describing teaching behaviour observations

Table 6.7 Mixture probabilities and mean (standard deviation) for glucose, insulin and SSPG levels for cases of each component.

Component	1	2	3
Mixing probabilities	56.2%	22.3%	21.5%
Glucose	91.4 (7.87)	105.2 (46.08)	219.4(51.78)
Insulin	358.8 (12.38)	517.0 (80.71)	1041.4 (161.48)
SSPG	166.2 (79.65)	320.8 (349.29)	98.3 (76.76)
Label	Normal	Chemical	Overt

Table 6.8 Classification table (%).

Clinical classification	Classification based on finite mixture model		
	Normal	Chemical	Overt
Normal	97.4	2.6	0
Chemical	22.2	72.2	5.6
Overt	0	15.2	84.4

made on 468 teachers. The parameter estimates obtained by maximum likelihood via the EM algorithm for the two- and three-class models are shown in Table 6.9. For the two-class model, the response probabilities marked [a] show large differences between the classes, indicating systematic differences in behaviour on these items for teachers in the two groups. For the three-class model the response probabilities for classes 1 and 2 are very close to those for the corresponding classes in the two-class model (though in most cases more widely separated), and the response probabilities for class 3 are mostly between those for classes 1 and 2. Aitkin *et al*. interpreted the groups they found in terms of 'formal' and 'informal' teaching styles.

A further interesting application of latent class analysis is reported in Pickering and Forbes (1984), where the method is used to study how to allocate neonatal resources throughout Scotland. By classifying individuals into groups with similar medical characteristics, the resources required by each area can be estimated from the prevalence of each type ('casemix'). The data for analysis consisted of 11 categorical variables observed on 45 426 individuals. The variables contained clinical and diagnostic information extracted from the Scottish Neonatal Discharge Record. Two-, three- and four-class models were fitted using maximum likelihood methods. The results are shown in Table 6.10. Of the solutions shown, that involving four groups was considered by the investigators as the most interesting and relevant. One class was identified as healthy infants, two others were associated with moderately ill infants requiring various types of special neonatal care, and the fourth class contained severely ill infants of very low birth weight. How the classification was finally used is discussed in detail in the original paper.

6.9.4 Application of a mixture model with different component densities

As an example of the application of a finite mixture density in which the components have different forms, we shall briefly describe a study reported in Everitt and Bullmore (1999). This involved an investigation of how to identify brain regions activated when a subject is performing a particular task. There is already an extensive literature on statistical methods for the analysis of functional magnetic resonance images of the brain; see, for example, Rabe-Hesketh *et al.* (1997, 1998); Ford and Holmes (1998), Hartvig and Jensen (2000); Woolrich *et al.* (2005) and Xu

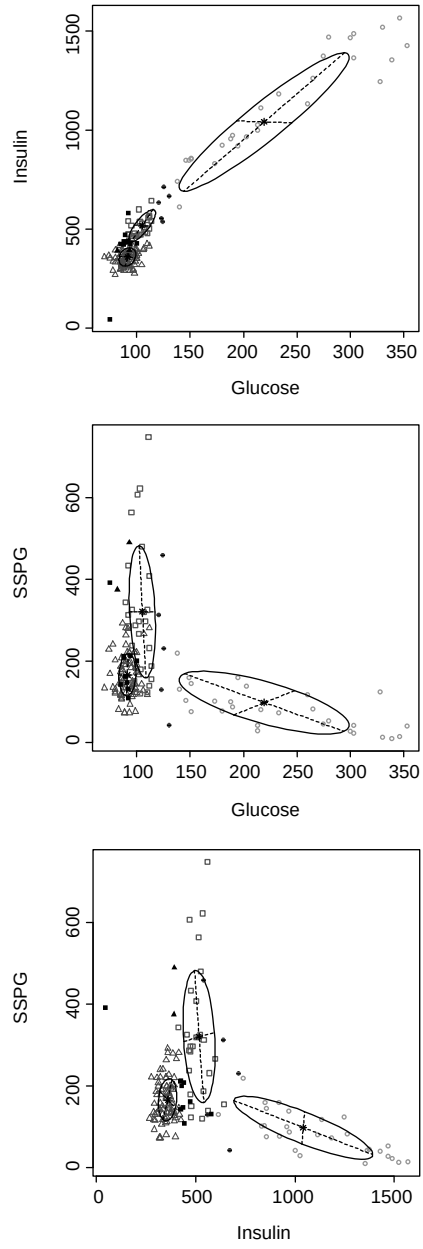


Figure 6.10 *Bivariate scatterplot matrix with misclassified cases plotted as filled symbols.*

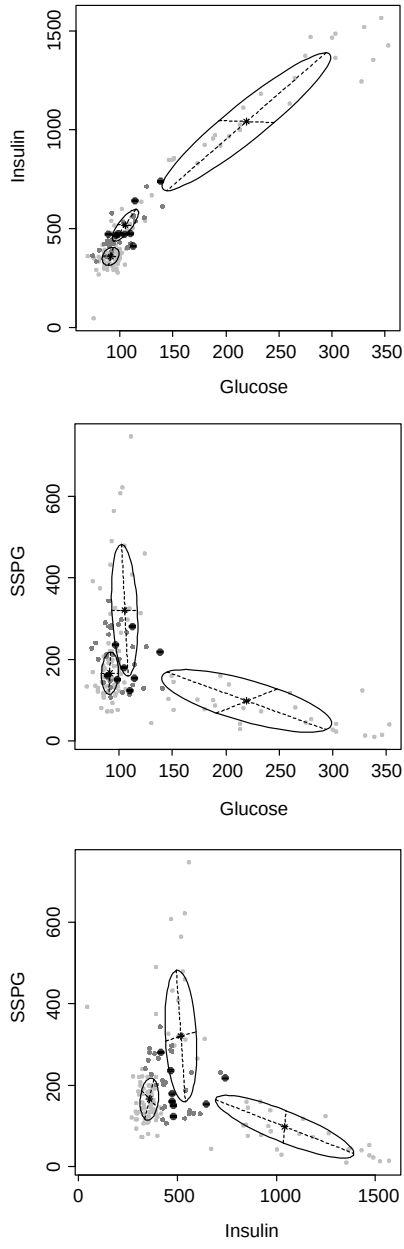


Figure 6.11 *Bivariate scatterplot matrix showing classification uncertainty. Larger symbols indicate more uncertainty in classification.*

Table 6.9 Two and three latent class parameter estimates (100 $\hat{\theta}_{ij}$) for teacher data (reproduced with permission from Aitkin *et al.*, 1981).

		Two-class model		Three-class model		
	Item	Class 1	Class 2	Class 1	Class 2	Class 3
1	Pupils have choice in where to sit	22	43	20	44	33
2	Pupils sit in groups of three or more	60	87[a]	54	88	79
3	Pupils allocated to seating by ability	36	23	36	22	30
4	Pupils stay in same seats for most of day	91	63[a]	91	52	89
5	Pupils not allowed freedom of movement in classroom	97	54[a]	100	53	74
6	Pupils not allowed to talk freely	89	48[a]	94	50	61
7	Pupils expected to ask permission to leave room	97	76[a]	96	69	95
8	Pupils expected to be quiet	82	42[a]	92	39	56
9	Monitors appointed for special jobs	85	67	90	70	69
10	Pupils taken out of school regularly	32	60	33	70	35
11	Timetable used for organizing work	90	66[a]	95	62	77
12	Use own materials rather than textbooks	19	49	20	56	26
13	Pupils expected to know tables by heart	92	76	97	80	75[b]
14	Pupils asked to find own reference materials	29	37	28	39	34
15	Pupils given homework regularly	35	22	45	29	12[b]
16 (i)	Teacher talks to whole class	71	44	73	37	62
(ii)	Pupils work in groups on teacher tasks	29	42	24	45	38
(iii)	Pupils work in groups on work of own choice	15	46[a]	13	59	20
(iv)	Pupils work individually on teacher tasks	55	37	57	32	50

(*Continued*)

Table 6.9 (*Continued*)

		Two-class model		Three-class model		
	Item	Class 1	Class 2	Class 1	Class 2	Class 3
(v)	Pupils work individually on work of own choice	28	50	29	60	26[b]
17	Explore concepts in number work	18	55[a]	14	62	34
18	Encourage fluency in written English even if inaccurate	87	94	87	95	90
19	Pupils' work marked or graded	43	14[a]	50	16	20
20	Spelling and grammatical errors corrected	84	68	86	64	78
21	Stars given to pupils who produce best work	57	29	65	30	34
22	Arithmetic tests given at least once a week	59	38	68	43	35[b]
23	Spelling tests given at least once a week	73	51	83	56	46[b]
24	End of term tests given	66	44	75	48	42[b]
25	Many pupils who create discipline problems	09	09	07	01	18[b]
26	Verbal reproof sufficient	97	95	98	99	91[b]
27 (i)	Discipline – extra work given	70	53	69	49	67
(ii) (16)	Smack	65	42	64	33	63
(iii)	Withdrawal of privileges	86	77	85	74	85
(iv)	Send to head teacher	24	17	21	13	28[b]
(v) (17)	Send out of room	19	15	15	08	27[b]
28 (i) (18)	Emphasis on separate subject teaching	85	50[a]	87	43	73
(ii)	Emphasis on aesthetic subject teaching	55	63	53	61	63[b]
(iii) (19)	Emphasis on integrated subject teaching	22	65[a]	21	75	33
p	Estimated proportion of teachers in each class	0.538	0.462	0.366	0.312	0.322

[a] An item with large differences in response probability between classes 1 and 2.
[b] An item in which class 3 is extreme.

Table 6.10 Classification of Scottish infants. Parameter estimates for one- to four-class models using 1980 data and complete cases only.

			1 class	2 classes		3 classes			4 classes			
Variable[a]	No. of levels[b]		i	iia	iib	iiia	iiib	iiic	iva	ivb	ivc	ivd
1	4	2001–2500 g	0.05	0.01	0.48	0.00	0.18	0.79	0.00	0.20	0.45	0.03
		1501–2000 g	0.01	0.00	0.15	0.00	0.26	0.09	0.00	0.32	0.09	0.00
		$\leq$1500 g	0.01	0.00	0.08	0.00	0.21	0.01	0.00	0.25	0.01	0.00
2	2	<10th centile	0.10	0.07	0.43	0.07	0.19	0.62	0.07	0.21	0.62	0.10
3	2	<7	0.02	0.01	0.12	0.01	0.26	0.01	0.00	0.21	0.01	0.32
4	3	Intermediate[c]	0.09	0.08	0.19	0.08	0.26	0.13	0.07	0.25	0.13	0.31
		By intubation	0.03	0.02	0.17	0.02	0.33	0.04	0.01	0.29	0.00	0.52
5	2	Present	0.01	0.00	0.10	0.00	0.25	0.00	0.00	0.29	0.00	0.01
6	2	Present	0.01	0.00	0.06	0.00	0.17	0.00	0.00	0.20	0.00	0.00
7	2	Present	0.30	0.28	0.58	0.28	0.67	0.49	0.28	0.71	0.49	0.32
8	2	Present	0.00	0.00	0.03	0.00	0.07	0.00	0.00	0.07	0.00	0.01
9	2	Present	0.03	0.01	0.30	0.01	0.60	0.10	0.01	0.67	0.10	0.05
10	2	Present	0.00	0.00	0.05	0.00	0.13	0.00	0.00	0.15	0.00	0.00
11	3	4–10 days	0.80	0.84	0.34	0.83	0.09	0.53	0.83	0.04	0.50	0.84
		>11 days	0.05	0.03	0.61	0.03	0.79	0.45	0.03	0.82	0.46	0.14
Frequency of class			1.00	0.92	0.08	0.92	0.03	0.05	0.89	0.03	0.05	0.04

[a] 1 = birthweight; 2 = birthweight for gestation age; 3 = apgar at 5 min; 4 = resuscitation; 5 = assisted ventilation after 30 min; 6 = recurrent apnoea; 7 = jaundice (>86 μmol/litre bilirubin); 8 = convulsions; 9 = in tube feeding; 10 = dead at discharge; 11 = age at discharge.

[b] Parameters for all levels of a variable sum to 1; the first level is omitted without loss of information.

[c] Mask with intermittent positive pressure ventilation; drugs only; other.

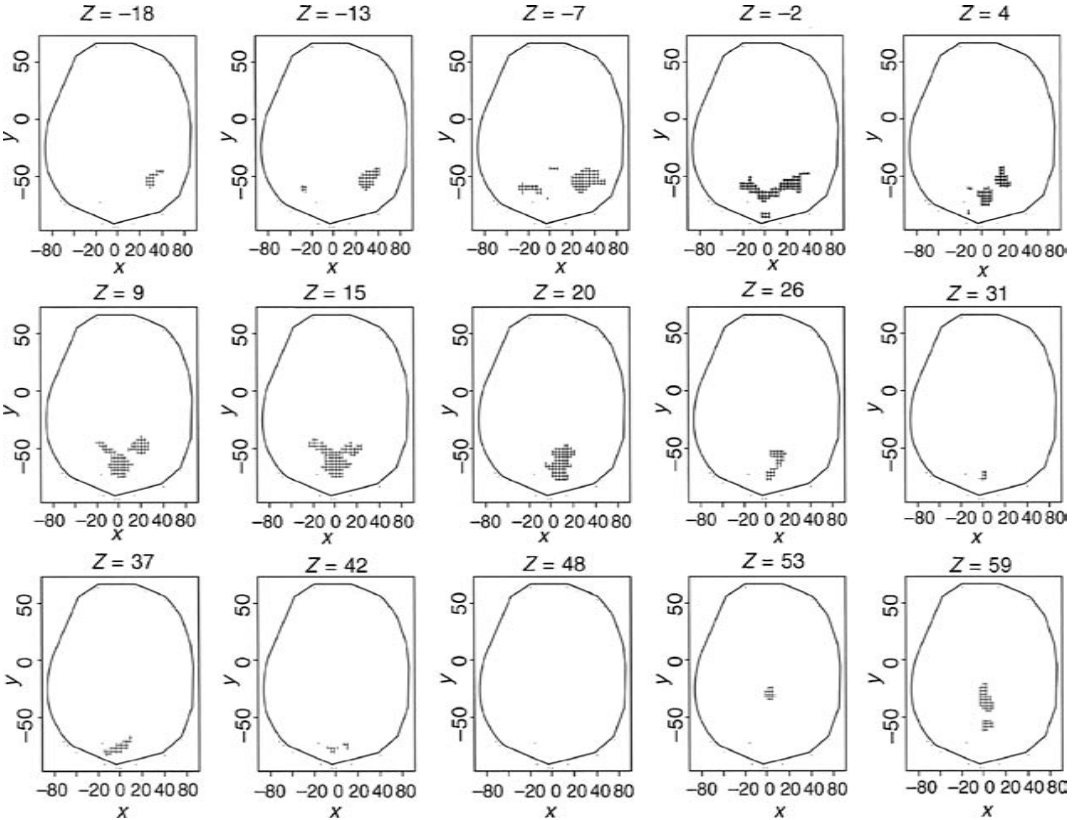


Figure 6.12 *Mixture model activation map of visual simulation data derived from estimated posterior probabilities of activation for the 26 535 voxels. Threshold for posterior probabilities = 0.5. Each slice of data is displayed in the standard anatomical space of Talairach and Tournoux (1988).*

et al. (2009). The last three studies applied spatial Bayesian mixture models to analyse fMRI data. Here we deal only with the use of a relatively straightforward finite mixture approach.

In the experiment of interest, fMRI data was collected from a healthy male volunteer during a visual simulation experiment (for details see Everitt and Bullmore, 1999). A measure of the experimentally determined signal at each voxel in the image was calculated as described in Bullmore *et al.* (1996). Under the null hypothesis of no experimentally determined signal change (no activation), the derived statistic has a chi-squared distribution with two degrees of freedom. Under the presence of an experimental effect (activation), however, the statistic has a noncentral chi-squared distribution. Consequently it follows that the distribution of the statistic over all voxels in an image, both activated and non-activated, can be modelled by a mixture of those two component densities. So if p denotes the proportion of non-activated voxels in an image comprising n voxels in total, the mixture distribution to be assumed is

$$f(x;\lambda,p) = pf_1(x) + (1-p)f_2(x;\lambda), \tag{6.23}$$

where f_1 is a chi-squared distribution with two degrees of freedom and f_2 is a noncentral chi-squared distribution with noncentrality parameter λ. In essence, λ will be a measure of the experimental effect. Specifically

$$f_1(x) = \tfrac{1}{2}\, e^{-\frac{1}{2}x} \tag{6.24}$$

$$f_2(x) = \tfrac{1}{2} e^{-\frac{1}{2}(x+\lambda)} \sum_{r=0}^{\infty} \frac{\lambda^r x^r}{2^{2r}[r!]^2}. \tag{6.25}$$

Given the n observed values of the statistic, $x_1, x_2, \ldots, x_n$, the parameters p and λ can be estimated by maximizing the log-likelihood. Voxels can then be classified as activated or non-activated on the basis of the maximum values of the estimated posterior probabilities.

For the visual simulation data this procedure led to estimates $\hat{p} = 0.96$ and $\hat{\lambda} = 12.04$. Voxels were classified as activated if their posterior probability of activation was greater than 0.5. Figure 6.12 shows the 'mixture model activation map' of the visual simulation data for selected slices of the brain (activated voxels indicated).

Mixture models are not only used for classifying areas of fMRI brain images in activating, deactivating or not-activating brain regions, but are also applied in the statistical segmentation of images in general (see for example Yang and Krishnan, 2004; Wehrens *et al.*, 2004).

6.10 Summary

In this chapter we have introduced finite mixture modelling as a more formal approach to cluster analysis than the methods described in previous chapters. Finite

mixture models assume that the population from which our sample data arise consists of subpopulations, each of which can be described by a different multivariate probability distribution. Finite mixture densities often provide a sensible statistical base for performing cluster analysis. Such models can be compared and assessed in an objective way with, in particular, the problem of determining the number of clusters reducing to a model selection problem for which objective procedures exist. A large number of possible distributions may be employed to model the mixture components; this provides an extremely flexible approach to clustering and so finite mixture models are increasingly used for cluster analysis.

Despite the many advantages of the finite mixture approach to cluster analysis, there are still some unresolved issues, especially concerning choosing the number of clusters, and future research is needed. One disadvantage of the approach is that large sample sizes may be necessary to get reasonably precise parameter estimates.

7

Model-based cluster analysis for structured data

7.1 Introduction

In the previous chapter we described how finite mixture models could be used as the basis of a sound statistical approach to cluster analysis. In this chapter we stay with the model-based framework but consider the implications for finite mixture models for clustering data where the subpopulation means and covariance matrices can be described by a *reduced* set of parameters because of the special nature of the data. This reduction in number of parameters achieved by exploiting the structure of the data helps in two ways:

(i) It may lead to more precise parameter estimates and ultimately more informative and more useful cluster analysis solutions.
(ii) It may be possible to convincingly fit finite mixture models to smaller data sets; the unstructured models introduced in the last chapter often contain a very large number of parameters, with the consequence that it may only be possible to fit the models using very large samples.

Figure 7.1 illustrates this second point. The general finite mixture model assumes that the variables in each subpopulation follow a multivariate distribution, with different mean vectors and, possibly, different covariance matrices. In the figure the subpopulation is indicated by a *latent* cluster variable C (latent in the sense that it is not known *a priori*). The arrows pointing from C to observed variables $x_1, x_2, \ldots, x_p$, which are the basis of the clustering, indicate parameters that describe differences in means between the clusters. The variables $x_1, x_2, \ldots, x_p$

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

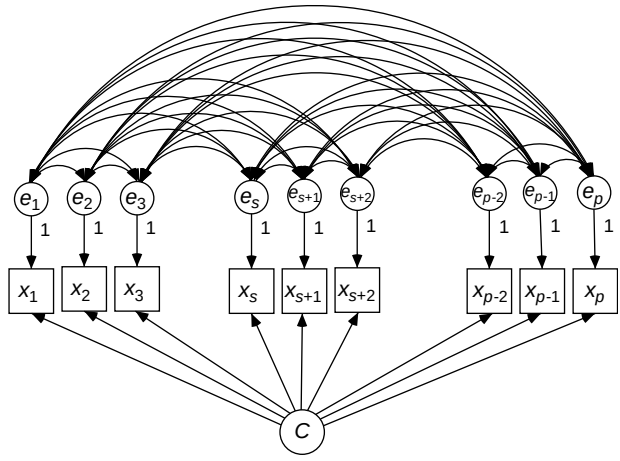


Figure 7.1 *Illustration of unstructured finite mixture model. Squares indicate observed variables and circles latent variables. C indicates cluster membership, x the observed variables and e residual terms with mean zero. Cluster membership is allowed to affect all free parameters in the model; that is, means of the observed variables as well as the covariance matrix of the residual terms. The effects of the errors on the manifest variables have been set to 1.*

may be assumed to be generated by adding zero-mean disturbance terms $e_1, e_2, \ldots, e_p$ respectively to the cluster means. The errors may correlate between variables, and the resulting correlation parameters are indicated by the double-headed arrows connecting disturbance terms. It is apparent that the number of covariance matrix parameters within a cluster increases drastically as the number of variables (p) increases. (There are $p(p+1)/2$ variance or covariance parameters.) If the covariance matrix varies with cluster, this number has to be multiplied by the number of clusters (c). Even with a moderate number of variables the total number of parameters to be estimated then becomes very large.

The most common types of data that impose a structure on the means and covariance matrices are where the same variable is observed under a number of different conditions (*repeated measures data*), or simply at a number of different time points (*longitudinal data*), or where different indicator variables scored on the same scale follow a known *factor model*. Such data can be represented by the usual $n \times p$ multivariate data matrix, $\mathbf{X}$, where the rows contain the objects' multiple measurements made on the same scale. In addition, *structured data* (a term introduced previously in this book first in Chapter 1 and then again in Chapter 3 in the context of defining proximity definitions for such data) are further characterized by a p-dimensional reference vector $\mathbf{r}$, whose kth entry r_k is used to code the nature of the kth variable column in $\mathbf{X}$. For example, in a set of longitudinal data where measurements are taken at times, $t_1, t_2, \ldots, t_p$, the reference vector would simply be $\mathbf{r}' = (t_1, t_2, \ldots, t_p)$. Later in the chapter we will

show that by using appropriate models for the covariance matrix of structured data, we can 'lose the umbrella' in Figure 7.1 and so drastically reduce the number of parameters to be estimated.

Concrete examples will help to clarify the type of data that will be of central concern in this chapter. First consider a longitudinal study investigating how children develop over time, and in which their heights are measured monthly for several years. Here the data are structured, with the data matrix containing repeated height measurements and accompanied by a time reference vector. Because of the temporal ordering, developmental studies focus on describing growth profiles or trajectories over time. Were we to cluster the children based on their growth data, then the longitudinal structure would help us to describe the within-cluster mean trajectories and covariances between repeated assessments by a reduced set of parameters, instead of expressing every mean and covariance by a separate parameter.

As an example of more general repeated measures data, consider the data generated by microarray experiments. Typically in such experiments, gene expressions are obtained over a range of experimental conditions with hybridizations replicated under each condition. Thus, for every object (gene) we have repeated measures of the same variable (expression level) referenced by a vector that codes the experimental condition under which the measure was taken. Under such a study design, interest focuses on describing the mean expression level (over replicates) for each condition. If we wish to cluster the genes based on their expression levels across experiments, then the repeated measures structure can help us to estimate the within-cluster expression means per condition and describe the covariances between repeated measures by a reduced set of parameters.

Finally, structured data arise when the variables can be assumed to follow a known *factor model*. Such models assume that the observed or *manifest* variables are *indicators* of a relatively small number of *latent variables* or *factors*, and it is the relationships of observed to latent variables that account for the observed correlations between the manifest variables. Under a so-called *confirmatory factor analysis model*, each indicator variable or item can be allocated to one of a set of the underlying factors or concepts which cannot be observed directly. Many questionnaires employed in the behavioural and social sciences produce multivariate data of this type. For example, recall the Hospital Anxiety and Depression Scale (HADS; Zigmond and Snaith, 1983) which assesses patients on 14 items (each scored on a four-point Likert scale), with half the items targeting the unobserved concept of 'depression' and the remainder targeting 'anxiety'. Data generated under such models are structured, as all items are measured on the same scale, and the structure can be identified by a reference vector whose entries code the factors that have been targeted. Due to the known factor structure, questionnaire data is typically summarized by means or totals over items in subscales that target a specific factor. As we shall see, were we to cluster subjects based on their questionnaire item scores, then the categorical factor reference variable can be used in the same way as a categorical condition reference variable

to estimate the within-cluster factor means and describe the covariances between items by a reduced set of parameters.

The modelling described in this chapter represents a more formal statistical alternative to the two-stage approach for clustering structured data alluded to in Section 3.5. Then, we sought to exploit the structural information to define suitable summary measures of an object's variable profile in a first stage which effectively reduced the number of variables in the data matrix that is subject to clustering. To cluster objects, a suitable method would then be applied to the matrix of summary measures at a separate second stage. The data matrix used in the second stage is devoid of structure, and so any suitable method for hierarchical or optimization clustering, including finite mixture modelling as discussed in Chapters 4, 5 and 6, may be used at this stage. In contrast, the approach detailed in this chapter is a single-stage modelling approach where the process that generates the original data is modelled explicitly and the structural information is used to provide efficient estimates of parameters describing object profiles within clusters.

In the next section we outline the general model-based clustering procedure that will be central to the rest of the chapter, before looking in more detail at how the model can be applied to different types of structured data, in particular longitudinal data and data generated by a factor model. In-depth descriptions of a variety of applications of model-based clustering to structured data appear towards the end of the chapter.

7.2 Finite mixture models for structured data

We begin by assuming the same finite mixture model as in the previous chapter; that is

$$f(\mathbf{x}; \mathbf{p}, \boldsymbol{\theta}) = \sum_{j=1}^{c} p_j g_j(\mathbf{x}; \boldsymbol{\theta}_j), \tag{7.1}$$

where $\mathbf{x}$ is a vector containing an object's p variable values, $\mathbf{p}' = (p_1, \ldots, p_{c-1})$, the probabilities of belonging to subpopulation or cluster $j = 1, \ldots, c$ ($\sum_{j=1}^{c} p_j = 1$), and g_j the cluster probability distributions parameterized by vectors $\boldsymbol{\theta}_j$. Since our focus is on variables that have been measured on the same scale, we now not only assume that the multivariate distributions from the different clusters belong to the same family but also that each univariate distribution belongs to the same family. Typical choices of families are Gaussian or t-distributions (continuous variables), Poisson distribution (count variables) or binomial distributions (binary variables).

Due to the structured nature of the data, is it often possible to model the mean or covariance matrix of an object's variable vector $\mathbf{x}$ by fewer parameters than used previously for the general finite mixture model as described in Chapter 6. The mean or the covariance matrix may be expressed as functions of sets of parameters and

the reference vector $\mathbf{r}$ (e.g. time, experimental condition, factor), whose entries code the nature of the kth variable column in the data matrix $\mathbf{X}$. Letting $\boldsymbol{\beta}_j$ denote a vector containing the parameters needed to describe the mean vector for an object in the jth cluster, and $\boldsymbol{\gamma}_j$ a vector containing the parameters needed to describe the cluster's covariance matrix, and setting $\boldsymbol{\theta}_j' = (\boldsymbol{\beta}_j', \boldsymbol{\gamma}_j')$, we can write

$$\mathrm{E}(\mathbf{x}) = h(\boldsymbol{\beta}_j, \mathbf{r}) \tag{7.2}$$

and

$$\mathrm{Cov}(\mathbf{x}) = q(\boldsymbol{\gamma}_j, \mathbf{r}), \tag{7.3}$$

where h and q are functions of the reference vector and the respective parameters. For example, in the case of longitudinal data $\mathbf{x}' = (x_1, x_2, x_3, x_4)$ observed at four equally spaced time points, so that the reference vector is $\mathbf{r}' = (1,\ 2,\ 3,\ 4)$, were we to assume that variable means changed linearly over time, that variances remained constant over time and that the covariances between any two time points were constant across time points, we would model the mean vector and the covariance matrix as follows:

$$\mathrm{E}(\mathbf{x}) = \begin{pmatrix} \beta_{1j} + 1\beta_{2j} \\ \beta_{1j} + 2\beta_{2j} \\ \beta_{1j} + 3\beta_{2j} \\ \beta_{1j} + 4\beta_{2j} \end{pmatrix} = \beta_{1j} + \beta_{2j}\mathbf{r}. \tag{7.4}$$

and

$$\mathrm{Cov}(\mathbf{x}) = \begin{pmatrix} \gamma_{1j} & \gamma_{2j} & \gamma_{2j} & \gamma_{2j} \\ \gamma_{2j} & \gamma_{1j} & \gamma_{2j} & \gamma_{2j} \\ \gamma_{2j} & \gamma_{2j} & \gamma_{1j} & \gamma_{2j} \\ \gamma_{2j} & \gamma_{2j} & \gamma_{2j} & \gamma_{1j} \end{pmatrix} = \gamma_{1j}\mathbf{I}_4 + \gamma_{2j}(\mathbf{L}_4 - \mathbf{I}_4), \tag{7.5}$$

where $\mathbf{I}_4$ is a 4×4 identity matrix and $\mathbf{L}_4$ is a 4×4 matrix, the elements of which are all ones. So here the function h is a linear function of time, and $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j})'$ contains the two parameters required to describe the mean vector of the four observed variables in cluster j. The simple *compound symmetry* covariance matrix structure assumed here does not depend on the reference vector and only requires a single variance and a single covariance parameter to describe it, so that $\boldsymbol{\gamma}_j = (\gamma_{1j}, \gamma_{2j})'$.

Using the known structure of the data often considerably reduces the number of parameters needed to model means and covariance matrices compared to the unstructured model (which requires as many parameters as number of variables, p, to model the mean vector, and $p(p+1)/2$ parameters to model the covariance

matrix). For example, under the assumptions of linear trend over four time points and a compound symmetry covariance structure, we require only four parameters to model the mean vector and the covariance matrix of each cluster, as opposed to $4 + 4 \times 5/2 = 14$ parameters under an unstructured model. The parameter savings per cluster can be substantial, especially when the data contain a large number of variables.

Except for using a reduced set of parameters to describe the subpopulation means and covariance matrices, finite mixture model-based clustering proceeds in the same manner as described for unstructured data in Chapter 6, and for a known number of clusters the parameters of the relevant finite mixture model can be estimated by maximum likelihood. If necessary the number of clusters can be determined empirically by comparing competing models using likelihood or information theoretic methods (see Section 6.5). Having estimated the parameters, the maximum (estimated) *a posteriori* probability (MAP) is typically used to allocate objects to clusters.

7.3 Finite mixtures of factor models

In this section we will consider finite mixture models for cluster analysis where the multivariate observations within each cluster can realistically be assumed to have a factor analysis (FA) structure. Factor analysis has a long history, having first been suggested by Spearman as early as 1904. The method is commonly applied when q ($q \ll p$) latent or unobservable variables, often referred to as factors, are thought to underlie and be responsible for the covariances of the p manifest or measured variables. Factor analysis is popular in the behavioural and social sciences, where concepts such as 'cognition' and 'functioning' cannot be assessed directly and are instead measured by a number of items on a standardized scale whose expected average value is the 'truth', but which are also subject to measurement error.

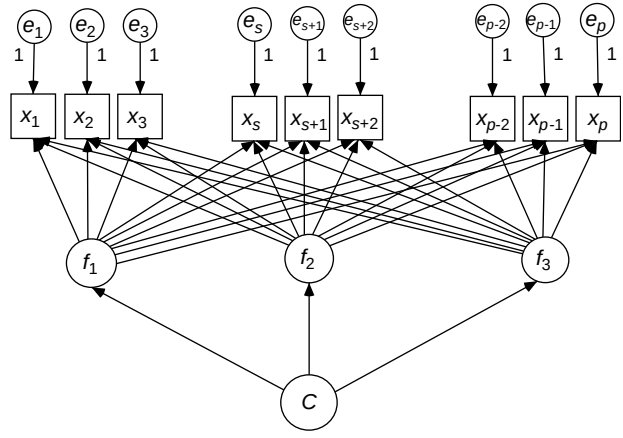
The classical factor analysis model states that a p-dimensional observation $\mathbf{x}$ from cluster j arises from the model

$$\mathbf{x} = \boldsymbol{\mu}_j + \boldsymbol{\Lambda}_j \mathbf{f} + \boldsymbol{\varepsilon}_j \tag{7.6}$$

with $\boldsymbol{\Lambda}_j$ a $p \times q$ matrix of so-called *factor loadings*. The q factors in $\mathbf{f}$ are independent of each other and follow standard normal distributions: $\mathbf{f} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_q)$. The p variables in $\boldsymbol{\varepsilon}_j$ are independent noise terms and follow normal distributions: $\boldsymbol{\varepsilon}_j \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_j)$, where $\boldsymbol{\psi}_j = \mathrm{diag}(\psi_{1j}, \ldots, \psi_{pj})$. The term ψ_{ij} is known as the *unique variance* of the ith observed variable. The factors and the noise terms are assumed independent of each other. (Model-based principal component analysis, PCA, is a special case of the factor model in which the noise terms are assumed to have the same distribution across variables, so that $\boldsymbol{\psi}_j = \psi_j \mathbf{I}_p$.)

From the classical factor analysis model it follows that the marginal distribution of an observation from cluster j is multivariate normal with the specific

(a) Mixture of exploratory factor models



(b) Mixture of confirmatory factor models

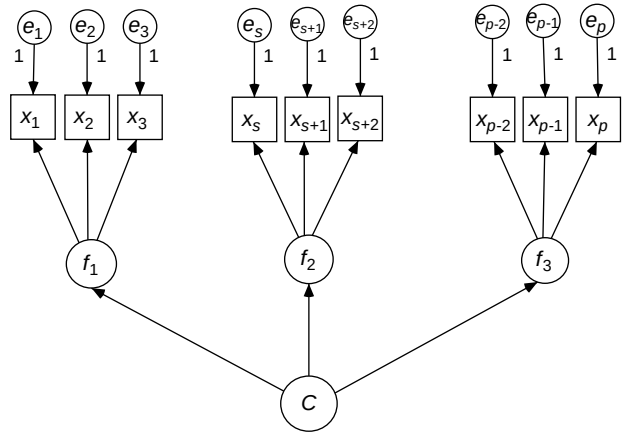


Figure 7.2 *Illustration of finite mixture of factor models. (a) Mixture of exploratory factor models; (b) mixture of confirmatory factor models. The f indicate latent factors with unit variance conditional on cluster, and zero mean in a reference cluster. Cluster membership is allowed to affect all free parameters in models (a) and (b); that is, means of the observed variables (x), factor loadings and variances of the residual terms (e).*

parameterization $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}'_j + \boldsymbol{\psi}_j)$. Figure 7.2(a) provides a graphical illustration of the classical factor analysis model (ignoring the cluster variable C for now). The figure illustrates how a larger number of manifest variables (the x_k) is driven by a smaller number of (latent) factors (the f_l), with the factor loadings and the unique variances determining the relative contributions of the underlying latent concepts.

Before extending the factor analysis model to allow for distinct subpopulations and defining an appropriate finite mixture density, we need to distinguish between the two major uses of factor analysis models. *Exploratory factor analysis* (EFA) is used to uncover the underlying structure of the relatively large set of manifest variables. The number of factors is not necessarily known. Any factor may affect a manifest variable and no restrictions are therefore placed on the factor loadings (Figure 7.2(a)). In contrast, *confirmatory factor analysis* (CFA) starts with the premise that each factor (the number of which is known, perhaps from a preliminary EFA) is associated with a specific set of indicator variables, for example a set of items from a questionnaire (Figure 7.2 (b)). A number of methods have been suggested for estimating parameters modelling the way in which latent variables affect manifest variables (here termed the 'factor structure'), namely the factor loadings and unique variances, including maximum likelihood estimation. For more details on fitting factor models see for example Bollen (1989) or Loehlin (2004).

Due to the constraints imposed by a confirmatory factor analysis model, it is possible to model the mean and covariance matrix of the observed variables by relatively few parameters. The sets of indicator variables can be referenced by a p-dimensional categorical variable $\mathbf{r} = (1, \ldots, 1, 2, \ldots, 2, \ldots, q, \ldots, q)'$ whose levels identify the underlying concepts, or alternatively by q corresponding p-dimensional dummy vectors assembled in the matrix $\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_q)$. We may restrict the means and the measurement error variances to be constant within levels of $\mathbf{r}$; that is:

$$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}_j = \mathbf{D}(\beta_{1j}, \ldots, \beta_{qj})' \tag{7.7}$$

and

$$\boldsymbol{\psi}_j = \mathrm{diag}[\mathbf{D}(\tau_{1j}, \ldots, \tau_{qj})']. \tag{7.8}$$

A confirmatory factor model further stipulates that entries in a row of the loadings matrix $\boldsymbol{\Lambda}_j$ are set to zero except for the single column that corresponds to the level in the respective row of $\mathbf{r}$. (Or in other words paths are deleted in Figure 7.2(b) so that each manifest variable is only associated with a single factor.) Consider the case of four indicator variables with the first two variables loading on factor 1 and the third and fourth on factor 2, with the factors being statistically independent. Then

$$\mathrm{E}(\mathbf{x}) = \begin{pmatrix} \beta_{1j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{2j} \end{pmatrix}, \ \boldsymbol{\psi}_j = \mathrm{diag}\left[\begin{pmatrix} \tau_{1j} \\ \tau_{1j} \\ \tau_{2j} \\ \tau_{2j} \end{pmatrix}\right], \ \boldsymbol{\Lambda}_j = \begin{pmatrix} \lambda_{1j} & 0 \\ \lambda_{2j} & 0 \\ 0 & \lambda_{3j} \\ 0 & \lambda_{4j} \end{pmatrix} \text{ and}$$

$$\mathrm{Cov}(\mathbf{x}) = \begin{pmatrix} \lambda_{1j}^2 + \tau_{1j} & \lambda_{1j}\lambda_{2j} & 0 & 0 \\ \lambda_{1j}\lambda_{2j} & \lambda_{2j}^2 + \tau_{1j} & 0 & 0 \\ 0 & 0 & \lambda_{3j}^2 + \tau_{2j} & \lambda_{3j}\lambda_{4j} \\ 0 & 0 & \lambda_{3j}\lambda_{4j} & \lambda_{4j}^2 + \tau_{2j} \end{pmatrix}.$$

Thus here $q < p$ parameters are used to model the mean, and $p + q$ parameters to model the covariance matrix. Due to the reduced number of parameters, it is now feasible to introduce further parameters and allow the factors to be correlated with each other (i.e. introduce double-headed arrows between the factors in Figure 7.2(b)). The purpose of classical CFA then is to determine whether such a restricted model is supported by sample data.

To introduce the factor analysis model into the finite mixture model for cluster analysis giving what has been termed a *mixture of factor analysers model*, we let the parameters of the factor model vary with cluster membership. This is demonstrated in Figure 7.2 by the inclusion of the latent cluster variable C. When the (manifest) variables can be explained by a smaller set of underlying factors, cluster membership can be modelled as driving the underlying factors and the independent noise terms rather than the manifest variables directly (as done for unstructured data, cf. Figure 7.1).

In the most general mixture of factor analysers model, all model parameters may vary with cluster; in other words both the parameters needed to model the mean vector *and* the parameters modelling the factor structure are allowed to vary with cluster. The number of parameters needed in a mixture of EFA models is reduced compared to the unstructured finite mixture model when the number of factors (q) is much smaller than the number of manifest variables (p). The number of parameters needed in a mixture of CFA models tends to be markedly reduced compared to the unstructured case because the loadings matrices are constrained. However, in particular in an EFA context we may wish to consider further parameter savings by using yet more parsimonious parameterizations. Due to the parameterization employed in a factor analysis model, the following options suggest themselves for restricting the covariance matrix:

(i) restrain both factor loadings and unique variances to be the same across clusters;
(ii) vary only the unique variances while constraining the factor loadings to be the same;
(iii) vary only the factor loadings while the unique variances remain constant (McLachlan *et al*., 2003);
(iv) vary both aspects (McLachlan and Peel, 2000; McLachlan *et al*., 2003).

If we further consider restricting the error variances to be constant across variables as in model-based PCA, then there are, altogether, eight covariance matrix parameterizations (including the Tipping and Bishop (1999) model) which

Table 7.1 Parsimonious covariance structures suitable for mixtures of factor analysers models.

Model name	Loading matrix	Error variance	PCA or FA model?	Resulting covariance matrix	Number of covariance parameters
CCC	Constrained	Constrained	Constrained	$\mathbf{\Lambda\Lambda}' + \boldsymbol{\psi}\mathbf{I}_p$	$[pq - q(q-1)/2] + 1$
CCU	Constrained	Constrained	Unconstrained	$\mathbf{\Lambda\Lambda}' + \boldsymbol{\psi}$	$[pq - q(q-1)/2] + p$
CUC	Constrained	Unconstrained	Constrained	$\mathbf{\Lambda\Lambda}' + \psi_j\mathbf{I}_p$	$[pq - q(q-1)/2] + c$
CUU	Constrained	Unconstrained	Unconstrained	$\mathbf{\Lambda\Lambda}' + \boldsymbol{\psi}_j$	$[pq - q(q-1)/2] + cp$
UCC	Unconstrained	Constrained	Constrained	$\mathbf{\Lambda}_j\mathbf{\Lambda}_j' + \boldsymbol{\psi}\mathbf{I}_p$	$c[pq - q(q-1)/2] + 1$
UCU	Unconstrained	Constrained	Unconstrained	$\mathbf{\Lambda}_j\mathbf{\Lambda}_j' + \boldsymbol{\psi}$	$c[pq - q(q-1)/2] + p$
UUC	Unconstrained	Unconstrained	Constrained	$\mathbf{\Lambda}_j\mathbf{\Lambda}_j' + \psi_j\mathbf{I}_p$	$c[pq - q(q-1)/2] + c$
UUU	Unconstrained	Unconstrained	Unconstrained	$\mathbf{\Lambda}_j\mathbf{\Lambda}_j' + \boldsymbol{\psi}_j$	$c[pq - q(q-1)/2] + cp$

are summarized in Table 7.1. McNicholas and Murphy (2008) refer to these as a 'class of parsimonious Gaussian mixture models', since constraining covariance matrices presents an approach for arriving at finite mixture models with fewer parameters. Clearly the mixture models that assume a constant factor structure across clusters (CCC and CCU in Table 7.1) present the most parsimonious options, and should be employed when there is sufficient evidence to support such an assumption, for example when a well-validated questionnaire is being used. However, when there is doubt as to how latent variables affect manifest variables in different clusters, cluster-varying covariance matrices should be considered, and comparing the fit of the eight models in the McNicholas and Murphy (2008) class provides a way of arriving at the most parsimonious model without losing explanatory power. Empirical selection of both the number of factors and the covariance constraints in exploratory factor mixture modelling then provides an alternative model-based approach to the parameter reduction techniques for dealing with high-dimensional data mentioned in the last chapter (see Sections 6.2.2 and 6.6). (Mixtures of EFA models also open up the possibility of identifying variables that are irrelevant to the clustering by shrinking factor loadings; for more details see Galimberti *et al*., 2009.)

Exploratory factor mixture modelling has been found to perform well in terms of identifying known classes in a number of real data sets (McNicholas and Murphy, 2008; McNicholas *et al*., 2010; see later Section 7.5.3) and also in terms of identifying the correct number of clusters and factors in a simulation study assuming 'moderate' class separation (Lubke and Neale, 2006). Mixtures of exploratory factor models have received increasing attention in the analysis of microarray data, where they can be applied to simultaneously cluster individuals or cells as well as group large numbers of genes according to their function; see McLachlan *et al*. (2003). For more on confirmatory factor mixture models see Yung (1997), Jedidi *et al*. (1997), Dolan and Van der Maas (1998) and Arminger *et al*. (1999).

While most of the literature concentrates on mixtures of factor models assuming multivariate normality within clusters, other distributions can also be handled. Ordered categorical variables can be modelled by assuming that there are underlying continuous latent traits (or liabilities) which are converted into sets of binary variables by imposing threshold structures on the continuous scales. Depending on the distribution of the latent traits, response probabilities are modelled by logistic or probit link functions; see for example Muthén and Asparouhov (2006) or Lubke and Neale (2008). Applying mixtures of factor analysers modelling to binary data is sometimes referred to as *mixture latent trait modelling* or *mixture item response theory modelling* (mixture IRT modelling, in particular for models containing a single factor; see for example Maij-deMeij *et al*., 2008). Lubke and Neale (2008) present results of a simulation study for binary and ordinal variables, which shows that it is possible to discriminate between latent classes and continuous factors even if the manifest data are not multivariate normal.

The classical factor model represents a particular type of *structural equation model* (SEM), and hence finite mixtures of factor models are included under the umbrella term *structural equation mixture models* or SEMMs. (For more on this framework see Jedidi *et al*., 1997; Arminger and Stein, 1997; Dolan and Van der Maas, 1998; Muthén, 2002 and Bauer and Curran, 2004.) We have already met another finite mixture model that uses structural equations to describe within-cluster covariance matrices, namely the finite mixtures regression model in Section 6.7. The difference between this SEMM and the ones looked at in this chapter is that the latter invoke latent variables to describe the within-cluster relationships between variables. Such models can be fitted by maximum likelihood using specialist software for structural equation modelling which allows for categorical latent variables, for example `Mplus` (Muthén and Muthén, 2007), `Latent GOLD` (Vermunt and Magidson, 2000) or `gllamm` (a user-contributed program for `Stata` by Rabe-Hesketh *et al*., 2004), by using specially developed fitting programs for mixtures of factor analysers such as `EMMIX` (McLachlan *et al*., 2002, 2003; Ng *et al*., 2006) or by implementing an extension of the EM algorithm (see Section 6.2), a so-called alternating expectation–conditional maximization (AECM) algorithm (Meng and vanDyk, 1997) detailed by McNicholas and Murphy (2008) and McNicholas *et al*. (2010).

7.4 Finite mixtures of longitudinal models

In this section we will consider applying finite mixture models to the clustering of longitudinal data so that the observed variables arise from some particular model for such data, and where the parameters of this model may differ between clusters. There is a large literature on modelling longitudinal data – for example see Verbeke and Mohlenberghs (2000); Diggle *et al*. (2002); Everitt and Pickles (2004); Fitzmaurice *et al*. (2004) – but here we will concentrate on two commonly used classes of models for such data – *linear growth curve models* (e.g. Pan and Fang, 2002; Bollen and Curran, 2006) and *autoregressive models*

(e.g. Zimmerman and Nunez-Anton, 2010) – and extend these to become the basis for clustering for this type of data. We further focus on multivariate normal variables and only briefly mention extensions to distributions suitable for non-continuous data.

Longitudinal data are characterized by the same variable being recorded at multiple time points, and for simplicity in the following account we will assume that each object's observations are taken at the same time points, although this is not strictly necessary. Associated with an object's multivariate observation is therefore a p-dimensional reference variable $\mathbf{r} = (t_1, t_2, \ldots, t_p)'$, where t_k, $k = 1, \ldots, p$ is the time point at which the kth variable is recorded. When variables are recorded at equidistant time intervals the reference variable simply indicates serial order, that is, $\mathbf{r} = (1, \ldots, p)'$.

The classical linear growth curve model for longitudinal data states that a p-dimensional observation $\mathbf{x}$ from cluster j arises from the model

$$\mathbf{x} = \mathbf{\Lambda}_j \mathbf{u} + \boldsymbol{\varepsilon}_j, \tag{7.9}$$

where $\mathbf{u} = (u_1, u_2, \ldots, u_g)'$ is a vector of g latent variables that are assumed to arise from a multivariate normal distribution with mean vector $\boldsymbol{\beta}_j$ and covariance matrix $\mathbf{\Omega}_j$; that is, $\mathbf{u} \sim \mathrm{N}(\boldsymbol{\beta}_j, \mathbf{\Omega}_j)$. The values of latent variables in $\mathbf{u}$ represent object-varying random coefficients; u_1 describes the intercept of the object's growth trajectory over time and u_2 its slope, with the remaining latent variables extending the growth model to a gth-order polynomial. In practice, using growth curve models with g greater than three is rare. The random coefficients in $\mathbf{u}$ are often referred to as *growth factors*. The p variables in $\boldsymbol{\varepsilon}_j$ are mutually independent noise terms that follow normal distributions; that is, $\boldsymbol{\varepsilon}_j \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi}_j)$, where $\boldsymbol{\psi}_j = \mathrm{diag}(\psi_{1j}, \ldots, \psi_{pj})$. The growth factors and the noise terms are assumed independent of each other. The entries of the $p \times g$ matrix $\mathbf{\Lambda}_j$ determine the functional form of the time trend in the mean and are typically fixed. For example, if we wished to model a linear time trend in all clusters we could set $\mathbf{\Lambda}_j = (\mathbf{1}_p, \mathbf{r})$ so that the mean of the kth variable at time point t_k in cluster j becomes $\mathrm{E}(x_{t_k}) = [1, t_k]\boldsymbol{\beta}_j = \beta_{1j} + \beta_{2j} t_k$. It is possible to leave some of the entries in $\mathbf{\Lambda}_j$ unspecified and, instead of prescribing the shape of the growth curve, let the shape be determined from the data by estimating these parameters. Note that under this model both the mean $\mathrm{E}(\mathbf{x}) = \mathbf{\Lambda}_j \boldsymbol{\beta}_j$ and the covariance matrix $\mathrm{Cov}(\mathbf{x}) = \mathbf{\Lambda}_j \mathbf{\Omega}_j \mathbf{\Lambda}_j' + \boldsymbol{\psi}_j$ are functions of sets of parameters and the reference vector $\mathbf{r}$. Figure 7.3 provides a graphical illustration of the classical linear growth curve model (ignoring the cluster variable C for now). The figure illustrates how time-ordered manifest variables (the x_k) are driven by a few latent variables (the u_l) with the effects of the growth factors only partly unspecified.

To introduce the linear growth curve model into a finite mixture model for the clustering of growth curve data, we allow growth structure variability within each cluster and let the free mean and covariance parameters of the growth curve model vary with cluster membership. The result is what is often referred to as a *growth mixture model* (GMM; Muthén and Shedden, 1999). Such a model is illustrated in Figure 7.3 by the inclusion of the latent cluster variable C. When the (manifest)

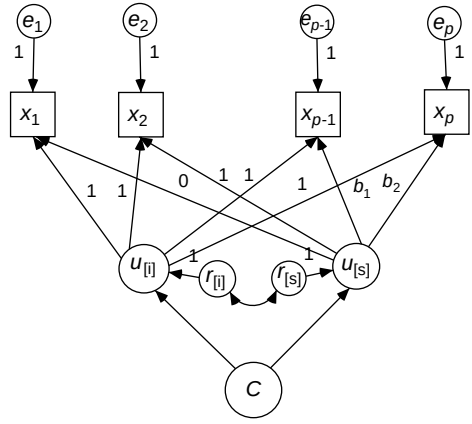


Figure 7.3 *Illustration of growth mixture model with linear growth trajectories. u indicate latent growth factors with [i] denoting the intercept and [s] the slope of a linear trajectory. r[i] and r[s] are residual terms with zero means accounting for within-cluster variability of the growth factors. b_1 and b_2 indicate free parameters. Cluster membership is allowed to affect all free parameters; that is, means of the growth factors, their within-cluster covariance matrix, any free factor loadings and variances of the residual terms (e).*

variables can be explained by a small set of growth factors, cluster membership can be modelled as driving these and the noise terms. (Growth mixture models are related to mixtures of linear mixed effects or hierarchical models for longitudinal data; for more see, e.g., De la Cruz-Mesia *et al.*, 2008.)

In contrast to growth models, autoregressive models explain individual trajectories by regressing a variable on itself at an earlier time point. The basic autoregressive model for longitudinal data states that an observation from cluster j arises from the model

$$\mathbf{x} = \mathbf{H}_j\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{7.10}$$

where $\mathbf{H}_j$ is a $(p-1) \times 2(p-1)$ matrix defined as $\mathbf{H}_j = [\mathbf{I}_{p-1}, \text{diag}(x_1, \ldots, x_{p-1})]$ for multivariate observations $\mathbf{x} = (x_2, \ldots, x_{p-1}, x_p)'$ ordered according to the time reference vector $\mathbf{r}$. $\boldsymbol{\beta}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\theta}_j)'$ is a parameter vector of length $2(p-1)$ containing intercept parameters in $\boldsymbol{\alpha}_j$ and so-called *autoregressive parameters* in $\boldsymbol{\theta}_j$. The $p-1$ variables in $\boldsymbol{\varepsilon}_j$ are mutually independent noise terms that follow normal distributions, $\boldsymbol{\varepsilon}_j \sim \text{N}(\mathbf{0}, \boldsymbol{\psi}_j)$, where $\boldsymbol{\psi}_j = \text{diag}(\psi_{2j}, \ldots, \psi_{pj})$. The noise terms are further assumed independent of the observations at earlier time points. The observation at the first time point (x_1) is treated as predetermined. From this follows that $\text{E}(\mathbf{x})$ is a function of all the parameters in $\boldsymbol{\beta}_j$ (and the reference vector $\mathbf{r}$), while the structure of $\text{Cov}(\mathbf{x})$ is described by the autoregressive parameters and the variances of the noise terms. Figure 7.4 provides a graphical illustration of the autoregressive model (ignoring the cluster variable C for now).

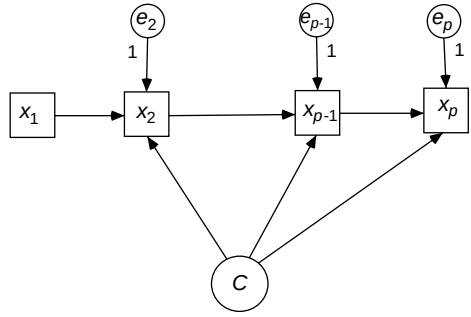


Figure 7.4 *Illustration of mixtures of first order autoregressive models. Values of the manifest variable at the previous time point are allowed to have a direct effect on the variable (x). Cluster membership is allowed to affect all free parameters; that is, intercepts and autoregressive parameters affecting x at all but the first time point, and variances of the residual terms (e).*

The figure illustrates how time-ordered manifest variables (the x_k) are driven by earlier values of the same variables. (Note that we have specified the commonly used autoregressive model of order 1 or AR(1) model; models can be extended to drive observed variables by several earlier values and are then referred to as AR(q) models.)

To extend the simple AR(1) model into the finite mixture model for cluster analysis we again allow for growth structure variability within each cluster and let all the model parameters vary with clusters. (A commonly used restriction to reduce the number of parameters for equidistant time points is the assumption that autoregressive parameters within a cluster are constant over time.) The AR mixture model is illustrated in Figure 7.4 by the inclusion of the latent cluster variable C. It is possible to define a general class of *autoregressive latent trajectory models* or ALT models for longitudinal data, which include autoregressive and linear growth models as special cases; see Bollen and Curran (2004). However, this generalization increases the number of parameters needed to specify the structural component and may require restrictions to ensure identifiability, so is not considered at this point.

While most of the longitudinal literature concentrates on models assuming multivariate normality within clusters, binary or ordinal data can be handled by assuming that there are underlying continuous latent traits which are converted into a set of binary variables by imposing a threshold structure on the continuous scale; for a general framework see Muthén and Shedden (1999) and Muthén (2002). However, typically so-called *latent class growth analysis* or LCGA (Nagin, 1999) is applied to binary longitudinal data which restrains the within-cluster variances and covariances of the growth factors to zero (and hence enforces the conditional independence assumption used in classical latent class analysis). Examples of such group-based trajectory modelling with binary data can be found in Nagin and Tremblay (1999), Croudace *et al.* (2003) and Kuss *et al.* (2006).

Growth mixture models and mixtures of autoregressive models are further examples of SEMMs and, in principle, that is assuming sufficient sample sizes and identifiable model parameters (a GMM with as many growth factors as time points cannot be identified), can be fitted using SEM packages provided the package allows for continuous and categorical latent variables. Current GMM fitting practice largely follows guidance provided by Muthén and Muthén (2000) using `Mplus`. Typical reports use the full range of BIC, AIC, entropy and the Lo–Mendell–Rubin likelihood ratio test (LMR LRT, Lo *et al.*, 2001; see also Section 6.5) to choose the optimal combination of number of clusters and number of growth factors. Once a model has been selected, objects can be allocated to clusters using the MAP. Clusters are best described by their mean trajectory curves over time. In addition, empirical Bayes predictions of individual growth trajectories can be obtained, and overlaying these in trajectory graphs may provide an impression of within-cluster heterogeneity.

All the above is based on the premise that the investigator is able to specify a model for the longitudinal data within a cluster, with the choice of longitudinal model made on substantive grounds. McNicholas and Murphy (2010) pursue an altogether different approach to model-based clustering of longitudinal data. They start with the premise that, while they wish to take account of the longitudinal nature of the data in the mixture modelling, they do not wish to pre-specify the parametric form of the covariance matrices. To achieve their aim they suggest fitting a series of more or less parsimonious models based on covariance structures that account for the relationships between variables at different time points. Specifically, under a multivariate normal model the covariance matrix of the jth cluster is expressed using the Cholesky decomposition

$$\text{Cov}(\mathbf{x}) = \mathbf{T}_j \mathbf{\Delta}_j \mathbf{T}_j', \tag{7.11}$$

where $\mathbf{\Delta}_j$ is a unique $p \times p$ diagonal matrix with strictly positive diagonal entries and $\mathbf{T}_j$ a unique lower triangular matrix with diagonal elements of ones. This decomposition is helpful here, since the values of $\mathbf{T}_j$ and $\mathbf{\Delta}_j$ have interpretations of generalized autoregressive parameters and noise variances, respectively (Pourahmadi, 1999). This decomposition then provides the option of constraining the $\mathbf{T}_j$ or the $\mathbf{\Delta}_j$ or both to be equal across clusters. A further possible restriction is to enforce constant noise variances across variables: $\mathbf{\Delta}_j = \delta_j \mathbf{I}_p$. The resulting family of eight covariance parameterizations suitable for longitudinal data is described in Table 7.2. Note that the nomenclature is intended to be consistent with that used by the MCLUST family (Fraley and Raftery, 1999, 2003; see also Section 6.2.2). McNicholas and Murphy (2010) suggest that this family can be extended by including yet more parsimonious covariance structures in which certain subdiagonals of $\mathbf{T}_j$ are constrained to be 0.

Note that in this section we have been concerned with modelling the within-cluster structure of repeated measures of a single variable. Extensions of factor analysis models, linear growth models and autoregressive models to simultaneously model repeated measures of several variables are described in, for example,

Table 7.2 Parsimonious covariance structures suitable for mixture modelling of longitudinal data.

Model name	Lower triangular matrix	Diagonal matrix	Noise variances restrained to be constant?	Resulting covariance matrix	Number of covariance parameters
EEI	Constrained	Constrained	Constrained	$\delta\mathbf{T}\mathbf{T}'$	$p(p-1)/2+1$
EEA	Constrained	Constrained	Unconstrained	$\mathbf{T}\boldsymbol{\Delta}\mathbf{T}'$	$p(p-1)/2+p$
EVI	Constrained	Unconstrained	Constrained	$\delta_j\mathbf{T}\mathbf{T}'$	$p(p-1)/2+c$
EVA	Constrained	Unconstrained	Unconstrained	$\mathbf{T}\boldsymbol{\Delta}_j\mathbf{T}'$	$p(p-1)/2+cp$
VEI	Unconstrained	Constrained	Constrained	$\delta\mathbf{T}_j\mathbf{T}_j'$	$c[p(p-1)/2]+1$
VEA	Unconstrained	Constrained	Unconstrained	$\mathbf{T}_j\boldsymbol{\Delta}\mathbf{T}_j'$	$c[p(p-1)/2]+p$
VVI	Unconstrained	Unconstrained	Constrained	$\delta_j\mathbf{T}_j\mathbf{T}_j'$	$c[p(p-1)/2]+c$
VVA	Unconstrained	Unconstrained	Unconstrained	$\mathbf{T}_j\boldsymbol{\Delta}_j\mathbf{T}_j'$	$c[p(p-1)/2]+cp$

Muthén (2002), Curran and Hussong (2003), Pickles and Croudace (2010), Beath and Heller (2009) and Villarroel *et al.* (2009). Extensions to the multivariate setting either assume appropriate SEMs for each repeated univariate variable and then relate the processes solely at the level of the factors driving these SEMs (for examples see Hix-Small *et al.* (2004), Elliott (2007) or Putter *et al.* (2008)), or model the multivariate variable for each repeated measure and then formulate an appropriate SEM for the factors driving these models. Whichever approach is followed, latent classes are then introduced as driving these combined models and potentially affecting all resulting parameters.

7.5 Applications of finite mixture models for structured data

In this section we provide detailed examples of how finite mixture models are used in practice for clustering data having one or other of the types of structure described above. We present three quite different applications of finite mixture models for factor analysis structured data, and two applications of finite mixture models to data with growth curve structure.

7.5.1 Application of finite mixture factor analysis to the 'categorical versus dimensional representation' debate

Recent years have seen what might be termed an 'explosion' of applications of mixtures of factor analyser models aimed at addressing the question 'Is a mental health disorder best conceptualized as a distinct categorical subtype or as an extreme of a continuous trait?' Examples investigating achievement data, substance use disorders, tobacco dependency, attention deficit/hyperactivity disorder

(ADHD) and psychosis can be found in Lubke and Muthén (2005), Muthén (2006), Muthén and Asparouhov (2006), Lubke *et al*. (2007) and Shevlin *et al*. (2007), respectively.

Here we describe a study by Lubke *et al*. (2009) which applies factor mixture models to add to the ongoing debate of whether ADHD should be conceptualized as a categorical disorder or as an extreme of a continuous trait. Current diagnostic rules, as defined in DSM-IV-TR (American Psychiatric Association, 2000), categorise patients as 'no ADHD', 'combined ADHD type', 'predominantly inattentive type' or 'predominantly hyperactive/impulsive type'. The authors studied Dutch male twins whose parents voluntarily registered with the Netherlands Twin Registry. Mothers rated their sons on the Child Behaviour Checklist (CBCL) which consists of more than 100 items each rating behaviours on a three-point Likert scale (0 = behaviour not true, 1 = sometimes true, 2 = often true). Ratings were obtained at age 7 years (wave 1, $n = 8079$ children), 10 years (wave 2, $n = 5278$) and 12 years (wave 3, $n = 3139$). The study focused on 11 items from the attention problem (AP) syndrome scale of the CBCL and analysed data from each of the three waves using mixtures of multivariate normal densities.

The confirmatory factor model hypothesized is illustrated in Figure 7.5. The first factor is largely defined by symptoms relating to hyperactivity/impulsivity, the second factor affects indicators of inattentiveness/dreaminess and the third factor explains two items related to nervous behaviour. The factors are allowed to correlate. For each wave, seven mixture models were fitted: models 1, 2 and 3 assume that a multivariate observation within a cluster arises from a three-factor model with paths stipulated by Figure 7.5 and consider two-, three- and four-cluster solutions respectively. In contrast, models 4, 5, 6 and 7 assume uncorrelated items within clusters (a type of multivariate normal mixture modelling sometimes referred to as *latent profile analysis*) and estimate similar numbers of parameters by considering an increased number of clusters, namely three, four, five and six clusters, respectively. All respective mean vectors and covariance matrices are allowed to vary with cluster. The choice of models in the first set (models 1–3) versus the second (4–7) is motivated by the presence of factors suggesting sample correlations between variables partly due to underlying continuous scales (the continuous trait view), while the latter models are consistent with the observed sample correlations wholly due to partitioning of the population into distinct groupings (the categorical disorder view; see also Lubke and Muthén (2005), Lubke and Neale (2006)). Model fitting was carried out using maximum likelihood in `Mplus`. Data from all twins were used, but a 'sandwich-type' estimator which is robust against non-normality and nonindependence of observations (option MLR in `Mplus`) was employed to obtain standard errors. Model comparisons were performed based on BIC, amongst other model fit indices.

For each of the waves, comparisons across models showed that factor mixture models fitted the data better (BIC lower) than finite mixture models with uncorrelated observations within clusters (BIC higher). At the first wave (children aged 7 years), BIC supports the three-factors and three-clusters model. At the

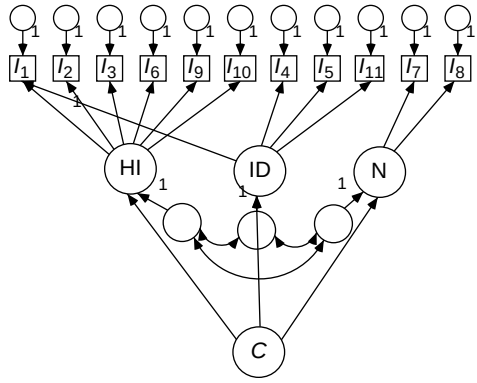


Figure 7.5 *Confirmatory factor analysis model for AP data employed in Lubke et al. (2009). Key:*

Child Behaviour Checklist items:

I_1 = *acts young*

I_2 = *cannot concentrate*

I_3 = *cannot sit still*

I_6 = *impulsive*

I_9 = *poor school*

I_{10} = *clumsy*

I_4 = *confused*

I_5 = *daydreams*

I_{11} = *stares blankly*

I_7 = *nervous*

I_8 = *twitching*

Factor labels:

HI = *hyperactivity/impulsiveness*

ID = *inattention/dreaminess*

N = *nervousness.*

second and third wave (children aged 10 and 12 years, smaller sample size), BIC favours a three-factor, two-cluster model, though the reduced sample size may have led to insufficient power to detect the third cluster (see Table 7.3). Figure 7.6 describes the three-factor, three-cluster solution for each of the waves by means of profiles of probabilities of scoring 'very true'. Cluster 1 (approximately 20% of the children) has the largest probabilities for most variables, cluster 3 (approximately 60%) the lowest probabilities and cluster 2 (the remaining 20%) is generally of intermediate severity in terms of probabilities. This ordering of the clusters combined with the evidence of the existence of underlying factors lead the authors to support the view that the AP syndrome exists on a severity continuum, with similar cluster structure across the developmental period of ages 7 to 12 years.

Table 7.3 Comparative fit of seven models considered in Lubke *et al.* (2009).

Model	No. of model parameters	Wave 1 (8079 children)		Wave 2 (5278 children)		Wave 3 (3139 children)	
		LL	BIC	LL	BIC	LL	BIC
F3C2	66	−48556	97707	−32433	65432	−17992	36516
F3C3	95	−48385	97625	−32312	65439	−17913	36592
F3C4	124	−48286	97688	−32238	65540	−17858	36713
LPAC3	68	−49619	99850	−33064	66711	−18320	37187
LPAC4	91	−49265	99349	−32873	66527	−18165	37063
LPAC5	114	−49026	99079	−32709	66396	−18074	37067
LPAC6	137	−48834	98901	−32551	66277	−18000	37102

LL = Log-likelihood; F3Cx = mixture of factor analysers, 3 factors and x clusters; LPACx = latent profile analysis with x clusters.

7.5.2 Application of finite mixture confirmatory factor analysis to cluster genes using replicated microarray experiments

Ng *et al.* (2006) demonstrate their software `EMMIX WIRE` (EM-based MIXture analysis WIth Random Effects) for multivariate normal mixtures by clustering genes using expression data from microarrays. In the yeast galactose study (Ideker *et al.*, 2001), expression levels of $n = 205$ genes were quantified for $q = 20$ tissue samples using cDNA experiments. Hybridizations for each cDNA array experiment were replicated four times ($p = 20 \times 4 = 80$). The microarray data set therefore contains repeated measurements per gene. To allow for correlation between expression levels due to 'true' latent tissue expression level being measured repeatedly, the confirmatory factor analysis model illustrated in Figure 7.7 was assumed. Observed tissue expression means were assumed constant

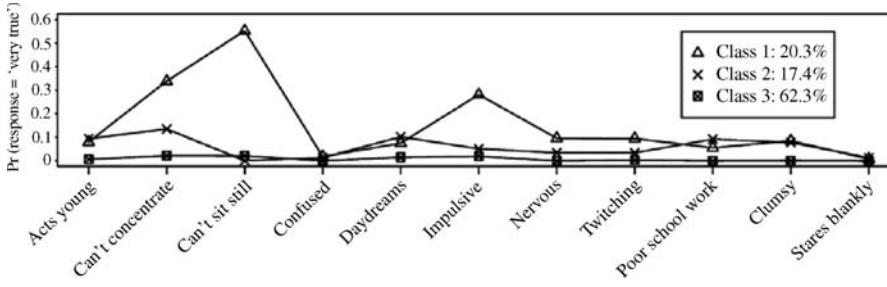


Figure 7.6 *Response pattern of the clusters obtained by fitting a three-factor, three-cluster mixture of factor analysers to the AP data of the seven-year-olds in the first wave (taken from Lubke et al., 2009).*

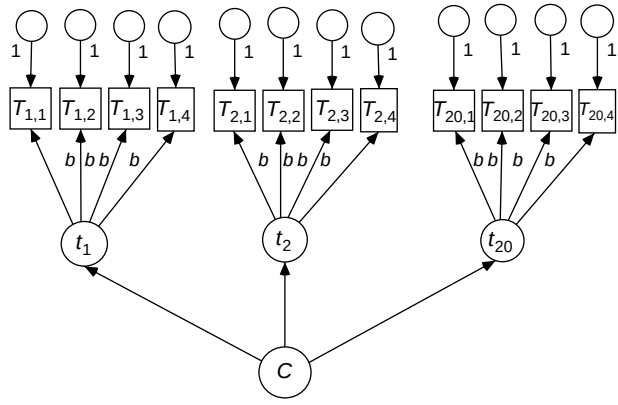


Figure 7.7 *Confirmatory factor analysis model for gene expression data assumed in Ng et al. (2006).*
Key:
$T_{x,y}$ = *Observed expression level for tissue x in yth replication study*
t_x = *true (latent) expression level for tissue x.*

across replicates; that is, $\boldsymbol{\mu}_j = (\mathbf{I}_{20} \otimes \mathbf{1}_4)(\beta_{1j}, \ldots, \beta_{20j})'$ in Equation (7.7). (Here $\otimes$ denotes the Kronecker product which simply multiplies every entry in the identity matrix $\mathbf{I}_{20}$ with the vector of ones $\mathbf{1}_4$.) Similarly, measurement error variances were assumed to vary only with tissue; that is, $\boldsymbol{\psi}_j = \text{diag}[(\mathbf{I}_{20} \otimes \mathbf{1}_4)(\tau_{1j}, \ldots, \tau_{20j})']$ in Equation (7.8). Importantly, latent tissue expression was assumed to load only on respective observed tissue expression with all factor loadings set to a constant value b_j within cluster j; that is, $\boldsymbol{\Lambda}_j = b_j(\mathbf{I}_{20} \otimes \mathbf{1}_4)$, and latent expression levels of different tissues were assumed statistically independent. (Assuming constant factor loadings across the 20 factors makes sense here because each factor measures the same latent concept, 'true gene expression', albeit for different tissues.) Due to the constraints imposed on the factor loadings, this particular confirmatory factor model can be conceptualized as a linear mixed model with gene-varying random effects. Importantly the authors further extend their 'mixture model with random effects components' to include experiment-varying random effects within clusters, which introduces correlation between the gene expression profiles; for more see Ng *et al.* (2006).

`EMMIX WIRE` was used to cluster the genes into various numbers of clusters. Model fit was compared via BIC, a seven-cluster model selected as the best model, and genes allocated to clusters based on MAP. Table 7.4 cross-tabulates this cluster solution with an external classification of the 205 genes into four functional categories in the Gene Ontology (GO) listings (Ashburner *et al.*, 2000; Yeung *et al.*, 2003). This shows that data-driven clusters 1 and 2 more or less correspond to functional categories 2 and 4, respectively. Genes in functional category 1 are split into two new clusters (4 and 7), while those in functional category 3 are split into three new clusters (3, 5 and 6). The authors suggest that the subdivisions of the

Table 7.4 Distribution of genes over empirical seven-cluster solution detected by factor mixture modelling and external classification from GO listings (yeast galactose data, taken from Ng *et al.*, 2006).

Empirical cluster	GO functional category			
	1	2	3	4
1	0	13	0	0
2	0	0	0	14
3	0	3	44	0
4	38	0	0	0
5	0	0	17	0
6	0	0	32	0
7	45	0	0	0

functional categories could be relevant to some unknown function in the GO listings.

For further applications of model-based cluster analysis of structured gene expression data see McLachlan *et al.* (2002, 2003), Celeux *et al.* (2005) and Qin and Self (2006).

7.5.3 Application of finite mixture exploratory factor analysis to cluster Italian wines

McNicholas and Murphy (2008) employ mixtures of EFA models to cluster Italian wines. The wine data are described in Forina *et al.* (1986). Three types of wine (Barolo, Grignolino and Barbera) from the Piedmont region of Italy were sampled ($n = 178$) between 1970 and 1979. For each sample, $p = 27$ chemical properties were determined (Table 7.5). Parsimonious Gaussian mixture modelling was applied. All eight models in Table 7.1 were fitted to the data for number of clusters $c = 1, 2, \ldots, 5$ and number of factors $q = 1, 2, \ldots, 5$. The model fit of the resulting 200 models was compared using BIC. The best fit (lowest BIC) was achieved by the CUU model in Table 7.1, with $c = 3$ and $q = 4$. The latter model was selected, and wine samples allocated to clusters using MAP.

For comparison, competitor approaches for parsimonious modelling, that is model-based clustering using the MCLUST family (Fraley and Raftery (1999, 2003); see Section 6.2.2) and variable selection using `R` package `clustvarsel` (Raftery and Dean (2006); see Section 6.6), were also applied. When all variables were used, a $c = 3$ component mixture with a VVI covariance structure (Fraley and Raftery, 2002; for details see Table 6.1 in Chapter 6) was selected (lowest BIC of 12119.3). When variable selection was allowed (with the maximum number of clusters preset to 8), the 19 variables marked ^{a} in Table 7.5 were selected, and a VVI model with $c = 4$ clusters chosen as the best model. Both models were fitted and MAP used to allocate wine samples to three or four clusters, respectively.

Table 7.5 Chemical properties of Italian wines.

Variable	Property	Variable	Property	Variable	Property
1[a]	Alcohol	10[a]	Sugar-free extract	19	Fixed acidity
2[a]	Tartaric acid	11[a]	Malic acid	20[a]	Uronic acids
3	pH	12	Ash	21[a]	Alkalinity of ash
4	Potassium	13[a]	Calcium	22[a]	Magnesium
5[a]	Phosphate	14[a]	Chloride	23	Total phenols
6[a]	Flavonoids	15[a]	Nonflavonoid phenols	24	Proanthocyanins
7[a]	Colour intensity	16[a]	Hue	25[a]	OD_{280}/OD_{315} of dilutes wines
8	OD_{280}/OD_{315} of flavonoids	17	Glycerol	26[a]	2-3-butanediol
9[a]	Total nitrogen	18[a]	Proline	27[a]	Methanol

[a] = variable was selected by variable selection procedure `clustvarsel`.

Table 7.6 demonstrates the cluster distribution by true wine type for each of the three approaches. All three methods are good at discovering the group structure in the wine data. Both parsimonious Gaussian mixture modelling and model-based clustering via `mclust` detect the correct number of clusters. However, in this example Gaussian mixture modelling leads to better model fit, as measured by BIC, and better group recovery. (One can speculate that mixtures of EFA models perform best when a factor structure exists within clusters. In any case, as McNicholas and Murphy (2008) point out, different model-based methods can be compared using model selection criteria such as BIC, thus allowing for methods being used in conjunction with each other.)

7.5.4 Application of growth mixture modelling to identify distinct developmental trajectories

Connell and Frye (2006) demonstrate the use of GMMs for examining heterogeneity in patterns of development. Developmental studies often assess change over time

Table 7.6 Distribution of Italian wines over type of wine and cluster. Empirical clusters identified by parsimonious Gaussian mixture modelling, `R` procedure `mclust` and `R` procedure `clustvarsel` (taken from McNicholas and Murphy, 2008).

Type	Parsimonious Gaussian mixture cluster			`mclust` cluster			`clustvarsel` cluster			
	1	2	3	1	2	3	1	2	3	4
Barolo	59	0	0	58	1	0	52	7	0	0
Grignolino	0	70	1	4	66	1	0	17	54	0
Barbera	0	0	48	0	0	48	0	1	0	47

in samples from a reference population in order to provide a 'normative' pattern of symptom change. However, the use of such summaries may be limited if there are multiple subgroups following distinct developmental trajectories. The authors consider adolescent antisocial behaviour as an example. Given that most youths never experience clinically significant behavioural or emotional problems, studies that focus on the mean change over time in the sample will be distorted by the presence of a relatively small subgroup of youths who show extreme symptom levels.

Growth mixture modelling was applied to data from a longitudinal study of adolescent problem behaviours. Four hundred and ninety-eight youths were followed up from age 12 to 17 years, with measures taken at grades 6, 7, 8, 9 and 11 ($p = 5$). A continuous measure of antisocial behaviour was analysed under a multivariate normal model. The assumed GMM is illustrated in Figure 7.8. Two

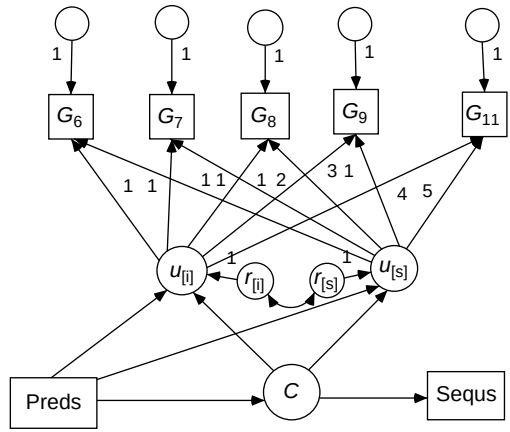


Figure 7.8 *Growth mixture model for adolescent antisocial behaviour data assumed in Connell and Frye (2006).*

Key:

Manifest variables:

G_6 = *antisocial behaviour score at grade 6*

G_7 = *antisocial behaviour score at grade 7*

G_8 = *antisocial behaviour score at grade 8*

G_9 = *antisocial behaviour score at grade 9*

G_{11} = *antisocial behaviour score at grade 11*

Preds = predictors of cluster membership:

Gender

Ethnicity

Deviant peers at grade 6

Substance use at grade 6

Parental monitoring at grade 6

Sequs = sequelae:

Arrest at grade 11.

growth factors were allowed for, a subject-varying random intercept and a subject-varying random slope. All growth factor parameters and residual parameters were allowed to vary with latent class membership. In addition, the model contained some predictors of latent class membership and growth factors (gender, ethnicity, deviant peers, substance use, parental monitoring assessed at grade 6) and sequelae of latent class membership (arrest at grade 11). The latter provided a test of whether or not arrest differed across trajectory classes.

Note that SEMMs are easily extended to include predictors and sequelae of the latent classes. Details can be found in Clogg, 1981; Dayton and Macready, 1988; Vermunt, 1997; Bandeen-Roche *et al.*, 1997 and Yamaguchi, 2000. Simultaneous modelling is preferable to traditional two-stage approaches which separate the clustering process and the evaluation of associations with clusters, since the latter can yield inefficient and inconsistent regression parameter estimates (Bolck *et al.*, 2004).

To identify the number of distinct trajectories (clusters), recommendations provided by Muthén and Muthén (2000) were followed and carried out in `Mplus`. One to five class models were allowed. Despite using multiple starting values, models with more than two classes did not converge, which was interpreted as an indication of poor model fit and led the authors to exclude such models. To choose between the single-class and two-class models, first indices of relative model fit were considered, with BIC minimization seen as the most important. Second, the LMR LRT was used to provide a statistical comparison of the fit of a given model with a model of one fewer class. Third, the quality of the classification was examined by means of entropy (a measure of the probability of membership in the class allocated to an individual under MAP: values range from zero to one, with values closer to one representing better classifications). Fourth, the usefulness and interpretability of the resulting latent trajectory classes were considered.

Table 7.7 shows the results of the comparison between the one-class and the two-class model. According to both the BIC and the LMR LRT, the two-class model fitted the data better than the one-class model. Entropy was acceptable (0.82). Under the two-class model, cluster 1 represented 37.7% of the sample and exhibited a significant level of antisocial behavioural and nonsignificant change over time, and was termed the 'chronic-high group'. Cluster 2 represented the majority of the

Table 7.7 Comparison of fit of one-class and two-class model for adolescent antisocial behaviour trajectories.

Number of clusters	Loglikelihood	BIC	Entropy	LMR LRT
1	−1544.29	3231.05	n.a.	n.a.
2	−938.95	2187.63	0.82	1216.45, $p < 0.0001$

n.a. = not applicable.

sample (62.3%) and also exhibited a significant initial level of antisocial behaviour and nonsignificant change over time, although the level of problems was substantially lower than in cluster 1, and so this cluster was termed the 'stable-low group'. Two predictors, namely deviant peer groups and parental monitoring, were found to significantly discriminate between the two clusters. Adolescent membership of cluster 1 was associated with more frequent affiliation with deviant peer groups and less frequent parental monitoring of their behaviour. A number of predictors were associated with the level of the growth factors within clusters (see Connell and Frye, 2006), and membership of cluster 1 significantly increased the chances of being arrested at age 17 years. The authors concluded that there was evidence to support the existence of two distinct developmental trajectories, and consider the trajectory of cluster 2 'normative'.

7.5.5 Application of growth mixture modelling to identify trajectories of perinatal depressive symptomatology

Mora *et al.* (2009) apply growth mixture modelling to describe heterogeneity in the timing and persistence of maternal depressive symptomatology. These authors analyse longitudinal data from a cohort study of 1735 low-income, multiethnic, inner-city women from Philadelphia, PA, recruited in pregnancy from 2000 to 2002. The women were followed up prospectively (one prenatal and three postpartum time points) in order to establish trajectories of depressive symptomatology from pregnancy to two years postpartum.

Growth mixture models were fitted to the $p = 4$ repeated measures in `Mplus`. To capture the two phases of development (phase 1: from pre- to postpartum time points 1 and 2; phase 2: during the postpartum period, time points 2, 3 and 4), piecewise trajectories with slope change at time point 2 were considered, which translated into three growth factors. The variance of the second growth factor (slope between time points 1 and 2) was constrained to be zero for all clusters to identify the model. The variance of the third growth factor (slope during postpartum period) had further to be constrained to be constant across clusters, to achieve model convergence. The number of clusters was allowed to vary between two and seven. Statistical indices (e.g. entropy, BIC and a bootstrap likelihood ratio test of k versus $k-1$ clusters) as well as interpretability of the model were used to select the optimal model (e.g. group prevalence >5% and interpretability of the groups).

On the basis of these criteria, a five-cluster solution was chosen (entropy = 0.79; for details of this decision-making process see Mora *et al.*, 2009). This solution added an extra cluster to an initially hypothesized four-cluster solution. The clusters were labelled according to their mean values for the three growth factors; here 'high' = CES-D depression score $\geq$16 (CES-D: Center for Epidemiologic Studies Depression scale – see Radloff, 1977):

- Cluster 1 (7%): 'Chronic' – persistently high level of depressive symptoms
- Cluster 2 (6%): 'Antepartum' – depressive symptomatology present only at the first prenatal visit

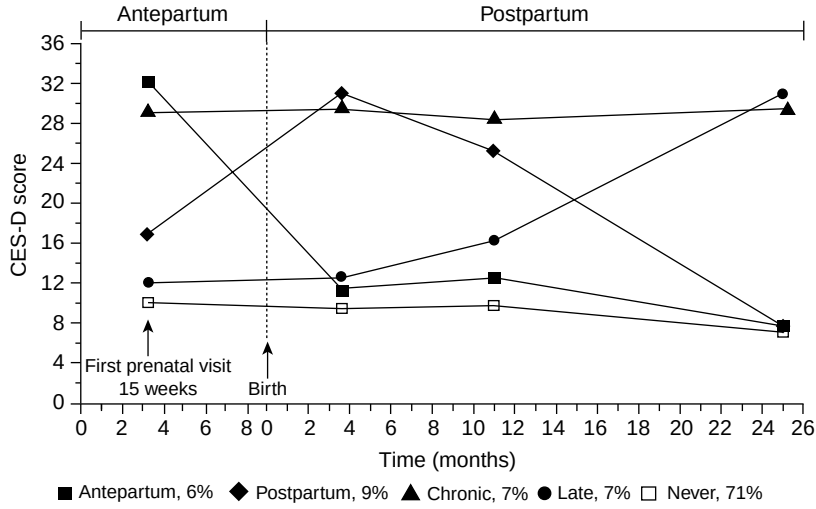


Figure 7.9 *Illustration of five trajectories of depressive symptomatology in pregnancy and postpartum found in Mora et al. (2009).*

- Cluster 3 (9%): 'Postpartum' – depressive symptoms present within six weeks of delivery, that subside over time
- Cluster 4 (7%): 'Late' – with low levels of depressive symptoms ante- and peripartum that increase in the second year postpartum
- Cluster 5 (71%): 'Never' – continuous low levels of depressive symptomatology.

A graphical illustration of the five trajectories is provided in Figure 7.9, which shows a plot of average depression scores by time point and allocated cluster. Note that the model assumes a constant slope over postpartum time points (time points 2, 3 and 4).

In a second step, after allocating women to clusters on the basis of MAP, predicted cluster membership was associated with a number of women's characteristics (demographic, socioeconomic, health and psychosocial) using multinomial logistic regression. A full list of external variables considered as predictors of cluster membership can be found in Mora *et al.* (2009). Cluster membership was predicted by demographic variables (nativity, education, race, parity), health, health behaviour and psychological characteristics (ambivalence about pregnancy and high objective stress). The authors conclude that their identified distinct trajectories are meaningful, had some level of external validation and should be considered when planning maternal depression programmes.

7.6 Summary

Most of the finite mixture models for structured data discussed in the chapter were structural equation mixture models, for example the mixture of factor analysers

model and the growth mixture model. They work on the basis that the covariance structure within a cluster is generated by a particular structural equation model, for example a confirmatory factor analysis model or a linear growth model. By doing so, they provide powerful tools for simultaneously clustering objects and modelling relationships between variables.

However, two notes of caution need to be mentioned. First, while making use of the structure in the data in this way can reduce the number of parameters and so increase precision compared to the unstructured model-based clustering approach discussed in the last chapter, large sample sizes may still be necessary to acquire parameter estimates with acceptable precision for complex structural models such as growth mixture models with many growth factors. Second, as Bauer and Curran (2003, 2004) point out, the flexibility of the SEMM to gain a fuller understanding of the phenomenon under study comes at the price of additional model assumptions. When these do not hold, that is, when the structural model is misspecified, the latent variables do not follow the assumed distribution, or nonlinear relationships exist between manifest and latent variables, then parameter estimates can suffer bias, and spurious clusters may be identified.

8

Miscellaneous clustering methods

8.1 Introduction

The methods described in the preceding chapters form the major part of the body of work on cluster analysis. Nevertheless, there remain a substantial number of other methods that do not fall clearly into any of the previous categories, and in this chapter an attempt is made to describe a number of these techniques. While a comprehensive review is impossible simply because of the vastness of the literature involved, this situation is less daunting than it first appears, since some of the apparently specialized techniques are, in essence, very similar to standard clustering techniques. For example, some of the techniques developed in genetic research entail applying hierarchical clustering (see Chapter 4) but using a specialized distance measure such as Jukes–Cantor or the *optimal matching* coefficient (see Chapter 3). A method patented by the Hewlett Packard Development Company for text clustering uses a recursive hierarchical technique, but considering parts of speech in order (nouns, then verbs, then adjectives, for example) – see Kettenring (2009). Similarly, some of the newer pattern recognition techniques in imaging are closely related to traditional methods. The Kohonen self-organizing map, for example, discussed in Section 8.8.2 as an example of a neural network, is similar in principle to the k-means clustering technique described in Chapter 5.

The methods to be discussed in this chapter can be categorized as follows:

- Density search and mode analysis, where clusters are assumed to be concentrated in relatively dense patches in a metric space.
- Methods which allow overlapping clusters, including pyramids.

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

- Direct clustering of data matrices, rather than proximity matrices, in order to cluster both variables and objects simultaneously.
- Constrained clustering, where the membership of clusters is determined partly by external information, for example spatial contiguity.
- Fuzzy methods, where objects are not assigned to a particular cluster but possess a membership function indicating the strength of membership to each cluster.
- Neural networks: pattern recognition algorithms that imitate the computational capabilities of large, highly connected networks such as the neurons in the human brain.

This categorization is mainly one of convenience, and some methods have characteristics of more than one of these categories. For example, direct clustering of data matrices may involve reordering rows and columns, so that sets of contiguous rows and columns form clusters of objects and variables, respectively. Having rearranged the matrix in this way it is clearly easy to obtain overlapping clusters. Additionally, overlapping or fuzzy clustering might be viewed as resulting from a relaxation of the usual constraints that some numerical measure of cluster membership, such as the conditional probability that an observation belongs to a particular cluster (see Chapter 6), should sum to 1 over clusters, or should take only the values 0 or 1 for, say, 'in cluster' and 'not in cluster', in a non-overlapping cluster solution.

8.2 Density search clustering techniques

If individuals are depicted as points in a metric space, a natural concept of clustering (see Chapter 1) suggests that there should be parts of the space in which the points are very dense, separated by parts of low density. Several methods of cluster analysis have been developed which search for regions of high density in the data, each such region being taken to signify a different group. The mixture approach described in Chapter 6 might be seen as a formal way of using this concept.

A number of these techniques have their origins in single linkage clustering (see Chapter 4), but attempt to overcome chaining, one of the main problems with this method. One such attempt is the *taxmap* method of Carmichael *et al.* (1968) and Carmichael and Sneath (1969). The clusters are formed initially in a way similar to single linkage, but criteria are used to prevent the addition of objects that are much further away from the last object admitted, such as the drop in average similarity on adding the candidate object. Objects that are rejected in this way initiate new clusters. Two other methods that rely on seeking dense regions will now be described in more detail.

8.2.1 Mode analysis

Mode analysis (Wishart, 1969) is a derivative of single linkage clustering which searches for natural subgroupings of the data by seeking disjoint density surfaces

in the sample distribution. The search is made by considering a sphere of some radius, R, surrounding each point, and counting the number of points falling in the sphere. Individuals are then labelled as *dense* or *non-dense* depending on whether their spheres contain more or fewer points than the value of the *linkage* parameter, K, which is preset at a value dependent on the number of individuals in the data set. (Some possible values of K for various values of n are suggested in Wishart, 1987.)

The parameter R is gradually increased and so more individuals become 'dense'. Four courses of action are possible with the introduction of each new dense point:

- The new point is separated from all other dense points by a distance that exceeds R. When this happens the point initiates a new cluster nucleus and the number of clusters is increased by one.
- The new point is within distance R of one or more dense points which belong to only one cluster nucleus. In this case, the new point is added to the existing cluster.
- The new point is within distance R of dense points belonging to two more clusters. If this happens, the clusters concerned are combined.
- At each 'introduction' cycle, the smallest distance, D, between dense points belonging to different clusters is found, and compared with a threshold value calculated from the average of the $2K$ smallest distance coefficients for each individual. If D is less than this threshold value, then the two clusters are combined. Sometimes only one cluster is produced (indicating a lack of cluster structure in the data), but usually the analysis reaches a point at which a maximum number of clusters is isolated. It is usually this solution which is taken as the most significant.

A difficulty with mode analysis is its failure to identify both large and small clusters simultaneously. A small radius R may distinguish two large, disjoint modes without finding a third smaller (but distinct) mode, because each of its individuals fails to qualify as a dense neighbourhood. Alternatively, if a larger R is specified, the small cluster might be found, but the two large clusters could possibly be merged. Such potential difficulties led Wishart (1973) to suggest an improved mode-seeking cluster method, in which the spherical neighbourhoods of two growing clusters may intersect at some large value of R, to the extent that they would have been fused in the original version of mode analysis; now the fusion level is merely noted, and the clusters are not united.

As an illustration of the use of mode analysis, it will be applied to the distance matrix for 11 forms of the bee *Hoplitis producta*, based on 23 variables (Michener, 1970). The proximity matrix and results are shown in Tables 8.1 and 8.2.

8.2.2 Nearest-neighbour clustering procedures

Wong (1982) and Wong and Lane (1983) describe a hierarchical clustering method which is similar in some respects to the method of mode analysis discussed in Section 8.2.1. The method is designed to detect what Hartigan (1975) defines as

Table 8.1 Matrix of distance coefficients (based on standardized data) for the forms of the bee *Hoplitis producta*[a].

	1	2	3	4	5	6	7	8	9	10	11
1	0										
2	0.940	0									
3	1.229	0.791	0								
4	1.266	0.847	0.303	0							
5	1.507	1.331	1.070	1.026	0						
6	1.609	1.306	0.778	0.573	1.175	0					
7	1.450	1.266	1.475	1.506	1.829	1.876	0				
8	1.239	1.286	1.510	1.540	1.908	1.832	1.665	0			
9	1.493	1.160	0.848	0.792	0.965	0.978	1.847	1.761	0		
10	1.494	1.396	1.497	1.528	1.724	1.687	1.954	1.733	1.721	0	
11	1.348	1.238	1.352	1.385	1.724	1.559	1.844	1.608	1.596	0.645	0

[a]The forms of *Hoplitis* are: 1, *H. gracilis;* 2, *H. subgracilis;* 3, *H. interior;* 4, *H. bernardina;* 5, *H. panamintana;* 6, *H. producta;* 7, *H. colei;* 8, *H. elongata;* 9, *H. uvularis;* 10, *H. grinelli;* 11, *H. septentrionalis*.
Source: Michener (1970).

high-density clusters, these being *maximal connected sets* of the form

$$\{x|f(x) \geq f^*\}, \tag{8.1}$$

where f is the population density of the observations, and f^* is some threshold value. Wong and Lane (1983) estimate the density at a point x by $f_n(x)$ given by

$$f_n(x) = k/[nV_k(x)], \tag{8.2}$$

where $V_k(x)$ is the volume of the smallest sphere centred at x containing k sample observations. A distance matrix arises from these density estimates according to the following two definitions:

Table 8.2 Results of applying mode analysis to the bee distance data in Table 8.1.

Stage	
1	Observation 3 initiates new cluster centre
2	Observation 5 joins observation 3
3	Observation 4 joins [3, 5]
4	Observation 6 joins [3, 5, 4]
5	Observation 7 joins [3, 5, 4, 6]
6	Observation 10 initiates new cluster centre
7	Observation 9 joins observation 10
8	Observation 2 joins [3, 5, 4, 6, 7]
9	Observation 8 joins [3, 5, 4, 6, 7, 2]
10	Observation 1 joins [3, 5, 4, 6, 7, 2, 8]
11	Observation 11 joins [10, 9]
Final results:	
Cluster 1: 3, 5, 4, 6, 7, 2, 8, 1	
Cluster 2: 10, 9, 11	

Definition 1: Two observations x_i and x_j are said to be neighbours if

$$d^*(x_i, x_j) \leq d_k(x_i) \text{ or } d_k(x_j), \tag{8.3}$$

where d^* is the Euclidean metric and $d_k(x_i)$ is the kth nearest-neighbour distance to point x_j.

Definition 2: The distance $d(x_i, x_j)$ between the observations x_i and x_j is

$$d(x_i, x_j) = \frac{1}{2}\left[\frac{1}{f_n(x_i)} + \frac{1}{f_n(x_j)}\right] \tag{8.4}$$

$$= \begin{cases} \dfrac{n}{2k}\left[V_k(x_i) + V_k(x_j)\right] & \text{if } x_i \text{ and } x_j \text{ are neighbours} \\ \infty & \text{otherwise.} \end{cases} \tag{8.5}$$

The single linkage clustering algorithm is then applied to this distance matrix to obtain the dendrogram of sample high-density clusters. The value of k controls the amount by which the data are 'smoothed' to give the density estimate on which the clustering procedure is based. There appears to be no unique recommendation concerning the choice of k, although Wong and Schaack (1982) derive empirical evidence for a rule of thumb of the form $k = 2\log_2 n$. Since the hierarchical clusterings obtained for different values of k can be very different, Wong and Lane (1983) suggest that several values around this value should be tried. To illustrate the operation of their proposed method, Wong and Lane (1983) apply it to the data shown in Figure 8.1. The dendrogram giving the hierarchical clustering obtained by the kth nearest-neighbour method with $k = 5$ appears in Figure 8.2. Two disjoint modal regions, corresponding to the crescentic clusters in Figure 8.1, can be identified in the dendrogram. Ling (1972) suggests another nearest-neighbour

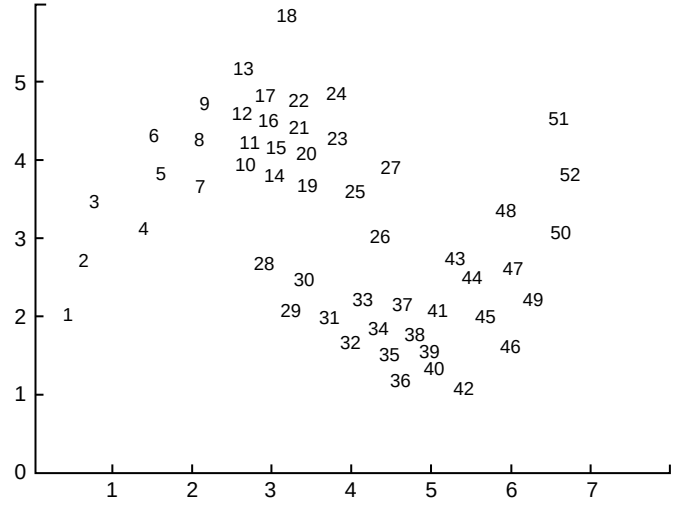


Figure 8.1 *Bivariate data containing crescentic clusters.* (Source: Wong and Lane, 1983.)

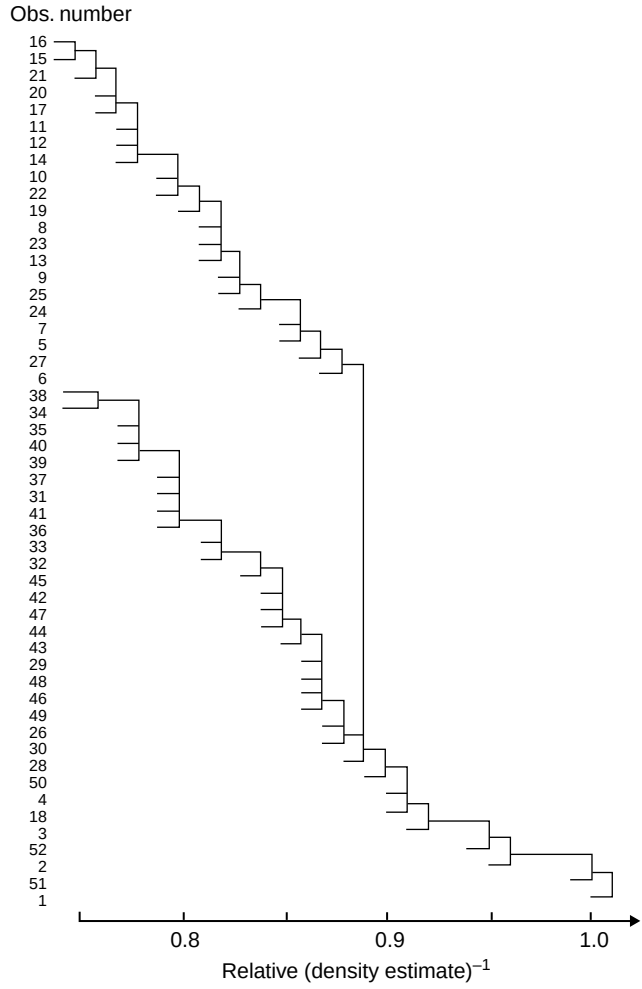


Figure 8.2 *Dendrogram obtained by applying Wong and Lane clustering to the data shown in Figure 8.1. (*Source: Wong and Lane, 1983.)

type clustering procedure, and Wong and Schaack (1982) propose a procedure for indicating the number of clusters when using the method outlined above. Other density-search-type clustering methods are described in Gitman and Levine (1970), Cattell and Coulter (1966) and Katz and Rohlf (1973).

8.3 Density-based spatial clustering of applications with noise

An approach that has had some success and has spawned a number of variants to cope with different scenarios is the DBSCAN – density-based spatial clustering of applications with noise (Sander *et al.*, 1998). The algorithm classifies objects as

clusters (dense regions) or noise (objects in low-density regions). Clusters are sets of at least M objects in a dense region with a radius R; M and R are defined by the user. Within the clusters, two types of object are defined: 'core' and 'non-core' objects. A core object has at least M objects contained within its neighbourhood with radius R, and forms an initial cluster. The neighbourhood (within a radius R) is examined and objects within the neighbourhood are assigned to the cluster. When the neighbourhoods of core objects overlap the clusters are merged. This algorithm defines some objects as 'border' objects; these are objects that are 'density reachable' from a core object. (Object p_1 is 'direct density reachable' from p_2 if it is within a radius R, and it is 'density reachable' from core object p_c if a chain of direct density reachable objects $p_1 \cdots p_n$ can be found). Border objects are not core (they do not have the requisite minimum number of points in their neighbourhood), but they are still sufficiently close (density reachable) to a core point to be included in its cluster. Objects that do not join any cluster at the end of the process are termed 'noise'. Euclidean distance is used as the similarity measure, but any suitable measure could be used in a similar way.

The basic algorithm of DBSCAN is easy to program and it copes with clusters of different shapes, although not so well with clusters of varying density. It is useful not only for clustering per se but where the detection of outliers is of interest. A disadvantage is that it may not work well for high-dimensional space (which is usually sparse). A variation of density searching which is more suitable in this situation is CLIQUE (Agrawal *et al*., 1998), where multivariate space is divided up into a grid of cells, which are classified according to density. Once the dimensions of high density are identified, clustering can proceed in subspaces of lower dimension.

Birant and Kut (2007) have adapted the DBSCAN algorithm to deal with spatially and temporally related data (ST-DBSCAN), and have illustrated it with an example from sea temperature and wave height in the Black Sea. Building-in constraints on clusters, for example based on geographical criteria, is described in Section 8.6. Here the spatial information is used as an additional clustering criterion which pre-specifies the acceptable radius for clusters, alongside the main clustering variables, which were sea characteristics such as temperature, wave height and surface height. Figure 8.3 shows the clusters found for sea surface

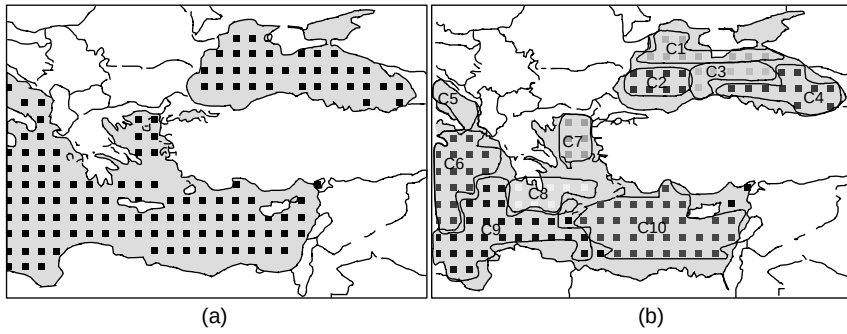


Figure 8.3 *Map of the Black Sea showing (a) the locations of 134 stations; (b) the results of ST-DBSCAN cluster analysis on sea surface height residual data (from Birant and Kut, 2007).*

height (which varies dues to currents, gravity and seawater temperature). The measurements were collected by satellite over five-day periods in years between 1992 and 2002 on a two-dimensional grid separated by one degree in latitude and longitude and stored in a 'data warehouse'.

8.4 Techniques which allow overlapping clusters

In the methods discussed so far, objects are assumed to belong to one particular cluster. However, in many applications it is plausible that objects should belong to more than one cluster simultaneously. An example might be market research where an individual person belongs to several consumer groups, or social network research where people belong to overlapping groups of friends or collaborators. In the latter type of application, it is often the individuals in the intersections of the clusters who are of most interest. An example of a type of study in which overlapping clusters would definitely be inappropriate would be archaeological provenancing, since an object can have only one specific origin. (It might well be that in such an application a single cluster could not be definitely assigned to each object, but this would suggest the use of fuzzy rather than overlapping clustering; see Section 8.7.)

Methods that directly reorder rows and columns of data matrices can be used to produce overlapping clusters, simply by extending the set of rows (columns) to cover a portion of adjoining clusters. These are described under the heading of direct data clustering (Section 8.5); here we discuss methods that have been more specifically designed to deal with overlap. In Section 8.4.1 two relatively early techniques, *clumping* and the B_k *technique*, are briefly discussed. In Section 8.4.2 a more recent and more widely used method, *additive clustering*, is described. Finally, a generalization of dendrograms that allow overlap, the *pyramid*, is discussed in Section 8.4.4.

8.4.1 Clumping and related techniques

Clumping techniques begin with the calculation of a similarity matrix, followed by the division of the data into two groups by a 'cohesion' function including a parameter controlling the degree of overlap. Needham (1967) considered a symmetric cohesion function $G_1(A)$ given by

$$G_1(A) = S_{AB}/S_{AA}S_{BB}. \tag{8.6}$$

Parker-Rhodes and Jackson (1969) suggest a modification, $G_2(A)$, given by

$$G_2(A) = \frac{S_{AB}}{S_{AA}}\left[\frac{n_A(n_A-1)}{S_{AA}} - \frac{S_{AA}}{bn_A(n_A-1)}\right], \tag{8.7}$$

where A and B refer to the two groups into which the data are divided, A being their putative clump. S_{AB} is the sum of the similarities between members of groups

A and B; that is

$$S_{AB} = \sum_{i \in A} \sum_{j \in B} s_{ij}, \tag{8.8}$$

where s_{ij} is an inter-individual similarity, n_A is the number of individuals in group A, and b is an arbitrary parameter which allows the investigator some control over the size of the clumps and the amount of overlap. Algorithms to minimize these functions proceed by successive reallocations of single individuals from an initial randomly chosen cluster centre (Jones and Jackson, 1967). By iterating from different starting points, many divisions into two groups may be found. In each case the members of the smaller group are noted and constitute a class to be set aside for further examination. The cohesion function $G_1(A)$ is designed to find good partitions of the set of individuals, whilst $G_2(A)$ allows the internal similarities of A and the separation of A from B to be adjusted relative to each other by the use of parameter b.

The B_k *technique* (Jardine and Sibson, 1968) is a hierarchical method in which individuals are represented by nodes on a graph, and pairs of nodes are connected which correspond to individuals having a similarity value above some specified threshold H. At each stage in the hierarchy, the set of *maximal complete subgraphs*, or *cliques*, is found. These are the largest sets of individuals for which all pairs of nodes are connected at some level of similarity. If all possible thresholds are considered, an unmanageable number of subgraphs are produced. Moon and Moser (1965) give formulae for the upper bounds to this number. For example, if the proximity matrix consisted of zeros and ones, the upper bound to the number of cliques would be 59 049 for $n = 30$. In the B_k technique, clusters are subsets of the cliques, and the number of clusters is restricted by choosing a value or range of values for k, such that a maximum of $k - 1$ objects belong to the overlap between clusters; any having more than $k - 1$ objects in common are amalgamated. An example is shown in Figure 8.4, using the distance matrix in Table 8.3.

Algorithms for applying this technique have been proposed by Jardine and Sibson (1968) and Rohlf (1975). Although the B_k method has been shown to have various favourable properties, such as stability and invariance under relabeling or monotonic transformation of the proximity matrix (Sibson, 1970), it has had little application other than in a study of acoustic confusion matrices (Morgan, 1973).

8.4.2 Additive clustering

The ADCLUS method (Shepard and Arabie, 1979) has found wider acceptance as a method for identifying overlapping clusters. In this approach, the similarity between two objects is a weighted sum of their common features. A model is fitted to an observed proximity matrix, **S**, such that the model proximity between any pair of objects is the sum of the weights (see below) of those clusters containing

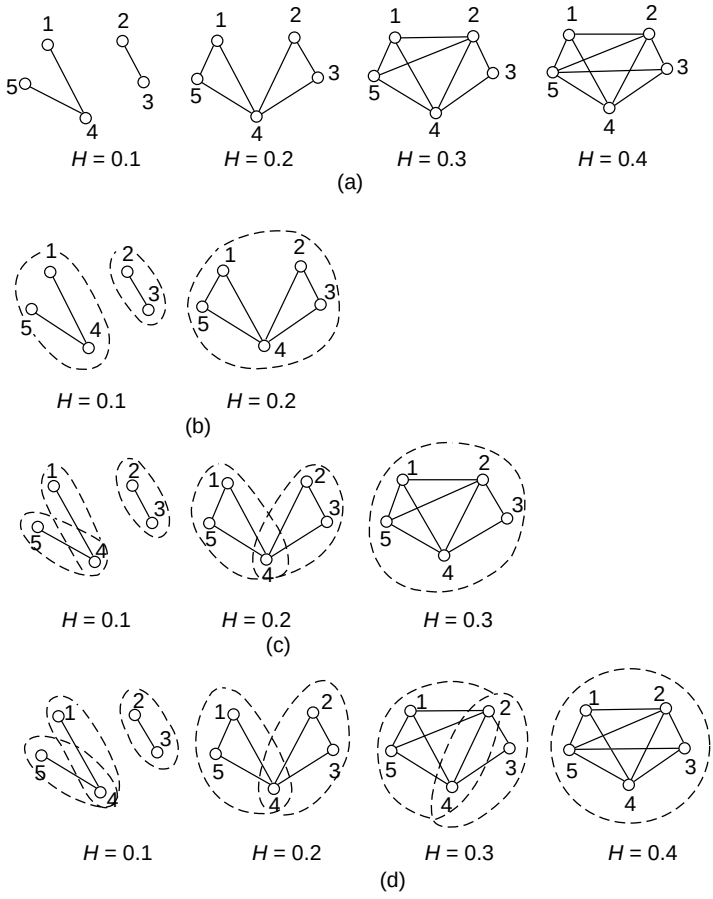


Figure 8.4 *Results of applying Jardine and Sibson's clustering method to the distance matrix in Table 8.3: (a) shows the maximal complete subgraphs for various thresholds H; (b)–(d) show any subsets found within these subgraphs that overlap by 0, 1 and 2 objects.*

Table 8.3 Hypothetical distance matrix to illustrate the B_k method.

Object	1	2	3	4
2	0.1			
3	0.4	0.1		
4	0.1	0.2	0.2	
5	0.2	0.3	0.4	0.1

that pair. The first step is to use the distinct entries s_{ij} in **S** as the possible thresholds for the maximal complete subgraphs; this defines all the potential clusters. These are then coded as columns in a binary $n \times m$ matrix, **P**, of cluster memberships, where n is the sample size and m is the number of potential clusters. The algorithm then fits a set of weights **W** to the clusters, and at the same time reduces the number of clusters. The model fitted is of the following form:

$$\hat{\mathbf{S}} = \mathbf{PWP}' + \mathrm{C}, \tag{8.9}$$

where **C** is a matrix with zeros in the diagonal and a constant in other entries, the constant being equal to the weight fitted to the complete sample; it is used to assess goodness of fit. A balance has to be struck between the goodness of fit and the number of clusters. Once the clusters (as defined by the cluster memberships **P** and the weights **W**) have been fitted to the data, it is hoped that each cluster can be labelled in terms of a discrete, latent common feature linking its constituent objects. The cluster weight indicates the proportion of the similarity between its objects that is attributable to this feature.

Even after fitting the model, an excessive number of cluster solutions may still be encountered, and a modified version in which the number of clusters is chosen in advance by the investigator has been developed, called MAPCLUS (Arabie and Carroll, 1980). An application of this is illustrated below. The technique has also been generalized (as 'INDCLUS') to a multi-observer situation, in which each observer has their own set of weights (Carroll and Arabie, 1983), and to two-mode data by De Sarbo (1982) and Both and Gaul (1986). Further developments have been described by Navarro and Griffiths (2008), who employ methods from nonparametric Bayesian statistics to obtain estimates of the number and importance of features in defining additive models.

8.4.3 Application of MAPCLUS to data on social relations in a monastery

The social structure of an American monastery during the mid 1960s, a period when there was considerable internal strife leading to the departure of many of the monks, has been studied by Sampson (1968). This data set has been analysed a number of times as an interesting example of data relating to a social group that was about to disintegrate. The social relations between 18 novice monks were assessed in 'sociograms', in which each monk rated (in order) his 4 colleagues who respectively stood out in terms of four positive qualities (like, esteem, influence and praise) and four negative qualities (antagonism, disesteem, negative influence, and blame). The novices had been separately classified, on the basis of Sampson's prose description and other methods of cluster analysis, as Young Turks (the newer arrivals), Loyal Opposition, and Outcasts, and also as Leaders or Followers.

The raw data are given by Fienberg *et al.* (1985), and a correlation matrix derived from them is given by Breiger *et al.* (1975). Both publications present alternative ways of analysing the data, which are shown in Table 8.4.

Table 8.5 shows the cluster solution provided by MAPCLUS, which accounts for 62% of the variance, with the weights of the clusters (the sums of the weights of the objects they contain) indicated. These weights are interpreted as the *psychological saliences* of the clusters. Figure 8.5 shows the overlapping clusters superimposed on a multidimensional scaling solution.

The authors observe that Outcasts 3, 17 and 18 were asked to leave the monastery, as was the leader of the Young Turks (2). All but one of the remaining Young Turks (12), the remaining Outcast (13) and only two of the Loyal Opposition (8 and 10) left voluntarily. These last two belong to the bipartisan clusters 7 and 8 that are the least heavily weighted and also have overlap between the two opposing factions. The sixth cluster links three Outcasts with two of the leading Young Turks and a Follower, indicating that the Outcasts' sympathies were with the Young Turks. On the basis of these and other observations, the authors conclude that the overlapping clustering was more realistic and informative in portraying the true situation than a partition-based model.

8.4.4 Pyramids

The pyramid, like the additive tree (see Figure 4.7), is another generalization of the dendrogram which can be used as a basis for clustering: a cut through the pyramid at any given height gives rise to a set of ordered, overlapping clusters. First developed by Diday (1986), the pyramid is less restrictive than the dendrogram, but more restrictive than general methods for overlapping clusters in that objects can belong to (at most) two clusters. It is constructed by means of an algorithm that is similar to those used in agglomerative hierarchical clustering (see Chapter 4), employing an aggregation index based on the dissimilarity matrix **D** to decide on which classes of objects to merge. However, in contrast to hierarchical clustering, it maintains certain order relationships that allow the borders of classes to retain objects in common. Figure 8.6 shows a typical pyramid, and illustrates how *classes* are linked to clusters by connected components of the graph, with objects at the borders of classes linking clusters.

Bertrand (1995) describes the mathematical properties of the pyramidal clustering model, showing that, if the entries in the corresponding cophenetic matrix (see Chapter 4) are the dissimilarities, they never decrease when moving away from the diagonal, so long as the order of the objects is *compatible* with the pyramid. This compatibility property is also known as *Robinsonian* after the developer of a method for seriating pottery assemblages in archaeology (Brainerd, 1951). The recognition of Robinsonian dissimilarities is discussed by Chepoi and Fichet (1997). Compatibility implies that

- partitions produced by cutting the pyramid at a given height are included within these orders;
- the pyramid could be represented with the objects in these orders along the bottom.

Table 8.4 Correlation matrix showing levels of effect, aggregated over four positive and four negative relations, among 18 monks.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1.0																	
0.23	1.0																
0.02	−0.07	1.0															
−0.33	−0.34	−0.06	1.0														
−0.29	−0.48	−0.10	0.15	1.0													
−0.08	−0.17	−0.23	0.41	−0.07	1.0												
0.12	0.24	−0.14	−0.42	−0.01	0.13	1.0											
−0.04	−0.28	−0.37	0.25	0.24	0.35	−0.09	1.0										
−0.19	−0.21	−0.15	0.44	0.26	0.15	−0.40	0.02	1.0									
−0.15	−0.34	−0.19	0.05	0.00	0.18	−0.02	0.21	0.00	1.0								
−0.35	−0.48	0.06	0.45	0.18	0.18	−0.17	−0.01	0.10	0.43	1.0							
0.13	0.19	−0.26	−0.25	−0.19	0.04	0.00	0.04	−0.17	−0.17	−0.25	1.0						
−0.06	−0.33	0.15	0.02	0.09	−0.23	−0.09	−0.05	0.04	0.00	0.04	−0.24	1.0					
0.10	0.31	−0.17	−0.17	−0.06	−0.13	−0.03	0.02	−0.04	−0.33	−0.39	0.19	−0.21	1.0				
0.26	0.38	−0.16	−0.41	−0.17	0.02	0.23	−0.12	−0.14	0.00	−0.33	0.17	−0.26	−0.01	1.0			
−0.12	0.31	−0.18	−0.24	−0.09	−0.28	−0.02	−0.16	−0.26	0.08	−0.18	0.17	0.10	−0.03	0.20	1.0		
0.11	−0.14	0.31	−0.43	−0.04	−0.24	0.12	−0.26	−0.17	−0.15	−0.22	0.05	0.19	0.11	−0.18	−0.10	1.0	
0.07	−0.15	0.25	−0.37	−0.05	−0.56	0.06	−0.27	−0.07	0.12	−0.09	−0.11	0.20	0.08	−0.06	−0.01	0.56	1.0

Source: Sampson (1968).

Table 8.5 Overlapping clusters of monks identified by additive clustering (based on correlation matrix in Table 8.4).

Cluster	Weight	Monks in the subset[a]
1	0.298	4 LOL, 6 LOL, 8 LOL, 5 LOF, 9 LOF, 10 LOF, 11 LOL
2	0.272	3 O, 17 O, 18 O, 13 LOF; O[b]
3	0.271	4 LOL, 11 LOL
4	0.261	4 LOL, 9 LOL
5	0.256	1 YTL, 2 YTL, 7 YTL, 12 YTL, 14 YTF, 15 YTF, 16 YTF
6	0.146	1 YTL, 2 YTL, 14 YTF, 3 O, 17 O, 18 O
7	0.134	5 LOF, 10 LOF, 13 LOF; O, 17 O, 18 O, 16 YTF
8	0.114	2 YTL, 12 YTL, 15 YTF, 6 LOL, 8 LOL

[a]LO, member of Loyal Opposition; YT, member of Young Turks; O, Outcast.
Suffix of L denotes Leader; F denotes Follower.
[b]Novice 13 was variously regarded as a Follower in the Loyal Opposition and as an Outcast.
With these eight clusters and an additive constant of 0.338, the variance accounted for was 62.4%.
Source:Arabie and Carroll (1989).

In Figure 8.6, compatible orders are $\{a, b, c, d, e\}$, as in the figure; $\{a, b, c, \boldsymbol{e}, \boldsymbol{d}\}$; $\{\boldsymbol{c}, \boldsymbol{b}, \boldsymbol{a}, d, e\}$; and $\{\boldsymbol{e}, \boldsymbol{d}, \boldsymbol{c}, \boldsymbol{b}, \boldsymbol{a}\}$ – bold indicating inversion of a cluster. The compatible orders are potentially important in subject matter interpretation (for example, they might indicate an archaeological or geological seriation). They also indicate clearly the objects responsible for overlaps, which themselves may be of particular interest.

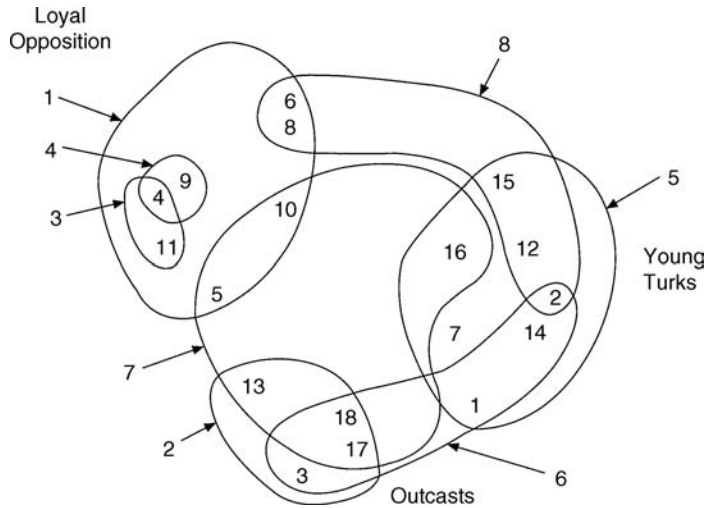


Figure 8.5 *Overlapping clusters of monks found by additive clustering, using* MAPCLUS*, superimposed on a multidimensional nonmetric scaling solution. Numbers with arrows indicate the rank of the clusters according to their weights; see also Table 8.5.* (Source: Arabie and Carroll, 1989.)

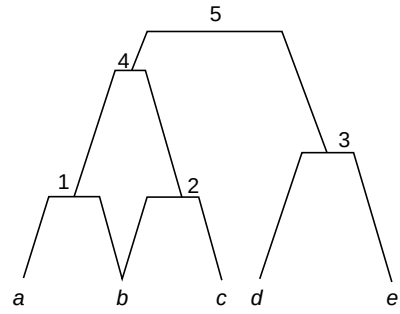


Figure 8.6 *Pyramidal representation with five classes. The heights of the nodes indicate the proximity of all the elements in a given class; b is an element which is on the border of {a, b} and {b, c} and forms an overlap; class 2 is a successor of class 4.* (Taken with permission of the publisher, Elsevier, from Aude et al., 1999.)

Before briefly outlining the algorithm for constructing a pyramid, two further relevant definitions are now given, beginning with the concepts of *order* (*before* and *after*) and *inclusion*.

Subset a is said to be *before* subset b if

$$(\min(a) < \min(b) \text{ and } \max(a) < \max(b)) \text{ or } a = b,$$

and b is said to be *included* in a if and only if

$$\min(a) < \min(b) \leq \max(b) < \max(a),$$

where for each class, min and max refer to the leftmost and rightmost singletons in the class, and $<$ ($>$) means 'to the left (right) of.

The *aggregation index* μ for a new class p, aggregated from classes a and b containing n_a and n_b objects respectively, with any other class q is

$$\mu(p, q) = [n_a \mu(a, q) + n_b \mu(b, q)/(n_a + n_b)]. \tag{8.10}$$

For singletons x and y,

$$\mu(x, y) = d(x, y). \tag{8.11}$$

The pyramid algorithm is initiated with an arbitrary order for the objects, and proceeds by aggregating classes p^* and q^* with the lowest aggregation index, subject to the following conditions:

- p^* is before q^*;
- either p^* or q^* is not included in a previously constructed class;
- if p^* and q^* belong to the same cluster they can be aggregated only if their intersection is not void;

- if p^* and q^* do not belong to the same cluster they can be aggregated only if p^* contains a border of $\min[C(p^*)]$; that is, if $\min[C(p^*)] = \min(p^*)$ or $\max[C(p^*)] = \max(p^*)$, and q^* contains a border of $C(q^*)$.

If, before aggregation, p^* and q^* do not belong to the same connected component, the objects in $C(q^*)$ are positioned before those of $C(p^*)$; this produces a compatible order. The process ends when all objects have been aggregated. Fitting pyramid structures with incomplete data is discussed by Gaul and Schader (1994).

8.4.5 Application of pyramid clustering to gene sequences of yeasts

Aude *et al.* (1999) give further details and a worked example of the algorithm outlined above, and a number of practical applications. One of these is concerned with 11 sequences from the yeast genome, which were analysed using average linkage (Figure 8.7(b)) and as a pyramid (Figure 8.7a)). The authors state that

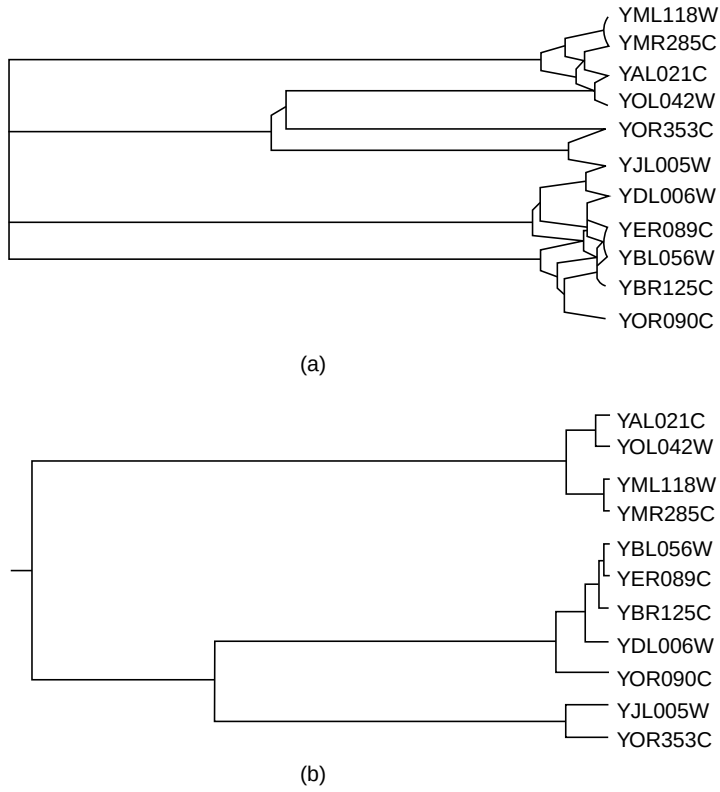


Figure 8.7 *(a) Pyramid and (b) average linkage representations of 11 sequences from the yeast genome.* (Taken with permission of the publisher, Elsevier, from Aude et al., 1999.)

'the complex relationship between these sequences ... are correctly mirrored by the pyramidal classification'. In particular, YJL005W and YAL021C are multi-domain proteins, a fact that is not reflected in the hierarchical classification but is reflected in the pyramid. The authors point out that the orders are not necessarily unique and this might pose problems for interpretation (see Figure 8.5, where four orders were compatible). However, they also point out that in practice it is unusual to find many different compatible orders.

8.5 Simultaneous clustering of objects and variables

In many applications, especially in the social and health sciences, the interpretation of cluster membership (objects) and of cluster characteristics (variables) is equally important. Clustering both simultaneously is known as *biclustering*. Methods that operate on a two-mode data matrix without recourse to a proximity matrix are known as *two-way, two-mode* or *direct clustering* methods. Such methods potentially provide more information than the constituent separate analyses, since they allow the interpretation of the (possibly overlapping) clusters of both objects and variables simultaneously, and also the associative relations between them.

One class of techniques reorders the rows (objects) and columns (variables) of the data matrix, and is sometimes known as *two-way joining*. Early clustering methods of this type include the *bond energy* approach, proposed by McCormick *et al.* (1972) and also discussed by Arabie and Hubert (1990). Here the bond energy of two matrix elements is defined as their product, and rows and columns are permuted, sequentially placing rows (columns) together according to their contribution to the total bond energy. The form of the data has to be carefully considered for valid operation of these reordering methods, and it may be necessary to scale data from different variables so that a comparable response is induced on every object–variable combination. Hartigan (1975) discusses this issue in relation to a number of direct data clustering algorithms. An example of a data set which is naturally in the correct form would be a subject $\times$ stimuli matrix, in which each entry measures the subject's liking for each stimulus; this can be regarded as a rectangular proximity matrix.

Another approach to direct clustering is to fit a tree structure (either ultrametric or additive) to data, rather than to proximity matrices derived from the data. De Soete *et al.* (1984a) have proposed a least-squares procedure for fitting tree structures to data matrices. The algorithm fits a matrix to the observed data such that it obeys a generalization of the ultrametric inequality to two-mode data. Standard clustering methods can then be used to obtain a dendrogram of both objects and variables. Espejo and Gaul (1986) have developed a two-mode variant of average linkage clustering, and De Sarbo (1982) has adapted the ADCLUS model (see Section 8.4.2) for two-mode data.

A recent increase in biclustering methods has resulted from the field of bioinformatics, especially gene expression data, where it may be required to cluster

both genes and samples simultaneously. A review of two-mode clustering methods has been published by Van Mechelen *et al.* (2004), who give references to applications, and information on software from the field of bioinformatics. Prelić *et al.* (2006) have reviewed a number of biclustering methods for such data, and compared them on real and simulated data sets. They concluded that biclustering was generally preferable to conventional hierarchical methods. A number of the more recent algorithms have been included in an `R` package, `Biclust` (Kaiser and Leisch, 2008).

We now describe in more detail two methods for direct clustering of data matrices. The first is the *hierarchical classes* method, which is appropriate for binary data and is an example of a matrix reordering technique. The second is the *error variance* technique, which is less restrictive than the hierarchical classes method in that the data need not be binary, and is an example of a standard hierarchical method applied to a two-mode matrix. Data must be scored or normalized so that larger entries indicate stronger relationships between the corresponding row and column elements.

8.5.1 Hierarchical classes

The *hierarchical classes* method of De Boeck and Rosenberg (1988) is appropriate for binary attribute data. Two hierarchical class structures are defined, one for objects and one for variables, by reordering the matrix. The objects are first grouped into classes with identical attributes, and the classes are then ordered to reflect subset/superset relations. This is repeated for the attributes, in terms of the objects. The object classes are sets of objects having identical attributes, and attribute classes are similarly defined; classes with no objects (attributes) are termed 'undefined'. The hierarchy is defined on the basis of subset/superset relations and may overlap.

Table 8.6 shows a (hypothetical) simple data set concerned with an individual's perception of self and others (Rosenberg *et al.*, 1996). In this, a particular person describes eight 'targets' (objects) using eight traits (attributes). The corresponding hierarchical classes are shown in Figure 8.8, the two hierarchical structures being

Table 8.6 Hypothetical matrix of a person's perception of targets (objects) by traits (attributes).

	Successful	Articulate	Generous	Outgoing	Hardworking	Loving	Warm	Shy
Father	1	1	0	0	1	0	0	0
Boyfriend	0	0	1	0	1	1	1	0
Uncle	0	0	0	1	0	1	1	0
Brother	0	0	0	1	0	1	1	0
Mother	0	0	1	1	1	1	1	0
Me Now	0	0	1	1	1	1	1	0
Ideal Me	1	1	1	1	1	1	1	0
Casual Friend	0	0	0	0	0	0	0	0

Source: Rosenberg *et al.* (1996).

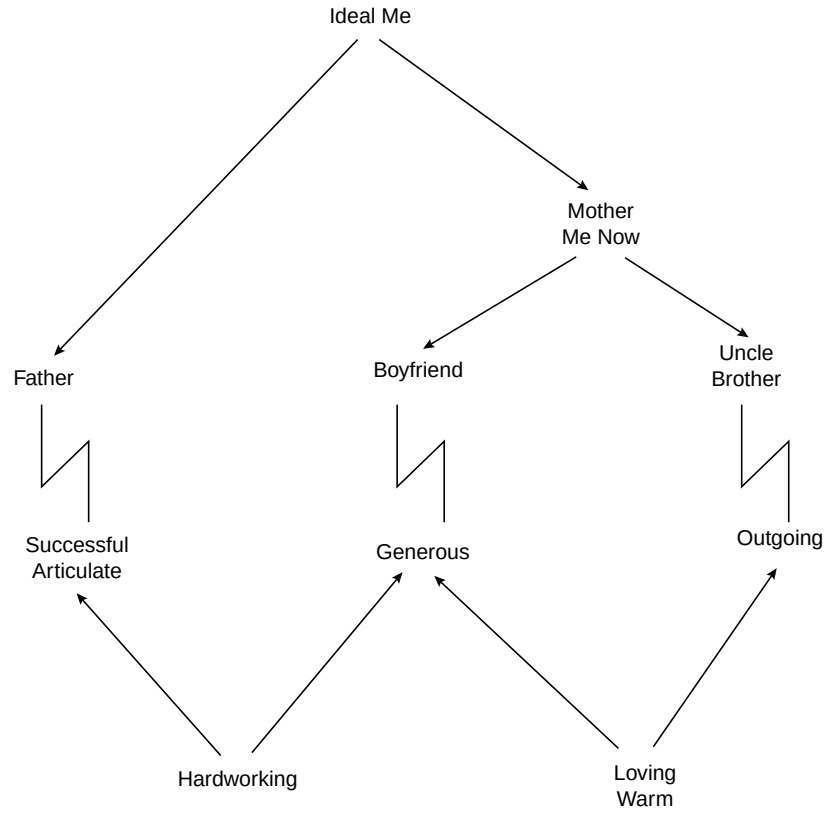


Figure 8.8 *Hierarchical classes representation of hypothetical data in Table 8.6.* (Source: Rosenberg et al., 1996.)

the upper and lower halves of the figure. The object *bundles* are the sets of objects associated with the same attributes (i.e. they are linked by a zigzag line), and similarly for the attribute bundles. All the objects in a bundle are associated with the same attribute bundle and vice versa. An example is the 'Ideal Me' and 'Father' object bundle, which is associated with the 'Hardworking', 'Successful' and 'Articulate' attribute bundle.

One hopes that, in fitting this model to real data, a small number of low-level clusters (bundles) will be found in which all members have exactly the same attributes. In practice, however, this ideal structure is usually impossible to achieve and there will be discrepancies between the model and the data.

The algorithm operates by fitting a structure to data such that the number of discrepancies between the structure obtained and the data is a minimum, using a Jaccard coefficient as the goodness-of-fit measure (see Chapter 3). The rank of

the model (the number of lowest-level clusters) is indicated by the number of zigzag lines in Figure 8.8. It has to be chosen as a trade-off with the goodness of fit, and typically the investigator chooses a rank where the goodness of fit changes little in comparison with the previous increases. In order to avoid local minima in the solution, an initial ordering of either rows or columns is found (e.g. through a conventional cluster analysis on the objects).

Leenen *et al.* (2008) have developed a new stochastic extension to the HICLAS model using Bayesian estimation. According to the authors, 'the benefit of the new extension is that the relation between the predicted values and the observed values is made explicit thanks to the inclusion of one or two error-probability parameters'. In addition to its other advantages over the original deterministic HICLAS algorithm, such as providing model checking and selection criteria, this model-based approach also potentially allows the fixing of any partial or full ordering which may be known *a priori.*

8.5.2 Application of hierarchical classes to psychiatric symptoms

In an application in psychiatry, Gara *et al.* (1998) described 1455 patients using primary care facilities at a community clinic in terms of 41 symptoms. These were grouped into eight body/organ systems (pseudoneurological (PN), gastro-intestinal (GI), genitourinary (GU), musculoskeletal (MS), female reproductive urinary (FR), cardiorespiratory (CR), headache (H) and other pain and skin). A hierarchical classes model was fitted, giving a good overall fit to the data ($\kappa = 0.73$). A Jaccard coefficient for comparing clusters and symptoms was also calculated: for example, the coefficient for blurred vision compared to its associated cluster was 0.387. Most coefficients were greater than 0.7, but pseudoneurological and skin complaints (which were rare) did not fit well. Figure 8.9 shows the symptom clusters. The bottom line of the diagram consists of symptom clusters which can be identified with patient clusters (e.g. F has patients with exclusively cardiorespiratory symptoms). Clusters A–E are super-sets describing patients with combinations of symptoms (e.g. patients in A have all symptom types). Note that the hierarchy applies to the symptoms (i.e. it is the attribute half of the model).

The authors discuss the relationship of this analysis to the results of a grade-of-membership analysis (a type of *fuzzy analysis:* see Section 8.7.1) on similar data (Swartz *et al*., 1986). They found a general convergence in the result, but preferred the hierarchical classes analysis as it was able to highlight the role of pseudoneurological symptoms, which always co-occur with all the other symptoms in the superset A, and were thus interpreted as evidence for somatization.

8.5.3 The error variance technique

A technique which combines features of both the direct clustering and tree fitting methods is the *error variance* approach of Eckes and Orlik (1993). The basic

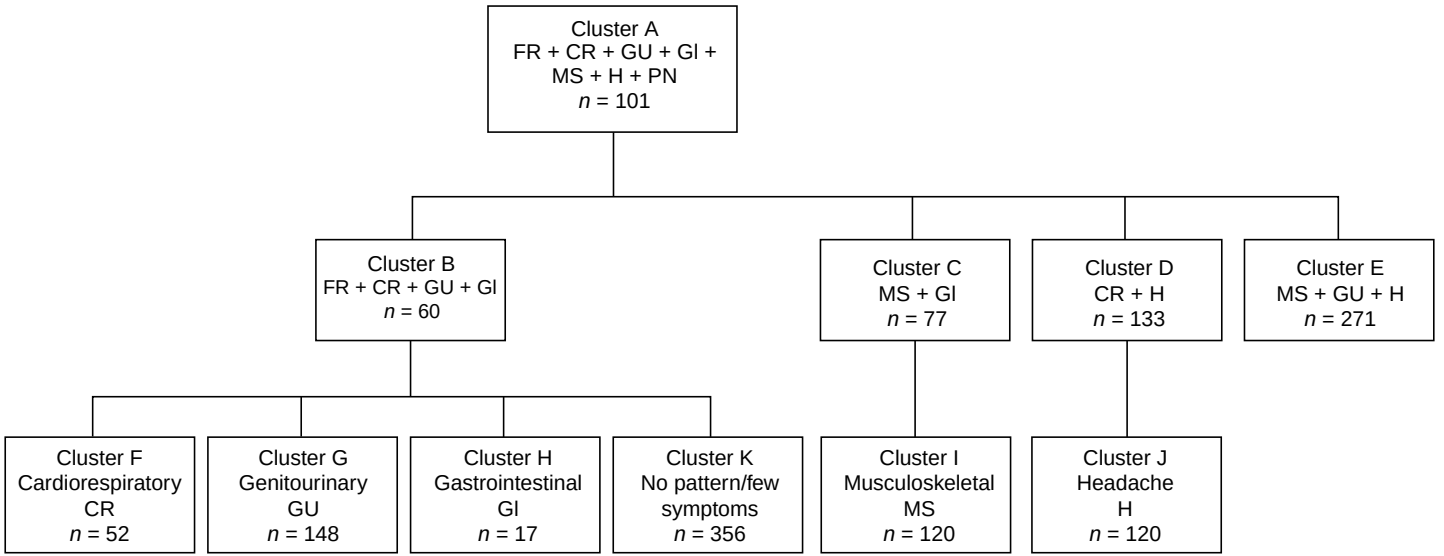


Figure 8.9 *Hierarchical classes analysis of 41 symptoms reported by 1455 patients using primary care facilities at a community clinic: pseudoneurological (PN), gastrointestinal (GI), genitourinary (GU), musculoskeletal (MS), female reproductive (FR), cardiorespiratory (CR), headache (H).* (Taken with permission of the publisher, Elsevier, from Gara et al., 1998.)

method is hierarchical and produces a dendrogram, although a modification allows overlapping clusters to be obtained. The method combines elements (either rows or columns or clusters obtained from them) in such a way as to minimize the increase in the internal heterogeneity of a two-mode cluster (the mean squared deviation or MSD), in a manner analogous to Ward's method for one-mode data (see Chapter 4).

Given a data matrix $\mathbf{X}$, the MSD for a cluster is the mean squared deviation of entries x_{tj} in the corresponding submatrix $\mathbf{X}_r$, from the maximum entry m in $\mathbf{X}$. The procedure involves merging one-mode clusters (either objects or variables) or two-mode clusters (containing both objects and variables), at each stage merging those with the smallest increase in MSD. In calculating this, only those cells representing new combinations are counted. This is illustrated in the example below, where, for example, in step 6, A1B2, A1B3, A2B2, A2B3, A3B1 and A4B1 are new combinations.

The MSD can be written as

$$MSD_r = s_r^2 + (\bar{x}_r - m)^2, \tag{8.12}$$

where s_r^2 is the variance of data in the rth cluster, and $\bar{x}_r$ is the mean. A measure of cluster cohesion is the *centroid effect ratio* (CER):

$$CER_r = \bar{x}_r^2/(s_r^2 + \bar{x}_r^2). \tag{8.13}$$

The CER reflects the contribution of the 'mean cluster effect size' to the 'total cluster effect size'. Clusters with a low CER have a low cohesion. If the CER is less than about 80%, the authors suggest excluding the cluster from further consideration.

The procedure is now illustrated on the data in Table 8.7. The steps in the procedure are as follows:

Step	CER	Two-mode cluster	Increase in MSD
1	1.00	{A2, B1}	$0 = (6.8 - 6.8)^2$
2	1.00	{A3, B3}	$0.49 = (6.1 - 6.8)^2$
3	1.00	{A4, B2}	$1.00 = (5.8 - 6.8)^2$
4	0.94	{A1, A2, B1}	$7.29 = (4.1 - 6.8)^2$
5	0.93	{A3, A4, B2, B3}	$11.60 = [(3.6 - 6.8)^2 + (3.2 - 6.8)^2]/2$
6	0.83	{A1, A2, A3, B1, B2, B3}	$19.63 = [(3 - 6.8)^2 + (3.1 - 6.8)^2 + (2.3 - 6.8)^2 + (2.4 - 6.8)^2 + (1.9 - 6.8)^2 + (1.7 - 6.8)^2]/6$

This example is strictly hierarchical, but overlapping clusters can be obtained by adding elements (rows or columns) to an existing cluster such that the increase in MSDs is below a threshold, and again a resulting CER of 80% would be a lower limit for including a new element.

Table 8.7 Hypothetical data to illustrate direct clustering of data matrices using the error variance method.

Object	Variable		
	B1	B2	B3
A1	4.1	1.9	3.1
A2	6.8	2.3	2.4
A3	1.7	3.6	6.1
A4	3.0	5.8	3.2

8.5.4 Application of the error variance technique to appropriateness of behaviour data

Eckes and Orlik (1993) give two examples of the application of this technique. One is concerned with brand-switching between soft drinks. The other examines the perceived appropriateness of various types of behaviour in different situations, rated by 52 people; the data had previously been analysed using two separate cluster analyses, one for variables and one for objects (Price and Bouffard, 1974), and are given in Table 8.8. Before clustering, each behaviour element was duplicated and multiplied by -1 to form a 15×30 matrix, so that inappropriateness as well as appropriateness of behaviours would be represented. The resulting, not altogether surprising, clusters of behaviours and situations are shown in Table 8.9. Elements have been added to the pre-existing clusters to form overlapping clusters, and the results show that 'job interview' allows for only one appropriate behaviour (talk), whereas 'own room' allows for any of the behaviours. The previous hierarchical analyses were not considered useful because they provided only separate classifications of the situations and behaviours, with no link between them.

8.6 Clustering with constraints

Clustering with constraints is necessary when the membership of clusters is to be restricted in some way, and often occurs when objects and clusters need to retain their spatial relationships. This situation is commonly encountered in geographical or image processing applications. Semi-supervised clustering methods should be mentioned as an important development which involves the use of constraints. To give just one example, in internet clustering of documents, users can feed back their response to the relevance of documents presented to them by a search engine (see Cohn, Caruana and McCallum, Chapter 2 *Semi-supervised clustering with user feedback* in Basu *et al.*, 2009). A recent wide-ranging review by Basu *et al.* (2008) gives the theoretical background and many examples of constrained clustering.

While spatial constraints are usually two-dimensional (or occasionally three- or four-dimensional if time is included), one-dimensional constraints arise in

Table 8.8 Perceived appropriateness of various types of behaviour in different situations, rated by 52 people[a].

Situation	Run	Talk	Kiss	Write	Eat	Sleep	Mumble	Read	Fight	Belch	Argue	Jump	Cry	Laugh	Shout
Class	2.52	6.21	2.10	8.17	4.23	3.60	3.62	7.27	1.21	1.77	5.33	1.79	2.21	6.23	1.94
Date	5.00	8.56	8.73	3.62	7.79	3.77	3.12	2.88	3.58	2.23	4.50	4.42	3.04	8.00	3.79
Bus	1.44	8.08	4.27	4.87	5.48	7.04	5.17	7.17	1.52	2.15	4.17	3.12	3.08	7.10	3.00
Family dinner	2.56	8.52	4.92	2.58	8.44	2.29	2.54	3.96	1.67	2.50	3.25	2.29	3.21	7.13	1.96
Park	7.94	8.42	7.71	7.00	8.13	5.63	5.40	7.77	3.06	5.00	5.06	7.42	5.21	8.10	6.92
Church	1.38	3.29	2.38	2.85	1.38	1.77	3.52	3.58	0.62	1.42	1.92	1.71	3.13	2.60	1.33
Job interview	1.94	8.46	1.08	4.85	1.73	0.75	1.31	2.48	1.04	1.21	1.83	1.48	1.37	5.88	1.65
Sidewalk	5.58	8.19	4.75	3.39	4.83	1.46	4.96	4.81	1.46	2.81	4.08	3.54	3.71	7.40	4.88
Movies	2.46	4.98	6.21	2.73	7.48	4.08	4.13	1.73	1.37	2.58	1.71	2.31	7.15	7.94	2.42
Bar	1.96	8.25	5.17	5.38	7.67	2.90	6.21	4.71	1.90	5.04	4.31	3.75	3.44	8.23	4.13
Elevator	1.63	7.40	4.79	3.04	5.10	1.31	5.12	4.48	1.58	2.54	2.58	2.12	3.48	6.77	1.73
Restroom	2.83	7.25	2.81	3.46	2.35	2.83	5.04	4.75	1.77	5.12	3.48	3.65	4.79	5.90	3.52
Own room	6.15	8.58	8.52	8.29	7.94	8.85	7.67	8.58	4.23	6.81	7.52	6.73	8.00	8.17	6.44
Dorm lounge	4.40	7.88	6.54	7.73	7.19	6.08	5.50	8.56	2.40	4.00	4.88	4.58	3.88	7.75	3.60
Football game	4.12	8.08	5.08	4.56	8.04	2.98	5.23	3.69	2.04	3.85	4.98	7.12	4.31	7.90	7.94

[a]The higher the score, the more appropriate the behaviour in the situation.
Source:Price and Bouffard (1974).

Table 8.9 Clusters of behaviours and situations from data in Table 8.8. Elements have been added to the pre-existing clusters to form overlapping clusters.

Cluster	Original elements (non-overlapping)		Added elements (overlapping)	
	Behaviours	Situations	Behaviours	Situations
A	*Fight, run*	Church, class, sidewalk, elevator, restroom, bus		Job interview, bar, movies, family dinner
B	*Sleep, kiss, belch, mumble, cry, jump, shout, eat, argue, read*	Job interview	Talk, *fight, run*	Church
C	Laugh, eat, kiss	Bar, movies, dorm lounge, family dinner, park, football, date	Talk, *fight*	Own room, sidewalk, bus, elevator
D	Sleep, talk, read, write, cry, mumble, argue, belch, jump, shout, run	Own room	Kiss, *laugh, eat*	Park

Note: behaviours in italics are considered inappropriate in the respective situations.

stratigraphy, in areas such as archaeology and geology. Applications that require one-dimensional temporal constraints include the indexing and retrieval of multimedia documents. These may contain excerpts from audio or video tapes, for example scenes from films, annotated with text describing their features. Clustering these annotations should maintain their correct temporal relationship (see Yeung *et al.*, 1996).

In some situations, constraints may not be spatially or temporally defined. In the globalization example in Section 4.5.3, cities were clustered according to measures of economic activity. If the purpose of the clustering had been, say, to locate a new company, the existence of good transport links might have been an appropriate constraint rather than geography. Constraints can also be applied, not to the clustering, but to the choice of cluster representative. Girgensohn and Boreczky (2000), for example, clustered videos of meetings in an unconstrained manner, but *keyframes* (frames considered to represent the clusters) were chosen to satisfy various constraints, for example to produce a fixed number of keyframes with an approximately even distribution in time.

Model-based methods can be adapted readily to the incorporation of constraints. Tree fitting methods, both ultrametric and additive, were briefly mentioned in Section 8.5, and the topologies of such trees (see Section 4.4.1) can be constrained as discussed by De Soete and Carroll (1996). These authors give an example of constraining an (additive) kinship tree so that lineal (e.g. father–son) and collateral (e.g. uncle–nephew) relationships are maintained in the appropriate paths. With spatial constraints, the actual distance between two objects can be incorporated into the metric computed from the clustering variables (see Jain and

Farrokhina, 1991, for example). One of the standard hierarchical or partitioning methods can then be employed, using this modified proximity matrix. A similar approach has been described for the DBSCAN algorithm. The difficulty with this approach is to weight the external contribution to the proximity with that defined on the basis of the clustering variables. A more commonly used method, especially for spatial constraints, involves the concept of contiguity, where a matrix is defined whose elements are 1 if two objects are contiguous and 0 otherwise. The contiguity matrix approach is now described in more detail.

8.6.1 Contiguity constraints

In contiguity-constrained clustering, standard distance measures and clustering methods can be employed, but the optimization procedure is constrained on the basis of the contiguity matrix. This is a binary matrix indicating which units are contiguous, most easily defined if the units to be clustered have a natural boundary in space such as an administrative department. Otherwise an unambiguous definition based on spatial relationships is used, such as a contiguity graph, for example the *Voronoi diagram.* In this, the plane in which the units to be clustered lie is subdivided into polygonal regions such that the points in each region are closer to the candidate points than to any other point. Points whose regions share a boundary are said to be contiguous. A mathematically equivalent representation of contiguity is *Delaunay triangulation,* in which contiguous points are joined by line segments.

Where there are very large numbers of candidate points, the spatial field in which the objects lie may be divided into units using a regular grid, for example pixels in image processing. In this situation point B is contiguous with point A if it is situated in one of the four or eight units which surround B. Figure 8.10 shows the Voronoi and grid definitions of contiguity.

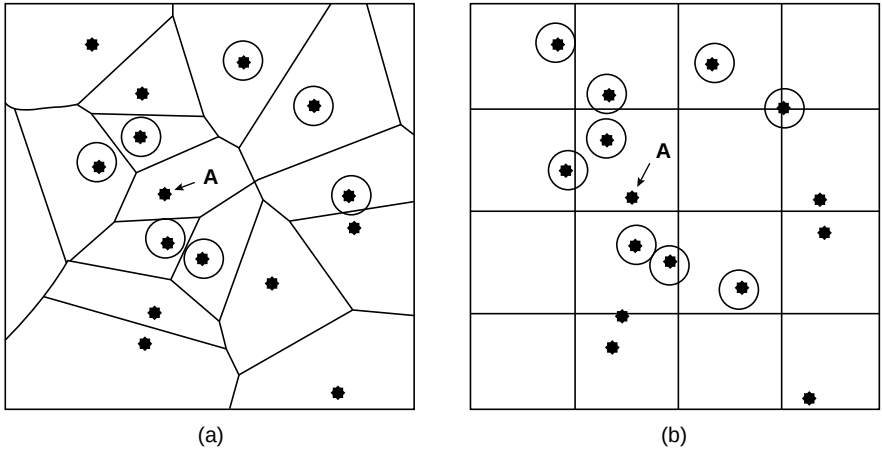


Figure 8.10 *Definitions of contiguity with object A, defined in terms of (a) the Voronoi diagram and (b) the eight surrounding cells in a grid. Circled points are contiguous with A.*

Contiguity graphs, and applications of constrained clustering making use of them, are described by Gordon (1999).

Once an appropriate contiguity matrix has been defined, standard partitioning or hierarchical methods can be applied with the appropriate modification to the algorithm so as to ensure that only contiguous points are clustered. Contiguity constraints can be applied, for example, in the package `Clustangraphics`. Maravalle *et al.* (1997) discuss computational issues in contiguity-constrained hierarchical clustering. The modified methods do not necessarily retain properties of their parent method, such as avoiding inversions (see Section 4.4.3). Ferligoj and Batagelj (1982) discuss constrained hierarchical methods and show that, of widely used methods, only constrained complete linkage does not give rise to inversions. More recent work by Murtagh (1995) also considered inversions, and showed that constrained single linkage also avoids them, so long as the *single* objects that are linked are contiguous (a technique known as *contiguity-constrained single link*). Two methods also discussed by Murtagh are the constrained centroid and minimum variance methods. These can produce inversions but have the advantage that they give rise to natural regional representatives (the exemplars or typical members).

One-dimensional constraints, where clustering is required to follow a given order, can be treated with special methods, as developed by Gordon (1973) for an application concerning palaeoecological samples from a vertical bore. Partitions of contiguous samples can be defined by (virtual) 'markers' placed between neighbouring samples; the number of possible placements and hence possible partitions is $(n-1)!/[(g-1)!(n-g)!]$, where g is the number of groups, and n the number of objects. Gordon (1980) suggests two approaches to finding the optimal partition. The first uses a divisive algorithm that begins by finding a single marker leading to minimum within-group sums of squares. Each of the two groups is then optimally divided, choosing the division that leads to maximum decrease in the sum of squares. The algorithm continues by successive division of existing groups.

The second procedure described by Gordon (1980) involves a dynamic programming algorithm. Let $s(i, j)$ denote the within-group sum of squares of objects i to j inclusive, and let $t(g, k)$ denote the total within-group sum of squares when objects 1 to k are optimally divided into g groups. We require $t(g, n)$ for $2 \leq g \leq n$, together with the corresponding markers. The solution is built up recursively, evaluating $\{t(g, k), k = g, g + 1, \ldots, n; g = 1, 2, \ldots, n\}$ by means of the following formulae:

$$\begin{aligned} t(1,k) &= s(1,k)(1 \leq k \leq n) \\ t(g,k) &= \min_{g-1 \leq i \leq k-1} [t(g-1,i) + s(i+1,k)](g \leq k \leq n; 2 \leq g \leq n). \end{aligned} \tag{8.14}$$

Equation (8.14) involves dividing the first k objects into two classes, the first class containing $g-1$ groups and the second class containing 1 group of objects. This is equivalent to placing the last marker at position i, and the algorithm finds the

optimal value of i. The complete set of $g-1$ markers is obtained using a *trace-back* procedure.

8.6.2 Application of contiguity-constrained clustering

An example of the use of a contiguity constraint in clustering is given in a study of the incidence of breast cancer in Argentina. Wojdyla *et al.* (1996) clustered administrative departments of Argentina to form regions, using a variant of the procedure described by Ferligoj and Batagelj (1982). Contiguous departments were clustered together according to their Euclidean distance, as computed from sociodemographic variables, with contiguity defined on the basis of shared administrative boundaries. Clustering was terminated, not according to a standard clustering criterion, but when regions had attained the minimum population regarded as sufficient to make valid statistical inferences regarding rates of breast cancer. Standardized mortality rates from breast cancer for each region are shown in Figure 8.11(a), and the regions found by clustering are shown in Figure 8.11(b). The authors concluded that, on the basis of the aggregated data, only two regions (Rosario and Córdoba, relatively underdeveloped regions) had significantly higher rates than the national rate. This is in contrast with the more irregular picture obtained before clustering.

8.7 Fuzzy clustering

In *fuzzy clustering*, objects are not assigned to a particular cluster: they possess a membership function indicating the *strength of membership* in all or some of the clusters. In most of the previous clustering techniques described in this text, 'strength of membership' has been either zero or one, with an object being either in or not in a cluster, except perhaps in the case of the mixture approach of Chapter 6, where it might be taken as the posterior probability of belonging to a cluster. In fuzzy clustering jargon, methods where strength of membership is zero or one are known as *crisp* methods.

Fuzzy clustering has two main advantages over crisp methods. Firstly, memberships can be combined with other information. In particular, in the special case where memberships are probabilities, results can be combined from different sources using Bayes' theorem. Secondly, the memberships for any given object indicate whether there is a 'second best' cluster that is almost as good as the 'best' cluster, a phenomenon which is often hidden when using other clustering techniques.

In a fuzzy cluster analysis, the number of subsets is assumed known, and the membership function of each object in each cluster is estimated using an iterative method, usually a standard optimization technique based on a heuristic objective function. In general, membership functions do not obey the rules of probability theory, although, once found, memberships can be scaled to lie between zero and one, and can then be interpreted as probabilities. (Mixtures methods, where

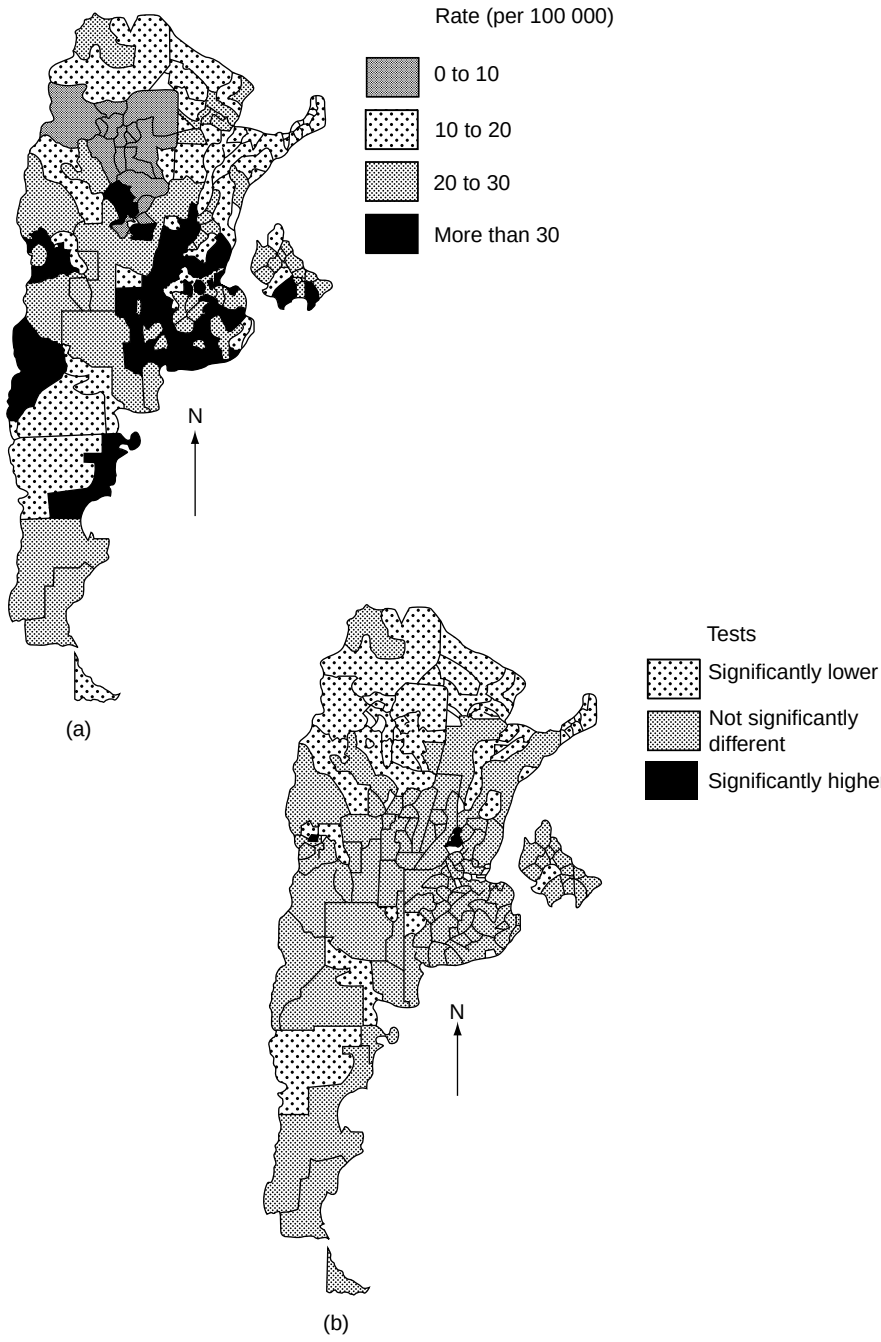


Figure 8.11 *Standardized mortality rates from breast cancer in the departments and regions of Argentina, (a) before and (b) after constrained clustering.* (Taken with permission of the publisher, John Wiley & Sons Ltd, from Wojdyla et al., 1996.)

the memberships *are* true probabilities, and where probability theory underlies the estimation method, have been discussed in Chapter 6.)

The concept of a membership function derives from *fuzzy logic*, an extension of Boolean logic in which the concepts of true and false are replaced by that of partial truth. Boolean logic can be represented by set theory, and in an analogous manner fuzzy logic is represented by fuzzy set theory. Such techniques were originally developed for the description of natural language (Zadeh, 1965). As an example of a fuzzy membership function, Figure 8.12 shows a possible function for the description of IQ.

The connection between fuzzy cluster analysis and fuzzy logic is usually only through the application of membership functions, and not the more comprehensive theory. However, an example in which the principles of fuzzy logic (as well as the membership functions) are used to derive a clustering algorithm is given by Zhang *et al.* (1998) in an application to a small data set concerned with monitoring mechanical equipment. Laviolette *et al.* (1995), along with a number of discussants in the same volume, compare fuzzy and probabilistic approaches in general, and among these contributions is a discussion of fuzzy cluster analysis (Rousseeuw, 1995).

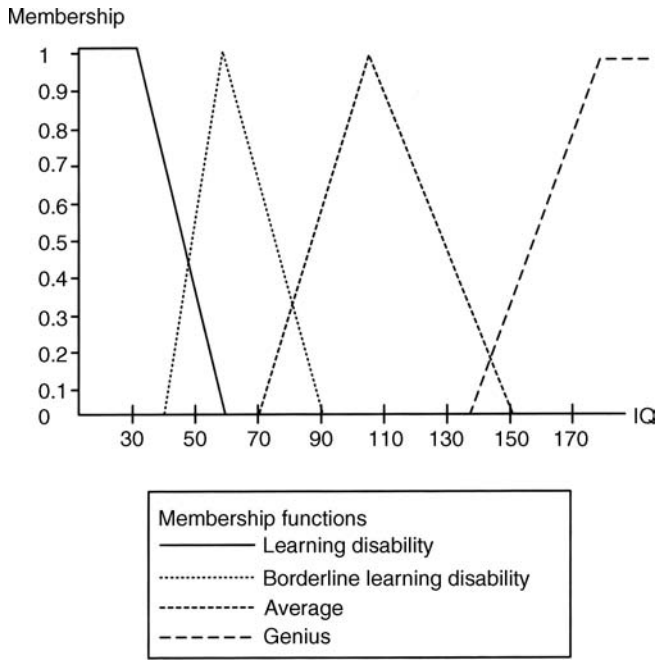


Figure 8.12 *A fuzzy membership function for the verbal description of intelligence quotient (IQ); an example is IQ 85, which has memberships of about 0.4 and 0.2 in the sets 'average' and 'borderline learning disability', respectively.*

The following subsection gives examples of three heuristic fuzzy methods: for binary data, for continuous data and for a proximity matrix.

8.7.1 Methods for fuzzy cluster analysis

Grade-of-membership (GOM) analysis (Woodbury and Manton, 1982; Manton *et al*., 2004) has been proposed for binary data, and has been used in health-related applications, for example psychiatry and genetics. Grade-of-membership analysis assigns to cases a 'grade of membership' in two or more latent classes, and is thus similar to latent class analysis (see Chapter 6). In GOM and latent class analysis, probabilities of cluster membership are provided and, in that sense, both methods are fuzzy. The difference is that in GOM analysis, the grades of membership are estimated as parameters as part of the clustering process, whereas in latent class analysis the probabilities of class membership are estimated once the latent class model has been estimated. Further details are given in Woodbury *et al.* (1994), and a recent application profiling doctors' practice styles in pain management is given by Maelzel *et al.* (2000).

For continuous data, a weighted sum-of-squares criterion leading to *fuzzy k*-means (also known as fuzzy *c*-means), clustering has been described by Bezdek (1974). For a set of n objects and g clusters this is, for data vectors $\mathbf{x}_i$,

$$\sum_{t=1}^{g}\sum_{i=1}^{n} u_{it}^{v}\, d^2(\mathbf{x}_i, \mathbf{m}_t), \tag{8.15}$$

where $\mathbf{m}_t$ is the centre of cluster t, $u_{it} \geq 0$ for all $i = 1,\ldots, n$ and $\Sigma_{t=1}^{g} u_{it} = 1$. The memberships u_{it} are unknown; the $d(\mathbf{x}_i, \mathbf{m}_t)$ are Euclidean distances between the data point and the cluster centres; v is called the *fuzzifier* and affects the final membership distribution; typically it is 2 (setting $v = 1$ leads to the crisp solution). The cluster centres are weighted cluster means, given for the kth variable by

$$m_t = \sum_{i=1}^{n} u_{it}^{v}\mathbf{x}_i \Big/ \sum_{i=1}^{n} u_{it}^{v}. \tag{8.16}$$

The clustering algorithm computes the optimal memberships by minimizing (8.15); see, for example, Hathaway and Bezdek (1988) for details of algorithms. If the u_{it} are restricted to zero or one, the usual k-means method is obtained (see Chapter 5). Kettenring (2009) illustrates the use of fuzzy k-means in the rapid valuation of portfolio assets in a study of the use of cluster analysis in patents. Nasibov and Ulutagay (2010) compare fuzzy k-means with fuzzy neighbourhood DBSCAN, a fuzzy version of DBSCAN, described earlier in Section 8.3.

As in traditional agglomerative or optimization methods, various other choices of d can be made, such as city block or Mahalanobis distance. Kaufman and Rousseeuw (2005) describe a method (called 'FANNY') in which the

following objective function is minimized:

$$\sum_{t=1}^{k}\left\{\sum_{i,j=1}^{n} u_{it}^2 u_{jt}^2 d(\mathbf{x}_i,\mathbf{x}_j)\right\Big/ 2\sum_{j=1}^{n} u_{jt}^2\right\}, \tag{8.17}$$

where $d(\mathbf{x}_i,\mathbf{x}_j)$ are dissimilarities between objects. Note that in this case the means do not enter the objective function, and are not squared. For this method only a proximity matrix is required since there is no need to estimate central values for the clusters. Furthermore, the inclusion of $d(\mathbf{x}_i,\mathbf{x}_j)$ rather than $d(\mathbf{x}_i,\mathbf{x}_j)^2$ means that the method is relatively robust to nonspherical clusters.

8.7.2 The assessment of fuzzy clustering

The *silhouette plot* (Rousseeuw, 1987), an example of which is given in Figure 5.5, is useful in connection with partitioning methods in general, but particularly so in the context of fuzzy clustering. It reflects the strength of a classification to the nearest crisp cluster, compared to the next best cluster. The width of each bar is the 'silhouette value', which is one if the object is well classified, zero if it is intermediate between the best and second best, and negative if it is nearest to the second-best cluster (see also Section 5.5).

Dunn's partition coefficient (Dunn, 1974) is a criterion for assessing the strength of membership specifically designed for fuzzy methods. When normalized to lie in the range [0,1], it has the form

$$\left(k\sum_{i=1}^{n}\sum_{t=1}^{k}\frac{u_{it}^2}{n}\right)\Big/(k-1), \tag{8.18}$$

and is equal to 1 for completely distinct clustering.

Dunn's coefficient and silhouette plots give information to allow a number of clusters to be chosen so that a balance can be struck in the degree of fuzziness in different clusters. However, like other internal methods of cluster validation, they do not provide a definitive guide to the number of clusters, and this is a subject of continuing research (Pal and Bezdek, 1995). It is usually advisable to use a low-dimensional plot such as principal components analysis in addition to the silhouette plots in order to assess the degree of fuzziness present in the data.

8.7.3 Application of fuzzy cluster analysis to Roman glass composition

The chemical composition of Roman glass found in Norway has been studied by Christie *et al.* (1979) with a view to identifying its origins. A subset of the data was reanalysed by Baxter (1994) in order to illustrate the results obtained by four standard methods of hierarchical cluster analysis. Baxter concluded that there were probably two main clusters, possibly with one outlier. This was also concluded by Christie *et al.* and is suggested by the principal components plot in Figure 8.13,

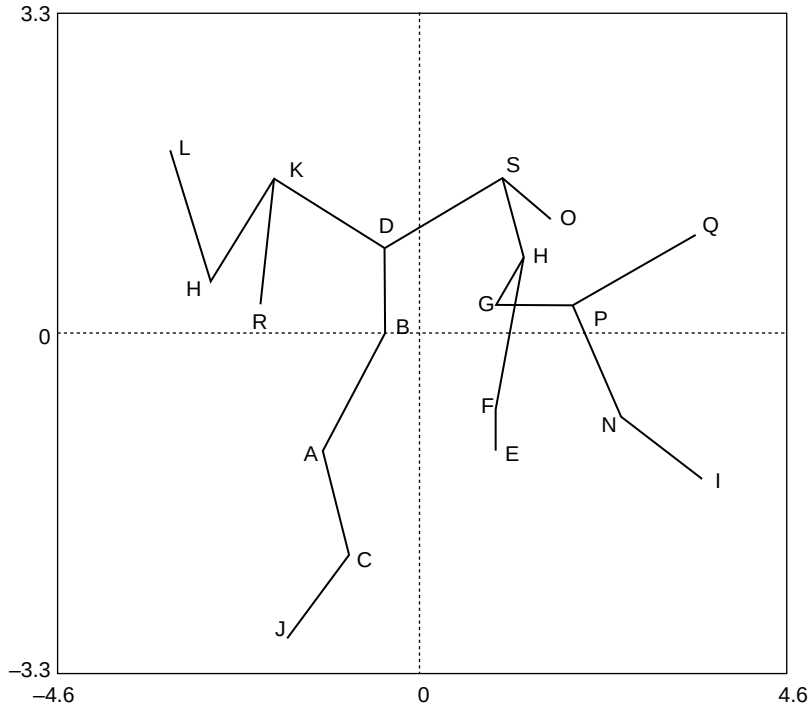


Figure 8.13 *Principal components plot of Roman glass composition (standardized to unit standard deviation), showing the minimum spanning tree; note the intermediate position of D, between two main clusters (see also Table 8.10).* (Source: Baxter, 1994.)

where there seem to be two 'clouds' of points, although the two main clusters are not very distinct. However, one point (object D) is intermediate between the two, and links the clusters in the minimum spanning tree.

The data, which are the percentage compositions of eight oxides in 19 specimens, are shown in Table 8.10, along with the average linkage clusters suggested by Baxter, the nearest crisp clusters in the six- and two-cluster case, and the fuzzy memberships for the two-cluster solution. The method used was that proposed by Kaufman and Rousseeuw (see Section 8.7.1). The two-cluster solution is shown in the silhouette plot (Figure 8.14).

The normalized Dunn coefficient is 0.06 for the two-cluster solution, which is close to zero, indicating a very high fuzziness. The silhouette plot shows that cluster 2 is less fuzzy (average width 0.41) than cluster 1 (average width 0.20) and that object D has a negative width indicating a bad classification; this is due to its intermediate position between the two clusters. If a six-cluster solution is chosen, the average width drops, and while cluster 3 is more compact, clusters 1 and 2 are more dispersed. Two singleton clusters J and L now become apparent and the Dunn coefficient is higher (0.18). The nearest crisp solutions to the fuzzy two-, three- and

Table 8.10 Data and fuzzy cluster analysis of Roman glass composition. Data are percentages of eight oxides.

Specimen	Ti	Al	Fe	Mn	Mg	Ca	Na	K	Fuzzy 6-cluster	Average linkage 2-cluster	Fuzzy 2-cluster	Memberships for fuzzy 2-cluster solution (%)	
A	0.10	2.0	0.8	1.5	1.18	6.3	18.0	0.58	1	1	1	58	42
B	0.10	2.0	0.5	1.4	1.16	6.4	18.4	0.43	2	1	1	56	44
C	0.10	2.0	1.0	1.2	0.77	7.0	19.0	0.61	2	1	1	55	45
D	0.20	2.0	0.7	1.2	0.90	6.1	19.3	0.36	2	2	1	56	44
E	0.09	1.8	0.95	1.0	0.70	6.2	16.2	0.45	3	2	2	42	58
F	0.09	1.8	1.1	0.9	0.68	6.0	16.1	0.44	3	2	2	43	57
G	0.08	1.7	0.6	1.4	0.71	6.35	17.6	0.37	4	2	2	34	66
H	0.08	1.7	0.6	1.3	0.70	6.2	17.2	0.32	4	2	2	32	68
I	0.05	1.5	0.2	0.02	0.53	6.2	18.9	0.45	2	2	2	43	57
J	0.30	1.8	1.0	1.4	1.01	8.8	18.1	0.53	5	0[a]	1	54	46
K	0.30	2.2	1.0	1.9	1.06	6.2	18.6	0.34	1	1	1	64	35
L	0.35	2.8	1.2	2.0	0.96	5.9	18.5	0.37	6	1	1	62	38
M	0.30	2.5	1.0	2.0	0.96	6.7	18.5	0.41	1	1	1	65	35
N	0.07	1.5	0.45	0.95	0.58	6.85	17.5	0.35	4	2	2	35	65
O	0.07	1.5	0.45	1.0	0.78	6.25	19.4	0.27	2	2	2	43	57
P	0.08	1.6	0.5	1.1	0.65	6.2	17.5	0.37	4	2	2	29	71
Q	0.06	1.3	0.3	0.85	0.50	5.9	16.8	0.29	4	2	2	37	63
R	0.35	2.2	1.0	1.5	1.20	6.5	18.0	0.40	1	1	1	65	35
S	0.07	2.0	0.4	1.2	0.80	6.0	18.0	0.30	2	2	2	40	60

[a]Interpreted as an outlier in the average linkage solution.

Source: Baxter (1994); original compositional data from Jackson (1994).

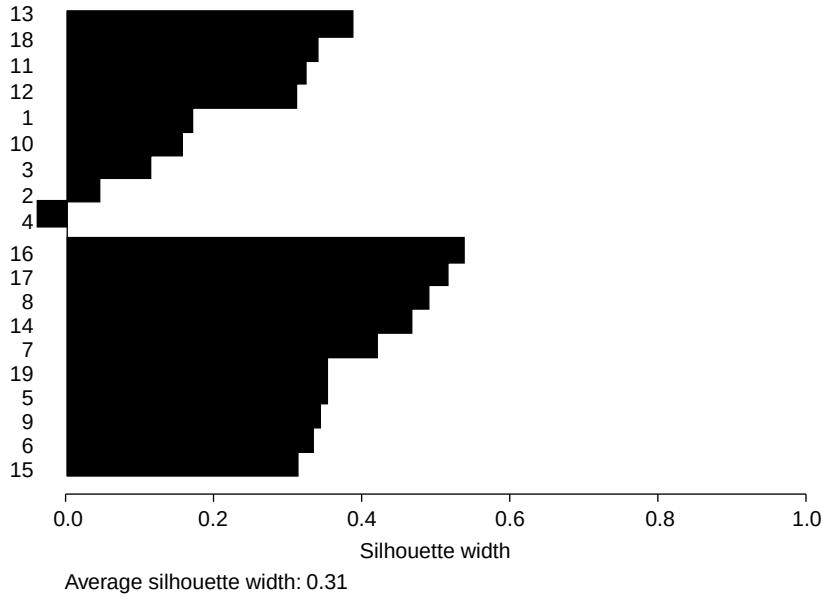


Figure 8.14 *Silhouette plot of Roman glass clusters, two-cluster solution (see also Figure 8.13 and Table 8.10). The width of the bars indicates the strength of clustering for each object. Negative bars indicate unsatisfactory classification, for example object D (no. 4)*

four-cluster solution are very similar to the average linkage solution (see Chapter 4 for a description of the latter technique).

8.8 Clustering and artificial neural networks

Neural networks have received a great deal of attention over the last few years. They have been used for a wide variety of applications, where conventional statistical methods such as regression and discriminant function analysis might normally be employed. But just what are neural networks? Essentially they are computing algorithms that attempt to imitate the computational capabilities of large, highly connected networks of relatively simple elements such as the neurons in the human brain. The interest in such techniques arises from the desire to mimic some of the desirable pattern-recognition type tasks for which the human brain is so well adapted. In the beginning, neural network models were intended as realistic models of neural activity in the human or animal brain, and this is still true in some areas of psychology and biology. But general interest now centres on the computational potential of neural network algorithms or computers without regard for their realism. Cheng and Titterington (1994) and Ripley (1994) provide surveys describing the relevance of neural networks in statistics. Ripley (1996) discusses

neural networks using a style and language familiar to statisticians, and Dewdney (1997, Chapter 5) gives a readable, if somewhat sceptical, general introduction.

8.8.1 Components of a neural network

The three essential features of a neural network are the basic computing elements usually referred to as *neurons*, the network architecture describing the connections between computing units, and the training algorithm used to find values of the network parameters for performing a particular task. A very simple neural network is the *single-unit perceptron* or McCulloch–Pitts neuron (McCulloch and Pitts, 1943). This is illustrated in Figure 8.15.

From a set of 'inputs' (predictors) $x_1, x_2, \ldots, x_p$ and weights $w_1, w_2, \ldots, w_p$, the neuron provides an 'output' (response) y given by

$$y = \text{sign}\left(w_0 + \sum_{i=1}^{p} w_i x_i\right), \tag{8.19}$$

where sign$(\cdot)$ equals 1 if its argument is positive and -1 otherwise. In this simple *binary thresholding* model, the neuron 'fires' or does not fire, depending on whether or not the summation is positive. Networks of artificial neurons such as these are constructed by forming interconnected banks of such elements with xs feeding into every such node, such as the circle in Figure 8.15, and the outputs of these nodes feeding into similar nodes at another level and so on. Such an arrangement is known as a *layered feed-forward neural network* or, equivalently, a *multilayer perceptron.*

Layers in between input and output are called *hidden layers* since they are not observable directly; Figure 8.16 depicts such a neural network with three inputs, a single layer of four hidden units and two outputs.

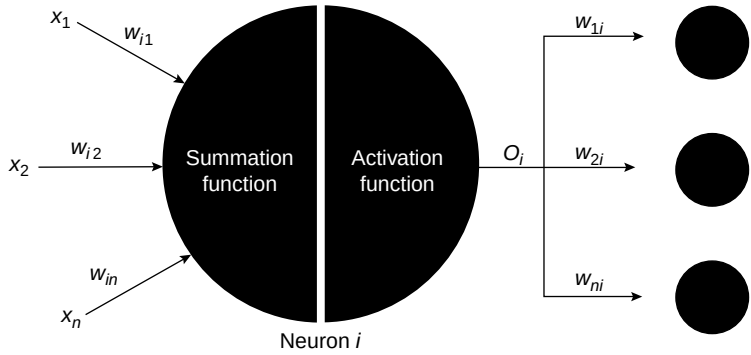


Figure 8.15 *Artificial neuron: weighted inputs are summed, processed by an activation function and output to the next layer of neurons. There is a layer of input nodes, a middle layer of hidden nodes and a layer of output nodes.* (Reprinted by permission of Sage Publications Ltd, from Garson, Neural Networks: An Introductory Guide for Social Scientists, 1998, Sage, London.)

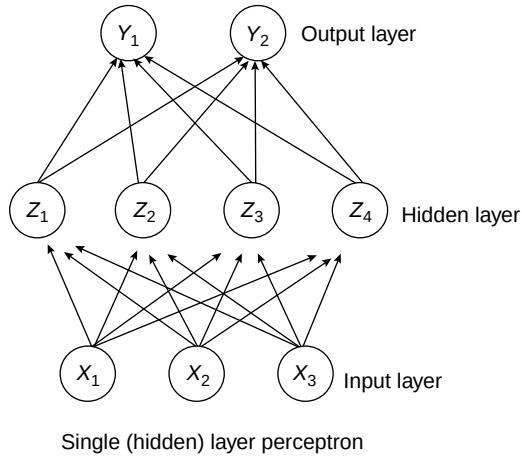


Figure 8.16 *A network diagram representing a single layer neural network with three inputs (predictors), four hidden units and two output (responses).* (Source: Hastie, 1998.)

In more traditional terms the model specified in Figure 8.16 can be written as follows:

$$z_j = \sigma(\alpha_{j0} + \boldsymbol{\alpha}'_j \mathbf{x}), j = 1, \ldots, 4, \tag{8.20}$$

$$\hat{y}_k = f_k(\beta_{k0} + \boldsymbol{\beta}'_k \mathbf{z}), k = 1, 2, \tag{8.21}$$

where $\mathbf{x}' = (x_1, x_2, x_3)$, $\mathbf{z}' = (x_1, x_2, x_3, x_4)$, $\boldsymbol{\alpha}'_j = (\alpha_{j1}, \alpha_{j2}, \alpha_{j3})$ and $\boldsymbol{\beta}'_k = (\beta_{k1}, \beta_{k2}, \beta_{k3}, \beta_{k4})$. The other terms in 8.20 and 8.21 are as follows:

- σ is known as the activation function and is used to allow a possible nonlinearity at the hidden layer; in the simple example above it was the *sign* function, but commonly it is taken to be the *sigmoid* function $\sigma(z) = 1/(1 + \mathrm{e}^{-z})$.
- The parameters α_{ji} and β_{ki} are the weights mentioned previously and define linear combinations of the input vector $\mathbf{x}$ and the hidden output vector $\mathbf{z}$, respectively.
- The intercept terms α_{j0} and β_{k0} are known here as *biases*.
- The function f_k is included to permit a final transformation of the output, for example the inverse logit function when responses should lie in [0,1].

The weights to be used in a neural network model are estimated from the training set data by least squares; for example, for the network described above, by minimizing

$$R(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_k (y_k - \hat{y}_k)^2, \tag{8.22}$$

a criterion that is nonlinear in the parameters. It is not always easy to minimize R, since it may have local minima and typically neural networks are overparameterized, often with more parameters than observations. (In the neural network literature this iterative estimation stage is often described as 'training' the network.)

8.8.2 The Kohonen self-organizing map

Most applications of neural networks have been in an assignment context, where the groups are known *a priori* (see Chapter 1). Details of such applications are given in Ripley (1996). But one well-known neural network method, the self-organizing map (SOM) due to Kohonen (1982), is an example of unsupervised learning because the assignment classes for the output vectors are not known *a priori* (see Kohonen, 1997, for further details and examples). Consequently, the network classifies the observations according to internally generated allocation rules; its performance can therefore be compared with more conventional approaches to cluster analysis.

The Kohonen model is illustrated in Figure 8.17. The network contains two layers:

- an input layer consisting of p-dimensional observations $\mathbf{x}$;
- an output layer (represented by a grid) consisting of k nodes for the k clusters, each of which is associated with a p-dimensional weight $\mathbf{w}$.

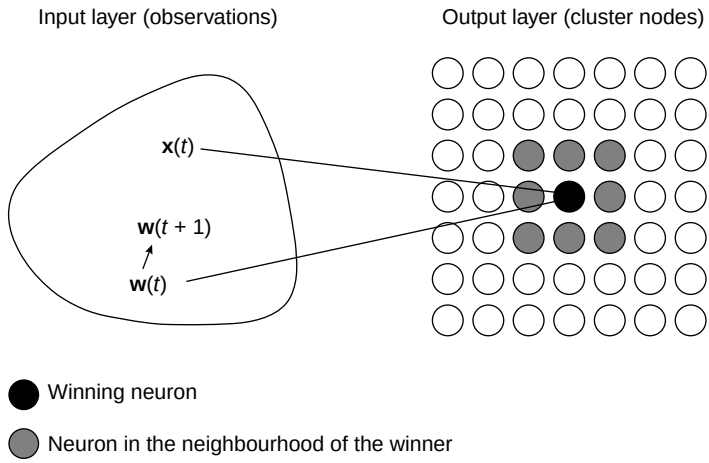


Figure 8.17 *A Kohonen self-organizing map, showing an observation* $\boldsymbol{x}$ *and its closest (winning) neuron at stage* t *in the iterative process. The weights* $\boldsymbol{w}$ *associated with winning neurons (and to a lesser extent the weights of those in their neighbourhoods) are moved towards the observation at stage* t + 1 (adapted from the webpage of the SOMLib Digital Library Project, Vienna University of Technology.)

Classification (clustering) occurs where an input vector is assigned to an output node. Operationally, each output node has a p-dimensional vector of synaptic weights $\mathbf{w}$. The output node is initially assigned a random weight; as the network learns, the input cluster points are provisionally assigned to clusters and the weights are modified. The iterative process eventually stabilizes with the weights corresponding to cluster centres in such a way that clusters that are similar to one another are situated close together on the map (the result being somewhat analogous to multidimensional scaling: see Chapter 2).

The SOM method thus makes the surface of the neurons recreate (i.e. change associated weight values) in accordance with the outside world as represented by the input vectors. In more mathematical terms the process can be described as follows.

- Consider p-dimensional weight vectors associated with neurons, each of the values of which is initially random and in the interval (0, 1).
- A p-dimensional observation, also scaled to be in (0, 1), is presented with the values of this weight vector.
- The Euclidean distance (or some other preferred distance measure) is calculated between the observation and the vector associated with each neuron.
- The neuron with the smallest distance (the 'winner') is then updated, as are a small neighbourhood of neurons around the 'winner'. The winner's weight vector $\mathbf{w}_{\text{old}}$ is brought closer to the input patterns $\mathbf{x}$ as follows:

$$\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + \alpha(\mathbf{x} - \mathbf{w}_{\text{old}}). \tag{8.23}$$

  The value of α is a small fraction, which decreases as learning takes place, as does the size of the neighbourhood. The excited neurons in the neighbourhood of the 'winner' are updated in an identical manner but with a smaller α.
- As the network 'learns', the weights are modified and the input observations are provisionally assigned to clusters.

It is clear from this description that the SOM procedure is in many respects similar to a standard iterative partitioning method such as k-means, as described in Chapter 5. One difference is that a number of parameters need to be initialized in the Kohonen algorithm at the start. Advantages of the SOM are that it produces a low-dimensional plot as a visual representation of the clustering, and that it can handle very large data sets.

A comparison of the neural network approach to clustering with a number of more conventional methods is reported in Waller *et al.* (1998). These authors applied their chosen methods to 2580 data sets with known cluster structure. Overall, the performance of the Kohonen network was similar to, or better than, the performance of the other methods. Further simplification can be achieved by making use of the topological relationships in the output layer, using methods of constrained clustering (Murtagh, 1995). Ambroise and Govaert (1996) have adopted a probabilistic approach, making use of the neighbourhood interaction

function as in the SOM, but in the context of the EM algorithm (see Chapter 6), which they term a 'Kohonen type EM'. An application involving a very large text database is given by Kohonen *et al.* (2000). See also Janasik *et al.* (2009) for a general description of SOM applied to qualitative data. Interesting astronomical and meteorological applications of neural networks used for cluster analysis are given in Murtagh and Hernández-Pajares (1995). A bibliography of earlier SOM papers is given by Kaski *et al.* (1998) and recent developments are described by Principe and Miikkulainen (2009).

8.8.3 Application of neural nets to brainstorming sessions

In research aimed at categorizing World Wide Web pages based on concepts (rather than keyword searches or hypertext browsing as used, for instance, by Lycos and Yahoo), Chen *et al.* (1996) developed a multilayered self-organizing map for 10 000 internet pages concerned with entertainment (the 'ET-Map'). They then assessed its success with a manually catalogued system. A smaller application was concerned with electronic brainstorming sessions, and this is reported here. In this application, 20 group participants were asked to comment on the most important information technology problems with respect to collaborative systems, using electronic keyboards in parallel. The group generated 201 comments in 30 minutes; the single two-dimensional representation of the layer SOM produced from these is shown in Figure 8.18.

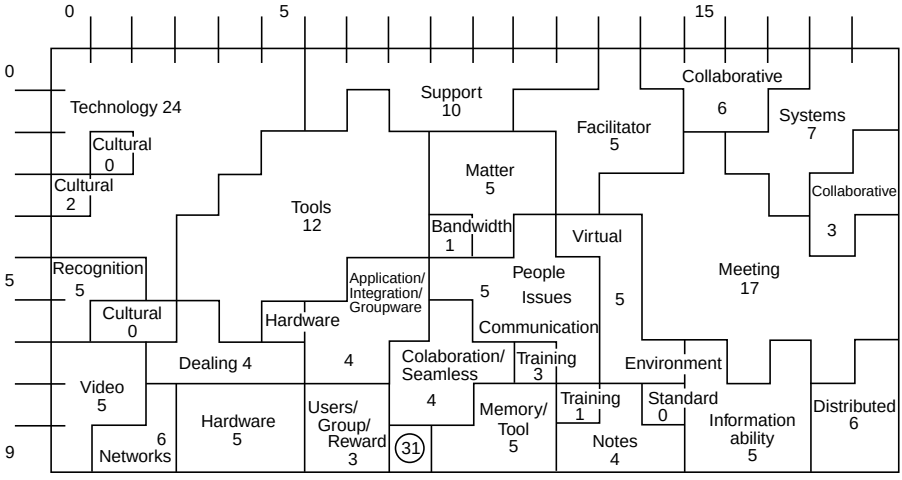


Figure 8.18 *A Kohonen self-organizing map of comments made by participants during an electronic brainstorming session on the future of collaborative systems. Clusters are represented by distinct areas on the map and are characterized by phrases such as 'Tools'.* (Figure from 'Internet categorization and search: A self-organizing approach', in *Journal of Visual Communication and Image Representation,* Volume 7, pp. 88–102, copyright © 1996 by Academic Press, reproduced by permission of the publisher.)

A small-scale experiment compared the performance of a human facilitator (who manually compiled topic lists) with the self-organizing map approach by asking a further eight individuals to revise the lists to provide a 'gold standard'. This showed that the facilitator had better *precision* than the SOM (i.e. a high proportion of the terms identified were gold-standard terms) at 81% compared to 55%. The *recall* (i.e. returning a high proportion of gold-standard terms) was similar at 89% compared to 81%. However, the time to produce the lists was very much in favour of the SOM: 4 minutes compared to 45 minutes. The authors concluded that the SOM output was an efficient algorithm for producing a 'straw man' list that could be used as a basis for improvement.

8.9 Summary

Standard clustering methods have been developed in many directions to encompass realistic situations, such as those involving constraints and overlapping clusters. Many of these developments allow for more comprehensive interpretation of clusters in terms of both objects and variables, or incorporate features such as fuzzy memberships. Application fields such as multimedia documentation, genetics and image analysis, combined with increasing computing power, have prompted some of these developments.

9

Some final comments and guidelines

9.1 Introduction

It should by now be obvious to most readers that the use of cluster analysis in practice does not involve simply the application of one particular technique to the data under investigation, but rather necessitates a series of steps, each of which may be dependent on the results of the preceding one. It is generally impossible *a priori* to anticipate what combination of variables, similarity measures and clustering techniques is likely to lead to interesting and informative classifications. Consequently, the analysis proceeds through several stages, with the researcher intervening if necessary to alter variables, choose a different similarity measure, concentrate on a particular subset of individuals, and so on. The final, extremely important stage concerns the evaluation of the clustering solutions obtained. Are the clusters real or merely artefacts of the algorithms? Do other solutions exist which are better? Can the clusters be given a convincing interpretation? A long list of such questions (which are full of traps for the unwary; see Dubes and Jain, 1979) might be posed.

It should also be clear from the preceding chapters that no one clustering method can be judged to be 'best' in all circumstances. The various studies that have compared a variety of clustering procedures on artificially generated data all point to the existence of a 'method $\times$ data type' interaction. In other words, particular methods will be best for particular types of data. Table 9.1 shows some possible method–data combinations.

In many applications it might be reasonable to apply a number of clustering methods. If all produce very similar solutions, the investigator might, justifiably

Cluster Analysis: 5th Edition Brian S. Everitt, Sabine Landau, Morven Leese and Daniel Stahl

Table 9.1 Overview of data types and applicable clustering methods.

Data type	Method	Notes
Continuous (Ordered data can be treated as continuous, possibly with standardization based on range)	Hierarchical agglomerative or partitioning methods	Ward's method, average linkage, and k-means or methods based on det($\mathbf{W}$) are popular choices; contiguity-constrained versions are possible
	Density search, mode analysis	
	Mixtures models	Especially for multivariate normal data; may need to restrict number of parameters fitted and/or take account of structure (such as repeated measures)
	Fuzzy k-means	Where clusters other than the 'best' are of interest
	Direct data clustering	Useful for data representing positive associations; both objects and variables clustered; need to consider scaling of data
	Kohonen self-organizing map (SOM)	Useful for large data sets; produces low-dimensional plot of clusters
Binary (Categorical data can be converted to binary)	Hierarchical or partitioning methods using appropriate proximity measure	Need to consider negative match issue; some special proximity measures available for categorical data
	Monothetic divisive method	Useful for developing diagnostic keys
	Latent classes or grade of membership (GOM)	These are 'fuzzy' methods; GOM often used in health applications
	Hierarchical classes (HICLAS)	Objects and variables clustered; overlapping clusters, useful for psychological data
Mixed mode	Hierarchical or partitioning methods using appropriate proximity measure or using ranks	Gower's similarity measure can be used
	Two step (`SPSS`)	Note possible problems with commensurability
	Model-based models for mixed data types	

Proximity matrix (Either computed from data or measured directly)	Hierarchical agglomerative or partitioning methods that do not require raw data	Examples are single, average and complete linkage, or partitioning around medoids (PAM)
	Additive clustering (ADCLUS and variants)	Overlapping clusters
	Kaufman and Rousseeuw fuzzy method (FANNY)	Similar to fuzzy k-means but raw data not required
	Tree-fitting methods	Dendrograms, additive trees; pyramids (limited overlap)
Large data sets	Data reduction methods such as principal components, spectral analysis	Reduction methods do not generally respect the cluster structure (so may be misleading)
	Clustering a subsample; using the clusters as seeds for further processing	Sampling process may need to be adapted to likely cluster structure

perhaps, have more confidence that the results are worthy of further investigation. Widely different solutions might be taken as evidence against any clear-cut cluster structure.

Most all-purpose statistical packages contain the 'standard methods' described in Chapters 4 and 5, and a few of the model-based methods described in Chapter 6. In `Stata`, *k*-means and *k*-medians, and hierarchical clustering methods are available using the `cluster` command or via a user-supplied dissimilarity matrix, using the `clustermat` command. Linkage methods are single, average, complete, weighted average, median, centroid and Ward's method. Dissimilarities for continuous variables include Euclidean and squared Euclidean, city block and a generalization of Euclidean and city block (Minkowski), Canberra, and similarity coefficients include correlation and angular separation. A large number of binary coefficients are available including simple matching and Jaccard similarities, and Gower's mixed data dissimilarity coefficient. `Stata` also includes the partitioning methods *k*-means and *k*-medians, and commands for the Calinski–Harabasz and the Duda–Hart indices are available. Clustering in `SPSS` can be obtained via the menu system or using `cluster` in syntax. Hierarchical methods are: between-groups linkage, within-groups linkage, nearest neighbour, furthest neighbour, centroid clustering, median clustering, and Ward's method, and the partitioning method *k*-means. The range of similarity and dissimilarity measures is similar to that in `Stata`, and slightly more extensive for binary variables, although Gower's method is not available. Both values and dissimilarity measures can be transformed in a number of ways (e.g. *z*-scores, or to a range of 0–1, and by case or by variable) within the command. Dendrograms and icicle plots can be obtained, but no stopping rules. The `twostep` method for mixed variable types is also available. Most of the other main statistical packages such as `SAS`, `Genstat`, `BMDP` and `Statistica` contain a similar range of methods. `S-PLUS` provides traditional hierarchical methods in `hclust` (single, complete, average, McQuitty, median, centroid, Ward's), but also some more unusual hierarchical methods such as `agnes`, and two divisive methods, DIANA and the monothetic method MONA, the partitioning method PAM as well as the model-based `mclust`. The specialized packages `Clustan` and `Clustan Graphics` are also available. There is an increasing number of specialized routines available in `R`.

There is no optimal strategy for either applying clustering or evaluating results, but some suggestions, which might be helpful in many situations, are discussed in this chapter, starting with an overview of the steps in a typical analysis. Those steps concerned with cluster validation and interpretation will be discussed in more detail in Sections 9.3–9.6, and two applications that illustrate many of the issues involved will be discussed in Section 9.7.

9.2 Using clustering techniques in practice

Milligan (1996) identifies seven steps that generally constitute a cluster analysis, based on findings from a number of studies, and gives a series of very useful tables

summarizing the results of these studies. The framework suggested by Milligan is outlined below, with some further comments. The first six steps have largely been covered in previous chapters. The topics mentioned in the final steps, interpretation, testing and replication, are discussed in more detail in this chapter, in sections on graphical methods for interpretation, testing for absence of cluster structure, comparing clusterings (both partitions and trees) and checks for robustness and stability.

The steps in a typical cluster analysis suggested by Milligan (1996) are as follows (the comments are a combination of Milligan's original suggestions and some additional points which we consider important):

(i) *Objects to cluster*: These should be representative of the cluster structure believed to be present, and they should be randomly sampled if generalization to a larger population is required. However, since cluster analysis is not an inferential technique, over-sampling of small populations and the placement in the sample of 'ideal types' (representatives of clusters suspected of being present) may be acceptable so long as generalization is not required.

(ii) *Variables to be used*: Variables should only be included if there is good reason to think they will define the clusters. Irrelevant or *masking* variables should be excluded if possible. A possible solution to the problem of masking variables is to employ the data matrix to suggest variable weights as described in Section 3.7, and an alternative solution is to use model-based variable selection as described in Section 6.6.

(iii) *Missing values*: Where the proportion of missing values is low, imputation of the raw data matrix may be acceptable, for example based on the clusters obtained in an initial pass, followed by re-clustering. Alternatively, the elements in a similarity or dissimilarity matrix can be imputed using only variables that are present, sometimes known as 'pairwise deletion'. For methods that use (raw) categorical data, for example latent class or grade of membership analysis, it is possible to include an additional response level to denote 'missing'. Model-based methods may be able to accommodate missing values as part of the expression of the likelihood.

(iv) *Variable standardization*: Standardization is not necessarily always indicated and can sometimes be misleading, as shown in Section 3.8. Standardization using the range showed good recovery of clusters in the simulations of Milligan and Cooper (1988) and should be considered as an alternative to the more usual standardization using standard deviations. Another solution to the problem of choosing an appropriate unit of measurement is to employ a cluster method that is invariant under scaling – see Chapter 6.

(v) *Proximity measure*: There are few general guidelines for this (see Section 3.9), but knowledge of the context and type of data may suggest suitable choices from those given in Chapter 3.

(vi) *Clustering method*: Methods should be: designed to recover the types of clusters suspected; effective at recovering them; insensitive to error; and

available in software. It is also advisable to consider data-generating processes, and this might suggest the application of a model-based method, as described in Chapters 6 and 7.

(vii) *Number of clusters*: One of the most difficult decisions to make is the number of clusters. Where different stopping rules suggest different numbers, the highest should also be considered for safety (unless external information from the subject matter suggests a suitable choice). An alternative would be to consider the possibility that there are no clusters present (see Section 9.3).

(viii) *Replication and testing*: Replication can involve cross-validation techniques, to investigate how far clusters identified in a subsample are still identifiable among the subsample of objects *not* used in the clustering. Another useful technique is the perturbation of the sample by omitting or slightly changing particular data points. Section 9.5 describes some of the techniques available for assessing internal validity. Goodness-of-fit statistics can be calculated to compare the clusters to the data used to derive them. Quality assessment may also involve comparing results between subsets, between the sample and a second sample or an external classification, using, for example, the Rand index for comparing partitions, or the cophenetic correlation for comparing dendrograms (see Section 9.4).

(ix) *Interpretation*: This may require graphical representation and descriptive statistics. Section 9.6 describes some graphical techniques helpful for cluster interpretation. It is important to note that standard statistical tests such as analysis of variance are inappropriate for comparing the *clustering* variables between clusters, since the clustering technique will have maximized between-cluster differences on these variables in some way.

The logical starting point for a cluster analysis would be a test for the absence of cluster structure. However, such tests are not usually employed in practical applications of clustering. This may be because the available tests are of limited usefulness. Their power is generally unknown and depends on the cluster structure, and so a test might simply not detect any departures from the null model due to lack of power. Nevertheless, the subject has some theoretical interest and therefore a short overview of this topic, based on the excellent review by Gordon (1998), is given in the next section. Subsequent sections deal with comparing partitions and dendrograms, measures of internal validity, and graphical methods for interpretation.

9.3 Testing for absence of structure

A test for the absence of cluster structure may not be necessary if the reason for clustering is practical (e.g. for organizational purposes). However, if it is aimed at detecting an unknown underlying structure then testing becomes more relevant. What is required is a model that describes the data-generating process in the

absence of clustering, and a test statistic which will reflect departures from the model.

The *Poisson model* assumes that, in the absence of cluster structure, the n individual p-variate observations arise from a uniform distribution over some region A of p-dimensional space. Equivalently, the underlying frequency distribution is assumed to have no mode. The (random) number of individuals observed within any subregion A_s follows a Poisson distribution with mean $\lambda/|A_s|$where λ is the constant intensity and $|A_s|$ denotes the p-dimensional volume of A_s. In the absence of theoretical results for finite samples, the null distribution of a test statistic can be generated by Monte Carlo simulation, by repeated sampling of p-variate observations from a uniform distribution over A. This hypothesis has been referred to as the *uniformity hypothesis* (Bock, 1985), as the *random position hypothesis* (Jain and Dubes, 1988) or, in the framework of spatial statistics, as the *complete spatial randomness hypothesis* (Diggle, 1983).

Departures from random positioning can be due to regularity or clustering. The number of interpoint distances below a specified threshold (Strauss, 1975; Kelly and Ripley, 1976; Saunders and Funk, 1977) and the largest nearest-neighbour distance within the set of individuals (Bock, 1985) can be used to assess departures from the Poisson model. The distances from randomly selected positions to the nearest points can be compared with distances between those points and their nearest neighbour (Cross and Jain, 1982; Panayirci and Dubes, 1983; Zeng and Dubes, 1985). Ripley (1981) and Diggle (1983) have generalized tests for more than two dimensions.

The *unimodal null hypothesis* assumes that the p-dimensional observations are generated from a frequency distribution with one mode. Most tests for this are limited to univariate data, but Bock (1985) assessed the distributions of the mean of all pairwise similarities and the minimum total within-group sum of squared distances when the data are partitioned into a fixed number of groups. Hartigan (1988) suggested a generalization of a one-dimensional test based on the difference between the empirical distribution function and the theoretical distribution function of the unimodal distribution nearest to it. Hartigan and Mohanty (1992) introduced a test for detecting bimodality based on single linkage clustering, and assessed the distribution of a relevant test statistic under both the Poisson and unimodal null models.

The *random dissimilarity matrix model* assumes that, under the null hypothesis of absence of cluster structure, all permutations of the ranks of the pairwise dissimilarities are equally likely. Hence the approach makes use only of the ranks of the dissimilarities and might be relevant when the dissimilarities are considered to be on an ordinal scale or when analysis by a clustering algorithm that only uses ranks is contemplated. In graph theory this model is also referred to as the *random graph hypothesis* (Jain and Dubes, 1988).

This model has some serious drawbacks. It generates an unrealistic distribution of any test statistic under the null hypothesis of absent cluster structure because it ignores existing relationships in the data. For example, if the dissimilarity between the ith and jth individuals, δ_{ij}, is small then the dissimilarities δ_{ik} and δ_{jk} would be

expected to have similar ranks (for further criticism see Jain and Dubes, 1988). However, Ling (1973) pointed out that the model could be used to obtain a lower limit for the p-value, and claimed that if the random dissimilarity model was not able to detect clustering then no other null model would.

9.4 Methods for comparing cluster solutions

Comparing partitions or trees, either with each other or with data, is a common requirement in cluster validation. For example, one might hope that different subsamples of the same data set, or different methods applied to the data, would give similar results. This might be considered as an aspect of robustness. It is also possible that an external classification is available and it is wished to investigate the similarity between this and the clustering, as an aspect of external validity. Sometimes it is more appropriate to compare proximity matrices, without clustering. A number of techniques are available to compare two partitions, two dendrograms or two proximity matrices. These are now discussed.

9.4.1 Comparing partitions

Two classifications may be represented as a $c_1 \times c_2$ matrix $\mathbf{N} = n_{ij}$, where n_{ij} is the number of objects in group i of partition 1 ($i = 1, \ldots, c_1$) and group j of partition 2 ($j = 1, \ldots, c_2$). The two classifications might be two different partitions of the same data set based on different clustering methods, or one might be a clustering and the other might be some externally defined classification. The labellings of the two partitions are arbitrary. If the number of clusters in the two partitions is the same and agreement is good, the correspondence of labels is usually obvious from inspection, and one partition can be relabelled to match the other. Partitions with equal numbers of clusters can, after relabelling, be compared using a simple percentage agreement, or the kappa coefficient (see Cohen, 1960).

However, if the number of clusters differs between the two partitions, the *Rand index* (Rand, 1971) can be used, since it is based on the agreement or otherwise of every pair of n objects rather than the simple cross-tabulation of frequencies. The index computes the proportion, of the total of $\binom{n}{2}$ object pairs, that agree; that is, are either (i) in the same cluster according to partition 1 *and* the same cluster according to partition 2, or (ii) in different clusters according to 1 *and* in different clusters according to 2. The index is defined as

$$I_{\mathrm{R}} = 2A/n(n-1), \tag{9.1}$$

where

$$A = \binom{n}{2} + 2\sum_{i=1}^{C_1}\sum_{j=1}^{C_2} n_{ij}^2 - \left(\sum_{i=1}^{C_1} n_{i.}^2 + \sum_{j=1}^{C_2} n_{.j}^2\right),$$

in which

$$n_{i.} = \sum_{i=1}^{C_1} n_{ij}, n_{.j} = \sum_{j=1}^{C_2} n_{ij}. \tag{9.2}$$

As Fowlkes and Mallows (1983) point out, this index tends to increase as the number of clusters increases, and the possible range of values is quite narrow. To counter these problems, Hubert and Arabie (1985) developed the *adjusted* (or *corrected*) *Rand index*; this has been recommended for general use by Milligan and Cooper (1986). This adjusted coefficient, I_{HA}, is analogous to the κ coefficient, since it measures the agreement over and above that expected by chance.

An alternative formulation for the (unadjusted) Rand index when $c_1 = c_2 = c$ is as follows:

$$I_{\mathrm{R}} = \left[T_c - \tfrac{1}{2} P_c - \tfrac{1}{2} Q_c + \binom{n}{2} \right] \Big/ \binom{n}{2}, \tag{9.3}$$

where

$$T_c = \sum_{i=1}^{c} \sum_{j=1}^{c} n_{ij}^2 - n,$$

$$P_c = \sum_{i=1}^{c} n_{i.}^2 - n,$$

and

$$Q_c = \sum_{i=1}^{c} n_{j.}^2 - n.$$

To illustrate these concepts, the values of the Rand and adjusted Rand indices are calculated for the matrix of n_{ij} values obtained from two normal mixtures classifications of the famous Fisher (1936) data on sepal and petal widths and lengths of irises. One solution assumes equal covariances, and the other allows for different covariance matrices. The two classifications are shown in Table 9.2. The agreement between the two classifications is clearly very good (Rand index 0.87), and well above that expected by chance (adjusted Rand index 0.72).

9.4.2 Comparing dendrograms

Hierarchical classifications can also be compared with each other using measures such as the cophenetic correlation or Goodman and Kruskal's γ (see Section 4.4.2).

Table 9.2 Two alternative classifications of irises into three groups, assuming equal and unequal covariance matrices.

		Clusters assuming equal covariance matrices			
		1	2	3	Total
Clusters assuming unequal covariance matrices	1	0	0	50	50
	2	2	47	0	49
	3	36	15	0	51
	Total	38	62	50	150

Cell entries are the numbers of objects classified into each of three clusters by the two methods.

Random tree models can be used to generate appropriate null distributions, in the context of both comparing dendrograms with proximity matrices and, as discussed here, comparing two dendrograms. Such models assume that all possible trees based on the n objects are equally likely, the universe of possible trees depending on whether the tree is binary, labelled or ranked (see Section 4.4.1). Where n is large, the number of possible trees generated in a Monte Carlo simulation will be enormous, and a random sample of trees may be used. A problem with this method is that each type of clustering produces a distinct type of tree, incorporating, for example, the chaining effect in single linkage. Ideally, then, the null distribution should be based on random sets of, say, single linkage trees. Hubert (1974) considers the degree of distortion which single linkage and complete linkage impose on data generated under the three null models discussed in Section 9.3 (Poisson, unimodal and random dissimilarity). Further information on tree generation is given by Gordon (1998) and Furnas (1984).

Lapointe and Legendre (1995) compared three methods of randomization – (i) labels; (ii) labels and topology; and (iii) labels, topology and heights – as used to assess the statistical significance of the cophenetic correlation. The first of these is the well-known test of Mantel (1967). The authors show that the Mantel test is too conservative and conclude that the test based on all three, the *double permutation test*, is optimal in the sense that the universe of dendrograms sampled in this way is the most comprehensive. Published tables (Lapointe and Legendre, 1992) can be used to assess the significance of correlations without actually performing the permutations. Section 8.6.1 describes an application of these methods.

The B_c coefficient of Fowlkes and Mallows (1983) is an alternative to the Rand index, also for the case $c_1 = c_2 = c$, and is defined as follows (see Equation (9.3) for definitions):

$$B_c = T_c / \sqrt{P_c Q_c}. \tag{9.4}$$

The coefficient can be used in conjunction with dendrograms by plotting its value against the number of clusters; that is, plotting the pairs (c, B_c), $c = 2, \ldots,$

$n - 1$, for each pair of partitions containing c clusters obtained from two dendrograms. A series of Monte Carlo studies reported by Fowlkes and Mallows reveal the potential of this procedure for comparing classifications. Additionally, the plots appear to have the potential for selecting the appropriate number of clusters. However, if a hierarchical method has been used but only certain partitions are of interest, it may be more natural to compare these particular partitions, using the adjusted Rand index, than to compare the complete hierarchies.

9.4.3 Comparing proximity matrices

Arabie and Hubert (1996) point out that analysts sometimes inappropriately compare the clustering output (for example, dendrograms) when they are in fact interested in the correlation between the input proximity matrices. This would typically be the case if the aim of the analysis is to establish whether clustering according to one proximity measure corresponds to clustering according to another. In this case the cophenetic matrices can be thought of as containing simplified information about the underlying cluster structures, the full information about which is contained in the proximity matrices. Schneider and Borlund 2007a, 2007b) review methods for comparing proximity matrices.

Applications which investigate the association between two proximity matrices, derived from the same cases, can be found in epidemiology. An example might be when a disease is suspected to be aetiologically associated with an infectious agent, so that temporal dissimilarities may be associated with spatial dissimilarities (*time-space clustering*). Tests for association are typically derived according to the techniques suggested by Mantel (1967), Ederer *et al.* (1966) and Knox (1964); for an application to Hodgkin's disease, see Smith and Pike (1974). Chen *et al.* (1984) assessed the power of tests for time–space clustering under three alternative models for the distribution, transmission and development of Hodgkin's disease.

9.5 Internal cluster quality, influence and robustness

Internal cluster quality can be taken to refer to the extent to which clusters meet the requirements for good clusters, as defined by Cormack (1971), namely isolation and cohesion. *Robustness* refers to the effects of errors in data or missing observations, and changes in the data or methods. Similar solutions should be obtained from different data, methods or subsets of variables when the data are clearly structured. Milligan (1980) gives an example of how this can be achieved through error perturbation. For k-means and other hill-climbing techniques, different seeds for the initial clusters should not affect the cluster solutions. *Influence* refers to the deletion of a particular point and consequent changes to the cluster assignments of the other elements. In an extension of this idea, deletion of a small number of variables from the analysis should not, in most cases, greatly alter the clusters found, if these clusters are 'real' and not mere artefacts of the particular technique being used. These three topics are discussed in the following sections.

9.5.1 Internal cluster quality

Numerical measures of isolation and cohesiveness (or compactness) are generally based on indices reflecting the relative magnitudes of intra- and inter-cluster similarity. Some of these have been discussed earlier, in connection with the problem of the number of clusters (Section 5.5). The *silhouette index* was introduced earlier as a measure of cluster quality, in connection with the silhouette plot. Bailey and Dubes (1982) define indices for a particular cluster on the basis of (i) the numbers of edges in a graph between cluster members, and (ii) the number of edges between cluster members and non-members. The probabilities of these indices, given a null random graph model (see Section 9.3), are then used to produce a *cluster validity profile* for a given cluster. The uncertainty in each of the individual clusters in hierarchical cluster analysis can also be examined using the bootstrapping approach of Suzuki and Shimodaira (2006). This is available in `R` as the function `Pvclust`, which calculates probability values for each cluster using bootstrap resampling techniques. Two types of probabilities are available: approximately unbiased (AU) probability and bootstrap probability (BP) value. Multiscale bootstrap resampling is used for the AU probability, which has lower bias than the BP value calculated by the ordinary bootstrap resampling. The probability (*'p-value')* is the proportion of bootstrapped samples that contain the cluster so that larger p-values indicate more support for the cluster. Kapp and Tibshirani (2007) develop a measure of cluster quality, the IGP (in-group proportion), which quantifies how often points near each other are predicted to belong to the same group (when classified to their nearest cluster). They compare the method with a number of other measures of cluster quality. Software in `R`, `clusterRepro`, is available through http://cran.r-project.org.

Cohen *et al.* (1977) describe a number of other useful graphical techniques for evaluating cluster analysis solutions. The first of these involves consideration of the relative tightness of a k-point group (a potential cluster) compared to other k-point neighbourhoods in the data. The $k-1$ closest neighbours of each data point are found, and then the average interpoint distance, d_i, among these k individuals over all $k(k-1)/2$ pairs is determined. Data points contained in 'real' k-point clusters should give d_i values substantially smaller than data points not in such groupings. Cohen *et al.* suggest comparing the d_is for a given k by means of a normal probability plot. A tight cluster of size j should produce j points with nearly equal d_is which are well separated from the others at the bottom of the plot for $k=j$. Figure 9.1 shows such plots for a data set, using $k=2$, 5, 9 and 14. The behaviour at the lower ends of these plots suggests the existence of a group of size of approximately $k=9$ in these data.

A further plot described by Cohen *et al.* (1977) is that of squared distances from certain cluster centroids to individuals that are near the centroid. This is useful for examining the internal cohesiveness of a cluster. Figure 9.2 shows an example of such a plot. The symbol plotted corresponds to the cluster in which the individual resides. From this plot it is clear that there is a large distance between the members of cluster A and the closest individuals that are not assigned to A. Gnanadesikan

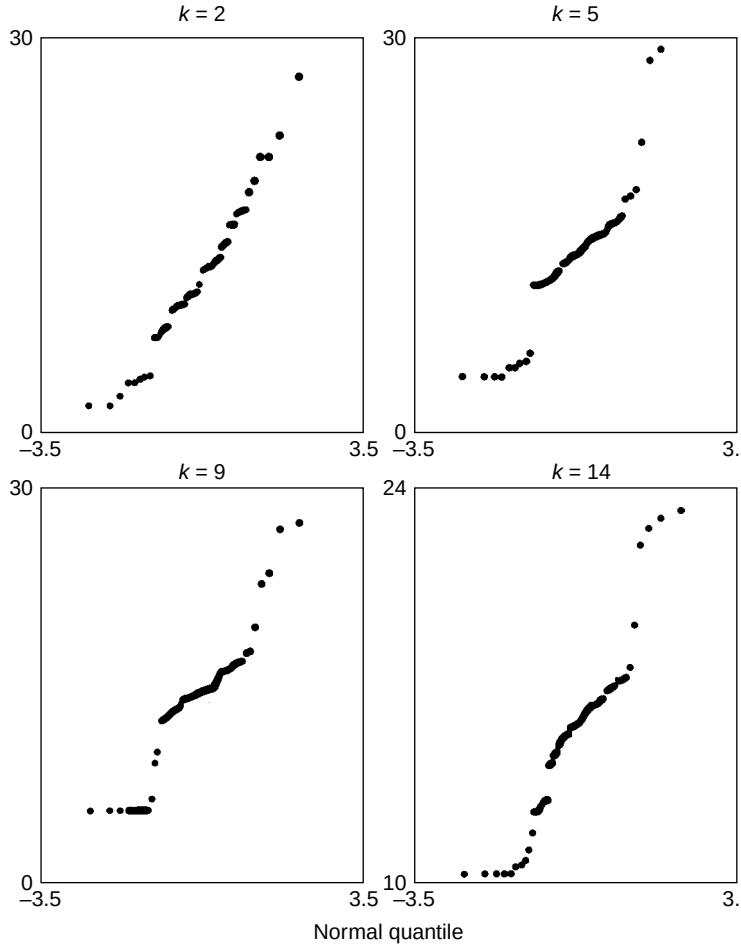


Figure 9.1 *Quantile–quantile plots to display clusters. (*Source: Cohen et al., 1977.)

et al. (1977) suggest a similar method in which the horizontal plotting positions are slightly perturbed, so that objects which would otherwise be coincident are distinguishable.

9.5.2 Robustness – split-sample validation and consensus trees

One approach to assessing the effects of perturbations of the data is the 'split-sample' method, randomly dividing the data into two subsets and performing an analysis on each subset separately. A scheme for performing split-sample cross-validation was proposed by McIntyre and Blashfield (1980), and was shown to give good results using Monte Carlo simulation. Their method involves the following steps:

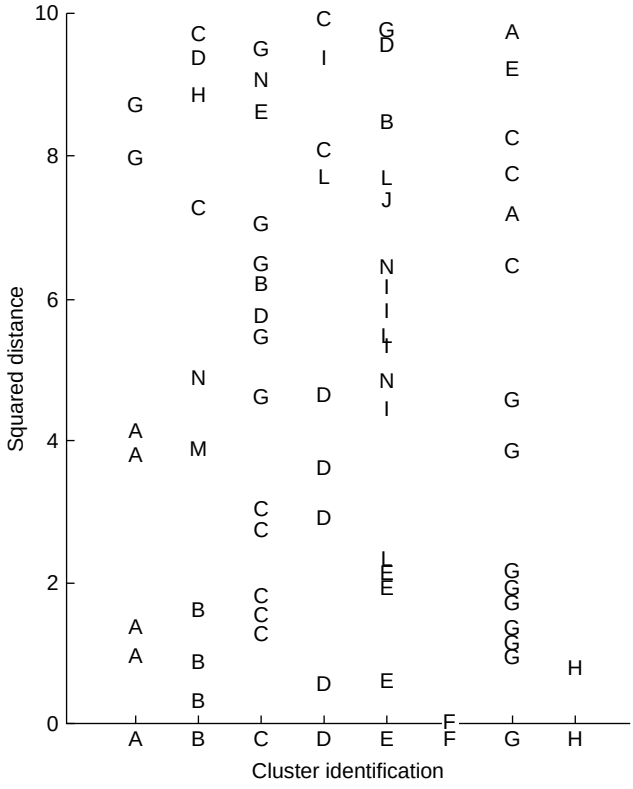


Figure 9.2 *Plot of squared distances of selected individuals from cluster centroids. (*Source: Cohen et al., 1977.)

- Divide the sample in two, and perform a cluster analysis on one of the samples, having a fixed rule for the number of clusters.
- Determine the centroids of the clusters, and compute proximities between the objects in the second sample and the centroids, classifying the objects into their nearest cluster.
- Cluster the second sample using the same methods as before, and compare these two alternative clusterings for the second sample, using, for example, the adjusted Rand index.

Breckenridge (1989) proposed a variation on this procedure in which a nearest-neighbour criterion was used to classify the second sample.

Such techniques have been criticized by Krieger and Green (1999), particularly when used for suggesting the number of clusters present, as proposed by Overall and Magee (1992). Krieger and Green showed through simulations of univariate data that validation improved for larger sample sizes, independently of the number

of clusters fitted, and also that the parity of the number of clusters could influence the level of validity inferred. For multivariate data, they concluded that the performance in determining the number of clusters degraded for unequal sized clusters and/or highly correlated data.

Where robustness is assessed by fitting different clusterings to the same data (rather than, as above, by using the same method on different subsets), it may be reasonable to synthesize the results as a *consensus clustering*. An alternative (and simpler) approach for clusterings based on different variable subsets is to reanalyse using the complete set of variables (De Querioz, 1993). Dendrograms may be combined into either a *strict consensus tree,* where each subset in the consensus tree is in every constituent tree (Sokal and Rohlf, 1981), or a *majority consensus tree* where each subset in the consensus tree is in a majority of the constituent trees (Margush and McMorris, 1981). Further types of consensus tree allow partial agreement between the constituent trees and the consensus tree. An example of a *consensus graph*, in which different clusterings are indicated by vertices in a graph, with edges joining those with a minimum level of agreement, is given in Section 9.7.1. Such graphs can indicate which trees might reasonably be combined. The formation of consensus trees is discussed in more detail by Gordon (1999).

9.5.3 Influence of individual points

Gnanadesikan *et al.* (1977) and Jolliffe *et al.* (1988) studied the deletion of selected individual points. The latter also considered the influence of the individual points on complete single and complete linkage dendrograms. Cheng and Milligan (1996) extended this approach to define a number of measures of the influence of individual points. The first part of this work compared the cluster solution of simulated data with external criterion clusters. The value of the adjusted Rand index for a clustering compared to an external grouping was used to assess what they call *cluster recovery*, with the clustering and assessment based on all n objects. Where the recovery is measured for $n-1$ observations (but based on clustering all n objects) it is termed the *adjusted cluster recovery* (relating to the depleted point).

The differences between the adjusted recovery index and that based on clustering the $n-1$ points can be used to measure the impact of the point. Where a positive difference is found, the point is regarded as a *facilitator*, whereas if it is negative it is considered as an *inhibitor*. Once an influential point is found, if it is an inhibitor, then the suggestion is to omit it. As Edelbrock (1979) has argued, in many applications there is no need to assign every point. This would usually be the case where the sample was purposive (see step (ii), Section 9.2), since the composition of the sample to be clustered is arbitrary.

Table 9.3 illustrates the calculation of these indices for a small example with one candidate influential point. In this example the influence of the point is $0.9183 - 0.7652 = 0.1531$ (a facilitator to the clustering method).

In applications where the true clustering is unknown, the comparisons must be internal using HA_{ii}. This contains no information about whether a point is a facilitator or inhibitor. For such cases, Goodman and Kruskal's γ (Section 4.4.2)

Table 9.3 Four types of influence index for a small example.

		Frequency tables			Adjusted Rand index	Cheng and Milligan name and interpretation
		Clustering method (n)				
		1	2	Total		
1 External criterion clusters	1	25	0	25	0.9200	HA_{cr} Cluster recovery
	2	1	24	25		
	Total	26	24			
		Clustering method (n)				
		1	2	Total		
2 External criterion clusters	1	25	0	25	0.9183	HA_{acr} Adjusted cluster recovery
	2	1	23	$24 = 25 - 1$		
	Total	26	$23 = 24 - 1$			
		Clustering method ($n-1$)				
		1	2	Total		
3 External criterion clusters	1	23	2	25	0.7652	HA_{ei} External influence
	2	1	23	$24 = 25 - 1$		
	Total	24	25			
		Clustering method ($n-1$)				
		1	2	Total		
4 Clustering method (n)	1	22	4	26	0.5611	HA_{ii} Internal influence
	2					
		2	21	$23 = 24 - 1$		
	Total	24	25			

1. Complete sample used for both clustering and index.
2. Complete sample used for clustering; candidate point omitted from index.
3. Candidate point omitted from clustering and index.
4. Two clusterings: (i) based on complete sample (method n) and (ii) omitted candidate point (method $n-1$); index computed from reduced sample. (Taken with permission of Springer-Verlag, from Cheng and Milligan, 1996.)

and the *point biserial correlation index* were suggested by Cheng and Milligan as indices for determining if an influential point is a facilitator or an inhibitor. The point biserial index is based on the correlation between the elements of the proximity matrix and a dummy variable indicating, for each pair of points, whether they were placed in the same or different groups by the clustering method. A positive (negative) value of either type of index indicates a facilitator (inhibitor). A further index, *internal influence*, compares the clustering obtained with and without a candidate point.

9.6 Displaying cluster solutions graphically

Before applying any clustering method, some graphical representation of the data should be obtained, and a number of possibilities were discussed in Chapter 2. Classical principal components analysis is commonly employed to obtain a low-dimensional mapping of the data, although this is not guaranteed to reflect any clustering present. Other, potentially more useful, ordinations may be obtained from other methods, for example those described by Sammon (1969), and the *projection pursuit* techniques discussed by Jones and Sibson (1987). In addition, multidimensional scaling techniques may be used to extract similar visual displays from a calculated proximity matrix. Some authors suggest that if these displays produce no evidence of clustering in the data, then more formal clustering procedures are not required; see, for example, Chatfield and Collins (1980).

Specialized techniques specifically designed for cluster analysis can be used. Examples of the latter have been given in earlier parts of the book: for example, banner plots (Figure 4.11) and silhouette plots (Figures 5.5 and 8.13). To illustrate the process of hierarchical clustering, as distinct from particular partitions, dendrograms are used (Figures 4.10, 4.13 and 4.14). The Kohonen self-organizing map (e.g. Figure 8.18) can be regarded as a clustering method which provides at the same time a complete visual representation of the clusters. Techniques that are designed to cluster both variables and objects with binary data may have their own specialized graphs (Figure 8.8).

Once the cluster analysis has been performed, the partition(s) found can be added to low-dimensional plots (Figure 5.4), and the minimum spanning tree (see Section 4.4.5) can indicate possible distortions in this. *Correspondence analysis* (Greenacre, 1984) can be used to produce a low-dimensional plot that shows how the clustering obtained corresponds to another classification (either a second clustering or an externally defined classification). It is generally important for interpretation to be able to associate the values of particular variables with the clusters, a simple approach being to describe the clusters by profiles, bar charts or scatterplots of pairs of variables. These three graphical approaches are illustrated by Stopford *et al.* (1991) in an application concerned with the chemical composition of decorated medieval tiles. This paper gives an example of (i) a principal components plot illustrating the compositional affinities of tile clusters; (ii) a correspondence plot showing how the compositional clusters relate to externally defined production groups; and

(iii) profile plots of the chemical compositions of selected tile clusters and related clay sources.

Leisch (2009) describes several graphical displays that can be used to visualize solutions from *k*-means-type clustering methods. The basis of a number of these graphics is the *shadow value*, $s(\mathbf{x})$ of each multivariate observation, $\mathbf{x}$, defined as follows:

$$s(\mathbf{x}) = \frac{2d[\mathbf{x}, c(\mathbf{x})]}{d[\mathbf{x}, c(\mathbf{x})] + d(\mathbf{x}, \tilde{c}(\mathbf{x})]}, \tag{9.5}$$

where $d[\mathbf{x}, c(\mathbf{x})]$is the distance of the observation $\mathbf{x}$ from the centroid of its own cluster, and $d[\mathbf{x}, \tilde{c}(\mathbf{x})]$ is the distance of $\mathbf{x}$ from the second-closest cluster centroid. If $s(\mathbf{x})$is close to zero then the observation is close to its cluster centroid; if $s(\mathbf{x})$is close to one then the observation is almost equidistant from the two centroids (a similar approach is used in defining silhouette plots – see Chapter 5.) The average shadow value of all observations where cluster i is closest and cluster j is second closest can be used as a simple measure of cluster similarity:

$$s_{ij} = n_i^{-1} \sum_{\mathbf{x} \in A_{ij}} s(\mathbf{x}), \tag{9.6}$$

where n_i is the number of observations which are closest to the centroid of cluster i and A_{ij} is the set of observations for which the centroid of cluster i is closest and the centroid of cluster j the second closest. The denominator of s_{ij} is taken to be n_i rather than n_{ij}, the number of observations in the set A_{ij}, to prevent inducing large cluster similarity when n_{ij} is small and the set of observations consists of poorly clustered points with large shadow values.

For a cluster solution derived from bivariate data, a *neighbourhood graph* can be constructed using the scatterplot of the two variables and where two cluster centroids are joined if there exists at least one observation for which these two are closest and second closest, with the thickness of the joining lines being made proportional to the average value of the corresponding s_{ij}. When there are more than two variables in the data set, the neighbourhood graph can be constructed on some suitable projection of the data into two dimensions; for example, the first two principal components of the data could be used. Such plots may help to establish which clusters are 'real' and which are not, as we will try to illustrate with two examples.

The first example uses some two-dimensional data generated to contain three clusters. The neighbourhood graph for the *k*-means five-cluster solution from the application of *k*-means clustering is shown in Figure 9.3. The thicker lines joining the centroids of clusters 1 and 5 and clusters 3 and 4 strongly suggest that both pairs of clusters overlap to a considerable extent and are probably each divisions of a single cluster.

For the second example we return to the pottery data previously met in Chapter 2. From previous analysis it is clear that these data contain three clusters; Figure 9.4 shows the neighbourhood plot for the *k*-means *four*-cluster solution in the space of the first two principal components of the data. The very thick line joining the centroids of clusters 2 and 4 suggests that the pottery in these two clusters really belongs in a single cluster.

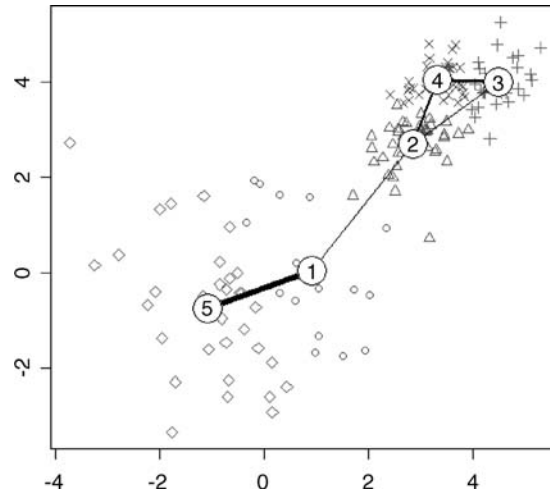


Figure 9.3 *Neighbourhood plot of k-means five-cluster solution for bivariate data containing three clusters.*

A simple graphical aid for evaluating clustering solutions, again suggested by Cohen *et al.* (1977), can be used for examining the clusters in terms of either variables used to form the clusters or other variables of interest. Here the clusters are again identified along the *x*-axis, and above each label the values of a particular variable are plotted, for each individual in the cluster. The median of the cluster is also plotted. The plot can then be used to compare individuals in the same cluster

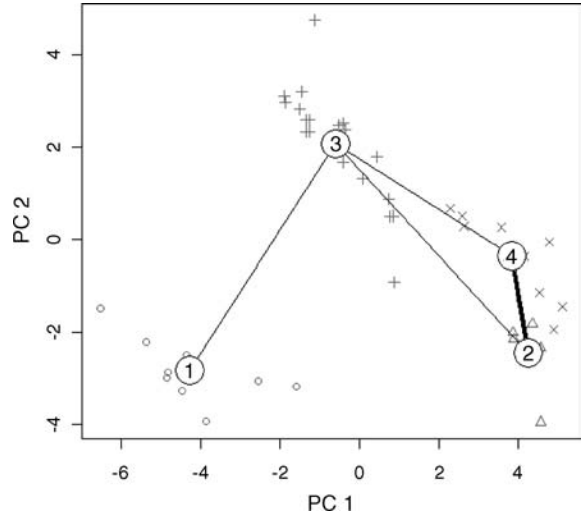


Figure 9.4 *Neighbourhood plot for the k-means four-cluster solution on the pottery data (see Chapter 2); the plot is shown in the space of the first two principal components of the data.*

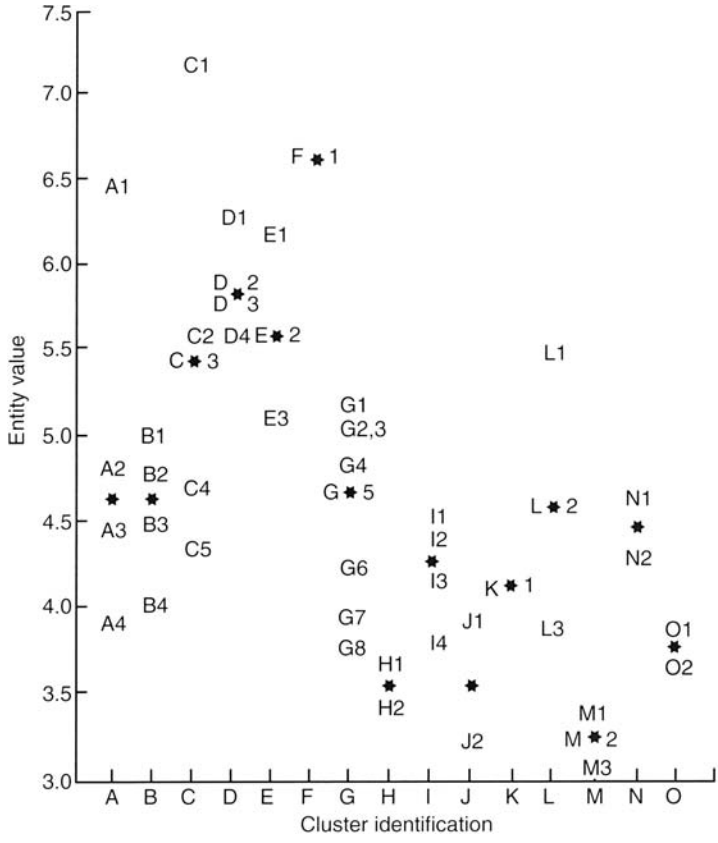


Figure 9.5 *Plot of values of a single variable for selected individuals in various clusters;* **denotes the median of a cluster.* (Source: Cohen et al., 1977.)

on the variable in question, and to make multiple comparisons across clusters. An example of such a plot appears in Figure 9.5. This shows that clusters A and B are quite similar on the variable apart from one individual, Al in cluster A. Cluster E tends to contain individuals with large values on this variable and these have moderate spread, whilst cluster M has much smaller values with small spread.

A graphic for displaying clustering solutions, similar to the one in Figure 9.5 in certain respects, is suggested by Leisch (2009). Known as a *stripes plot*, this graphic is a simple but often effective way of visualizing the distance of each point from its closest and second-closest cluster centroids. For each cluster, $k = 1, \ldots, K$, a stripes plot has a rectangular area which is vertically divided into K smaller rectangles, with each smaller rectangle, i, containing information about distances of the observations in cluster i from the centroid of that cluster, along with the corresponding information about observations that have cluster i as their second-closest cluster. The explanation of how the plot is constructed becomes more transparent if we look at an actual example, and Figure 9.6(a) shows a stripes plot for a five-cluster solution on a set of data generated to contain five relatively distinct clusters. Looking first at the

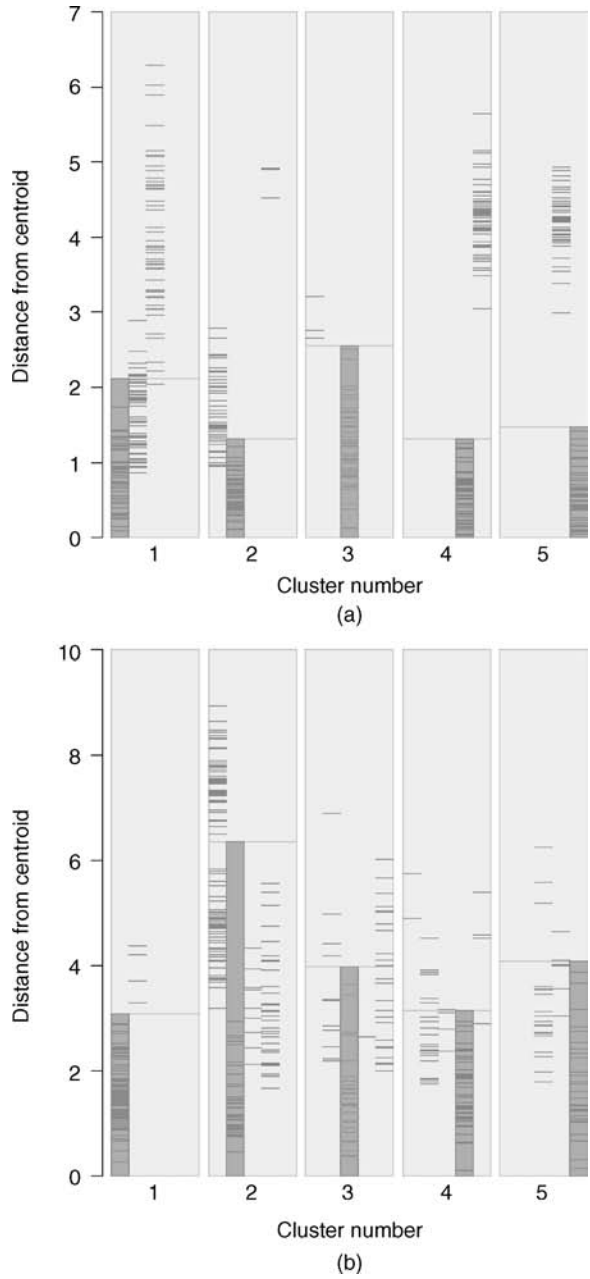


Figure 9.6 *Stripes plot for k-means five-group solution on (a) a simulated data set with five relatively distinct clusters; and (b) a second data set, where the plot suggests that the five-group solution is not appropriate in this case.*

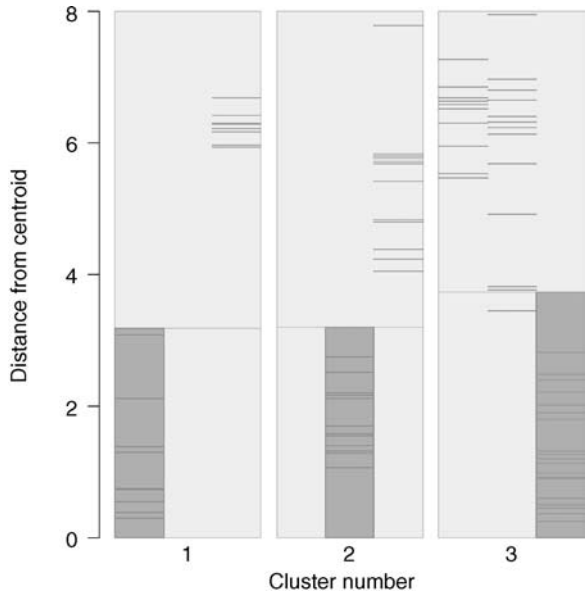


Figure 9.7 *Stripes plot for k-means three-group solution on pottery data.*

rectangle for cluster 1, we see that observations in clusters 2 and 3 have the cluster 1 centroid as their second closest. These observations form the other two stripes within the rectangle. Observations in cluster 3 are further away from cluster 1, but a number of observations in cluster 3 are at a similar distance from the centroid of cluster 1 as those observations that belong to cluster 1. Overall though, the stripes plot in Figure 9.6(a) suggests that the five-cluster solution matches quite well the actual structure in the data. The situation is quite different in Figure 9.6(b), where the stripes plot for the *k*-means five-group solution suggests that the clusters in this solution are not well separated, implying perhaps that the five-group solution is not appropriate for the data in this case. Lastly, the stripes plot for the *k*-means three-group solution on the pottery data is shown in Figure 9.7. The graphic confirms the three-group structure of the data.

All the information in a stripes plot is also available from a neighbourhood plot, but the former is dimension independent and may work well even for high-dimensional data where projections to two-dimensions lose a lot of information about the structure in the data.

Neither neighbourhood graphs nor stripes plots are infallible, but both offer some help in the often-difficult task of evaluating and validating the solutions from a cluster analysis of a set of data.

9.7 Illustrative examples

In this section, some further illustrative examples are given. The first two show how data matrices, proximity matrices and alternative hierarchical clusterings

can be compared. The third is an application using a model-based technique, which illustrates how a number of steps can be applied to reduce the complexity of data. Finally, an example shows how a combination of internal and external evidence can be used to choose between methods in genetics applications.

9.7.1 Indo-European languages – a consensus tree in linguistics

Atkinson *et al.* (2005) examined divergence time estimates for various Indo-European languages, with a focus on the date of the original parent language. Swadesh lists of words (see Chapter 3) were employed, but characters (grammatical and phonological features) were also considered in the analyses as well as words.

Two types of model were developed using ideas from evolutionary biology; this approach contrasts with previous research in glottochronology, which was based on standard hierarchical clustering methods. One type was based on the so-called 'finite sites' model (where changes in a fixed number of characters are considered); the other was based on a stochastic model of language creation, loss and splitting. MCMC sampling and Bayesian analysis were used to compare the performance of the models with different parameters and on various types of data set, including synthetic data. The overall conclusions from these two new types of model were surprisingly consistent. One of the benefits of this new analysis is the ability to assign uncertainties to estimates, since the Bayesian analysis allowed the quantification of phylogenetic uncertainty in date estimates. A simplified consensus tree is shown in Figure 9.8, in which the posterior probabilities for each branch (the degree of support for the branch) is given as the percentage of time that it appeared in the Bayesian MCMC sample.

A summary of recent developments, a discussion of the differences between this statistical analysis and other analyses from linguistic palaeontology, and an example of another type of figure, the consensus network, can be found in Nicholls (2008).

9.7.2 Scotch whisky tasting – cophenetic matrices for comparing clusterings

Lapointe and Legendre (1994) applied a hierarchical clustering method to binary characteristics of 109 Scotch whiskies, the 68 variables describing feature types such as colour, nose, body, and so on, derived from an expert's description (Jackson, 1989). Jaccard's similarity measure was used in conjunction with Ward's method, and results were compared with those from two other methods. A spatially constrained k-means approach was also used (see Section 8.5). Here we concentrate on the use of cluster comparison, rather than the clustering methods as such.

Comparisons were made between raw data matrices, distance matrices, both derived from the whisky features and geographic distances, and the cophenetic matrix representing the complete clustering. From the clustering with 2, 3, 6 and 12 group partitions, binary matrices were computed, with entries equal to 1 when two whiskies were in the same partition, and 0 otherwise. Since Scotch whiskies

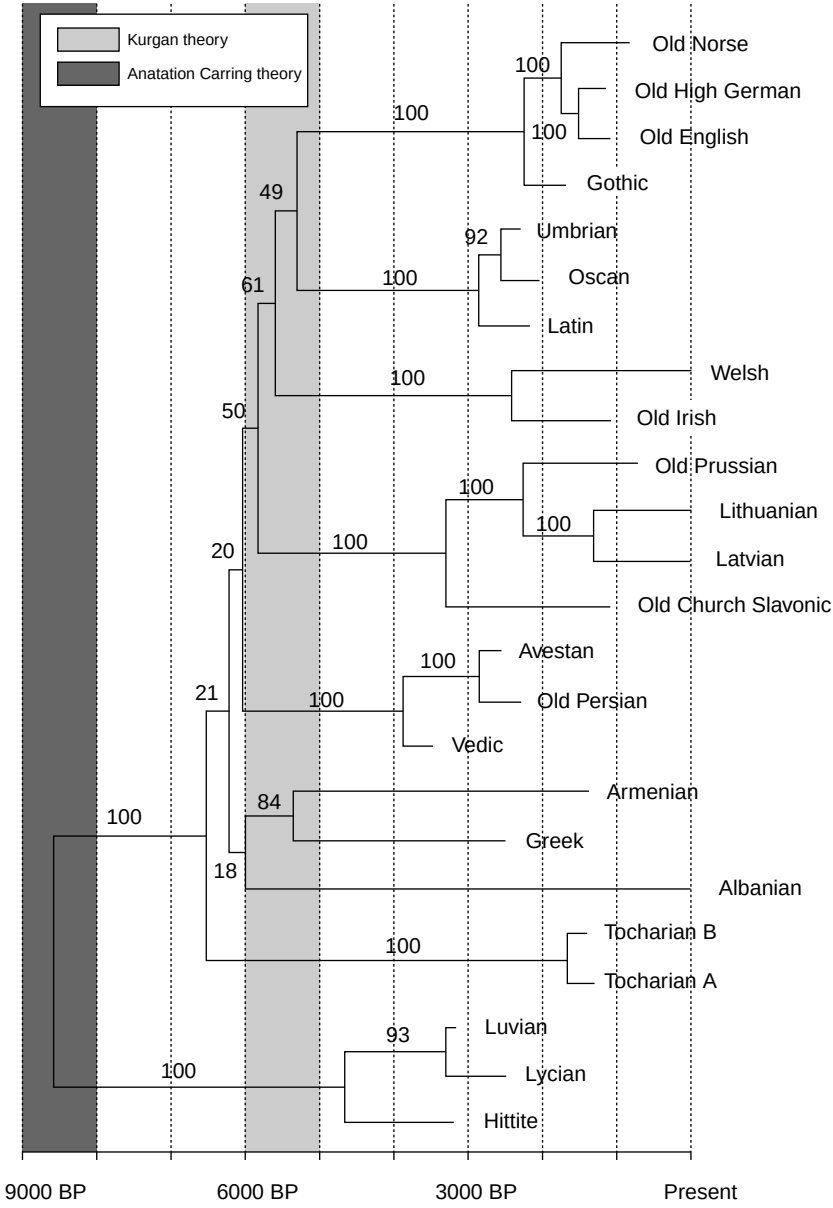


Figure 9.8 *A consensus tree of Indo-European languages (from Atkinson, et al., 2005). The shaded bars represent two hypotheses about the time of common origin of the languages. Figures above branches indicate the degree of support for that branch.*

are known to derive some of their characteristics from water, soil, air quality, and so on, which vary by region, two geographic matrices were also computed. The first of these was a matrix of distances between the distilleries based on map coordinates. The second was a proximity matrix containing binary entries: 1 if two distilleries were in the same region (Highland, Lowland or Islay) and 0 otherwise.

The authors computed cophenetic correlations to compare the clustering with the geographical classification. The significance of this was tested using both random permutation of labels (Mantel test) and the more realistic double permutation of labels, topology and fusion levels (see Section 9.4.2). A further detailed analysis of the feature types (colour, nose, body, palate and finish) was undertaken to assess the congruence of the five features. The analysis involved comparing raw data tables (for which canonical correlations and associated permutational test statistics were derived), distance matrices and dendrograms. Table 9.4 shows the results of these tests of significance. Entries in the tables are correlations (standardized Mantel statistics) or coefficients based on canonical correlations, indicated as significant at $p = 0.01$ or 0.001 according to the appropriate test. In Table 9.4(a), the authors interpreted the significant results obtained for the 6- or 12-group partition as indicating that geographic information seems more relevant in defining smaller clusters of closely related whiskies.

The remaining subtables relate to feature types. Some discrepancies between the comparison tests among feature types are explained in terms of the loss of information when moving from data to distance matrix, and from distance matrix to dendrogram. Figure 9.9 shows the significant and nonsignificant results in Table 9.4 (b)–(d) as a consensus graph, and from this it is evident that only three comparisons were congruent at all levels: palate–nose; nose–colour and colour–body; finish was always left unconnected with other features.

9.7.3 Chemical compounds in the pharmaceutical industry

Guttiérrez *et al.* (1999) described the application of a model-based hierarchical method to a large data set of binary 'fingerprints' identifying the molecular structure of chemical compounds used in the pharmaceutical industry. The fingerprints are an abstract representation of molecular patterns in the form of a sequence of bits of length 1024. Their account of the clustering process illustrates some of the steps described in Section 9.2

- The choice of objects (molecules), variables (binary fingerprints) and an appropriate proximity measure, followed by multidimensional scaling to reduce the dimensionality of the data.
- The choice of statistical model, and hence method of analysis – in this case normal mixtures.
- Consideration of the number of clusters.
- The assessment of the stability of the clusters by comparing the results for different MDS solutions; the use of another similar clustering method to check

Table 9.4 Correlations and comparison tests between clustering of Scotch whiskies and geography, and between different whisky feature groups.

(a) Clustering versus geographic distance matrix based on map coordinates

		Partitions into:			
Standardized Mantel statistics	Complete dendrogram	2 groups	3 groups	6 groups	12 groups
Mantel test	0.031**	−0.027	0.007	0.065**	0.064**
Double permutation test	0.031*				

(b) Feature types (comparisons between raw data matrices)

Redundancy coefficient	Nose	Body	Palate	Finish
Colour	0.147*	0.121*	0.151	0.152
Nose		0.121*	0.164**	0.153
Body			0.105	0.116
Palate				0.129

(c) Feature types (comparisons between distance matrices)

Standardized Mantel statistics	Nose	Body	Palate	Finish
Colour	0.032*	0.067**	0.027	−0.011
Nose		0.042*	0.074*	0.009
Body			0.054*	0.018
Palate				−0.012

(d) Feature types (comparisons between cophenetic matrices)

Standardized Mantel statistics	Nose	Body	Palate	Finish
Colour	0.048*	0.064**	0.046*	−0.025
Nose		0.021	0.071**	−0.001
Body			0.0510**	0.008
Palate				0.002

* $0.001 < p < 0.01$.
** $p < 0.001$.
(Taken with permission of the publisher, Blackwell, from Lapointe and Legendre, 1994.)

for robustness; comparison against previous findings and assessment of the usefulness of the solution.

The Jaccard coefficient (the proportion of bits present in either compound which were common to both) was used as a measure of similarity between two compounds (see Section 3.2), and this was subsequently converted to a distance measure. Because of the size of the distance matrix (460 320 elements), metric

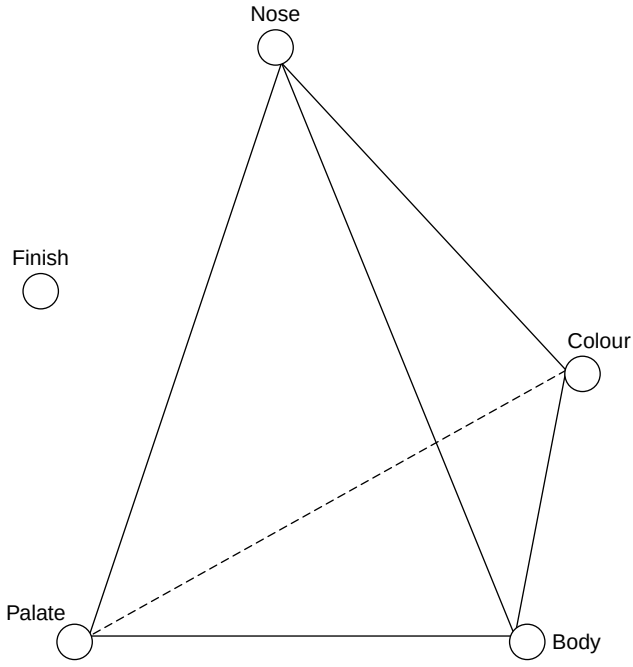


Figure 9.9 *Consensus graph of whisky feature types. Thick edges depict congruent relationships at all three levels (raw data tables, distance matrices and dendrograms); unbroken thin edges are congruent at two levels, whereas the broken edge indicates a relationship which is significant at one level only; see also Table 9.4.* (Taken with permission of the publisher, Blackwell, from Lapointe and Legendre, 1994.)

scaling was performed on a number of subsamples and then applied to the rest of the data to produce coordinates in a low dimension (five in this case). Before clustering, the data transformation procedure was as follows: binary data → similarity → distance → continuous data. Each scaling (from a different subsample) produced slightly different sets of continuous data. However, they all, when displayed on the first two (of the five) axes, showed an obvious clump of points in the middle, with ellipses emanating from the centre (see Figure 9.10); plots on other axes were similar.

The elliptical nature of the apparent clusters indicated the advisability of using a method that can identify such clusters. The large sample sizes and low dimensionality meant that normal mixtures models with different orientations and sizes (but the same shape; criterion S^* in Table 6.1) could be considered without computational problems. They were fitted using the method of Banfield and Raftery (1993) which suggested eight groups, whichever set of scaled data was used. Further examination of the data suggested the subdivision of the largest group into two, to give a final nine-cluster solution.

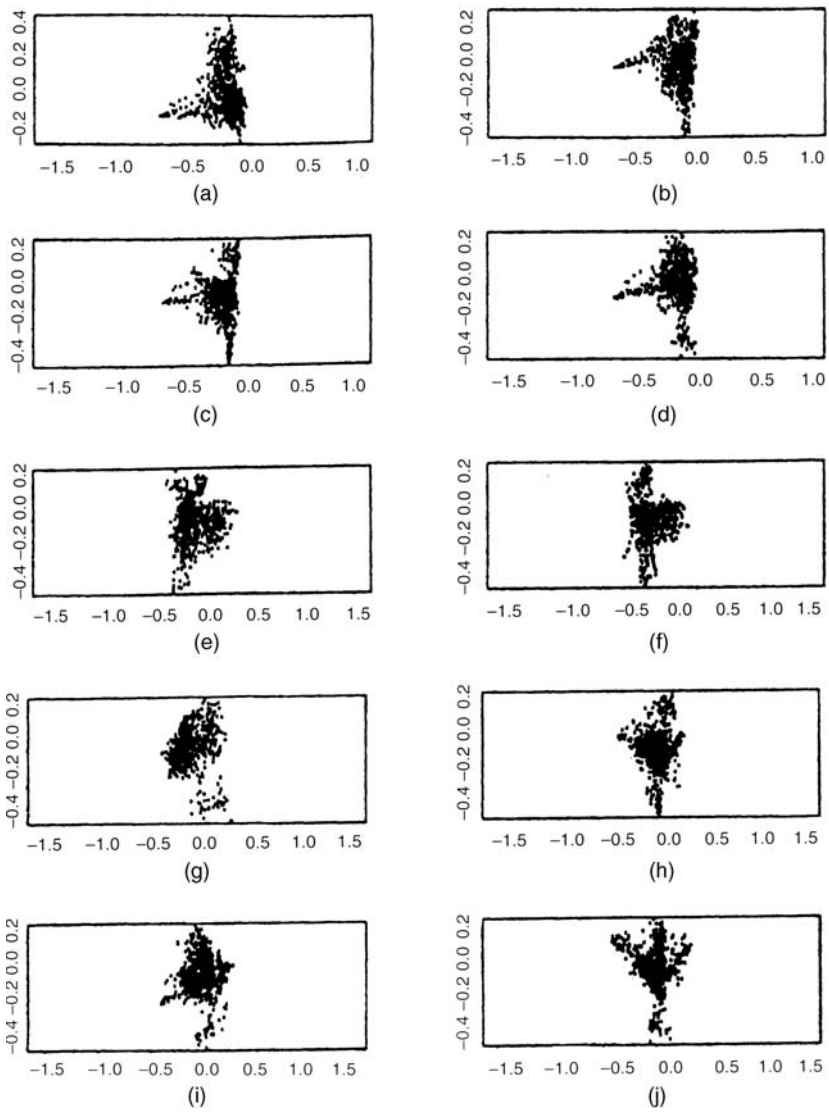


Figure 9.10 *Pairwise scatterplots of pharmaceutical data, showing elliptical clusters; axes 1 vs. 2–5; 2 vs. 3–5; 3 vs. 4–5; 4 vs. 5 in order (a) to (j) in a five-dimensional metric scaling based on a subsample of the data.* (Taken with permission of the publisher, Blackwell, from Guttiérrez et al., 1999.)

Table 9.5 shows the eight-cluster solutions after MDS using two different subsamples. There appears to be a good correspondence between these results, showing robustness to small changes in the input data. The agreement was 75% (after making the obvious correspondence of clusters), the Rand index was 0.79 and the adjusted Rand index was 0.51.

Table 9.5 Model-based clustering using MDS based on two subsamples of pharmaceutical data[a]).

Clusters when scaling based on subsample 2	Clusters when scaling based on subsample 1							
	1	2	3	4	5	6	7	8
1	5	0	0	5	0	0	0	**15**
2	9	6	**15**	0	0	2	2	0
3	0	0	6	**38**	0	2	0	0
4	0	**74**	0	0	0	0	0	0
5	59	0	0	0	**0**	0	0	0
6	1	0	1	0	0	0	**158**	3
7	40	0	0	0	0	**40**	0	0
8	**380**	8	5	0	40	12	11	23

[a]Corresponding clusters shown in bold.

The data were also analysed using a mixtures method, developed by Cheeseman and Stutz (1995), that can handle large data sets; in five dimensions the method worked well and gave reasonable agreement with the previous analysis, although the number of clusters suggested was higher for this second analysis.

Many of the clusters were confirmed by subject matter specialists to correspond to already known compounds, and the results proved to be useful in identifying groups of similar compounds for further pharmaceutical testing.

9.7.4 Evaluating clustering algorithms for gene expression data

In a study by Datta and Datta (2006), six clustering algorithms, UPGMA (average linkage hierarchical), *k*-means (partitioning), DIANA (divisive hierarchical method), FANNY (a fuzzy method), model-based (mixtures of Gaussians) and SOM (self-organizing map: a type of neural network – see Chapter 8) were evaluated on two publicly available genetic data sets, one on breast cancer patients and one from microarray data on yeast. The methods are those that might be considered reasonable choices for this type of data and have been mentioned earlier in this book. Two types of validation measures are proposed (each with two variants), the first 'statistical' (VS1 and VS2) which are based on statistical stability when a unit is deleted from the gene expression profile, and the second 'biological' (VB1 and VB2), where it is assumed that gene pairs with similar biological functions can be identified, and should appear in the same cluster. The two variants 1 and 2 use proportion of overlap of different clusters containing similar gene pairs, and average distance between clusters with and without the deleted gene, respectively. The aim was to assess the performance on statistical and biological validity, both separately and together, for different numbers of clusters. An illustration is given in Figure 9.11.

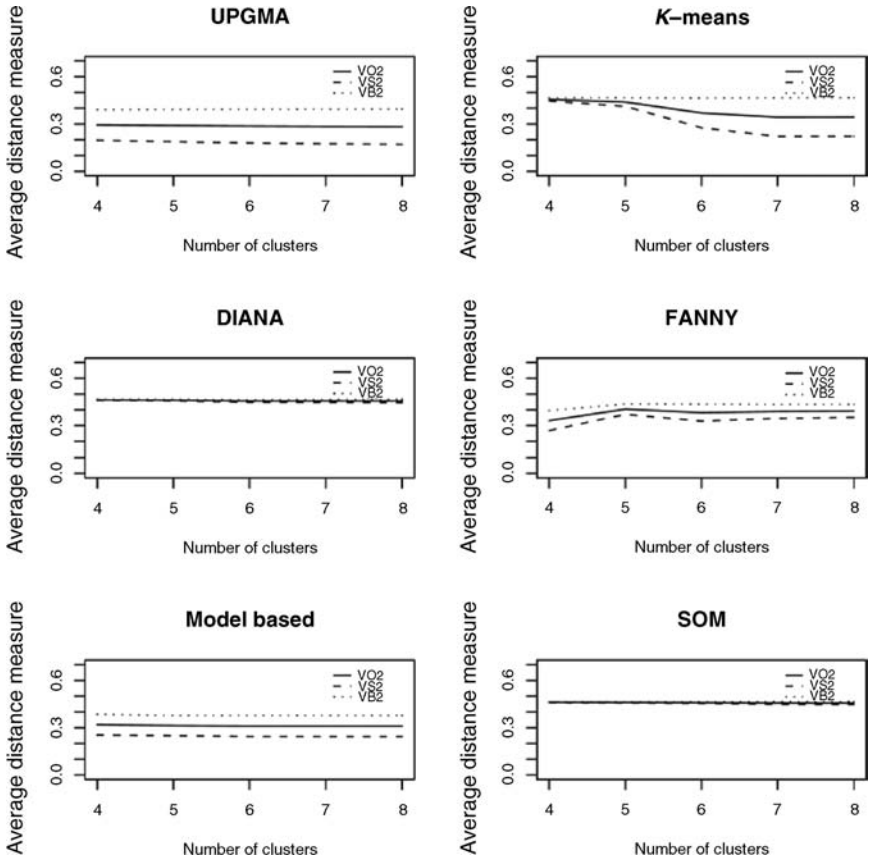


Figure 9.11 *The performance for various clustering methods and numbers of clusters on the cancer data, based on the average distance between the clusters containing similar gene pairs (dotted – VB2), and with and without each unit (dashed – VS2), and VO2 (solid – the average of VB2 and VS2). From Datta and Datta (2006).*

The model-based mixtures method and average linkage did well for this data set and measure, but taking into account both data sets and types of measures, no method was an overall winner. Although the results were inconclusive, the authors show that a fairly simple and systematic approach is possible in assessing clusters. It should be noted, however, that the 'biological' validation measure depends on additional information on which genes share functionality, which is not always available, and also that the relative weight of the statistical and biological measures is subjective (here they were equally weighted).

A much larger study (Souto *et al.*, 2008) did, however, come to firmer conclusions. The authors compared 'classic' methods and those designed to take advantage of the specific nature of gene expression data on 35 publicly available data sets (see http://algorithmics.molgen.mpg.de/Supplements/CompCancer/).

One characteristic of gene expression data is, as the authors point out, the high dimension (compared to say clustering genes themselves, which typically are described by a lower number of states). The criterion here was the recovery of known cancer types – thus using a 'biological' rather than 'statistical' criterion (in the terms used by Datta and Datta). Methods considered were hierarchical clustering with single, complete and average linkage, *k*-means and mixture of multivariate Gaussians, and more recent methods of spectral clustering (Ng *et al.*, 2002) and a nearest-neighbour-based method (Ertoz *et al.*, 2002). Proximity measures considered were Pearson's and Spearman's correlations, cosine and Euclidean distance, the latter with three types of pre-processing: standardization, scaling and ranking. The adjusted Rand coefficient was used to assess agreement with the best partition and the partition obtained by assuming that the correct number of clusters was known. The conclusion was that a mixture model, followed by *k*-means, was optimal in recovering the known clusters.

9.8 Summary

The methods of cluster analysis can be valuable tools in the exploration of multivariate data. By organizing such data into subgroups or clusters, clustering may help the investigator discover the characteristics of any structure or pattern present. Applying the methods in practice, however, requires considerable care if over-interpretation of the solutions obtained is to be avoided. Much attention needs to be given to questions of cluster validity, although such questions are rarely straightforward and are full of traps for the unwary. Simply applying a particular method of cluster analysis to a data set and accepting the solution at face value is in general not adequate.

Bibliography

Abbott, A. (1995) Sequence analysis: new methods for old ideas. *Annual Review of Sociology*, **21**, 93–113.

Abbott, A. and Tsay, A. (2000) Sequence analysis and optional matching methods in sociology. *Sociological Methods and Research*, **29**, 3–33.

Adachi, K. (2002) Nonmetric multidimensional scaling with clustering of subjects. *Japanese Psychological Research*, **42**, 112–122.

Agrawal, R., Gehrke, J., Gunopulos, D. *et al.* (1998) Automatic subspace clustering of high dimensional data for data mining applications, in *Proceedings of the ACM-SIGMOD'98 Int. Conf. Management of Data. Seattle, Washington, June 1998*, 94–105.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251–262.

Aitkin, M. (1999) Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, **18**, 2343–2351.

Aitkin, M. and Rubin, D. B. (1985) Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society Series B*, **47**, 67–75.

Aitkin, M., Anderson, D. and Hinde, J. (1981) Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society A*, **144**, 419–448.

Akaike, H. (1973) Information theory as an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (B. N. Petrov and F. Csaki, eds) 267–281. Akademiai Kiado, Budapest.

Allison, D. B., Gadbury, G. L., Heo, M. *et al.* (2002) A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis*, **39**, 1–20.

Ambroise, C. and Govaert, G. (1996) Constrained clustering and Kohonen self-organising map. *Journal of Classification*, **13**, 299–313.

American Psychiatric Association (2000) *Diagnostic and Statistical Manual of Mental Disorders* (4th edn, text rev.). APA, Washington, DC.

Anderberg, M. R. (1973) *Cluster Analysis for Applications*. Academic Press, New York.

Anderson, T. W. and Bahadur, R. R. (1962) Classification into two multivariate normal populations with different covariance matrices. *Annals of Mathematical Statistics*, **33**, 420–431.

Andrews, R. L. and Currim, I. S. (2003) A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, **40**, 235–243.

Arabie, P. and Carroll, J. D. (1980) MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, **45**, 211–235.

Arabie, P. and Carroll, J. D. (1989) Conceptions of overlap in social structure, in *Research Methods in Social Network Analysis* (L. C. Freeman, D. R. White and A. K. Romney, eds) 367–392. George Mason University Press, Fairfax, VA.

Arabie, P. and Hubert, L. J. (1990) The bond energy algorithm revisited. *IEEE Transactions on Systems, Man, and Cybernetics*, **2**, 268–274.

Arabie, P. and Hubert, L. J. (1996) An overview of combinatorial data analysis, in *Clustering and Classification* (P. Arabie, L. J. Hubert and G. De Soete, eds) 5–63. World Scientific, Singapore.

Arminger, G. and Stein, P. (1997) Finite mixtures of covariance structure models with regressors: loglikelihood function, minimum distance estimation, fit indices, and a complex example. *Sociological Methods and Research*, **26**, 148–182.

Arminger, G., Stein, P. and Wittenberg, J. (1999) Mixtures of conditional mean- and covariance-structure models. *Psychometrika*, **64**, 475–494.

Art, D., Gnanadesikan, R. and Kettenring, J. R. (1982) Data based metrics for cluster analysis. *Utilitas Mathematica*, **21a**, 75–99.

Ashburner, M., Ball, C. A., Blake, J. A. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Atkinson, Q. Nicholls, G., Welch, D. and Gray, R. (2005) From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society*, **103**, 193–219.

Aude, J. C., Diaz Lazcoz, Y., Codani, J. J. and Risler, J. L. (1999) Applications of the pyramidal clustering method to biological objects. *Computers and Chemistry*, **23**, 303–325.

Babu, G. J. and Feigelson, E. D., eds (1996) *Statistical Challenges in Modern Astronomy II*. Springer, New York.

Bacher, J., Vogler, M. and Wenzig, K. (2004) SPSS TwoStep Cluster – a First Evaluation. Arbeits- und Diskussionspapiere 2004-2. Lehrstuhl für Soziologie Nürnberg. www.soziologie.wiso.uni-erlangen.de/publikationen/a-u-d-papiere/a_04-02.pdf.

Bailey, T. A. and Dubes, R. (1982) Cluster validity profiles. *Pattern Recognition*, **15**, 61–83.

Baker, F. B. and Hubert, L. J. (1975) Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, **70**, 31–38.

Balakrishnan, V. and Sanghvi, L. D. (1968) Distance between populations on the basis of attribute data. *Biometrics*, **24**, 859–865.

Ball, G. H. and Hall, D. J. (1967) A clustering technique for summarizing multivariate data. *Behavioural Science*, **12**, 153–155.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L. and Rathouz, P. J. (1997) Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, **92**, 1375–1386.

Banfield, C. F. and Bassill, L. C. (1977) Algorithm AS 113. A transfer algorithm for non-hierarchical classification. *Applied Statistics*, **26**, 206–210.

Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Bansal, A. K. and Sharma, S. (2003) A model for clustering longitudinal data sets of infant mortality rates in India. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, **9**, PH 1–6.

Basu, S., Davidson, I. and Wagstaff, K. (2008) *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman and Hall/CRC, London.

Basu, S., Davidson, I. and Wagstaff, K. (2009) *Constrained Clustering*. Chapman and Hall/CRC, London.

Bauer, D. J. and Curran, P. J. (2003) Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, **8**, 338–363.

Bauer, D. J. and Curran, P. J. (2004) The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological Methods*, **9**, 3–29.

Baulieu, F. B. (1989) A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, **6**, 223–246.

Baxter, M. J. (1994) *Exploratory Multivariate Analysis in Archaeology*. Edinburgh University Press, Edinburgh.

Beale, E. M. L. (1969a) Euclidean cluster analysis, in *Bulletin of the International Statistical Institute: Proceedings of the 37th Session (London)* Book 2, 92–94. ISI, Voorburg, Netherlands

Beale, E. M. L. (1969b) *Cluster Analysis*. Scientific Control Systems, London.

Beath, K. J. and Heller, G. Z. (2009) Latent trajectory modelling of multivariate binary data. *Statistical Modelling*, **9**, 199–213.

Beaverstock, J. V., Smith, R. G. and Taylor, P. J. (1999) A roster of world cities. *Cities*, **16**, 445–458.

Becker, R. A. and Cleveland, W. S. (1994) *S-PLUS Trellis Graphics User's Manual*. Mathsoft, Seattle, WA.

Becker, T., Knapp, M., Knudsen, H. C. *et al.* (1999) The EPSILON study of schizophrenia in five European countries: design and methodology for standardising outcome measures and comparing patterns of care and service costs. *British Journal of Psychiatry*, **175**, 514–521.

Behboodian, J. (1970) On the modes of a mixture of two normal distributions. *Technometrics*, **12**, 131–139.

Belbin, L. (1987) The use of non-hierarchical allocation methods for clustering large sets of data. *Australian Computer Journal*, **19**, 32–41.

Benitez, R. and Nenadic, Z. (2008) Robust unsupervised detection of action potentials with probabilistic models. *IEEE Transactions on Biomedical Engineering*, **55**, 1344–1354.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Berchtold, A. (2004) Optimization of mixture models: comparison of different strategies. *Computational Statistics*, **19**, 385–406.

Bertrand, P. (1995) Structural properties of pyramidal clustering, in *Partitioning Data Sets* (I. Cox, P. Hansen and B. Julesz, eds) DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 19, 35–53. American Mathematical Society, Providence, RI.

Bezdek, J. C. (1974) Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology*, **1**, 57–71.

Biernacki, C., Celeux, G. and Govaert, G. (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, **41**, 561–575.

Biernacki, C., Celeux, G., Govaert, G. and Langrognet, F. (2006) Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, **51**, 587–600.

Birant, D. and Kut, A. (2007) ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, **60**, 208–221.

Blashfield, R. K. (1976) Mixture model tests of cluster analysis. Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, **83**, 377–385.

Bock, H. H. (1985) On some significance tests in cluster analysis. *Journal of Classification*, **2**, 77–108.

Böhning, D. (2000) *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping and Others*. Chapman and Hall/CRC, Boca Raton.

Bolck, A., Croon, M. A. and Hagenaars, J. A. P. (2004) estimating latent structure models with categorical variables: one-step versus three-step estimators. *Political Analysis*, **12**, 3–27.

Bollen, K. A. (1989) *Structural Equations with Latent Variables*. John Wiley & Sons, Inc., New York.

Bollen, K. A. and Curran, P. J. (2004) Autoregressive latent trajectory (ALT) models – a synthesis of two traditions. *Sociological Methods and Research*, **32**, 336–383.

Bollen, K. A. and Curran, P. J. (2006) *Latent Curve Models: A Structural Equation Perspective*. John Wiley & Sons, Inc., Hoboken, NJ.

Bonner, R. E. (1964) On some clustering techniques. *International Business Machines Journal of Research and Development*, **8**, 22–32.

Both, M. and Gaul, W. (1986) Ein Vergleich zweimodaler Clusteranalyseverfahren. *Methods of Operations Research*, **57**, 593–605.

Bouguila, N. and Amayri, O. (2009) A discrete mixture-based kernel for SVMs: application to spam and image categorization. *Information Processing and Management*, **45**, 631–642.

Bowman, A. W. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, Oxford.

Boyce, A. J. (1969) Mapping diversity. A comparative study of some numerical methods, in *Numerical Taxonomy* (A. J. Cole, ed.) Academic Press, New York.

Bradley, P. and Fayyad, U. (1998). Refining initial points for K-means clustering, in *Machine Learning. Proceedings of the Fifteenth International Conference (ICML'98)*, 91–99.

Brainerd, G. W. (1951) The place of chronological ordering in archaeological analysis. *American Antiquity*, **16**, 301–313.

Brame, R., Nagin, D. S. and Wasserman, L. (2006) Exploring some analytical characteristics of finite mixture models. *Journal of Quantitative Criminology*, **22**, 31–59.

Branchaud, E. A., Cham, J. G., Nenadic, Z. *et al.* (2010) A miniature robot for autonomous single neuron recordings, in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 1920–1926.

Brandeau, M. L. and Chiu, S. S. (1988) Parametric facility location in a tree network with an L_p norm cost function. *Transportation Science*, **22**, 59–69.

Breckenridge, J. N. (1989) Replicating cluster analysis: method, consistency and validity. *Multivariate Behavioural Research*, **24**, 147–161.

Breiger, R. L, Boorman, A. and Arabie, P. (1975) An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, **12**, 328–383.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*. CRC, Boca Raton, FL.

Brinsky-Fay, C., Kohler, U. and Luniak, M. (2006) Sequence analysis with Stata. *Stata Journal*, **6**, 435–461.

Brusco, M. J. (2006). A repetitive branch-and-bound procedure for minimum within-cluster sums of squares partitioning. *Psychometrika*, **71**, 347–363.

Brusco, M. J. and Cradit, J. D. (2001) A variable-selection heuristic for *K*-means clustering. *Psychometrika*, **66**, 249–270.

Brusco, M. J. and Kohn, H. F. (2008) Optimal partitioning of a data set based on the p-median model. *Psychometrika*, **73**, 89–105.

Brusco, M. J. and Kohn, H. F. (2009) Exemplar-based clustering via simulated annealing. *Psychometrika*, **74**, 457–475.

Brusco, M. J. and Steinley, D. (2007) A comparison of heuristic procedures for minimum within-cluster sums of squares partitioning. *Psychometrika*, **72**, 583–600.

Bryant, P. and Williamson, J. A. (1978) Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika*, **65**, 273–282.

Bryant, P. and Williamson, J. A. (1986) Maximum likelihood and classification: a comparison of three approaches, in *Classification as a Tool of Research* (W. Gaul and M. Schader, eds) 35–45. Elsevier, Amsterdam.

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997) Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.

Bullmore, E., Brammer, M., Rouleau, G. *et al.* (1995) Computerized brain tissue classification of magnetic resonance images: a new approach to the problem of partial volume artifact. *Neuroimage*, **2**, 133–147.

Bullmore, E., Brammer, M., Williams, S. C. R. *et al.* (1996) Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, **35**, 261–277.

Burnham, K. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference*. Springer, New York.

Bushel, P. R., Wolfinger, R. D. and Gibson, G. (2007) Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. *BMC Systems Biology*, **23**, 1–15.

Cailliez, F. (1983) The analytical solution to the additive constant problem. *Psychometrika*, **48**, 305–308.

Cailliez, F. and Kuntz, P. (1996) A contribution to the study of metric and Euclidean structures of dissimilarities. *Psychometrika*, **61**, 241–253.

Calinski, R. B. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communications in Statistics*, **3**, 1–27.

Campbell, J. G., Fraley, C., Murtagh, F. and Raftery, A. E. (1997) Linear flaw detection in woven textiles using model-based clustering. *Pattern Recognition Letters*, **18**, 1539–1548.

Campbell, N. A. and Mahon, R. J. (1974) A multivariate study of variation in two species of rock crab of genus *Leptograpsus*. *Australian Journal of Zoology*, **22**, 417–425.

Canty, A. and Ripley, B. (2010). boot: Bootstrap R (S-Plus) Functions. R package version 1.2–42. http://cran.r-project.org/.

Carlin, B. P. and Chib, S. (1995) Bayesian model choice via Markov-chain Monte-Carlo methods. *Journal of the Royal Statistical Society Series B*, **57**, 473–484.

Carmichael, J. W. and Sneath, P. H. A. (1969) Taxometric maps. *Systematic Zoology*, **18**, 402–415.

Carmichael, J. W., George, L. A. and Julius, R. S. (1968) Finding natural clusters. *Systematic Zoology*, **17**, 144–150.

Carmone, F. J., Kara, A. and Maxwell, S. (1999) HINoV: A new model to improve market segment definition by identifying noisy variables. *Journal of Marketing Research*, **36**, 501–509.

Carroll, J. D. and Arabie, P. (1983) INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika*, **48**, 157–169.

Carroll, J. D. and Chang, J. J. (1973) A method for fitting a class of hierarchical tree structure models to dissimilarities data and its application to some 'body parts' data of Miller's. *Proceedings of the 81st Annual Convention of the American Psychological Association*, **8**, 1097–1098.

Cattell, R. B. and Coulter, M. A. (1966) Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. *British Journal of Mathematical and Statistical Psychology*, **19**, 237–269.

Celeux, G. and Govaert, G. (1992) A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, **4**, 315–332.

Celeux, G. and Govaert, G. (1993) Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation*, **47**, 127–146.

Celeux, G. and Govaert, G. (1995) Gaussian parsimonious clustering models. *Pattern Recognition*, **28**, 781–793.

Celeux, G. and Soromenho, G. (1996) An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195–212.

Celeux, G., Hurn, M. and Robert, C. P. (2000) Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95**, 957–970.

Celeux, G., Martin, O. and Lavergne, C. (2005) Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, **5**, 243–267.

Cerioli, A. and Zani, S. (2001) Exploratory methods for detecting high density regions in cluster analysis, in *Advances in Classification and Data Analysis* (S. Borra, R. Rocci, M. Vichi and M. Schader, eds). Springer, Berlin.

Chaddha, R. L. and Marcus, L. F. (1968) An empirical comparison of distance statistics for populations with unequal covariance matrices. *Biometrics*, **24**, 683–694.

Chae, S. S. and Warde, W. D. (2006) Effect of using principal coordinates and principal components on retrieval of clusters. *Computational Statistics and Data Analysis*, **50**, 1407–1417.

Chakrapani, C. (2004) *Statistics in Market Research*. Arnold, London.

Chang, W. C. (1983) On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, **32**, 267–275.

Chatfield, C. and Collins, A. J. (1980) *Introduction to Multivariate Analysis*. Chapman and Hall, London.

Cheeseman, P. and Stutz, J. (1995) Bayesian classification (Autoclass C): theory and results, in *Advances in Knowledge Discovery and Data Mining* (U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, eds) 61–83. American Association for Artificial Intelligence Press, Menlo Park, CA.

Cheetham, H. L. and Hazel, J. E. (1969) Binary (presence-absence) similarity coefficients. *Journal of Paleontology*, **43**, 1130–1136.

Chen, H., Schuffels, C. and Orwig, R. (1996) Internet categorization and search: a self-organizing approach. *Journal of Visual Communication and Image Representation*, **7**, 88–102.

Chen, M. H., Shao, Q.-M. and Ibrahim, J. G. (2000) *Monte Carlo Methods in Bayesian Computation*. Springer, New York.

Chen, R., Mantel, N. and Klingberg, M. A. (1984) A study of three techniques for time-space clustering in Hodgkin's disease. *Statistics in Medicine*, **3**, 173–184.

Chen, Z. and Van Ness, J. W. (1996) Space-conserving agglomerative algorithms. *Journal of Classification*, **13**, 157–168.

Cheng, B. and Titterington, D. M. (1994) Neural networks: a review from a statistical perspective. *Statistical Science*, **9**, 2–54.

Cheng, M.-Y. and Hall, P. (1998) Calibrating the excess mass and dip tests of modality. *Journal of the Royal Statistical Society Series B*, **60**, 579–589.

Cheng, R. and Milligan, G. W. (1996) Measuring the influence of individual data points in a cluster analysis. *Journal of Classification*, **13**, 315–335.

Chepoi, V. and Fichet, B. (1997) Recognition of Robinsonian dissimilarities. *Journal of Classification*, **14**, 311–325.

Chib, S. (1995) Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **90**, 1313–1321.

Chiu, T., Fang, D., Chen, J. *et al.* (2001) A robust and scalable clustering algorithm for mixed type attributes in large database environment, in *KDD '01: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Christie, O. H. J., Brennan, J. A. and Straume, E. (1979) Multivariate classification of Roman glasses found in Norway. *Archaeometry*, **21**, 233–241.

Cleveland, W. S. (1994) *The Elements of Graphing Data*. Nobart Press, Summit, NJ.

Cliff, A. D., Haggett, P., Smallman-Raynor, M. R. *et al.* (1995) The application of multidimensional scaling methods to epidemiological data. *Statistical Methods in Medical Research*, **4**, 102–123.

Clogg, C. C. (1981) New developments in latent structure analysis, in *Factor Analysis and Measurement in Sociological Research* (D. J. Jackson and E. F. Borgotta, eds) 215–246. Sage Publications, Beverly Hills.

Clogg, C. C. (1996) Latent class models, in *Handbook of Statistical Modeling for Social and Behavioral Sciences* (G. Arminger, C. C. Clogg and M. E. Sobel, eds) 311–359. Plenum, New York.

Cohen, A., Gnanadesikan, R., Kettenring, J. R. and Landwehr, J. M. (1977) Methodological developments in some applications of clustering, in *Applications of Statistics* (P. R. Krishnaiah, ed.) 141–162. North-Holland, Amsterdam.

Cohen, J. (1960) A coefficient of agreement for normal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Cohen, J. (1994) The earth is round ($p < .05$). *American Psychologist*, **49**, 997–1003.

Connell, A. M. and Frye, A. A. (2006) Growth mixture modelling in developmental psychology: overview and demonstration of heterogeneity in developmental trajectories of adolescent antisocial behaviour. *Infant and Child Development*, **15**, 609–621.

Cook, D. and Swayne, D. F. (2007) *Interactive and Dynamic Graphics for Data Analysis.* Springer, New York.

Cooper, P. W. (1964) Non supervised adaptive signal detection. *Pattern Recognition, Information and Control*, **7**, 416–444.

Cormack, R. M. (1971) A review of classification. *Journal of the Royal Statistical Society A*, **134**, 321–367.

Cortese, J. D. (2000) The array of today: biomolecule arrays become the 21st century test tube. *The Scientist*, **14**, 25.

Cross, G. C. and Jain, A. K. (1982) Measurement of clustering tendency, in *Proceedings of IFAC Symposium on Theory and Application of Digital Control (Volume 2) New Delhi 24–29*.

Croudace, T. J., Jarvelin, M. R., Wadsworth, M. E. J. and Jones, P. B. (2003) Developmental typology of trajectories to nighttime bladder control: epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology*, **157**, 834–842.

Cunningham, K. M. and Ogilvie, L. C. (1972) Evaluation of hierarchical grouping techniques: a preliminary study. *Computer Journal*, **15**, 209–213.

Curran, P. J. and Hussong, A. M. (2003) The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology*, **112**, 526–544.

Dai, X. F., Erkkila, T., Yli-Harja, O. and Lahdesmaki, H. (2009) A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data. *BMC Bioinformatics*, **10**, 165.

Dasgupta, A. and Raftery, A. E. (1998) Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**, 294–302.

Datta, S. and Datta, S. (2006) Evaluation of clustering algorithms for gene expression data. *BMC Bioinformatics*, **7**, S17.

Davenport, M. and Studdert-Kennedy, G. (1972) The statistical analysis of aesthetic judgement: an exploration. *Applied Statistics*, **21**, 324–332.

Day, N. E. (1969) Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

Day, W. H. E. (1996) Complexity theory: an introduction for practitioners of classification, in *Clustering and Classification* (P. Arabie, L. J. Hubert and G. De Soete, eds) 199–233. World Scientific, Singapore.

Dayton, C. M. and Macready, G. B. (1988) Concomitant-variable latent class models. *Journal of the American Statistical Association*, **83**, 173–178.

Dean, N. and Raftery, A. E. (2009) clustvarsel: Variable Selection for Model-Based Clustering, Manual. R package version 1.3. http://cran.r-project.org/.

Dean, N. and Raftery, A. E. (2010) Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, **62**, 11–35.

De Boeck, P. and Rosenberg, S. (1988) Hierarchical classes: model and data analysis. *Psychometrika*, **53**, 361–381.

Degerman, R. (1982) Ordered binary trees constructed through an application of Kendall's tau. *Psychometrika*, **47**, 523–527.

De la Cruz-Mesia, R., Quintanab, F. A. and Marshall, G. (2008) Model-based clustering for longitudinal data. *Computational Statistics and Data Analysis*, **52**, 1441–1457.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

De Querioz, A. (1993) For consensus (sometimes). *Systematic Biology*, **42**, 368–372.

De Sarbo, W. S. (1982) GENNCLUS: New model for general nonhierarchical cluster analysis. *Psychometrika*, **53**, 361–381.

De Sarbo, W. S., Carroll, J. D., Clark, L. A. and Green, P. E. (1984) Synthesized clustering: a method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, **49**, 57–78.

De Soete, G. (1984a) A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters*, **2**, 133–137.

De Soete, G. (1984b) Ultrametric tree representation of incomplete dissimilarity data. *Journal of Classification*, **1**, 235–242.

De Soete, G. (1986) Optimal variable weighting for ultrametric and additive tree clustering. *Quality and Quantity*, **20**, 169–180.

De Soete, G. and Carroll, J. D. (1996) Tree and other network modes of representing proximity data, in *Clustering and Classification* (P. Arabie, L. J. Hubert and G. De Soete, eds) 157–197. World Scientific, Singapore.

De Soete, G., De Sarbo, W. S., Furnas, G. W. and Carroll, J. D. (1984) The estimation of ultrametric and path length trees from rectangular proximity data. *Psychometrika*, **49**, 289–310.

Dewdney, A. K. (1997) *Yes, We Have No Neutrons: An Eye-Opening Tour through the Twists and Turns of Bad Science*. John Wiley & Sons, Inc., New York.

Dias, J. G. and Vermunt, J. K. (2008) A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, **23**, 643–659.

Diday, E. (1986) Orders and overlapping clusters by pyramids, in *Multidimensional Data Analysis* (J. De Leeuw, W. Heiser, J. Meulman and F. Critchley, eds) 201–234. DSWO Press, Leiden.

Diday, E. and Govaert, G. (1977) Classification automatique avec distances adaptives. *RAIRO Informatique Théorique et Applications*, **11**, 329–349.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B*, **56**, 363–375.

Diehr, G. (1985) Evaluation of a branch and bound algorithm for clustering. *SIAM Journal on Scientific and Statistical Computing*, **6**, 268–284.

Diggle, P. J. (1983) *Statistical Analysis of Spatial Point Pattern*. Academic Press, London.

Diggle, P. J., Heagerty, P., Liang K.-Y. and Zeger, S. (2002) *Analysis of Longitudinal Data* (2nd edn). Oxford University Press, Oxford.

Dimitriadou, E., Dolnicar, S. and Weingessel, A. (2002) An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, **67**, 137–159.

Dolan, C. V. and Van der Maas, H. L. J. (1998) Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, **63**, 227–253.

Dubes, R. and Jain, A. K. (1979) Validity studies in clustering methodologies. *Pattern Recognition*, **8**, 247–260.

Duda, R. O. and Hart, P. E. (1973) *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., New York.

Duflou, H. and Maenhaut, W. (1990) Application of principal component and cluster analysis to the study of the distribution of minor and trace elements in the normal human brain. *Chemometrics and Intelligent Laboratory Systems*, **9**, 273–286.

Dunn, J. C. (1974) Some recent investigations of a new fuzzy partitioning algorithm and its application to pattern classification problems. *Journal of Cybernetics*, **4**, 1–15.

Dunson, D. B. (2009) Bayesian nonparametric hierarchical modeling. *Biometrical Journal*, **51**, 273–284.

Dyen, I., Kruskal, J. and Black, P. (1992) An Indoeuropean classification, a lexicostatistical experiment. *Transactions of the American Philosophical Society*, **82**, 1–132.

Eckes, T. and Orlik, P. (1993) An error variance approach to two-mode hierarchical clustering. *Journal of Classification*, **10**, 51–74.

Edelbrock, C. (1979) Comparing the accuracy of hierarchical clustering algorithms: the problem of classifying everybody. *Multivariate Behavioral Research*, **14**, 367–384.

Ederer, F., Myers, M. H. and Mantel, N. (1966) A statistical problem in space and time: do leukemia cases come in clusters? *Biometrics*, **20**, 626–638.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965) A method for cluster analysis. *Biometrics*, **21**, 362–375.

Efron, B. (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

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

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, P. O. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA*, **95**, 14863–14868.

Elliott, M. R. (2007) Identifying latent clusters of variability in longitudinal data. *Biostatistics*, **8**, 756–771.

Ellis, T. E., Rudd, M. D., Harsan Rayab, M. and Wehrly, T. E. (1996) Cluster analysis of McMI scores of suicidal psychiatric patients: four personality profiles. *Journal of Clinical Psychology*, **52**, 411–422.

Ertoz, L., Steinbach, M. and Kumar, V. (2002) A new shared nearest neighbour clustering algorithm and its applications, in *Workshop on Clustering High Dimensional Data and its Applications*, 105–115.

Espejo, E. and Gaul, W. (1986) Two-mode hierarchical clustering as an instrument for market research, in *Classification as a Tool of Research* (W. Gaul and M. Schader, eds) 121–128. North-Holland, Amsterdam.

Estabrook, C. G. and Rodgers, D. J. (1966) A general method of taxonomic description for a computed similarity measure. *Bioscience*, **16**, 789–793.

Everitt, B. S. (1981) A Monte Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, **16**, 171–180.

Everitt, B. S. (1987) *Introduction to Optimization Methods and Their Application in Statistics*. Chapman and Hall CRC, London.

Everitt, B. S. (1988) A finite mixture model for the clustering of mixed mode data. *Statistics and Probability Letters*, **6**, 305–309.

Everitt, B. S. (2005) *An R and S-PLUS Companion to Multivariate Analysis*. Springer, New York.

Everitt, B. S. and Bullmore, E. T. (1999) Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping*, **7**, 1–14.

Everitt, B. S. and Dunn, G. (2001) *Applied Multivariate Data Analysis*. Arnold, London.

Everitt, B. S. and Hand, D. J. (1981) *Finite Mixture Distributions*. Chapman and Hall CRC, London.

Everitt, B. S. and Hothorn, T. (2009) *A Handbook of Statistical Analyses Using R* (2nd edn). Chapman and Hall, Boca Raton.

Everitt, B. S. and Merette, C. (1989) The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, **17**, 283–297.

Everitt, B. S. and Pickles, A. (2004) *Statistical Aspects of the Design and Analysis of Clinical Trials* (2nd edn). Imperial College Press, London.

Everitt, B. S. and Rabe-Hesketh, S. (1997) *The Analysis of Proximity Data*. Arnold, London.

Fairburn, C. G. and Cooper, Z. (1993) The eating disorder examination (12th edn) in *Binge Eating: Nature, Assessment and Treatment* (C. G. Fairburn and G. T. Wilson, eds) 317–360. Guilford Press, New York.

Farmer, A., McGuffin, P. and Spitznagel, E. L. (1983) Heterogeneity in schizophrenia: a cluster analytic approach. *Psychiatric Research*, **8**, 1–12.

Faúndez-Abans, M., Ormeno, M. I. and de Oliveira-Abans, M. (1996) Classification of planetary nebulae by cluster analysis and artificial neural networks. *Astronomy Astrophysics Supplement*, **116**, 395–402.

Feldman, J. (1995) Perceptual models of small dot clusters, in *Partitioning Data Sets* (I. Cox, P. Hansen and B. Julesz, eds) DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 19, 331–364. American Mathematical Society, Providence, RI.

Felsenstein, J. (1989) PHYLIP – Phylogeny Inference Package (Ver. 3.2). *Cladistics*, **5**, 165–166.

Ferligoj, A. and Batagelj, V. (1982) Some types of clustering with relational constraints. *Psychometrika*, **47**, 541–552.

Fienberg, S. E., Meyer, M. and Wasserman, S. (1985) Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, **80**, 51–67.

Fisher, L. and Van Ness, J. W. (1971) Admissible clustering procedures. *Biometrika*, **58**, 91–104.

Fisher, R. A. (1936) The use of multiple measurements on taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004) *Applied Longitudinal Analysis*. John Wiley & Sons, Inc., Hoboken, NJ.

Fleiss, L. L. and Zubin, J. (1969) On the methods and theory of clustering. *Multivariate Behavioral Research*, **4**, 235–250.

Florek, K., Lukaszewiez, L., Perkal, L. *et al.* (1951) Sur la liaison et la division des points d'un ensemble fini. *Colloquium Mathematicum*, **2**, 282–285.

Ford, I. and Holmes, A. P. (1998) Functional neuroimaging and statistics, in *Statistical Analysis of Medical Data: New Developments* (B. S. Everitt and G. Dunn, eds) 277–307. Arnold, London.

Forgy, E. W. (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classification. *Biometrics*, **21**, 768–769.

Forina, M., Armanino, C., Catino, M., and Ubigli, M. (1986) Multivariate data analysis as a discriminating method of the origin of wines. *Vitis*, **25**, 189–201.

Fowlkes, E. B. and Mallows, C. L. (1983) A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**, 553–569.

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. (1988) Variable selection in clustering. *Journal of Classification*, **5**, 205–228.

Fraley, C. and Raftery, A. E. (1998) How many clusters? Which clustering method? – Answers via model-based cluster analysis. *Computer Journal*, **41**, 578–588.

Fraley, C. and Raftery, A. E. (1999) MCLUST: Software for model-based cluster analysis. *Journal of Classification*, **16**, 297–306.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Fraley, C. and Raftery, A. E. (2003) Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *Journal of Classification*, **20**, 263–286.

Fraley, C. and Raftery, A. E. (2006) MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report No. 504, Department of Statistics, University of Washington.

Fraley, C. and Raftery, A. E. (2007) Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, **24**, 155–181.

Fraley, C. and Raftery, A. E. (2010) mclust: Model-Based Cluster Analysis. R package version 3.4.6. http://cran.r-project.org/web/packages/mclust/index.html.

Friedman, H. P. and Rubin, J. (1967) On some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159–1178.

Friedman, J. H. (1987) Exploring projection pursuit. *Journal of the American Statistical Association*, **82**, 249–266.

Friedman, J. H. and Tukey, J. W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, **23**, 881–889.

Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. Springer, New York.

Frühwirth-Schnatter, S. and Pyne, S. (2010) Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, **11**, 317–336.

Furnas, G. W. (1984) The generation of random, binary unordered trees. *Journal of Classification*, **1**, 187–233.

Gale, N., Halperin, W. C. and Costanzo, C. M. (1984) Unclassed matrix shading and optimal ordering in hierarchical cluster analysis. *Journal of Classification*, **1**, 75–92. (Erratum: **1**, 289.).

Galimberti, G., Montanari, A. and Viroli, C. (2009) Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics and Data Analysis*, **53**, 4301–4310.

Gammermann, D. and Lopez, H. F. (2006) *MCMC Statistical Simulation for Bayesian Inference* (2nd edn). Chapman and Hall, Boca Raton.

Ganesalingam, J., Stahl, D., Wijesekera, L. *et al.* (2009) Latent cluster analysis of ALS phenotypes identifies prognostically differing groups. *Plos One*, **4**, e7107.

Ganesalingam, S. (1989) Classification and mixture approaches to clustering via maximum likelihood. *Journal of the Royal Statistical Society Series C*, **38**, 455–466.

Gara, M. A., Silver, R. C., Escobar, J. I. *et al.* (1998) A hierarchical classes analysis (HICLAS) of primary care patients with medically unexplained somatic symptoms. *Psychiatry Research*, **81**, 77–86.

Garson, G. D. (1998) *Neural Networks: An Introductory Guide for Social Scientists*. Sage, London.

Gascuel, O. and McKenzie, A. (2004) Performance analysis of hierarchical clustering algorithms. *Journal of Classification*, **21**, 3–18.

Gaul, W. and Schader, M. (1994) Pyramidal classification based on incomplete dissimilarity data. *Journal of Classification*, **11**, 171–193.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003) *Bayesian Data Analysis* (2nd edn). Chapman and Hall, London.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Ghosh, D. and Chinnaiyan, A. M. (2002) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**, 275–286.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall CRC, London.

Gilks, W. R., Roberts, G. O. and Suhu, S. K. (1998) Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, **93**, 1045–1054.

Girgensohn, A. and Boreczky, J. (2000) Time-constrained keyframe selection technique. *Multimedia Tools and Applications*, **11**, 347–358.

Gitman, I. and Levine, M. D. (1970) An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique. *IEE Transactions on Computers*, **19**, 583–593.

Gnanadesikan, R. (1997) *Methods for Statistical Data Analysis of Multivariate Observations* (2nd edn). John Wiley & Sons, Inc., New York.

Gnanadesikan, R., Kettenring, J. R. and Landwehr, J. M. (1977) Interpreting and assessing the results of cluster analyses, in *Bulletin of the International Statistical Institute: Proceedings of the 41st Session (New Delhi)* Book 2, 451–463. ISI, Voorburg, Netherlands

Gnanadesikan, R., Kettenring, J. R. and Tsao, S. L. (1995) Weighting and selection of variables. *Journal of Classification*, **12**, 113–136.

Golub, T. R., Slonim, D. K., Tamayo, P. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Good, I. J. and Gaskins, R. A. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, **75**, 42–73.

Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Gordon, A. D. (1973) Classification in the presence of constraints. *Biometrics*, **29**, 821–827.

Gordon, A. D. (1980) *Classification* (1st edn). Chapman and Hall CRC, London.

Gordon, A. D. (1987) A review of hierarchical classification. *Journal of the Royal Statistical Society A*, **150**, 119–137.

Gordon, A. D. (1990) Constructing dissimilarity measures. *Journal of Classification*, **7**, 257–269.

Gordon, A. D. (1998) Cluster validation, in *Data Science, Classification and Related Methods* (C. Hayashi, N. Ohsumi, K. Yajima *et al.*, eds) 22–39. Springer-Verlag, Tokyo.

Gordon, A. D. (1999) *Classification* (2nd edn). Chapman and Hall/CRC, Boca Raton, FL.

Govaert, G. and Nadif, M. (1996) Comparison of the mixture and the classification maximum likelihood in cluster analysis with binary data. *Computational Statistics and Data Analysis*, **23**, 65–81.

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gower, J. C. (1967) A comparison of some methods of cluster analysis. *Biometrics*, **23**, 623–628.

Gower, J. C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–872.

Gower, J. C. (1974) Maximal predictive classification. *Biometrics*, **30**, 643–654.

Gower, J. C. (1985) Measures of similarity, dissimilarity and distance, in *Encyclopaedia of Statistical Sciences*, Vol. 5 (S. Kotz, N. L. Johnson and C. B. Read, eds) 397–405. John Wiley & Sons, Inc., New York.

Gower, J. C. (1990) Clustering axioms. *Classification Society of North America Newsletter*, July, 2–3.

Gower, J. C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, **5**, 5–48.

Gower, J. C. and Ross, G. J. S. (1969) Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, **18**, 54–64.

Green, P. E., Frank, R. E. and Robinson, P. J. (1967) Cluster analysis in test market selection. *Management Science*, **13**, 387–400.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Greenacre, M. (1984) *Theory and Applications of Correspondence Analysis*. Academic Press, Orlando, FL.

Grun, B. and Leisch, F. (2007) Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics and Data Analysis*, **51**, 5247–5252.

Guha, S., Rajeev, R. and Kyuseok, S. (1998) CURE: An efficient clustering algorithm for large databases, in *Proceedings of the ACM SIGMOD Conference on Management of Data, Seattle, USA*, 73–84.

Guttiérrez Toscano, P. and Marriott, F. H. C. (1999) Unsupervised classification of chemical compounds. *Applied Statistics*, **48**, 153–163.

Hagenaars, J. A. (1988) Latent structure models with direct effects between indicators – local dependence models. *Sociological Methods & Research*, **16**, 379–405.

Hagenaars, J. A. and McCutcheon, A. L. (2002) *Applied Latent Class Analysis*. Cambridge University Press, Cambridge.

Han, C. and Carlin, B. P. (2001) Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association*, **96**, 1122–1132.

Hand, D. J. (1981) *Discrimination and Classification*. John Wiley & Sons, Ltd, Chichester.

Hands, S. and Everitt, B. S. (1987) A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. *Multivariate Behavioral Research*, **22**, 235–243.

Hansen, P. and Jaumard, B. (1997) Cluster analysis and mathematical programming. *Mathematical Programming*, **79**, 191–215.

Hansen, P. and Mladenovic, N. (2001). J-MEANS: A new local search heuristic for minimum sum of squares clustering. *Pattern Recognition*, **34**, 405–413.

Hansen, P., Jaumard, B. and Sanlaville, E. (1994) Partitioning problems in cluster analysis: a review of mathematical programming approaches, in *New Approaches in Classification and Data Analysis* (E. Diday, Y. Lechevallier, M. Schader *et al.*, eds) 228–240. Springer-Verlag, Berlin.

Hansen, P., Jaumard, B. and Simeone, B. (1996) Espaliers: A generalization of dendrograms. *Journal of Classification*, **13**, 107–127.

Hardin, J., Mitani, A., Hicks, L. and VanKoten, B. (2007) A robust measure of correlation between two genes on a microarray. *BMC Bioinformatics*, **8**, 1–13.

Hartigan, J. A. (1967) Representation of similarity matrices by trees. *Journal of the American Statistical Association*, **62**, 1140–1158.

Hartigan, J. A. (1975) *Clustering Algorithms*. John Wiley & Sons, Inc., New York.

Hartigan, J. A. (1988) The span test for unimodality, in *Classification and Related Methods of Data Analysis* (H. H. Bock, ed.) 229–236. North-Holland, Amsterdam.

Hartigan, J. A. and Hartigan, P. M. (1985) The dip test of multimodality. *Annals of Statistics*, **13**, 70–84.

Hartigan, J. A. and Mohanty, S. (1992) The RUNT test for multimodality. *Journal of Classification*, **9**, 63–70.

Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136. A k-means clustering algorithm. *Applied Statistics*, **28**, 100–108.

Hartvig, N. V. and Jensen, J. L. (2000) Spatial mixture modeling of fMRI data. *Human Brain Mapping*, **11**, 233–248.

Hasselblad, V. (1966) Estimation of parameters for a mixture of normal distributions. *Technometrics*, **8**, 431–444.

Hasselblad, V. (1969) Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, **64**, 1459–1471.

Hastie, T. (1998) Neural networks, in *Encyclopedia of Biostatistics*, Vol. 4 (P. Armitage and T. Colton, eds) 2986–2989. John Wiley & Sons, Ltd, Chichester.

Hathaway, R. J. and Bezdek, J. C. (1988) Recent convergence for the fuzzy c-means clustering algorithms. *Journal of Classification*, **5**, 237–247.

Hawkins, D. M., Muller, M. W. and ten Krooden, J. A. (1982) Cluster analysis, in *Topics in Applied Multivariate Analysis* (D. M. Hawkins, ed.) Cambridge University Press, Cambridge.

Hay, P. J., Fairburn, C. G. and Doll, H. A. (1996) The classification of bulimic eating disorders: a community-based cluster analysis study. *Psychological Medicine*, **26**, 801–812.

Heinrich, I., O'Hara, H., Sweetman, B. and Anderson, J. A. D. (1985) Validation aspects of an empirically derived classification for non-specific low back pain. *The Statistician*, **34**, 215–230.

Henning, C. (2004) Breakdown points for maximum likelihood estimators of location–scale mixtures. *Annals of Statistics*, **3**, 1313–1340.

Hernández-Avilia, A. (1979) Problems in cluster analysis, D. Phil Thesis, University of Oxford.

Heyer, L. J., Kruglyak, S. and Yooseph, S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Research*, **9**, 1106–1115.

Hix-Small, H., Duncan, T. E., Duncan, S. C., and Okut, H. (2004) A multivariate associative finite growth mixture modeling approach examining adolescent alcohol and marijuana use. *Journal of Psychopathology and Behavioral Assessment*, **26**, 255–270.

Hodson, F. R. (1971) Numerical typology and prehistoric archaeology, in *Mathematics in the Archaeological and Historical Sciences* (F. R. Hodson, D. G. Kendall and P. A. Tautu, eds) 30–45. Edinburgh University Press, Edinburgh.

Hodson, F. R., Sneath, P. H. A. and Doran, J. E. (1966) Some experiments in the numerical analysis of archaeological data. *Biometrika*, **53**, 311–324.

Hubálek, Z. (1982) Coefficients of association and similarity, based on binary (presence–absence) data: an evaluation. *Biological Review*, **57**, 669–689.

Hubert, L. (1974) Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association*, **69**, 698–704.

Hubert, L. J. and Arabie, P. (1985) Comparing partitions. *Journal of Classification*, **2**, 193–218.

Hughes, T. R., Marton, M. J., Jones, A. R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Hunt, L. and Jorgensen, M. (2003) Mixture model clustering for mixed data with missing information. *Computational Statistics & Data Analysis*, **41**, 429–440.

Huth, R., Nemesova, I. and Klimperova, N. (1993) Weather categorization based on the average linkage clustering technique: an application to European mid-latitudes. *International Journal of Climatology*, **13**, 817–835.

Ichino, M. and Yaguchi H. (1994) Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, **24**, 698–708.

Ideker, T., Thorsson, V., Ranish, J. A. *et al.* (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.

Ismail, M. A. and Kamel, M. S. (1989) Multidimensional data clustering utilizing hybrid search strategies. *Pattern Recognition*, **22**, 75–89.

Jaccard, P. (1908) Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise de Sciences Naturelles*, **44**, 223–370.

Jackson, C. (1994) Appendix A, Tables A1 and A2, in *Exploratory Multivariate Analysis in Archaeology* (M. J. Baxter, ed.) 228–231. Edinburgh University Press, Edinburgh.

Jackson, J. J. (2003) *A User's Guide to Principal Components Analysis*. John Wiley & Sons, Inc., Hoboken, NJ.

Jackson, M. (1989) *Michael Jackson's Malt Whisky Companion: A Connoisseur's Guide to the Malt Whiskies of Scotland*. Dorling Kindersley, London.

Jain, A. and Dubes, R. C. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

Jain, A. and Farrokhina, F. (1991) Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, **24**, 1167–1186.

Jajuga, K. and Walesiak, M. (2000) Standardisation of data set under different measurement scales, in *Classification and Information Processing at the Turn of the Millennium* (R. Decker and W. Gaul, eds.) 105–112 Springer-Verlag, Heidelberg.

Jajuga, K., Walesiak, M. and Bak, A. (2003) On the general distance measure, in *Exploratory Data Analysis in Empirical Research* (M. Schwaiger and O. Opitz, eds.) 104–109. Springer-Verlag, Heidelberg.

Jambu, M. (1978) *Classification Automatique pour l'Analyse des Données*, Vol. 1. Dunod, Paris.

Janasik, N., Honkela, T. and Bruun, H. (2009) Text mining in qualitative research: application of an unsupervised learning method. *Organizational Research Methods*, **12**, 436–460.

Jancey, R. C. (1966) Multidimensional group analysis. *Australian Journal of Botany*, **14**, 127–130.

Janik, V. M. (1999) Pitfalls in the categorization of behaviour; a comparison of dolphin whistle classification methods. *Animal Behaviour*, **57**, 133–143.

Jansen, R. C. (1993) Maximum-likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics*, **49**, 227–231.

Jardine, C. J., Jardine, N. and Sibson, R. (1967) The structure and construction of taxonomic hierarchies. *Mathematical Biosciences*, **1**, 173–179.

Jardine, N. and Sibson, R. (1968) The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, **11**, 117–184.

Jardine, N. and Sibson, R. (1971) *Mathematical Taxonomy*. John Wiley & Sons, Ltd, Chichester.

Jaro, M. A. (1995) Probabilistic linkage of large public health data files. *Statistics in Medicine*, **14**, 491–498.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20**, 50–67.

Jasra, A., Stephens, D. A., Gallagher, K. and Holmes, C. C. (2006) Bayesian mixture modelling in geochronology via Markov chain Monte Carlo. *Mathematical Geology*, **38**, 269–300.

Jedidi, K., Jagpal, H. S. and Desarbo, W. S. (1997) Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, **16**, 39–59.

Jeffries, N. O. (2003) A note on 'testing the number of components in a normal mixture'. *Biometrika*, **90**, 991–994.

Jensen, R. E. (1969) A dynamic programming algorithm for cluster analysis. *Operations Research*, **17**, 1034–1057.

Jiao, S. and Zhang, S. P. (2008) The t-mixture model approach for detecting differentially expressed genes in microarrays. *Functional & Integrative Genomics*, **8**, 181–186.

Johnson, S. C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Jolliffe, I. T., Jones, B. and Morgan, B. J. T. (1988) Stability and influence in cluster analysis, in *Data Analysis and Informatics V* (E. Diday, ed.) 507–514. North-Holland, Amsterdam.

Jones, K. S. and Jackson, D. M. (1967) Current approaches to classification and clump finding at the Cambridge Language Research Unit. *Computer Journal*, **1**, 29–37.

Jones, M. C. and Sibson, R. (1987) What is projection pursuit? *Journal of the Royal Statistical Society A*, **150**, 1–38.

Jorgensen, M. and Hunt, L. A. (1999) Mixture model clustering using the MULTIMIX program. *Australian and New Zealand Journal of Statistics*, **41**, 153–171.

Jukes, T. H. and Cantor, C. (1969) *Evolution of Protein Molecules*. Academic Press, New York.

Kaiser, S. and Leisch, F. (2008) A Toolbox for Bicluster Analysis in R. Technical Report Number 28, Department of Statistics, University of Munich.

Kapp, A. V. and Tibshirani, R. (2007) Are clusters in one dataset present in another dataset? *Biostatistics*, **8**, 9–31.

Kaski, S., Kanga, J. and Kohonen, T. (1998) Bibliography of self-organising map (SOM) papers: 1981–1997. *Neural Computing Surveys*, **1**, 102–350.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Katz, J. O. and Rohlf, E. L. (1973) Function point cluster analysis. *Systematic Zoology*, **22**, 295–301.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding Groups in Data. An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., New York.

Kaufman, L. and Rousseeuw, P. J. (2005) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Inc., Hoboken, NJ.

Kelly, F. P. and Ripley, B. D. (1976) A note on Strauss's model for clustering. *Biometrika*, **63**, 357–360.

Keribin, C. (2000) Consistent estimation of the order of mixture. *Sankhya*, **62**, 49–66.

Kerr, M. K. and Churchill, G. A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of the USA*, **98**, 8961–8965.

Kettenring, J. R. (2009) A patent analysis of cluster analysis. *Applied Stochastic Models in Business and Industry*, **5**, 460–467.

Khalili, A., Huang, T. and Lin, S. L. (2009) A robust unified approach to analyzing methylation and gene expression data. *Computational Statistics & Data Analysis*, **53**, 1701–1710.

Kidd, K. K. and Sgaramella-Zonta, L. A. (1971) Phylogenetic analysis: concepts and methods. *American Journal of Human Genetics*, **23**, 235–252.

Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

King, W. M., Giess, S. A. and Lombardino, L. J. (2007). Subtyping of children with developmental dyslexia via bootstrap aggregated clustering and the gap statistic: comparison with the double-deficit hypothesis. *International Journal of Language & Communication Disorders*, **42**, 77–95.

Kirkpatrick, S., Gelatt, C. D., Jr. and Vecchi, M. P. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Kirriemuir, J. W. and Willett, P. (1995) Use of cluster analysis methods for analysing the outputs of multiple data-base searches, in *Electronic Library and Visual Information Research: Proceedings of the Second ELVIRA Conference* (M. Collier and K. Arnold, eds) 117–126. Aslib, London.

Klein, R. W. and Dubes, R. C. (1989) Experiments in projection and clustering by simulated annealing. *Pattern Recognition*, **22**, 213–220.

Knox, G. (1964) Epidemiology of childhood leukaemia in Northumberland and Durham. *British Journal of Preventive and Social Medicine*, **18**, 17–24.

Koehler, A. B. and Murphree, E. S. (1988) A comparison of the Akaike and Schwarz criteria for selecting model order. *Journal of the Royal Statistical Society Series C*, **37**, 187–195.

Kohn, H. F., Steinley, D. and Brusco, M. J. (2010). The p-median model as a tool for clustering psychological data. *Psychological Methods*, **15**, 87–95.

Kohonen, T. (1982) Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, **43**, 59–69.

Kohonen, T. (1997) *Self-Organizing Maps: Second Extended Edition*, Springer Series in Information Sciences, Vol. **30** Springer-Verlag, Berlin.

Kohonen, T., Kaski, S., Lagus, K. K. *et al.* (2000) Self organisation of massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, **11**, 574–585.

Koontz, W. L. G., Narendra, P. M. and Fukunaga, K. (1975) A branch and bound clustering algorithm. *IEEE Transactions on Computers*, **24**, 908–915.

Kraepelin, E. (1919) *Dementia Praecox and Paraphrenia*. Livingstone, Edinburgh.

Krause, E. F. (1975) *Taxicab Geometry*. Addison Wesley, Menlo Park, CA.

Krieger, A. M. and Green, P. (1999) A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika*, **64**, 341–353.

Krzanowski, W. J. (1988) *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, Oxford.

Kurczynski, T. W. (1969) Genetic drift in a human isolate. PhD Thesis, Case Western University.

Kurczynski, T. W. (1970) Generalised distance and discrete variables. *Biometrics*, **26**, 525–534.

Kurtz, A., Moller, H. J., Bavidl, G. *et al.* (1987) Classification of parasuicide by cluster analysis. *British Journal of Psychiatry*, **150**, 520–525.

Kuss, O., Gromann, C. and Diepgen, T. L. (2006) Model-based clustering of binary longitudinal atopic dermatitis disease histories by latent class mixture models. *Biometrical Journal*, **48**, 105–116.

Lance, G. N. and Williams, W. T. (1966) Computer programs for hierarchical polythetic classification. *Computer Journal*, **9**, 60–64.

Lance, G. N. and Williams, W. T. (1967) A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal*, **9**, 373–380.

Lance, G. N. and Williams, W. T. (1968) Note on a new information-statistic classificatory program. *Computer Journal*, **11**, 195.

Langfelder, P., Zhang, B. and Horvath, S. (2008) Defining clusters from a hierarchical cluster tree; the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–720.

Lapointe, F.-J. and Legendre, P. (1992) Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees. *Systematic Biology*, **41**, 378–384.

Lapointe, F.-J. and Legendre, P. (1994) A classification of pure malt Scotch whiskies. *Applied Statistics*, **43**, 237–257.

Lapointe, F.-J. and Legendre, P. (1995) Comparison tests for dendrograms: a comparative evaluation. *Journal of Classification*, **12**, 265–282.

Larson, R. C. and Sadiq, G. (1983) Facility locations with the Manhattan metric in the presence of barriers to travel. *Operations Research*, **31**, 652–699.

Laviolette, M., Seaman, J. W., Jr., Barrett, J. D. and Woodall, W. H. (1995) A probabilistic and statistical view of fuzzy methods (with discussion). *Technometrics*, **37**, 249–292.

Lawless, H. T. (1989) Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis. *Chemical Senses*, **14**, 349–360.

Lawrence, C. J. and Krzanowski, W. J. (1996) Mixture separation for mixed-mode data. *Statistics and Computing*, **6**, 85–92.

Lazarsfeld, P. L. and Henry, N. W. (1968) *Latent Structure Analysis*. Houghton Mifflin Co., Boston.

Lee, M. L. T., Kuo, F. C., Whitmore, G. A. and Sklar, J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 9834–9839.

Leenen, I., Van Mechelen I., Gelman A. and De Knop, S. (2008) Bayesian hierarchical classes analysis. *Psychometrika*, **73**, 39–64.

Lefkovitch, L. P. (1978) Cluster generation and grouping using mathematical programming. *Mathematical Biosciences*, **41**, 91–110.

Lefkovitch, L. P. (1980) Conditional clustering. *Biometrics*, **36**, 43–58.

Legendre, L. and Legendre, P. (1983) *Numerical Ecology*. Elsevier, Amsterdam.

Legendre, P. and Chodorowski, A. (1977) A generalisation of Jaccard's association coefficient for *Q*-analysis of multi-state ecological data matrices. *Ekologia Polska*, **25**, 297–308.

Leisch, F. (2006). A toolbox for K-centroids cluster analysis. *Computational Statistics & Data Analysis*, **51**, 526–544.

Leisch, F. (2009) Neighborhood graphs, stripes and shadow plots for cluster visualization. *Statistics and Computing*, **20**, 457–469.

Lerman, I. C. (1987) Construction d'un indice de similarité entre objets décrits par des variables d'un type quelconque. Application au problème du consensus en classification (1). *Revue de Statistique Appliquée*, **25**, 39–60.

Levenshtein, V. I. (1966) Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and control theory*, **10**, 845–848.

Levine, R. R. J. (1981) Sex differences in schizophrenia: timing or subtypes? *Psychological Bulletin*, **90**, 432–444.

Lewis, S. M. and Raftery, A. E. (1997) Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, **92**, 648–655.

Liang, F. M. and Wong, W. H. (2001) Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association*, **96**, 653–666.

Ling, R. F. (1972) On the theory and construction of *k*-clusters. *Computer Journal*, **15**, 326–332.

Ling, R. F. (1973) A probability theory for cluster analysis. *Journal of the American Statistical Association*, **68**, 159–164.

Lingoes, J. C. (1971) Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, **36**, 195–203.

Little, R. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., New York.

Littmann, T. (2000) An empirical classification of weather types in the Mediterranean Basin and their interrelation with rainfall. *Theoretical and Applied Climatology*, **66**, 161–171.

Liu, B. (2007) *Web Data Mining*. Springer, New York.

Liu, G. L. (1968) *Introduction to Combinatorial Mathematics*. McGraw Hill, New York.

Liu, S. and George, R. (2005) *Mining Weather Data using Fuzzy Cluster Analysis*. Springer, Berlin.

Lo, Y. T., Mendell, N. R. and Rubin, D. B. (2001) Testing the number of components in a normal mixture. *Biometrika*, **88**, 767–778.

Loehlin, J. C. (2004) *Latent Variable Models* (4th edn). Lawrence Erlbaum Associates, Mahwah, NJ.

Lubke, G. H. and Muthén, B. O. (2005) Investigating population heterogeneity with factor mixture models. *Psychological Methods*, **10**, 21–39.

Lubke, G. and Neale, M. C. (2006) Distinguishing between latent classes and continuous factors: resolution by maximum likelihood? *Multivariate Behavioral Research*, **41**, 499–532.

Lubke, G. and Neale, M. (2008) Distinguishing between latent classes and continuous factors with categorical outcomes: class invariance of parameters of factor mixture models. *Multivariate Behavioral Research*, **43**, 592–620.

Lubke, G. H., Muthén, B. O., Moilanen, I. K. *et al.* (2007) Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort. *Journal of the American Academy of Child and Adolescent Psychiatry*, **46**, 1584–1593.

Lubke, G. H., Hudziak, J. J., Derks, E. M. *et al.* (2009) Maternal ratings of attention problems in ADHD: evidence for the existence of a continuum. *Journal of the American Academy of Child and Adolescent Psychiatry*, **48**, 1085–1093.

MacNaughton-Smith, P., Williams, W. T., Dale, M. B. and Mockett, L. G. (1964) Dissimilarity analysis. *Nature*, **202**, 1034–1035.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (L. LeCam and J. Neymen, eds) Vol. 1, 281–297. University of California Press, Berkeley.

Maelzel, A., Johnson, S. H., Woodbury, M. and Bombardier, C. (2000) Use of grade membership analysis to profile the practice styles of individual physicians in the management of acute low back pain. *Journal of Clinical Epidemiology*, **53**, 195–205.

Magidson, J. and Vermunt, J. K. (2001) Latent class factor and cluster models, bi-plots, and related graphical displays. *Sociological Methodology*, **31**, 223–264.

Mahalanobis, P. C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Science, India*, **12**, 49–55.

Maij-de Meij, A. M., Kelderman, H., and van der Flier, H. (2008) Fitting a mixture item response theory model to personality questionnaire data: characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, **32**, 611–631.

Mallory-Greenough, J. M. and Greenough, J. D. (1998) New data for old pots: trace element characterization of Ancient Egyptian pottery using ICP-MS. *Journal of Archaeological Science*, **25**, 85–97.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Manton, K. G., Gu, X. L., Huang, H. and Kovtun, M. (2004) Fuzzy set analyses of genetic determinants of health and status disability. *Statistical Methods in Medical Research*, **13**, 395–408.

Maravalle, M., Simeone, B. and Naldini, R. (1997) Clustering on trees. *Computational Statistics and Data Analysis*, **24**, 217–234.

Marden, J. I. (2000) Hypothesis testing: From p values to Bayes factors. *Journal of the American Statistical Association*, **95**, 1316–1320.

Margush, T. and McMorris, F. R. (1981) Consensus *n*-trees. *Bulletin of Mathematical Biology*, **43**, 239–244.

Marin, J. M., Mengersen, K., and Roberts, C. P. (2005) Bayesian modelling and inferences on mixtures of distributions, in *Bayesian Thinking, Modeling and Computation*, (D. Dey and C. R. Rao, eds) 15840–15845. Elsevier, Amsterdam.

Maronna, R. and Jacovkis, P. M. (1974) Multivariate clustering procedures with variable metrics. *Biometrics*, **30**, 499–505.

Marriott, F. H. C. (1971) Practical problems in a method of cluster analysis. *Biometrics*, **27**, 501–514.

Marriott, F. H. C. (1982) Optimization methods of cluster analysis. *Biometrika*, **69**, 417–421.

Maugis, C., Celeux, G. and Martin-Magniette, M. L. (2009a) Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**, 701–709.

Maugis, C., Celeux, G. and Martin-Magniette, M. L. (2009b) Variable selection in model-based clustering: a general variable role modeling. *Computational Statistics & Data Analysis*, **53**, 3872–3882.

Maulik, U. and Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition*, **33**, 1455–1465.

McCormick, W. R. T., Schweitzer, P. J. and White, T. W. (1972) Problem decomposition and data reorganisation by a clustering technique. *Operations Research*, **20**, 993–1009.

McCrone, P., Leese, M., Thornicroft, G. *et al.* (2000) Reliability of the Camberwell Assessment of Need – European Version: Epsilon Study 6. *British Journal of Psychiatry*, **177** (Suppl. 39), s34–s40.

McCulloch, W. S. and Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–118.

McIntyre, R. M. and Blashfield, R. K. (1980) A nearest-centroid technique for evaluating the minimum variance clustering procedure. *Multivariate Behavioral Research*, **22**, 225–238.

McLachlan, G. J. (1987) On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics*, **36**, 318–324.

McLachlan, G. J. (2004) *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc., Hoboken, NJ.

McLachlan, G. J. and Basford, K. E. (1988) *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions* (1st edn). John Wiley & Sons, Inc., New York.

McLachlan, G. J. and Krishnan, T. (2008) *The EM Algorithm and Extensions* (2nd edn). John Wiley & Sons, Inc., Hoboken, NJ.

McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. John Wiley & Sons, Inc., New York.

McLachlan, G. J., Bean, R. W. and Peel, D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

McLachlan, G. J., Peel, D. and Bean, R. W. (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**, 379–388.

McLachlan, G. J., Bean, R. W. and Jones, L. B. T. (2006) A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, **22**, 1608–1615.

McNicholas, P. D. and Murphy, T. B. (2008) Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**, 285–296.

McNicholas, P. D. and Murphy, T. B. (2010) Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, **38**, 153–168.

McNicholas, P. D., Murphy, T. B., Mcdaid, A. F. and Frost, D. (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**, 711–723.

McQuitty, L. L. (1966) Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, **27**, 21–46.

McRae, D. J. (1971) Micka, a Fortran IV iterative *k*-means cluster analysis program. *Behavioural Science*, **16**, 423–424.

Meghani, S. H., Lee, C. S., Hanlon, A. L. and Bruner, D. W. (2009) Latent class cluster analysis to understand heterogeneity in prostate cancer treatment utilities. *BMC Medical Informatics and Decision Making*, **9**, 47.

Meng, X. L. and vanDyk, D. (1997) The EM algorithm – An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B*, **59**, 511–540.

Mengersen, K. and Robert, C. P. (1996) Testing for mixtures: a Bayesian entropic approach (with discussion), in *Bayesian Statistics 5* (J. Berger, J. Bernando, D. Lindley and A. Smith, eds) 255–276. Oxford University Press, Oxford.

Michailidou, C., Maheras, P. Arseni-Papadimititriou, A. *et al.* (2009) A study of weather types at Athens and Thessaloniki and their relationship to circulation types for the cold-wet period, part I: two-step cluster analysis. *Theoretical and Applied Climatology*, **97**, 163–177.

Michener, C. D. (1970) Diverse approaches to systematics. *Evolutionary Biology*, **4**, 1–38.
Milligan, G. W. (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, **45**, 325–342.
Milligan, G. W. (1981) A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research*, **16**, 379–407.
Milligan, G. W. (1989) A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification*, **6**, 53–71.
Milligan, G. W. (1996) Clustering validation: results and implications for applied analyses, in *Clustering and Classification* (P. Arabie, L. J. Hubert and G. De Soete, eds) 341–375. World Scientific, Singapore.
Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.
Milligan, G. W. and Cooper, M. C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, **21**, 41–58.
Milligan, G. W. and Cooper, M. C. (1988) A study of standardisation of variables in cluster analysis. *Journal of Classification*, **5**, 181–204.
Mirkin, B. (1996) *Mathematical Classification and Clustering*. Kluwer, Dordrecht.
Mojena, R. (1977) Hierarchical grouping methods and stopping rules: an evaluation. *Computer Journal*, **20**, 359–363.
Moon, J. W. and Moser, L. (1965) On cliques in graphs. *Israel Journal of Mathematics*, **3**, 23–28.
Mora, P. A., Bennett, I. M., Elo, I. T. *et al.* (2009) Distinct trajectories of perinatal depressive symptomatology: evidence from growth mixture modeling. *American Journal of Epidemiology*, **169**, 24–32.
Morgan, B. J. T. (1973) Cluster analysis of two acoustic confusion matrices. *Perception and Psychophysics*, **13**, 13–24.
Morgan, B. J. T. and Ray, A. P. G. (1995) Non-uniqueness and inversions in cluster analysis. *Applied Statistics*, **44**, 117–143.
Morris, C. N. (1983) Parametric empirical Bayes inference: theory and application. *Journal of the American Statistical Association*, **78**, 47–55.
Murtagh, F. (1985) *Multidimensional Clustering Algorithms*, COMPSTAT Lectures 4. Physica-Verlag, Vienna.
Murtagh, F. D. (1995) Contiguity-constrained hierarchical clustering, in *Partitioning Data Sets* (I. Cox, P. Hansen and B. Julesz, eds) DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 19, 143–152. American Mathematical Society, Providence, RI.
Murtagh, F. and Hernández-Pajares, M. (1995) The Kohonen self-organizing feature map method: an assessment. *Journal of Classification*, **12**, 165–190.
Murtagh, F. and Raftery, A. E. (1984) Fitting straight lines to point patterns. *Pattern Recognition*, **17**, 479–483.
Muthén, B. O. (2002) Beyond SEM: general latent variable modeling. *Behaviormetrika*, **29**, 81–117.
Muthén, B. O. (2006) Should substance use disorders be considered as categorical or dimensional? *Addiction*, **101**, 6–16.
Muthén, B. O. and Asparouhov, T. (2006) Item response mixture modeling: application to tobacco dependence criteria. *Addictive Behaviors*, **31**, 1050–1066.
Muthén, B. O. and Muthén, L. K. (2000) Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, **24**, 882–891.

Muthén, B. O. and Shedden, K. (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463–469.

Muthén, L. K. and Muthén, B. O. (2007) *Mplus User's Guide. Fifth Edition*. Muthén & Muthén, Los Angeles, CA.

Muthén, L. K. and Muthén, B. O. (2010) *Mplus User's Guide. Sixth Edition*. Muthén & Muthén, Los Angeles, CA.

Nagin, D. S. (1999) Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological Methods*, **4**, 139–157.

Nagin, D. S. (2010) *Group-Based Modeling of Development*. Harvard University Press, Boston.

Nagin, D. S. and Tremblay, R. E. (1999) Trajectories of boys' physical aggression, opposition, and hyperactivity on the path to physically violent and nonviolent juvenile delinquency. *Child Development*, **70**, 1181–1196.

Nagin, D. S. and Tremblay, R. E. (2005) Developmental trajectory groups: fact or a useful statistical fiction? *Criminology*, **43**, 873–904.

Nasibov, E. N. and Ulutagay, G. (2010) Comparative clustering analysis of bispectral index series of brain activity. *Expert Systems with Applications*, **37**, 2495–2504.

Navarro, D. J. and Griffiths, T. L. (2008) Latent features in similarity judgments: a nonparametric Bayesian approach. *Neural Computation*, **20**, 2597–2628.

Nazareth, I., Landau, S., Yardley, L. and Luxon, L. (2006). Patterns of presentations of dizziness in primary care – a cross-sectional cluster analysis study. *Journal of Psychosomatic Research*, **60**, 395–401.

Neal, R. M. (1996) Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, **6**, 353–366.

Needham, R. M. (1965) Computer methods for classification and grouping, in *The Use of Computers in Anthropology* (I. Hymes, ed.) 345–356. Mouton, The Hague.

Needham, R. M. (1967) Automatic classification in linguistics. *The Statistician*, **17**, 45–54.

Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, **48**, 443–453.

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002) On spectral clustering: analysis and an algorithm, in *Advances in Neural Information Processing Systems 14* (T. G. Dietterich, S. Becker and Z. Ghahramani, eds), 849–856. MIT Press, Cambridge, MA.

Ng, S. K., McLachlan, G. J., Wang, K. *et al.* (2006) A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, **22**, 1745–1752.

Nicholls, G. (2008) Horses or farmers? *The tower of Babel and confidence in trees. Significance*, **5**, 112–117.

Nickerson, R. S. (2000) Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, **5**, 241–301.

Nobile, A. and Fearnside, A. T. (2007) Bayesian finite mixtures with an unknown number of components: the allocation sampler. *Statistics and Computing*, **17**, 147–162.

Nolan, D. and Speed, T. P. (1999) Teaching statistics theory through applications. *The American Statistician*, **53**, 370–376.

Nylund, K. L., Asparouhov, T. and Muthen, B. O. (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, **14**, 535–569.

O'Shea, J. M. (1985) Cluster analysis and mortuary patterning: an experimental assessment. *Journal of the European Study Group on Physical, Chemical and Mathematical Techniques Applied to Archaeology*, **11**, 91–110.

Ohsumi, M. and Nakamara, N. (1989) Space distorting properties in agglomerative hierarchical clustering algorithms and a simplified method for combinatorial method, in *Data Analysis, Learning Symbolic and Numeric Knowledge* (E. Diday, ed.) 103–108. Nova Science Publishers, New York.

Overall, J. E. and Magee, K. N. (1992) Replication as a rule for determining the number of clusters in hierarchical cluster analysis. *Applied Psychological Measurement*, **16**, 119–128.

Pacheco, J. and Valencia, O. (2003) Design of hybrids for the minimum sum-of-squares clustering problem. *Computational Statistics & Data Analysis*, **43**, 235–248.

Pal, N. R. and Bezdek, J. C. (1995) On cluster validity for the fuzzy *c*-means model. *IEEE Transactions on Fuzzy Systems*, **3**, 370–379.

Pan, J.-X. and Fang, K.-T. (2002) *Growth Curve Models with Statistical Diagnostics*. Springer, New York.

Pan, W., Lin, J. and Le, C. T. (2003) A mixture model approach to detecting differentially expressed genes with microarray data. *Functional & Integrative Genomics*, **3** (3), 117–124.

Panayirci, E. and Dubes, R. C. (1983) A test for multidimensional clustering tendency. *Pattern Recognition*, **16**, 433–444.

Parker-Rhodes, A. F. and Jackson, D. M. (1969) Automatic classification in the ecology of the higher fungi, in *Numerical Taxonomy* (A. J. Cole, ed.). Academic Press, New York.

Paykel, E. S. (1971) Classification of depressed patients: a cluster analysis derived grouping. *British Journal of Psychiatry*, **118**, 275–288.

Paykel, E. S. and Rassaby, E. (1978) Classification of suicide attempters by cluster analysis. *British Journal of Psychiatry*, **133**, 42–52.

Payne, R. W. and Preece, D. A. (1980) Identification keys and diagnostic tables: a review. *Journal of the Royal Statistical Society A*, **143**, 253–292.

Pearson, K. (1894) Contribution to the mathematical theory of evolution. *Philosophical Transactions A*, **185**, 71–110.

Peel, D. and McLachlan, G. J. (1999) User's Guide to EMMIX: Version 1.3. Department of Mathematics, University of Queensland. www.maths.uq.edu.au/∼gjm/emmix/emmix.html.

Peel, D. and McLachlan, G. J. (2000) Robust mixture modelling using the *t*-distribution. *Statistics and Computing*, **10**, 339–348.

Piccarreta, R. and Billari, F. C. (2007) Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society A*, **170**, 1061–1078.

Pickering, R. M. and Forbes, J. F. (1984) A classification of Scottish infants using latent class analysis. *Statistics in Medicine*, **3**, 249–259.

Pickles, A. and Croudace, T. J. (2010) Latent mixture models for multivariate and longitudinal outcomes. *Statistical Methods in Medical Research*, **19**, 271–289.

Pilowsky, I., Levine, S. and Boulton, D. M. (1969) The classification of depression by numerical taxonomy. *British Journal of Psychiatry*, **115**, 937–945.

Pledger, S. and Phillpot, P. (2008) Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal*, **50**, 1022–1034.

Podani, J. (1989) New combinatorial SAHN clustering methods. *Vegetatio*, **81**, 61–77.

Posse, C. (1990) An effective two-dimensional projection pursuit algorithm. *Communications in Statistics. Simulation and Computation*, **19**, 1143–1164.

Posse, C. (1995) Projection pursuit exploratory data analysis. *Computational Statistics and Data Analysis*, **20**, 669–687.

Pourahmadi, M. (1999) Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, **86**, 677–690.

Prelić, A., Bleuler, S., Zimmermann, P. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Price, R. H. and Bouffard, D. L. (1974) Behavioral appropriateness and situational constraint as dimensions of social behavior. *Journal of Personality and Social Psychology*, **30**, 579–586.

Principe, J. C. and Miikkulainen, R. (2009) *Advances in Self-Organizing Maps: 7th International Workshop, WSOM 2009, St. Augustine, Florida, June 8–10 2009: Proceedings*. Springer, Berlin.

Priness, I., Maimon, O. and Ben-Gal, I. (2007) Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, **8**, 1–12.

Putter, H., Vos, T., de Haes, H. and van Houwelingen, H. (2008) Joint analysis of multiple longitudinal outcomes: application of a latent class model. *Statistics in Medicine*, **27**, 6228–6249.

Qin, L. X. and Self, S. G. (2006) The clustering of regression models method with applications in gene expression data. *Biometrics*, **62**, 526–533.

Qu, P. P. and Qu, Y. S. (2000) A Bayesian approach to finite mixture models in bioassay via data augmentation and Gibbs sampling and its application to insecticide resistance. *Biometrics*, **56**, 1249–1255.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing, Reference Index, Version 2.10.1*. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org.

Rabe-Hesketh, S., Bullmore, E. T. and Brammer, M. J. (1997) The analysis of functional magnetic resonance images. *Statistical Methods in Medical Research*, **6**, 215–237. 1997.

Rabe-Hesketh, S., Brammer, M. J. and Bullmore, E. T. (1998) Localizing brain activation in a single subject using functional magnetic resonance imaging. *Statistical Methods in Medical Research*, **6**, 215–237.

Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2004) GLLAMM Manual. UC Berkeley Division of Biostatistics Working Paper Series. Working Paper 160. www.bepress.com/ucbbiostat/paper160/.

Radloff, L. S. (1977) 'The CES-D scale: a self report depression scale for research in the general population' *Applied Psychological Measurement*, **1**, 385–401.

Raftery, A. E. (1995) Bayesian model selection in social research. *Sociological Methodology*, **25**, 111–163.

Raftery, A. E. and Dean, N. (2006) Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**, 168–178.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

Rasmussen, M. and Willett, P. (1989) Efficiency of hierarchic agglomerative clustering using the ICL Distributed Array Processor. *Journal of Documentation*, **45** (1), 1–24.

Ray, S. and Lindsay, B. G. (2008) Model selection in high dimensions: a quadratic-risk-based approach. *Journal of the Royal Statistical Society Series B*, **70**, 95–118.

Reaven, G. M. and Miller, R. G. (1979) Attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, **16**, 17–24.

Reboussin, B. A., Ip, E. H. and Wolfson, M. (2008) Locally dependent latent class models with covariates: an application to under-age drinking in the USA. *Journal of the Royal Statistical Society Series A*, **171**, 877–897.

Richardson, S. and Green, P. J. (1997) A Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, **59**, 731–792.

Ripley, B. D. (1981) *Spatial Statistics*. John Wiley & Sons, Inc., New York.

Ripley, B. D. (1994) Neural networks and related methods for classification (with discussion). *Journal of the Royal Statistical Society B*, **56**, 409–456.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rissanen, J. (1978) Modelling by shortest data description. *Automatica*, **14**, 465–471.

Robert, C. P. and Titterington, D. M. (1998) Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, **8**, 145–158.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Rogers, D. J. and Tanimoto, T. T. (1960) A computer program for classifying plants. *Science*, **132**, 1115–1118.

Rohlf, F. J. (1970) Adaptive hierarchical clustering schemes. *Systematic Zoology*, **19**, 58–82.

Rohlf, F. J. (1975) A new approach to the computation of the Jardine-Sibson B_k clusters. *Computer Journal*, **18**, 164–168.

Rosenberg, S., Van Mechelen, I. and De Boeck, P. (1996) A hierarchical classes model: theory and method with applications in psychology and psychopathology, in *Clustering and Classification* (P. Arabie, L. J. Hubert and G. De Soete, eds) 123–155. World Scientific, Singapore.

Rousseeuw, P. J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Rousseeuw, P. J. (1995) Fuzzy clustering at the intersection. *Technometrics*, **37**, 283–286.

Rubin, J. (1967) Optimal classification into groups: an approach for solving the taxonomy problem. *Journal of Theoretical Biology*, **15**, 103–144.

Rufo, M. J., Martin, J. and Perez, C. J. (2006) Bayesian analysis of finite mixture models of distributions from exponential families. *Computational Statistics*, **21**, 621–637.

Russell, R., Meadows, L. A. and Russell, R. R. (2008) *Microarray Technology in Practice*. Academic Press, San Diego, CA.

Sammon, J. W. (1969) A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, **18**, 401–409.

Sampson, S. F. (1968) A novitiate in a period of change: an experimental and case study of social relationships. Doctoral dissertation, Cornell University.

Sander, J., Ester, M., Kriegel, H. *et al.* (1998) Density-based clustering in spatial databases: the algorithm DBSCAN and its applications. *Data Mining Knowledge Discovery*, **2**, 169–194.

Sarkar, D. (2008) *Lattice: Multivariate Visualization with R*. Springer, New York.

Sarstedt, M. and Schwaiger, M. (2008) Model selection in mixture regression analysis – a Monte Carlo simulation study, in *Data Analysis, Machine Learning and Applications* (C. Presiach, H. Burkhardt, H. Schmidt-Thieme and R. Decker, eds) 61–68. Springer, Berlin.

Saunders, R. and Funk, G. M. (1977) Poisson limits for a clustering model of Strauss. *Journal of Applied Probability*, **14**, 776–784.

Scheibler, D. and Schneider, W. (1985) Monte Carlo test of the accuracy of cluster analysis algorithms – a comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research*, **20**, 283–304.

Scherer, S. (2001) Early career patterns: a comparison of Great Britain and West Germany. *European Sociological Review*, **17**, 119–144.

Schlattmann, P. (2003) Estimating the number of components in a finite mixture model: the special case of homogeneity. *Computational Statistics and Data Analysis*, **41**, 441–451.

Schlattmann, P. (2005) On bootstrapping the number of components in finite mixtures of Poisson distributions. *Statistics and Computing*, **15**, 179–188.

Schlattmann, P. (2009) *Medical Applications of Finite Mixture Models*. Springer, Berlin.

Schlattmann, P. and Höhne, J. (2009) CAMAN: Finite Mixture Models and Metaanalysis Tools – Based on C.A.MAN. R package version 0.64. www.charite.de/biometrie/schlattmann/book.

Schneider, J. W. and Borlund, P. (2007a) Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology*, **58**, 1586–1595.

Schneider, J. W. and Borlund, P. (2007b) Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by the use of the Mantel and Procrustes statistics. *Journal of the American Society for Information Science and Technology*, **58**, 1596–1609.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Sclove, S. L. (1987) Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, **52**, 333–343.

Scott, A. J. and Symons, M. J. (1971) Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387–398.

Scott, D. W. (1992) *Multivariate Density Estimation*. John Wiley & Sons, Inc., New York.

Seidel, W., Mosler, K. and Alker, M. (2000) A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, **52**, 481–487.

Selim, S. Z. and Asultan, K. (1991) A simulated annealing algorithm for the clustering problem. *Pattern Recognition*, **24**, 1003–1008.

Selinski, S. and Ickstadt, K. (2008) Cluster analysis of genetic and epidemiological data in molecular epidemiology. *Journal of Toxicology and Environmental Health*, **71**, 835–844.

Shannon, W., Culverhouse, R. and Duncan, J. (2003) Analyzing microarray data using cluster analysis. *Pharmacogenomics*, **4**, 41–52.

Shepard, R. N. (1974) Representation of structure in similarity data: problems and prospects. *Psychometrika*, **39**, 373–421.

Shepard, R. N. and Arabie, P. (1979) Additive clustering: representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, **86**, 87–123.

Shevlin, M., Adamson, G., Vollebergh, W. *et al.* (2007) An application of item response mixture modelling to psychosis indicators in two large community samples. *Social Psychiatry and Psychiatric Epidemiology*, **42**, 771–779.

Sibson, R. (1970) A model for taxonomy. *Mathematical Biosciences*, **6**, 405–430.

Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society B*, **43**, 97–99.

Silverman, B. W. (1983) Some properties of a test for multimodality based on kernel density estimates, in *Probability, Statistics and Analysis* (J. F. C. Kingman and G. E. H. Reuter, eds) 248–259. Cambridge University Press, Cambridge.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall CRC, London.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. Springer-Verlag, New York.

Singleton, R. C. and Kautz, W. (1965) *Minimum Squared Error Clustering Algorithm*. Stanford Research Institute, Stanford, CA.

Sjőlander, K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170–179.

Skrondal, A. and Rabe-Hesketh, S. (2004) *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman and Hall/CRC, Boca Raton, FL.

Smith, G. P. and Pike, M. C. (1974) Case clustering in Hodgkin's disease: a brief review of the present position and report of work in Oxford. *Cancer Research*, **34**, 1156–1160.

Sneath, P. H. A. (1957) The application of computers to taxonomy. *Journal of General Microbiology*, **17**, 201–226.

Sneath, P. H. A. and Sokal, R. R. (1973) *Numerical Taxonomy*. W. H. Freeman, San Francisco.

Sokal, R. R. and Michener, C. D. (1958) A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, **38**, 1409–1438.

Sokal, R. R. and Rohlf, F. J. (1980) An experiment in taxonomic judgement. *Systematic Botany*, **5**, 341–365.

Sokal, R. R. and Rohlf, F. J. (1981) Taxonomic congruence in the Leptopodomorpha re-examined. *Systematic Zoology*, **20**, 209–225.

Sokal, R. R. and Sneath, P. H. (1963) *Principles of Numerical Taxonomy*. Freeman, London.

Sorensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter*, **5**, 1–34.

Soromenho, G. (1993) Comparing approaches for testing the number of components in a finite mixture model. *Computational Statistics*, **9**, 65–78.

Souto, M. C. P., Costa, I. G., de Araujo, D. S. A. *et al.* (2008) Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **9**, 497. doi: 10.1186/1471-2105-9-497.

Späth, H. (1985) *Cluster Dissection and Analysis*. Ellis Horwood, Chichester.

Spearman, C. (1904) The proof and measurement of association between two things. *American Journal of Psychology*, **15**, 72–101.

Sriram, N. and Lewis, S. (1993) Constructing optimal ultrametrics. *Journal of Classification*, **10**, 241–268.

Steinley, D. (2003) Local optima in K-means clustering: What you don't know may hurt you. *Psychological Methods*, **8**, 294–304.

Steinley, D. (2006a) K-means clustering: a half-century synthesis. *British Journal of Mathematical & Statistical Psychology*, **59**, 1–34.

Steinley, D. (2006b) Profiling local optima in K-means clustering: developing a diagnostic technique. *Psychological Methods*, **11**, 178–192.

Steinley, D. (2008) Stability analysis in K-means clustering. *British Journal of Mathematical & Statistical Psychology*, **61**, 255–273.

Steinley, D. and Brusco, M. J. (2007) Initializing K-means batch clustering: a critical evaluation of several techniques. *Journal of Classification*, **24**, 99–121.

Steinley, D. and Brusco, M. J. (2008a) A new variable weighting and selection procedure for *K*-means cluster analysis. *Multivariate Behavioral Research*, **43**, 77–108.

Steinley, D. and Brusco, M. J. (2008b) Selection of variables in cluster analysis. *Psychometrika*, **73**, 125–144.

Stephens, M. (2000) Label switching in mixture models. *Journal of the Royal Statistical Society B*, **62**, 795–809.

Stopford, J., Hughes, M. J. and Leese, M. N. (1991) A scientific study of medieval tiles from Bordesley Abbey, near Redditch (Hereford and Worcester). *Oxford Journal of Archaeology*, **10**, 349–360.

Strauss, D. J. (1975) A model for clustering. *Biometrika*, **62**, 467–475.

Strauss, J. S., Bartko, J. J. and Carpenter, W. T. (1973) The use of clustering techniques for the classification of psychiatric patients. *British Journal of Psychiatry*, **122**, 351–540.

Su, Y., Shan, S., Chen, X. and Gao, W. (2008) Classification based optimal discriminatory projection pursuit. *Computer Vision and Pattern Recognition*, **23**, 1–7.

Sun, J. (1998) Projection pursuit, in: *Encyclopedia of Statistical Sciences*, 2nd edn (S. Kptz, C. Read, D. Banks and N. Johnson, eds) 554–560. John Wiley & Sons, Inc., New York.

Sun, L.-X., Xie, Y.-L., Song, X.-H. *et al.* (1994) Cluster analysis by simulated annealing. *Computers and Chemistry*, **18**, 103–108.

Sutton, M. Q. and Reinhard, K. J. (1995) *Journal of Archaeological Studies*, **22**, 741–750.

Suzuki, R. and Shimodaira, H. (2006) pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinfomatics*, **22**, 1540–1542.

Swartz, M., Blazer, D., Woodbury, M. *et al.* (1986) Somatization disorder in a US southern community: use of a new procedure for analysis of medical classification. *Psychological Medicine*, **6**, 595–609.

Swartz, M. D., Mo, Q. X., Murphy, M. E. *et al.* (2008) Bayesian variable selection in clustering high-dimensional data with substructure. *Journal of Agricultural Biological and Environmental Statistics*, **13**, 407–423.

Swayne, D. F., Temple Lange, D., Buja, A. and Cook, D. (2003) GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics and Data Analysis*, **43**, 423–444.

Symons, M. J. (1981) Clustering criteria and multivariate normal mixtures. *Biometrics*, **39**, 35–43.

Tadesse, M. G., Sha, N. and Vannucci, M. (2005) Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**, 602–617.

Tajima, F. (1993) Unbiased estimation of evolutionary distance between nucleotide sequences. *Molecular Biology and Evolution*, **10**, 677–688.

Talairach, J. and Tournoux, P. (1988) *A Coplanar Stereotaxic Atlas of the Human Brain*. Thieme-Verlag, Stuttgart.

Tan, W. Y. and Chang, W. C. (1972) Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association*, **67**, 702–708.

Tanner, M. A. and Wong, W. H. (1988) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Tarpey, T., Yun, D. and Petkova, E. (2008) Model misspecification: finite mixture or homogeneous? *Statistical Modelling*, **8**, 199–218.

Thode, H. C., Finch, S. J. and Mendell, N. R. (1989) Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics*, **44**, 1195–1201.

Thorndike, R. L. (1953) Who belongs in a family? *Psychometrika*, **18**, 267–276.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B*, **63**, 411–423.

Tipping, M. E. and Bishop, C. M. (1999) Mixtures of probabilistic principal component analyzers. *Neural Computation*, **11**, 443–482.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Inc., New York.

Tofighi, D. and Enders, C. K. (2007) Identifying the correct number of classes in growth mixture models, in *Advances in Latent Variable Mixture Models* (G. R. Hancock and K. M. Samuelsen, eds) 317–341. Information Age, Charlotte.

Tsai, H.-R., Horng, S.-J., Lee, S.-S. *et al.* (1997) Parallel hierarchical clustering algorithm on processor arrays with a reconfigurable bus system. *Pattern Recognition*, **30**, 801–815.

Tubb, A., Parker, A. J. and Nickless, G. (1980) The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry*, **22**, 153–171.

Tufte, E. R. (1983) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

van Hattum, P. and Hoijtink, H. (2009) Market segmentation using brand strategy research: Bayesian inference with respect to mixtures of log-linear models. *Journal of Classification*, **26**, 297–328.

Van Mechelen, I., Bock, H.-H. and De Boeck, P. (2004) Two-mode clustering methods: a structural overview. *Statistical Methods in Medical Research*, **13**, 363–394.

van Os, B. J. and Meulman, J. J. (2004) Improving dynamic programming strategies for partitioning. *Journal of Classification*, **21**, 207–230.

Venables, W. N. and Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS* (2nd edn). Springer-Verlag, New York.

Verbeke, G. and Mohlenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer, New York.

Verdinelli, I. and Wasserman, L. (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, **90**, 614–618.

Vermunt, J. K. (1997) *Log-linear Models for Event Histories*. Sage Publications, Thousand Oaks.

Vermunt, J. K. and Magidson, J. (2000) *LatentGOLD*. Statistical Innovations, Belmont, MA.

Villarroel, L., Marshall, G. and Baron, A. E. (2009) Cluster analysis using multivariate mixed effects models. *Statistics in Medicine*, **28**, 2552–2565.

Wallace, C. R. and Boulton, D. M. (1968) An information measure for classification. *Computer Journal*, **11**, 185–194.

Wallace, M. and Denham, C. (1996) *The Office of National Statistics Classification of Local and Health Authorities of Great Britain*. HMSO, London.

Waller, N. G., Kaiser, H. A., Illian, J. B. and Manry, M. (1998) Cluster analysis with Kohonen neural networks. *Psychometrika*, **63**, 5–22.

Wand, M. P. and Jones, M. C. (1995) *Kernel Smoothing*. Chapman and Hall CRC, London.

Wang, H. X., Zhang, Q. B., Luo, B. and Wei, S. (2004) Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters*, **25**, 701–710.

Wang, K., Ng, S. K. and McLachlan, G. J. (2009) Multivariate skew t mixture models: applications to fluorescence-activated cell sorting data, in: *Proceedings of DICTA 2009 (Conference of Digital Image Computing: Techniques and Applications, Melbourne)* (H. Shi, Y. Zhang, M. J. Bottema et al.eds) 526–531. IEEE Computer Society, Los Alamitos, CA.

Ward, J. H. (1963) Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

Watanabe, M. and Yamaguchi, K. (2004) *The EM Algorithm and Related Statistical Models*. CRC Press, Boca Raton.

Wedel, M. (2001) GLIMMIX: Software for estimating mixtures and mixtures of generalized linear models. *Journal of Classification*, **18**, 129–135.

Wedel, M. (2002) Concomitant variables in finite mixture models. *Statistica Neerlandica*, **56**, 362–375.

Wedel, M. and Desarbo, W. S. (1994) A review of recent developments in latent class regression models, in *Advanced Methods of Marketing Research* (R. P. Bagozzi, ed) 352–388. Blackwell, Cambridge.

Wedel, M. and Desarbo, W. S. (2002) Mixture regression models, in *Applied Latent Class Analysis*, (J. A. Hagenaars and A. L. McCutcheon, eds) 366–382. Cambridge University Press, Cambridge.

Wehrens, R., Buydens, L. M. C., Fraley, C. and Raftery, A. E. (2004) Model-based clustering for image segmentation and large datasets via sampling. *Journal of Classification*, **21**, 231–253.

Williams, W. T. and Lambert, J. M. (1959) Multivariate methods in plant ecology, 1. Association analysis in plant communities. *Journal of Ecology*, **47**, 83–101.

Williams, W. T., Lambert, J. M. and Lance, G. N. (1966) Multivariate methods in plant ecology, V. Similarity analysis and information analysis. *Journal of Ecology*, **54**, 427–445.

Willse, A. and Boik, R. J. (1999) Identifiable finite mixtures of location models for clustering mixed-mode data. *Statistics and Computing*, **9**, 111–121.

Winkler, W. E. (1999) *The State of Record Linkage and Current Research Problems*, Internal Revenue Service Publication R99/04. Statistics of Income Division, Washington, DC. www.census.gov/srd/papers/pdf/rr99-04.pdf.

Wishart, D. (1969) Mode analysis, in *Numerical Taxonomy* (A. J. Cole, ed.) Academic Press, New York.

Wishart, D. (1973) An improved multivariate mode-seeking cluster method. Paper presented at Royal Statistical Society Conference on Multivariate Analysis and its Applications, Hull, England.

Wishart, D. (1987) Clustan User Manual (4th edn). Computing Laboratory, University of St Andrews.

Wishart, D. (1999) ClustanGraphics3: Interactive graphics for cluster analysis, in *Classification in the Information Age* (W. Gaul and H. Locarek-Junge, eds) 268–275. Springer-Verlag, Berlin.

Witten, D. M. and Tibshirani, R. (2010) Supervised multidimensional scaling for visualization, classification and bipartite ranking. *Computational Statistics and Data Analysis*, **55**, 789–801.

Wojdyla, D., Poletto, L., Cuesta, C. *et al.* (1996) Cluster analysis with constraints: its use with breast cancer mortality rates in Argentina. *Statistics in Medicine*, **15**, 741–746.

Wolfe, J. H. (1970) Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, **5**, 329–350.

Wolfe, J. H. (1971) A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. Technical Bulletin, STB 72-2, Naval Personnel and Training Research Laboratory, San Diego, CA.

Wong, M. A. (1982) A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association*, **77**, 841–847.

Wong, M. A. and Lane, T. (1983) A kth nearest neighbour clustering procedure. *Journal of the Royal Statistical Society B*, **45**, 362–368.

Wong, M. A. and Schaack, C. (1982) Using the *k*th nearest neighbor clustering procedure to determine the number of sub populations. *Proceedings of the Statistical Computing Section, American Statistical Association*, 40–48.

Woodbury, M. A. and Manton, K. G. (1982) A new procedure for the analysis of medical classification. *Methods of Information in Medicine*, **21**, 210–220.

Woodbury, M. A., Manton, K. G. and Tolley, H. D. (1994) A general model for statistical analysis using fuzzy sets: sufficient conditions for identifiability and statistical properties. *Information Sciences*, **1**, 149–180.

Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F. and Smith, S. M. (2005) Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. *IEEE Transactions on Medical Imaging*, **24**, 1–11.

Wright, C., Burns, T., James, P. *et al.* (2003) Assertive outreach teams in London: models of operation. *British Journal of Psychiatry*, **183**, 132–138.

Xu, L., Johnson, T. D., Nichols, T. E. and Nee, D. E. (2009) Modeling inter-subject variability in fMRI activation location: a Bayesian hierarchical spatial model. *Biometrics*, **65**, 1041–1051.

Yamaguchi, K. (2000) Multinomial logit latent-class regression models: an analysis of the predictors of gender-role attitude. *The American Journal of Sociology*, **105**, 1702–1740.

Yang, C. C. (2006) Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, **50**, 1090–1104.

Yang, C. C. and Yang, C. C. (2007) Separating latent classes by information criteria. *Journal of Classification*, **24**, 183–203.

Yang, X. and Krishnan, S. M. (2004) Image segmentation using finite mixtures and spatial information. *Image and Vision Computing*, **22**, 735–745.

Yeung, K. Y. and Ruzzo, W. L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.

Yeung, K. Y., Medvedovic, M. and Bumgarner, R. E. (2003) Clustering gene-expression data with repeated measurements. *Genome Biology*, **4**, R34.

Yeung, M., Yeo, B.-L. and Liu, B. (1996) Extracting story units from long programs for video browsing and navigation, in *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS 1996)* Tokyo, 296–305.

Yung, Y. F. (1997) Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, **62**, 297–330.

Yusuf, S., Peto, R., Lewis, J. *et al.* (1985) Beta-blockade during and after myocardial-infarction – an overview of the randomized trials. *Progress in Cardiovascular Diseases*, **27**, 335–371.

Zadeh, L. A. (1965) Fuzzy sets. *Information and Control*, **8**, 338–353.

Zahn, C. T. (1971) Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*, **20**, 68–86.

Zeng, G. and Dubes, R. C. (1985) A test for spatial randomness based on k-NN distances. *Pattern Recognition Letters*, **3**, 85–91.

Zhang, T., Ramakrishnan, R. and Livny, M. (1996) Birch: An efficient data clustering method for very large databases, in *Proceedings of the ACM SIGMOD Conference on Management of Data*, Montreal, Canada, 103–114.

Zhang, Y., Zhang, Z. P. and Zhou, Y. C. (1998) Research on the fuzzy and dynamic monitoring of the diesel engine operating conditions. *Proceedings of the Institution of Mechanical Engineers*, **212**, 421–426.

Zhu, H. T. and Lee, S. Y. (2001) A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, **66**, 133–152.

Zigmond, A. S. and Snaith, R. P. (1983) The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, **67**, 361–370.

Zimmerman, D. L. and Nunez-Anton, V. A. (2010) *Antedependence Models for Longitudinal Data*. Chapman and Hall/CRC, Boca Raton, FL.

Zupan, J. (1982) *Clustering of Large Data Sets*. Research Studies Press, Chichester.

Index