Optimal set covering for biological classification

$$\Sigma$$

$$\min \{ c^T x : Ax \geq 1, x \in \{0,1\}^m \}$$

Agriculture Canada

Numérisé par
Éditions et Services de dépôt,
Travaux publics et Services
gouvernementaux Canada - 2014

Digitized by
Publishing and Depository Services,
Public Works and Government Services
Canada - 2014

Numéro de catalogue / Catalogue Number: A53-1893/1993E-PDF

ISBN 978-0-660-20290-7

Publications du gouvernement du Canada / Government of Canada Publications
publications.gc.ca

Optimal set covering for biological classification

Optimal set covering for biological classification

L.P. Lefkovitch

Research Branch
Agriculture Canada

Publication 1893/E
1993

Available in Canada through
Associated Bookstores
and other booksellers
or by mail from

Canada Communication Group—Publishing
Ottawa, Canada K1A 0S9
Catalog No. A53-1893/1993E
ISBN 0-660-14821-8

Canadian Cataloguing in Publication Data

Lefkovitch, L.P. (Leonard P.)

Optimal set covering for biological classification /
L.P. Lefkovitch. –

(Publication ; 1893/E)

"Produced by Research Program Service."
Includes bibliographical references and index.
Cat no. A53-1893/1993E
ISBN 0-660-14821-8

1. Numerical taxonomy.
2. Cluster analysis. I. Canada. Agriculture Canada. Research Program Services Section. II. Title. III. Series: Publication (Canada. Agriculture Canada). English ; 1893/E.

QH83.L3 1993 574'.012 C93-099002-1

Produced by Research Program Service

Staff Editor: Jane T. Buckley

"Es gibt allerdings Unaussprechliches"

L. von Wittgenstein

Contents

Preface

The work described here had its origins in some early studies, using classical methods, in the taxonomy of beetles. While interesting to me, decisions were made about species' limits and generic content based on my subjective reactions to the available specimens. During those studies, discussions with many people reinforced my feelings of dissatisfaction with the subjectivity. This monograph is the result of my attempts to minimize its consequences.

I wish to thank in particular, C.E. Dyte, my colleague from 1956 to 1966 at the then Pest Infestation Laboratory, Slough, U.K. I also thank J.C. Gower, then at Rothamsted Experimental Station, Harpenden, U.K., with whom I enjoyed at first regular but now infrequent meetings since 1955. I have appreciated the comments made by participants at the Numerical Taxonomy and Classification Society meetings; by my colleagues in Agriculture Canada, the Canadian Museum of Nature, Carleton University, and in Rothamsted Experimental Station; and by reviewers of papers submitted to journals.

I am most grateful to Mark Wolynetz, who read the first draft I was prepared to show, and who drew my attention to ambiguities in notation and other oversights. Perhaps the real responsibility for any errors should be given to my family, who claimed that they understood what I had to say, even when I was uncertain of its meaning.

The spelling in this book follows Webster's Third New International Dictionary, the standard reference for scientific publication within the department, and is consistent with the Americanization of practically everything!

Abstract

The problem addressed in this book, how to reveal the natural groups contained within a set of n biological or other objects (Chapter I), is shown to be equivalent to optimal set covering, namely,

$$\text{choose } \mathbf{x} \text{ to minimize } (\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{1};\ \mathbf{x} \in \{0,1\}^m)$$

where $\mathbf{A}$ is a $n \times m$ incidence matrix of n objects in m subsets, and $\mathbf{c}$ is a measure of the "cost" of including each subset in the solution (Chapter II). It is emphasized that the solution is to be regarded as a hypothesis for further investigation.

Most of the book is devoted to methods for generating the subsets, among which are the following: directly from qualitative attributes (Chapter IV), from vector dissimilarity (Chapter V), and from single and multiple ordered attributes (Chapter VI). From scalar pairwise dissimilarity coefficients (Chapter VII), the key principle for generating subsets is the recursive principle of conditional clustering: for S_i a subset, S_{i+1} consists of all objects whose average distance to the members of S_i does not exceed the maximum among them (Chapters VI and VIII).

The costs, $\mathbf{c}$, are obtained from the logical relationships among the rows and columns of $\mathbf{A}$ based on the following observations: the more subsets to which an object belongs, the less information given by that object about the elements of $\mathbf{x}$; and the more objects a subset contains, the more likely is that subset to be part of the optimal solution (Chapter II).

Some special applications are discussed (mainly Chapter IX); a series of case studies (Chapter X) illustrates most of the procedures and principles.

Résumé

Le thème que l'on traite dans ce livre, la façon de révéler les groupes naturels contenus dans un ensemble de n objets biologiques ou autres (chapitre I), équivaut à un recouvrement d'ensemble optimal, à savoir

$$\text{choisir } \mathbf{x} \text{ de façon à minimiser } (\mathbf{c}^T\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{1};\ \mathbf{x} \in \{0,1\}^m)$$

où **A** est une matrice d'incidence $n \times m$ de n objets dans m sous-ensembles et **c** est une mesure du coût de l'inclusion de chaque sous-ensemble dans la solution (chapitre II). On insiste sur le fait que la solution doit être considérée comme une hypothèse pour des études plus poussées.

La plus grande partie du livre est consacrée aux méthodes d'obtention des sous-ensembles, notamment les suivantes : obtention directe à partir d'attributs qualitatifs (chapitre IV), à partir de la dissimilitude des vecteurs (chapitre V) et à partir d'attributs classés uniques et multiples (chapitre VI). Avec des coefficients de dissimilitude couplés scalaires (chapitre VII), le principe fondamental de l'obtention de sous-ensembles est le principe récursif de la mise en grappes conditionnelle : l'ensemble S_t étant donné, S_{t+1} comprend tous les objets dont la distance moyenne aux membres de S_t ne dépasse pas le maximum entre eux (chapitres VI et VIII).

Les coûts **c** sont tirés des relations logiques entre les lignes et les colonnes de **A** à partir des observations suivantes : plus le nombre de sous-ensembles auxquels appartient un objet est grand, et moins la quantité d'information que cet objet fournit à propos des éléments de **x** est grande; plus le nombre d'objets qu'un sous-ensemble contient est grand, plus la probabilité que ce sous-ensemble fasse partie de la solution optimale est élevé (chapitre II).

On examine quelques applications spéciales (chapitre IX surtout). Une série d'études de cas (chapitre X) illustre la plupart des procédures et des principes.

I Introduction

Consider a set of individuals (such as animals or plants) about which there is no doubt where each begins and ends. It is not a colonial animal or plant, and no one, even the least informed in biology, doubts the integrity of the unit. Such an individual can be described by its size, shape, color, physiology, behavior, geographical distribution, and so on. Each individual may belong in one of several classes, but which individuals belong in the same class, as well as the number of classes or whether these classes have diagnostic attributes, are not known.

To try to solve the three problems implied by this ignorance, morphological and other attributes of the individuals are studied and described. Although these attributes are verbally well defined, in practice they are arbitrary subdivisions of the individuals. Where, for example, does an antennal segment of an adult beetle begin and end, so that its length can be measured? And why choose that subdivision rather than another of the almost countless possibilities? The attributes (usually hundreds rather than thousands) are chosen often for no better reason than ease of access. Their states also have no clear limits, and so have their own subjective component. Even though the objects of study are well-defined individuals, their attributes are subjectively chosen and demarcated. In this example, it is apparent that I believe that only the individuals exist, and that attributes are man-made abstractions.

In a second example, consider an area of land where plant species seem to be nonuniformly distributed. Here, the problem is to define homogeneous subareas, the process that geographers call regionalization. One method is to place quadrats, of defined size and shape, in some pattern over the region of interest and to census the plants according to their species. From these data,

contiguous regions of plant associations may be defined. The objects of study here are the quadrats, but because they are defined arbitrarily, in the sense that the ecologist chooses their placing, size and shape, they are equivalent to attributes of the first example rather than to the individuals. This distinction between objective and subjective units, and also between subjective and objective attributes, is important because clustering without recognizing the status of the components only adds neglect to the ignorance that requires the clustering in the first place.

In the taxonomic and ecological examples, both parts of the two-way structure (objects-by-attributes) are important. In a third example, suppose a set of observations (discrete or continuous measurements of several variables) is made on a set of independent objects that are independent samples from some population. To reduce the number of variables for future observations, the variables can be grouped into subsets from each of which one or two can be chosen to represent the whole. Presumably, the variables in each subset should be highly mutually correlated and show some independence (complete if possible) from those of another subset. In this familiar problem in statistics, the set of objects are really no more than a means to an end. This model lends itself to well-understood statistical formulations such as principal components analysis; however, they are *not* the subject of this monograph. A multi-authored article edited by Gnanadesikan and Kettenring (1989) reviewed some of the large literature on clustering from a statistical aspect.

A recent advertisement for a book on cluster analysis reads:

> This book is intended for investigators ... who are involved in multivariate data analysis. The clustering procedures developed and analyzed in much detail demonstrate methods of exploratory data analysis. From the analyses, ... an investigator should be able to pick those clustering procedures best suited to his or her needs.

The last sentence expresses nicely what I consider to be the questionable component of clustering methodology as currently practised. Eclecticism has its place in science and has been important in the development of numerical taxonomy; but it should now be replaced by something more organized. Although the dangers of eclecticism are obvious, they have not always been understood by the practitioners of clustering. One explanation is to be seen in human behavior:

> if an agent knows that one of its actions will lead to one of its goals, then the agent will select that action.

It is not necessary that the agent can give reasons for what is done; I claim that many practitioners of cluster analysis choose the methods that (consciously or unconsciously) lead to results that they already know, or expect, or hope for, or can explain. As such, what they achieve is some aspect of themselves, which has little to do with the subject of the study. Often, they invoke the auxiliary principle

> for given knowledge, if action A and action B both lead to goal G, then both actions are selected.

Indeed "it did not matter which method was used; they all resulted in the same classification" paraphrases remarks made by many authors. Sometimes, several goals are postulated coupled with the argument that to achieve their common ground is desirable; this common ground represents yet another implicit principle, namely:

> for given knowledge, if goal G_i has the set of selected actions A_{ij}, then the effective set of selected actions is their intersection $\bigcap A_{ij}$.

In essence, this intersection principle leads to a functional definition of knowledge as follows:

> whatever can be ascribed to an agent such that its behavior can be computed according to the principle of rationality.

A *structural* definition of knowledge is somewhat different and is usually concerned with issues about certainty as in "knowledge is justified true belief." The latter expresses how the philosopher defines knowledge but is not the goal in the empirical world of clustering.

To establish some organization and unification in the clustering world requires that several points be raised, that some questions be asked (and answered, if possible), that some problems be stated (if they exist), and that some guiding principles be determined. I hope that this book will help to clarify some of the issues and describe some useful procedures.

Domain of the study

Two distinct problems have been labeled as "cluster analysis." The first can be stated as follows: suppose there exists a set of objects that belong to one or more groups, **Z**. On these objects, some measurements, **Y**, can be made. If **Z** is known, in principle at least one operator, F, acting from the topological space corresponding with **Z** into a topological space corresponding with **Y**, can be found such that

$$\mathbf{Y} = F(\mathbf{Z}).$$

If **Z** is not fully known, but there is some set of restrictions on it, such as that the groups are disjoint, or that the number of groups has some specified value, or that the interrelationship of the

measurements has some specified form, usually one or more solutions for F that give added detail about $\mathbf{Z}$ can be found. In cluster analysis, this class of problem leads to solution methods that include many variants of "k-means." These methods are valuable for circumstances in which our aim is to *form* the objects into groups.

The second problem labeled as cluster analysis is the inverse of the first, namely, given $\mathbf{Y}$, determine $\mathbf{Z}$, which can be formalized as

$$\mathbf{Z} = F^*(\mathbf{Y}).$$

This inverse problem is said to be well posed if the operator F^* satisfies three conditions.

(1) A solution $\mathbf{z}$ is unique, i.e., if $F^*(\mathbf{y}_1) = F^*(\mathbf{y}_2)$, then $\mathbf{y}_1 = \mathbf{y}_2$, $\mathbf{y}_j \in \mathbf{Y}$.

(2) A solution $\mathbf{z}$ exists for any $\mathbf{y} \in \mathbf{Y}$.

(3) A solution is stable, i.e., if $\check{\mathbf{y}}$ is an approximation to $\mathbf{y}$, and $\check{\mathbf{z}} = F^*(\check{\mathbf{y}})$ is to $\mathbf{z} = F^*(\mathbf{y})$, it follows that as $\check{\mathbf{y}} \to \mathbf{y}$, then $\check{\mathbf{z}} \to \mathbf{z}$.

Although this book is largely concerned with the inverse problem, the problem of biological clustering fails to satisfy these three conditions unless some further conditions are imposed. But in so doing, a dilemma—the need to reveal the "true" classification without imposing a solution by inappropriate restraints—is created from which there seems to be no escape. Bock (1989) wrote that:

> In order to call a subset [...] a cluster, common sense will generally combine various plausible criteria e.g. that all objects [in a cluster] must either (a) share the same or

> closely related 'properties', or (b) show small mutual [...] dissimilarities, [...] or (c) have 'contacts' or 'relations' with all [or] many [or] at least one other object [in the same cluster], or (d) are clearly distinguishable from the [non-members of the cluster], etc.

By implication, Bock listed a partial set of criteria by which subsets can be evaluated and converted into "clusters." The wide choice implied by the list, makes it almost certain that any competent practitioner can combine some criteria and obtain subsets to satisfy the needs of a practical problem, such as matching the solution with some preconceived classification. Does this have relevance to the problem of revealing the true groups? I claim that any numerical procedure must be miserly with added conditions so that solutions are dominated by the data and not by restrictions imposed either on **Z** or **Y**, or both.

It is important, therefore, to establish the domain of the present study. There are three premises.

(1) The starting point is a **set of real objects** for which all permutations of any tags assigned to them (using any system of tagging or accession listing) are equivalent. The tags convey no information beyond that of an accession label, i.e., tags are exchangeable.

(2) An object "acquires" an informative label by virtue of being assigned to a **labeled subset**. It is subsequently known as a member of that subset and is referred to as an individual of the assigned name of the subset. The tags of individual members of such a subset are **locally** exchangeable, in the sense of (1) above. By being assigned to a labeled subset, the object can be said to **join** the subset.

(The operation of "join" is clarified in mathematical terms in Appendix 1.)

(3) Subsets are either **labeled** or **unlabeled**. A labeled subset may contain just a single object. Empty labeled subsets are conceptually possible and may be useful for hypothetical objects. Unlabeled subsets are arbitrary assemblages of the objects and are of no particular interest.

With respect to the tags in (1) above, it is readily admitted that biological objects have provenance and other concomitant data that may be extensive. It is important to use these data fully, recognizing that clustering should be adopted only after these data are found to be unsatisfactory or of dubious quality.

The establishment of labeled subsets forms the primary focus of this study. This process, often called **clustering** and in biology usually called **phenetics**, is concerned with arranging the (biological) facts.

Consistency

This book considers a given set, N, of n objects, each individual of which is described by a set of m attributes, M. The arguments are presented as if the whole population consists just of N. If this is true, then either little more than a convenient and possibly concise description of the objects can be achieved, or some model of their relationships can be generated. However, in the taxonomic problem underlying this study, the N are a sample of the many individuals belonging to the (unknown) taxa. If the groups formed by the procedures described here do not apply to all members of the taxa, it would be wasted effort. As a result, three consistency requirements are needed. Let $k(N,M)$ be the number of labeled subsets identified using the empirical data.

REQUIREMENT I.1. *For a fixed set of attributes, M,*

$$\lim_{n\to\infty} k(N,M) \to K < \infty,$$

i.e., the number of labeled subsets has a finite limit independent of n. Arguments to support this requirement come from considering the reductions described in Chapter II. Without going into details, including more objects identical with those previously included should not change the solution. If the new objects introduce further unknown true groups, the value of K tends to increase.

REQUIREMENT I.2(a). *For a fixed set of objects, N,* $lim_{|M|\to\infty} k(N,M) \to K < \infty$,

i.e., the number of labeled subsets has a finite limit independent of M;

REQUIREMENT I.2(b). *Further, the membership of the labeled subsets becomes stable.*

Arguments in support of I.2(a) and I.2(b) come from the fact that although the attributes may be chosen independently, each set belonged to a viable individual, and so are mutually interdependent. If the choice of objects to be studied and the selection of attributes to describe them are independent, these two requirements can be combined to give:

REQUIREMENT I.3. $lim_{n\to\infty,|M|\to\infty} k(N,M) \to K < \infty$.

It seems likely that the rate of approach to K is slower for increasing $|M|$ than for increasing $|N|$. Unless requirements I.1 and I.2, or possibly I.3 alone, are satisfied, solutions to the biological clustering problem can have no generality. In Chapter VIII, I attempt to show that these requirements can be satisfied.

Several other assumptions, common to large areas of classification and clustering, are implicit in the previous remarks. They arise tacitly when an object is subdivided into different

components, some of which it has in common with other members of the group to which it truly belongs; others are unique to the individual, including the consequences of the interaction with the environment. These assumptions are as follows:

- Each object is characterized by underlying attributes.

For example, an animal, plant, or some artifact produced by them reflects its genetic make-up. Only those attributes common to the true group can be called the group parameters.

- Attributes depend only on the group parameters, the special parameters of the individual, and on chance.

An example of this subdivision is the gene pool common to a species, the particular combination of alleles possessed by one individual, and the response of these to the effects of the set of external conditions to which the organism was exposed. The special parameters and the chance phenomena are usually no more than of secondary interest and can be grouped together.

- Measures of dissimilarity depend only on the group parameters.

They do not depend on the special parameters of the individual, i.e., dissimilarity between two groups should not depend on which sets of individuals are chosen to represent the groups.

- The values of the group parameters, i.e., the "event" for which an object is seen, can be represented by continuous distribution functions.

The group parameters need not be directly observable nor need they coincide with the attributes.

- The objects are stochastically independent given the local group parameters.

Colonial and clonal organisms illustrate circumstances where the individuals are *not* independent in this sense.

- The estimated statistics, obtained by clustering or other methods, must contain all the relevant data about the group parameters in the study.

The validity of these assumptions in any context should be examined with special care. An organism is an integrated entity, not a collection of attributes. If in its evolutionary history an attribute has changed, this change is unlikely to have occurred without incurring changes in other attributes. An organism is atomized into attributes almost always for our own convenience; rarely, if ever, is it a property of the organism.

Grouping, phylogeny, classification, and identification[1]

Only when unlabeled objects are formed into several distinct, labeled subsets can we consider either identification, i.e., how to assign further objects to these subsets, or to infer what may have been their phylogenetic relationships, i.e., to consider the labeled subsets in a **cladistic** framework. Cladistic practice is considered only peripherally in this monograph, although evolutionary theory is not ignored. Unfortunately, the different aims of the three distinct objectives—division into labeled subsets, identification, and phylogenetic reconstruction—have often been confused, resulting in much polemic. Sometimes they have not been confused but linked. Gower (1974) argued that "the best classification ...[is]

[1] The reader uninterested in my opinions can omit this section.

the one that predicts the most characteristics correctly" basing this contention on reasoning quoted from J.S.L. Gilmour:

> [if] a system of classification is ... natural [,] the more propositions there are that can be made regarding its constituent classes.

Gower recognized that the reverse, i.e., the more propositions that can be made that are correct, the more natural is the system of classification, is not necessarily true. This link, however, can result in groups exhibiting attribute uniformity and so enables efficient identification. In consequence, such classifications have high empirical value, but because the relationship among members can be incidental, they are inconsistent with the objective of phylogenetic reconstruction.

Another link advocated is between group formation and phylogenetic reconstruction, for which there are two premises.

(1) For the true phylogeny, there is a root.

(2) The true phylogenetic history is ultrametric with respect to time.

This second premise means that if the time since any two entities separated measures the amount of nonresemblance, the three times (i.e., among the entities and their most recent common ancestor) form an isosceles triangle. The problem is that, not knowing when separation occurred, a surrogate, e.g., morphological dissimilarity, is used in its place. The use of this surrogate is sometimes supported by a (disputed) claim that there is a

> uniform average rate [of evolutionary change which] is nothing more (or less) than the inevitable result of

> averaging over billions of nucleotides and millions of years. (Sibley and Alquist 1984)

This statement implies that differences in morphology (at least for nucleotides) are proportional to time. In data considered for clustering, which normally do not include the nucleotide patterns, usually nothing links the objects with evolutionary time. The material consists of objects collected within the recent past (most specimens in museums, for example, are not more than 300 years old); furthermore, their study is provoked by information consciously or otherwise observed on living material.

The first major question that arises about a link between group formation and phylogenetic inference is to determine if morphological difference is relevant. To explain in more detail: evolution tells us

> species that diverged recently are generally more alike than those that diverged earlier.

But the converse is known to be false (consider plant hybridization followed by polyploidy). Not only can recent separation fail to imply high similarity, but also it can be false that species that are alike diverged recently. Unfortunately, the only data that can be used to measure the degrees of resemblance are a subset of all possible attributes either treated separately or integrated in some way, such as into a dissimilarity coefficient or the amount of DNA hybridization. Thus the only thing that we can be certain to achieve, if anything is ever certain in this area, is that

> clustering based on morphological differences identifies groups of objects that resemble each other more than they resemble members of other groups with respect to the available attributes.

This conclusion identifies a dilemma. Taxonomists seek a classification that reflects the evolutionary history of objects understudy (or the entities they represent). Yet they have no guarantee of success, because many complex issues arise from evolutionary parallelism and convergence. The equating of an ultrametric structure based on morphology, physiology, and so on with an ultrametric temporal arrangement to represent the phylogeny cannot be defended, not the least arising from the fact that evolution is a continuing process, so that not all distinctions are sharp.

This dilemma cannot be resolved without making further assumptions. One such is to adopt Occam's maxim (the parsimony principle), to say that evolution proceeded by the shortest pathway (e.g., Moore 1976, Wagner 1981). This assertion not only seems arbitrary, because it rules out parallelism and convergence, but also there seems to be no palaeontological evidence for it (Hoffmann 1989). The ontogenic implication of the assumption is also hard to accept; how does an evolutionary lineage "know" where it is going? Crisci (1982) discussed other weaknesses in it as a principle of evolutionary inference.

There are also some other dubious components within current methods of *numerical* cladistics. Current methods are based on an assumed (but unknown) hierarchy, within which reticulation, and the consequent polyphyly, cannot be accommodated naturally. It is well known that hybridization followed by polyploidization is common in plants and often leads to reproductive isolation from the parent species. This isolation implies that organization in nature can be reticulated as well as hierarchical, i.e., taxa need not be monophyletic. Usually at the start of cladistic analysis, entities considered to be bastard species are excluded but subsequently incorporated when a cladogram is established (Bremer and Wanntorp 1979, Humphries 1983). The effects of this selection

can be to negate precisely those properties that make an acceptable cladogram (Mossbrugger 1989).

Further a number of mathematical and numerical problems exist in current numerical cladistics. The most challenging is how to accommodate *continuity* within the same framework, which results from the passage of time, and *branching*, which is basic to phylogenetic models. Most authors adopt a somewhat cavalier attitude in their treatment of continuous attributes, either asserting that there are sufficient discrete attributes to omit those that are continuous, or proposing that continuous attributes be categorized. Discarding attributes, or their often ad hoc categorization seem dubious.

After careful study, Sokal (1985) summarized the situation as follows:

> Numerically estimated cladograms are not good estimates of the true phylogeny of a group of organisms. The shortest trees are not necessarily closest to the true tree. Differences between true cladograms and phenograms or between phenograms and estimated cladograms can be explained as the results of homoplasy or divergence. Estimated cladograms are affected almost as much by homoplasy as are phenograms. As the number of characters decreases or the number of [objects] increases, phenograms become better estimates of the true cladogeny than estimated cladograms. [...] Even the inclusion of fossils in the data matrix does not substantially increase the quality of the estimate of the phylogeny. The topology of the true tree is a critical factor in determining the quality of its estimate. Such results are not causes for optimism for those who wish to estimate phylogenies. (p. 746)

I do not think the situation has changed since then. McNeill (1982) expressed a similar point of view based on different arguments. In 1983 Gower wrote that

> ... both cladistic and phenetic classification are worthwhile pursuits but ... it should be recognised that cladistic classification, which pertains solely to biological material, needs many more assumptions than do phenetic and predictive classification. Thus character-states have to be ordered from least to most primitive, outgroups may need specifying and decisions have to be taken about reversals and compatibility. Intermediate states of characters and/or unknown intermediate taxa are postulated. All this seems ... to make any proposed cladistic classification more speculative than predictive classification especially when it is recalled that the data to support it rarely, if ever, are complete or entirely reliable.

This quotation from Gower does not offer any arguments against numerical cladistics but nicely summarizes the additional information required to bring about a solution. In a different context, Wilson (1985) wrote that

> [the] biogeographic theory [of P.J. Darlington] dwarfs the spiritless mechanics of the extreme cladistic school that was to follow, ... method substituted for theory, technique confused with science.

Sneath (1983) gave a thoughtful and rancor-free discussion of the contrasts and similarities between phenetics and cladistics.

Many authors have argued that clustering without the explicit objective of establishing the evolutionary history of a group is vacuous; in one sense, I agree. However, to adopt parsimony as a principle guiding numerical procedures in

phylogenetic reconstruction, and then to assert that the structures so formed give (almost) correct pictures of the phylogeny, seems presumptive or even self-deceptive. It certainly shows circular reasoning. One argument adduced to support the parsimony principle is that no other yet proposed appears reasonable and leads to an operational principle for extracting information from data. However, another long-established class that may satisfy these two needs is based on compromise between conflicting principles. This model is that the source of changes in an evolutionary lineage can be ascribed to two main causes.

(1) **Increase in variation** caused by mutation, errors in DNA replication, chromosome recombination and so on, i.e., **genetic variation results in a process of entropy increase.**

(2) **Decrease in variation** as a consequence either of inbreeding, which can be regarded as being passive, or of selection in the sense that noneffective extremes do not survive, which is an active process naïvely equivalent to the "survival of the fittest," i.e., **selection is an entropy-reducing process.**

Thus in this model of evolutionary change, rather than parsimony, there is balance between

an entropy-generating process,

presumably genetic in origin, and

an entropy-reducing process,

presumably environmental in part. This division is an informal presentation of the fundamental theorem of R.A. Fisher (1958), in

which natural selection is considered to be the only force responsible for optimizing biological systems, and which leads to a monotonic increase of population mean fitness. It is well recognized, however, that where there is mutation and random drift, this theorem does not hold, and population fitness may decrease over time. This model requires further development before it can be used for modeling macroevolution, although inbreeding and random drift have been argued as providing a sufficient mechanism. However, its value as a model of microevolution have long been established.

The reader will find that the procedures described in this book for group formation also consist of a compromise between two processes—maximum entropy and maximum information (minimum entropy)—**analogous** to the model described above. Maximum entropy (e.g., maximizing variance) serves to generate (probability) measures for subsets of the objects. Maximum information (e.g., minimizing variance) is used for choosing particular subsets. Nevertheless, these procedures should *not* be understood to have established anything beyond resemblance among the individuals of members of these subsets. I do *not* assert that the groupings obtained by the procedures advocated in this book need have any phylogenetic significance. The fact that one model is described as being analogous to another does not represent an assertion that one deductively explains the other. Analogy can only be supported by its effectiveness; it reduces a theory into as many others as there are analogies; it is heuristic.

Complexity, simplicity, descriptions, and clustering

Because even the simplest organisms are complex, it is a challenge to find a simple representation of them; consider for example the DNA sequence of a virus. It may be of some value to discuss complexity and simplicity from a formal simplified standpoint. Suppose x is a finite string of zeros and ones, and that $|x|$ denotes

its length (the number of zeros and ones). Intuitively, a string is simple if it can be described in a few words (e.g., the string of a thousand ones), complex if it cannot be described so simply, and random if it follows no rule and, to describe it, its elements must be listed.

A device, T, which can interpret descriptions such as "the string of a thousand ones" is called a decoder. A program, p, can be considered to be "a description of x" if, when p is input to T, the output is x, i.e.,

$$T(p) = x,$$

assuming (for simplicity) that no extra information besides p is needed to obtain x. If $p \in \{0,1\}^*$, the unconditional Kolmogorov descriptional complexity, $K(x)$, of x relative to decoder T is defined by

$$K(x) = \min\{|p| : p \in \{0,1\}^* \ \& \ T(p) = x\},$$

where T is assumed to be the universal Turing machine. If $K(x) \geq |x|$, the binary string x is incompressible; such strings pass the usual tests for randomness.

Suppose there exists a set of transformations, $C(p)$, of p that produce a program p' such that $T(p') = T(C(p)) = x$, and for which $K(p') < K(p)$; then it can be asserted that p' is less complex, i.e., simpler, than p. If p' is unique, the transformation system is called Church–Rosser after the two logicians who formalized theorems on the subject.

Suppose a biological object, **i**, can be represented by a binary string i of length $|i|$; it can be assumed that even if i is incompressible, $|i|$ is a huge number for all practical purposes (it may even be infinite). The problem in biological objects is that each is represented by an empirical description, x_i, a subset of i.

Thus $|x_i| \leq |i|$, and because biological objects are empirically observed, it must be true with probability unity that x_i is a proper subset of i and that strict equality can never apply. It follows that $K(x_i) < K(i)$, the inequality also being strict for the same reason.

It also follows that even if a program p_i' can be obtained that is simpler than p_i, it refers to x_i, and only indirectly to object **i**; and even if p_i' is unique with respect to x_i, it need not be unique with respect to i.

The conclusions of this line of reasoning are as follows:

(1) The only possibility to describe an object simply is to find the $p'(x_i)$ such that $T(p'(x_i)) = x_i$.

(2) That even if $p'(x_i) = p'(x_j)$, or even if $p(x_i) = p(x_j)$, it does not follow that object $\mathbf{i} = \mathbf{j}$, where the last equality means identical in all respects.

(3) That a clustering consists of a series of transformations, C, either of
(a) the complete set of descriptions, $\{x_i, i = 1...n\}$,
or
(b) the complete set of programs, $\{p(x_i), i = 1...k\}$,
or
(c) the complete set of programs $\{p'(x_i), i = 1...k\}$,

to obtain a new program, P, such that

(i) $T(P) < \sum T(p(x_i))$, and

(ii) $T(P)$ correctly assigns any input to its correct output.

Translating these three into practical issues, a choice has to be made on how to represent the set of objects, i.e., to determine the

$\{x_i\}$, what are the p_i' for each object, which may be conditional on the set of objects under study, and what are the transformation rules to obtain P.

Numerical clustering

Because there is no reason to assume that organisms resembling each other closely are recently diverged, this study is confined to empirically measured resemblance. Further study of each group may eventually reveal unsuspected heterogeneity, some of which may have resulted from parallelism, convergence, or just from previously inadequate description. At some stage, a taxonomist may consider that there are sufficient data to propose a structure representing the phylogeny of the established groups. This step generally ends an investigation rather than providing the model for group formation. Thus the surrogate for establishing evolutionary relationships is the forming of groups of objects that resemble each other. Each group should then be examined further for evidence against a hypothesis of homogeneity.

Phenetics, as here studied, is equivalent to having a (finite) set of objects, each of which has the same (prior) probability of membership to any of a large family of unlabeled subsets. The process of establishing labeled subsets, as already described, ideally changes the probability of membership of each object to one of the subsets to unity, and to the remaining subsets to zero. Thus the process of establishing labeled subsets changes our initial knowledge of the objects, which has no structure and therefore has maximum entropy, to a (relatively) complete structure, which therefore has maximum information (i.e., minimum entropy; maximum negentropy). There is no need for the solution to be a hierarchy, nor for the labeled subsets even to form a partition.

The general objective, therefore, is to generate a subset system (a family of subsets) from the elements of a set. This process is called either **clustering** or **cluster analysis** analysis, to

meet the demands of language. The elements, the units of clustering, are called **objects**; usually, an object has **attributes**, but what constitutes an attribute and what constitutes an object depend strongly on the material under study. This monograph focuses largely on the problem described in the opening paragraph, which is the continuing problem of taxonomy and classification. The ecological problem is also considered.

Although one reason for clustering is to group the objects to satisfy some need for homogeneity, it does not (or should not) pretend to give a final answer to the question about the existence of "true" groups within the objects of interest. The most that clustering should be expected to do is to generate one or more hypotheses, each consisting of a **subset system**, which are groupings of potential interest. These groups can then be examined and provisionally accepted or rejected based either on the same data or, preferably, on more data which may themselves generate hypotheses of interest. These hypotheses should be compared with others, e.g., classifications based on traditional procedures, or on the definition of diagnostic attributes (in taxonomic studies) or on some known diagnostic species (in ecology) or geographical contiguity (or both). The concordance (or otherwise) of the newly proposed hypotheses with the previously accepted classification can often be examined by formal statistical procedures. A new clustering becomes really useful **iff** (if, and only if) there is *no* consistency with previous classifications, because it challenges the clusterer to defend the new arrangement. Remember that even if numerical clustering appears to be objective, it is not, because the attributes in taxonomy are subjectively chosen and defined, as are the sizes, positions, and shapes of quadrats in ecology. Perhaps the only criteria available to assess a classification are based on two practical needs. First, the labeled subsets should be stable with respect to the attributes that are chosen, i.e., for different attributes, the same subsets (or at least, a consistent family) are obtained. Second, the essential differences among the labeled

subsets are not changed by including more objects. If in an empirical study the three consistency assumptions (requirements I.1–3) are not satisfied, the classification achieved should at best be considered as tentative, but preferably rejected.

Mathematical and nonmathematical classification

Classification, in a mathematical sense, is a set of transformations, a program of logical tautologies, which convert a set of clearly defined facts into their unique simplest form such that interrogation of the latter answers all questions about the former. Golomb (1961) considered

> that classification is the most fundamental objective in mathematics [and because it] is a familiar process to the layman ... the fundamental classificatory problem can be described in non-technical language.

Golomb considered the following three problems:

- enumerating how many categories there are
- listing invariants, which asks for properties common to all members of the same category (their completeness, independence, and legitimacy are crucial) so that equivalence classes can be defined
- finding representative assemblies of objects (having or knowing about one of each category is the collector's ambition); given a precisely defined equivalence class, the "canonical" representatives follow at once.

From another point of view, in classification, given objects i and j, together with a statement about the relationship between them, iRj, which can be recognized unequivocally as being true or false,

find classes x and y such that a statement $xR'y$ implies iRj and also that iRj implies $xR'y$.

Beyond the mathematical world, these three problems are neither distinct nor do they have precisely understood or definable rules. For example, constraints need to be imposed on enumeration, partly because of the size of the task (this size has tended to dominate clustering but not classification) and partly because of the properties of the supposed invariants. The latter, however, cannot be defined until either the categories are known or the representative assemblies are recognized. In clustering, both R and R' are empirical and are neither true nor false; they are measurements whose accuracy is important but only incidently to the problem of clustering. In clustering, everything is fuzzy; the facts are not clearly defined, the transformations have an intuitive component, the simplest form is unknown or unrecognizable, and the number of distinct questions is infinite. Classification is not identical with clustering, although there are many points in common.

Many authors on clustering have pointed out that the problem of disjoint group formation is equivalent to the mathematical concept of set partitioning. The biological problem became restated as the need to find the partition that maximizes the combined within-group homogeneity, or, equivalently, maximizes the between-group heterogeneity. Of several difficulties in translating these formulations into practical algorithms, the most important is that the number of partitions of more than a few objects is huge, making infeasible (even with today's computers) a consideration of them all. Several heuristic solutions to this problem have been advocated; for example, examining a large number of partitions either by choosing at random from the set of all, and retaining the best by whatever criterion is chosen, or by choosing some starting assignment (random or informed) followed by some relocation scheme to see if there is any improvement. A recent and interesting proposal for this class of procedure by Klein

and Dubes (1989), which is based on simulated annealing, shows some promise for a restricted set of circumstances (large data sets, tight Gaussian clusters).

These heuristic attempts to find an optimal partition offer no guarantee that the solution obtained is within some reasonably small region of the global optimum. Furthermore, the optimality criteria themselves are not usually an integral property of the objects but tend to be based on statistical principles, such as an assumption of multivariate normality. Thus the solutions found are approximations to the problem of partitioning the objects into a number of subsets having common (or occasionally different) within-group covariance matrices, differing in means, and so on. Often there is a need to specify the number of subsets, as well as a significance test. If the objective is to **form** the objects into subsets satisfying these statistical constraints, then this class of methods offers acceptable solutions. If the objective is to **recognize** natural groups, the statistical models are far too restrictive. Why should the range and pattern of variation in a species be the same as those of other under consideration, or for that matter, even be multivariate normal?

Even if there are good answers to these questions, there seems to be a flaw somewhere. Perhaps it lies in the framework, even though the definitions are often clear and verifiable. One source of confusion perhaps is in the ambiguous use of such common terms as cluster analysis. For some authors it is confined to obtaining partitions; for others it implies hierarchically structured partitions (i.e., dendrograms). Others miss the real distinction between the objective, if specified, and a numerical procedure that divides objects into subsets. An example of clearly stated conditions made to obtain a tractable formulation of the partitioning problem follows:

> A typical problem in cluster analysis is the following: let X be a finite set of points in some Euclidean space,

> $X \subset R^d$, and let $\rho : X \times X \rightarrow \mathbf{R}$ be a function defined on pairs of points, expressing their unsimilarity. [...] The problem is to find a partition $X = S_1 \cup S_2 \cup \ldots S_p$ of the base set into p groups (p is fixed) such that some objective function $h(S_1,\ldots,S_p)$ is minimized. (Boros and Hammer 1989)

Questions I ask here for the biological clustering problem include

- Why is there a constraint to a Euclidean space?
- Is there a reasonable dissimilarity function?
- Why does the solution have to be a partition?
- From where should the value of p be taken?

Another set of clearly stated conditions appears to have been made to allow use of what has become known as the greedy algorithm. This sequential class of procedures gives solutions that are locally optimal but can be far from so globally. An example of this class is to find the pair of objects that are most alike (or have something in common) and join them; then to replace them by some compound, regarded as a new object; finally to repeat this process until either some termination criterion is satisfied or all objects are in one group. The history of this joining process produces a dendrogram, which has the same form as a phylogenetic tree. Those dendrogram methods that use a final criterion either to produce a forest rather than one tree, or to cut such a tree after it is formed, or to prune and graft branches, also need to define an appropriate criterion. Unfortunately, many of the plausible criteria proposed for these purposes tend to be global in nature. In so doing, they result in groups differing in position but otherwise having essentially the same within-group variation patterns in some mathematical space. For reasons similar to the random start procedure, although this class of methods may be adequate for forming groups, for recognizing them it is dubious.

Usually, whole-object measures of resemblance are used, but one class of techniques uses individual attributes with or without a specified ordering of their states, forming a dendrogram by some compromise or consensus among them. This class of methods is also dubious, because the attributes tend to be arbitrarily defined, yet cannot be independent because they belong (or belonged) to viable organisms; the attribute states are also arbitrarily defined (and sometimes sequenced) by the scientist. An arrangement of attributes, each treated independently, does not imply to me a valid arrangement of the individuals unless it coincides with an arrangement in which the individuals are either treated as wholes or can be shown to be equivalent. The distinction between the two approaches can be represented by the categorical model in Fig. I.1. The abstract space for $\Xi(.)$ is of the attributes, and the objects are points in this space; the abstract space for $\Upsilon(.)$ is of the objects, and the attributes are points within it.

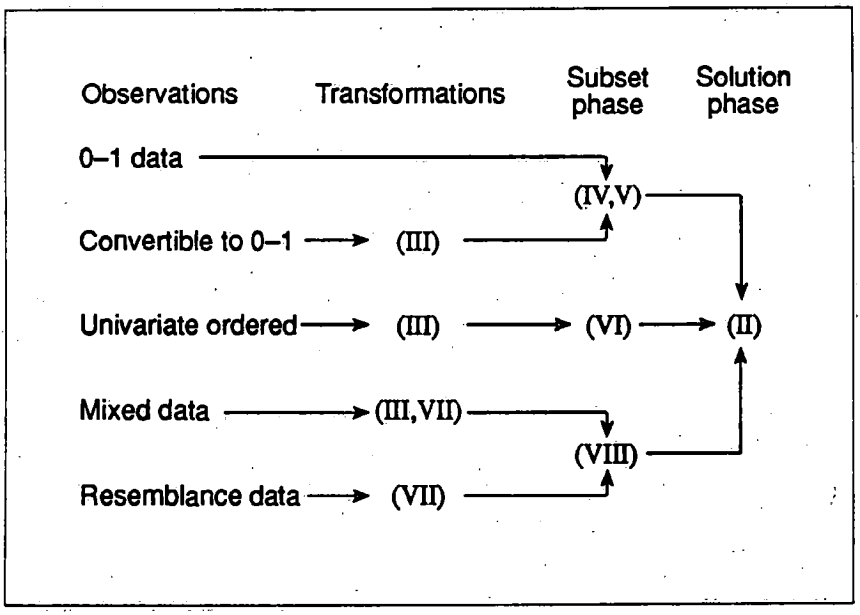


Fig. I.1 Chapter sequence relevant for grouping given different initial circumstances.

The common ground to the numerical methods for partitioning and dendrogram formation is a local heuristic criterion, which is

> consider the current situation, and make the best local change consistent with some specified criterion.

This general procedure can certainly be effective for clustering, and in fact, the coincidence of dendrogram structure with that of a phylogenetic tree has sometimes been claimed to solve the evolutionary reconstruction problem. Although many published studies have adopted this class of methods, it is also true that many sets of data processed by these methods have yielded groupings that those who know about the organisms find no reason to support. Consequently, the resulting classifications do not get beyond the computer screen. Although eclecticism in science, especially in the context of incomplete knowledge of the external world, stimulates the search for unification, this selectivity of local criteria and definition of compound objects, as well as lack of published reports on failed clustering attempts, point to an unsatisfactory situation.

Outline of the monograph

Origins

After attempts to remedy some of these problems, including how to obtain groups from dendrograms, how to avoid the need for compound objects, and how to obtain each group independently of the others, it became clear that a new criterion is needed. Although this monograph presents arguments for this new criterion as if it emerged logically from first principles, I did not recognize it in that way. Before stating this criterion, it may be of interest to outline how it emerged.

The first idea appeared in some unpublished notes made in 1972. It was that the nearest neighbors of an object are the subsets of objects with which it forms a maximal clique with respect to some threshold dissimilarity. There followed a need to find the best threshold (Lefkovitch 1975). In turn I recognized this threshold as being deficient in that it was necessarily global. What at first appeared to be a difficulty, the fact that maximal cliques were not always disjoint, also helped me to recognize that the range of variation within one disjoint subset of the objects had little **necessary** relationship with that of another. I also realized that different subsets of the objects in a given dissimilarity space may be contained in minimal (convex or other) hulls, which not only differ in position but also differ in size, shape, and orientation.

In some ways, the second and perhaps most important contribution came from the elementary mathematical principle of linear ordering: in grouping ordered events (e.g., observed at different times), candidate groups can be defined by pairs of events, to include those events occurring between. This simple principle of betweenness (W.D. Fisher 1958) was really where the ideas of this monograph began. Their extension led to the definition of **neighborhoods** in higher dimensional space and ultimately to the procedures redescribed in Chapter VIII. The resulting general principle, that of **conditional clustering**, forms the basis of this study. In its most general form it is that

> if some objects have been grouped together, determine the others also belonging in the same group.

When this principle is applied to a set of data describing some objects, a series of subsets are independently formed. The logical relationships among these subsets generates a covering, from which classifications are obtained.

Having established the principle and expressed it mathematically, it then became important to strip away as much of the unessential detail as possible. In so doing, I began to identify which parts are logical consequences of the neighborhoods, and which parts are included, as it were, *de novo*. In this way, those parts essential to solutions of the clustering problem became recognized. This search for the essentials is in the spirit of the following:

> Occams Devise ist naturlich keine willkürliche, oder durch ihren praktischen Erfolg gerechtfertige, Regel: Sie besagt, daß unnötige Zeichenheiten nichts bedeuten. [Tr.: Occam's maxim of course is neither an arbitrary rule, nor is it justified only by its practical success; it signifies that unnecessary symbols have no meaning.] (L. von Wittgenstein, *Tractatus Logico-Philosophicus*, §5.47321)

Further simplifications beyond those considered in Chapter II may be possible, perhaps in developing measures of the properties of subsets and in the consequent clustering criterion, although it is unclear from where these generalizations may come. One note of caution: no matter how plausible are the arguments in this monograph, it is important that *we* use the procedures, and *not that the procedures use us*. Thus we should not choose the data to satisfy the models but seek the model that is consistent with the data.

The group formation model that is described in this book I call **conditional clustering**. Its sequence of operations is

- assembling or generating subsets
- reductions
- probability and information calculations
- least-cost set-covering
- muster formation.

Structure

The sequence of chapters does not parallel the sequence of operations exactly; a notation, the reductions, probability, covering and muster phases are described in Chapter II, while the assembling of subsets, which is spread over chapters III–IX, reflects the many ways objects may be described. Chapter III describes methods for converting attribute descriptions into numerical code appropriate for subset formation. Chapter IV describes an application of the reduction rules of Chapter II to clustering without any formal measure of pairwise resemblance (i.e., the $\Xi(.)$ transformations; Fig. I.2). Chapter V describes a pairwise vector dissimilarity between objects and introduces a further transformation rule, based explicitly on the conditional clustering principle. In this chapter, the possibilities for the $\Upsilon(.)$ transformations (Fig. I.2) are begun. Chapter VI considers a special case, namely, clustering univariate data, which has potential application in the grouping of means after analyses of variance and deviance. Chapter VII describes some dissimilarity measures and discusses what their properties may be, while Chapter VIII shows how these dissimilarities may be used.

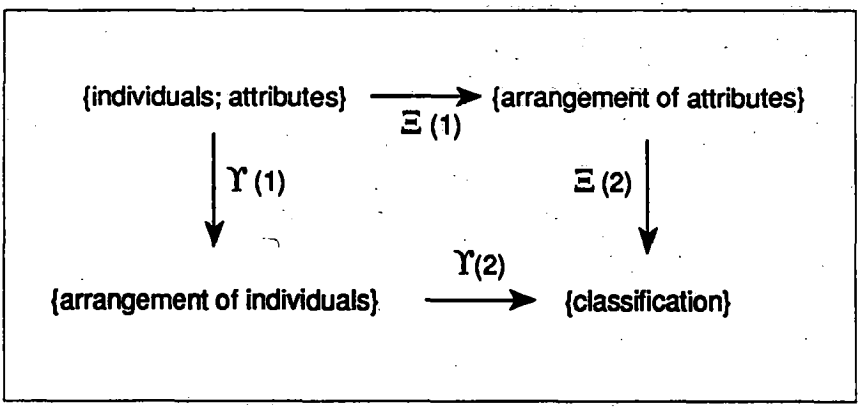


Fig. I.2 Categorical model for deriving classifications of objects, where $\Xi(.)$ represents the transformations of the attributes into a classification, and $\Upsilon(.)$ represents the transformation of the set of individuals into a classification.

Chapter IX describes a number of illustrative special applications. Chapter X collects together a number of numerical examples and case studies. It is important for the reader to refer to these examples when cited in earlier chapters, at least on a first reading. Appendixes 1–6 include material that provides background support. For a limited reading appropriate to special classes of data, see Fig. I.1.

II Set covering

In this chapter, the clustering problem is expressed in a formal mathematical notation, and some standard symbols are introduced. Details of definitions, i.e., how some empirical concepts are distinguished and represented, are postponed to later chapters.

Definition of A

Assume that a finite set, N, of cardinality $n = |N|$ objects is under study, and that a system of m subsets is either given or obtained in some way as described in later chapters. This **subset system** can be represented by a $n \times m$ incidence matrix, **A**, with elements a_{ik} defined as

$$a_{ik} = \begin{cases} 1 \text{ if object } i \text{ is an element of the } k^{\text{th}} \text{ subset,} \\ 0 \text{ otherwise.} \end{cases}$$

Without loss of generality, the ordering of rows of **A** representing the objects is in some fixed but arbitrary sequence. The i^{th} object is denoted by the m-element vector **i**, the k^{th} subset by the n-element vector **k**. A clustering of the n objects is the family of subsets of the objects indicated by a selection of some columns of **A**. The major constraint on the choice of columns is one so obvious that it is generally unstated but here forms the springboard for this study; it is that

> the choice of columns must be such that each object belonging to N must be assigned to at least one subset.

This requirement can be represented by any m-element (binary) column vector $\mathbf{x} \in \{0,1\}^m$ that satisfies the ordinary matrix algebra relationship

$$\mathbf{Ax} \geq \mathbf{1},$$

where $\mathbf{1} = \mathbf{1}_n$ is a n-element column vector whose elements are all equal to unity. This requirement implies that $\mathbf{A1}_m \geq \mathbf{1}_n$. Any $\mathbf{x}$ satisfying this constraint is called a **covering** of the n objects. If the subset system represented by $\mathbf{A}$ is such that no $\mathbf{x}$ satisfies these constraints, the problem is infeasible, i.e., some objects under study do not belong to any subset represented in $\mathbf{A}$ (presumably, they are isolated objects). If the constraint is the strict equality, $\mathbf{Ax} = \mathbf{1}$, those subsets indicated by $x_k = 1$ together form a **partition** of the n objects. It is quite possible that no $\mathbf{x}$ satisfies this constraint, i.e., no partition is consistent with $\mathbf{A}$, yet a covering can exist. (The covering constraint can also be expressed as the **Boolean** matrix equality $\mathbf{Ax} = \mathbf{1}$, but because this expression does not distinguish a partition from a covering, it is less convenient for some of the subsequent development.) A covering is called **irredundant** if the replacing of any element equal to unity in $\mathbf{x}$ by zero results in infeasibility; otherwise it is called **redundant**. Partitions are always irredundant, but coverings need not be. The arguments against requiring strict equality are twofold.

(1) All partitions are coverings, but not the reverse;

thus if the objective of the study is to form a partition of the objects, and inspection of the optimal covering solution shows that it satisfies the strict equality, the obtained partition can be regarded as having a greater support than if the only admissible solutions are partitions.

(2) An empirically obtained **A** may be such that there is no partition.

The first argument is perhaps the more important, because *requiring* a solution from a restricted class is consistent with the desire to *form* subsets rather than the need to reveal them for taxonomic purposes or evolutionary considerations. For example, nondisjoint subsets may suggest either subspeciation or hybridization. By contrast, that two subsets may have some objects in common need not mean that two true groups overlap; the non-null intersection may solely reflect of the inadequacy of the empirical observations to represent the underlying structure. Thus any clustering procedure claiming "to partition [the] observations into mutually exclusive groups" (Binder 1981), i.e., whose mathematical apparatus is constructed in such a way that the consequence is necessarily a partition, is here rejected as being inappropriate for the taxonomic problem.

As described in Chapter I, to determine the best partition of a set of objects appears to require the generation of all the partitions, followed by an investigation of some criterion for choosing among them. If $\lambda \log_e \lambda = n$, the number of partitions of n objects is given by the Bell number

$$B_n = e^{-1} \Sigma(k^n/k!) \approx \lambda^{n+\frac{1}{2}} e^{\lambda-n-1}/n^{\frac{1}{2}}$$

(Riordan 1958), which is very large even for small n. The number of coverings of a set is clearly much larger, because it includes the partitions, and is of order 2^{2^n} (Clarke 1990). However, if **A** is defined to be the incidence array of the power set of N, i.e., of $P(N)$, taken in some standard lexicographical order, then each partition and covering can be indicated by a unique vector $\mathbf{x} \in \{0,1\}^{2^n}$. The complete power set is of no great interest; for

example, neither the subset $\mathbf{k} = \mathbf{0}$ nor $\mathbf{k} = \mathbf{1}$ conveys information about clustering, and there may be others of no interest.[1] Thus the problem is to find a set of rules restricting **A** to some very small subset of $P(N)$, which include the subsets of the objects belonging to the clustering(s) of interest.

The following chapters describe some of these rules, proceeding from the least restrictive to the most. The restrictions depend primarily on the data but may also be chosen by the taxonomist for intuitive or mathematical reasons. For the rest of this chapter, assume that an incidence matrix, **A**, for a subset system has been obtained by some method, and that it represents the only available data. The objective is to exploit as many as possible of the implications of this array to determine an **x** without the major computational effort required to examine all the 2^m elements of $\mathbf{x} \in \{0,1\}^m$ satisfying $\mathbf{Ax} \geq \mathbf{1}$.

Permutations

Because there is no logical ordering of the rows of **A**, all their permutations are equivalent; this equivalence is equally true of the columns. Suppose there exists a permutation of the rows, **R**, and columns, **C**, of **A** such that **RAC** is block diagonal, where the diagonal blocks (either scalars, or square or rectangular submatrices) include some unities and all off-diagonal blocks are strictly zero, the clustering problem can then be simplified into subproblems each consisting of the subset of objects belonging to one such block. That such a set of blocks forms a partition of the objects may of itself provide a sufficient clustering for them. Although the search for optimum permutations can be

[1] An example: if a scalar measure of pairwise resemblance exists among the members of N, then that subset consisting of *just* the two objects least alike is not going to participate in any clustering of interest (see Chapter VIII).

computationally expensive, it is "noninvasive" in the sense that for all **R** and **C**, the matrix **RAC** is completely equivalent to the original **A**. This proposal resembles the search for cliques in undirected graphs, briefly reviewed by Arabie et al. (1978), but differs in that the rows and columns of **A** have different classification sets.

A simple procedure for obtaining an approximate solution to the block-diagonal problem follows. Let α_i denote the integer formed by regarding the entries in the i^{th} row of **A** in order as its bits, and β_j the integer formed similarly by the entries in the j^{th} column. Then,

> *step 1: form the α_i from the current A; sort the α_i in descending order, permuting the rows of A to correspond;*
>
> *step 2: form the β_j from the current A; sort the β_j in descending order, permuting the columns of A to correspond; and*
>
> *step 3: repeat steps 1 and 2 until there are no more changes.*

This procedure can be illustrated with the following two examples:

Example II.1. The matrix has a block-diagonal form.

	12345
a	11...
b	..1..
c	11..1
d	..11.
e	.1...

original

	12345
c	11..1
a	11...
e	.1...
d	..11.
b	..1..

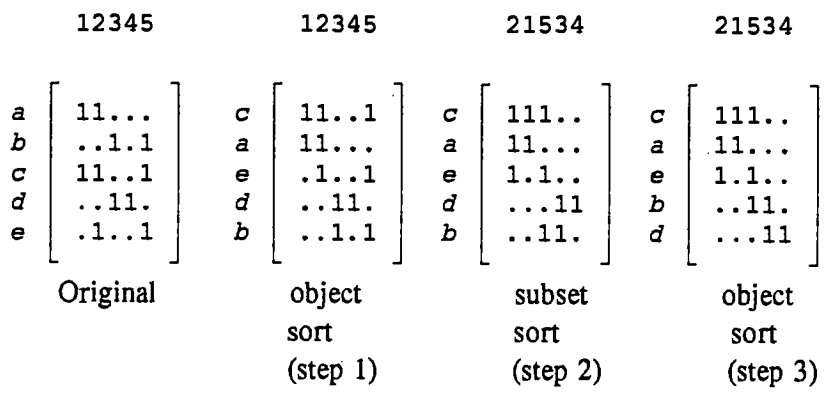
object sort (step 1)

	21534
c	111..
a	11...
e	1....
d	...11
b	...1.

subset sort (step 2)

Here, there are two distinct blocks, and a minimal cover (in fact, a partition) is given by the original columns 2 and 3; there is no other irredundant cover.

Example II.2. The matrix has no block-diagonal form.



Here, there are no distinct blocks, and although columns 2 and 3 form a covering (in fact, a partition), there are several other irredundant solutions. Further information may need to be extracted from the array to obtain a grouping.

Because the objects corresponding to a block belong to the same component, an alternative procedure is to find the connected components of **A**, i.e., a subset of the columns of **A** in which no column is disjoint from the union of the remaining columns in the component. This procedure may be carried out as follows and is clearly related to that described for obtaining a block-diagonal structure:

step 1: consider all columns are unmarked;

step 2: choose an arbitrary unmarked column of A and mark it (initialize a new component);

step a: find all unmarked columns with which it overlaps and mark them;

step b: select one of the newly marked columns not previously selected, find all unmarked columns with which it overlaps, and mark them;
step c: repeat steps a and b until no new unmarked columns are found (a component found); and

step 3: go to step 2 if there are columns not part of a component.

Another procedure for the same problem is as follows. For any Boolean vector, α, let the designation number, α, be the integer corresponding with the string of bits in α. Then,

step 1: sort the columns of A in decreasing designation numbers;

step 2: sort the rows of A in decreasing designation numbers; and

step 3: repeat steps 1 and 2 mn times or until A stabilizes.

It is necessary to rearrange the labels to match the sorting. Other rearrangement schemes have been described by Fulkerson and Gross (1965), Hartigan (1975), Dewdney (1979), Marcotorchino (1987), and Streng (1991).

If there is no block-diagonal structure, some insight may be gained in attempting to permute the array to achieve a seriation; Gourlay (1979) described such a procedure.

Bases for A

Since **A** is the incidence matrix of a system of subsets of n objects, and since a covering consists of the subsets selected so that every object is included in at least one of them, it is important to *exclude*

from candidate solutions those arising purely from mathematical concepts, however useful they may be for other purposes. An important concept among these is that of a minimum basis for **A**, which is

> a minimum subset of columns of **A** such that the element by element Boolean sum of subsets of the basis generates all columns of **A**.

Two things are obvious:

- if no row of **A** is null, the choice forms a covering
- a row basis can be defined in an analogous way.

It is easy to show by example that a column basis need not be a minimal covering; in

$$\begin{bmatrix} 1 & 1 & . \\ . & 1 & 1 \end{bmatrix}$$

a minimal cover is given by the second column, but a minimal column basis requires columns 1 and 3. This example indicates that a column basis generally consists of those columns that each cover a few objects; if a basis is assumed to coincide with the unknown true groups, the solution will often consist of many small subsets.

Structural implications of A for set covering

This section develops the implications of **A** in two ways. The first considers the immediate implications of the arrangement of unities within **A** and exploits their arrangement to simplify the problem. The second develops a measure for each row and column of **A**

based on the same starting point. Exploiting both developments jointly leads to the procedures advocated in this book.

Reductions

Elementary logical considerations can reduce the size of the problem so that the solution to the reduced problem also solves the original. The objective is to eliminate from $\mathbf{Ax} \geq \mathbf{1}$ any redundant constraints.

Suppose object i belongs to precisely one subset, $\mathbf{k}$; it is a consequence of $\mathbf{Ax} \geq \mathbf{1}$ that the k^{th} subset must be in *every* solution to the clustering problem, i.e., $x_k, = 1$. Further consequences are

(1) The i^{th} row of $\mathbf{A}$ can be deleted.

(2) The k^{th} column of $\mathbf{A}$ can be deleted.

(3) *All* rows corresponding to unities in the k^{th} column of $\mathbf{A}$ can be deleted,

because the k^{th} subset of the objects may include more than object i, say $\{j \in \mathbf{k}\}$, for *all* of which $\mathbf{Ax} \geq \mathbf{1}$ is now satisfied.

These three considerations can be further generalized into a fourth consequence. Suppose object i belongs to $s > 1$ subsets. Then *at least* one of these subsets is required to satisfy $\mathbf{Ax} \geq \mathbf{1}$, leading to the rule

(4) if object j belongs to at least the same subsets as object i, the j^{th} row of $\mathbf{A}$ can be deleted.

This action is easily justified; any choice of $\mathbf{x}$ that satisfies the constraints for object i will also satisfy them for object j.

These four operations can be repeated in any order until there are no more deletions and settings of elements of **x** to unity; at any stage, any empty column of **A** can be deleted. An **A** unchanged by the reduction process is called **irreducible** and **reducible** otherwise; it is **fully reducible** if it is emptied. Where **A** is fully reducible, $\mathbf{Ax} \geq \mathbf{1}$ is satisfied by the **x** formed during the reduction process, i.e., **x** has a unique solution. Fully reducible subset systems are rare, but it is not difficult to show that any solution for **x** for the reduced **A**, together with the elements set to unity during the reductions, will give a solution to the original problem. These reductions have been discussed in a general way by Garfinkel and Nemhauser (1972), and in a quite different context by Goodman (1971) and Malvestuto (1989). The reduction algorithm described by Malvestuto gives the same result as here if applied to the complement of **A**.

The above reduction rules can be summarized as:

ALGORITHM II.1. *Covering reductions to form a reduced A*:

RULE 0: *Delete duplicated subsets.*

RULE 1: *Suppose i is such that it consists of zeros, except for a single element* i_k. *Then in all solutions to the problem,* x_k *must be unity to satisfy the constraints, i.e., set* $x_k = 1$ *and delete object i and subset k from A. Furthermore, if subset k also contains object j, delete j from A, since all solutions containing i will also contain j.*

RULE 2: *Consider i and j, and suppose that object i belongs to the same subsets as object j (object j may belong to other subsets); then delete row j from A, since all solutions containing i will also contain j.*

RULE 3: *Delete from A any subset that becomes emptied by rules 1 and 2.*

Rules 1-3 of Algorithm II.1 are applied repeatedly in any order until no further changes are possible. Note that if **A** is given by the complete $P(N)$, other than the deletion of the empty subset, no reductions are possible. If the completely reduced **A** is not empty, its structure is such that no row is a subset of any other, i.e., the reduced **A** forms a clutter of subsets. Its columns, however, do not necessarily form a clutter. For examples of the reduction process, see "Butterflies"[2] and "Beetles" in Chapter X.

The reductions just described are based first on the inequality in the constraint set $\mathbf{Ax} \geq \mathbf{1}$, and second, that there should be no redundancy in $\mathbf{x}$ (to be considered below in more detail); they need to be modified if strict equality (i.e., a partition) is required. An argument against requiring a strict partition is given above.

The objects remaining after all reductions, together with those deleted by Rule 1 of Algorithm II.1, form a representation (transversal) of the subsets. There may be more than one transversal, but other than permutations of rows and columns, the structure of the final array is unique.

A subset probability measure

No properties of the individual subsets, such as some measure of the heterogeneity of the members, have been involved in the reduction rules. The reductions have depended only on the relationships implied by the requirement that $\mathbf{Ax} \geq \mathbf{1}$, and an as

[2] Names given in quotation marks refer to the case studies in Chapter X.

yet unconsidered notion of irredundancy. A requirement for a partition necessarily implies irredundancy, which is not a property of a covering; irredundancy in coverings is considered later in this chapter.

The constraint set, $\mathbf{Ax} \geq \mathbf{1}$ is now exploited further in a way somewhat different from the reductions, based on the fact that the set of $\mathbf{x}$ from which a choice is to be made forms a Borel σ-field, and that a probability measure can be obtained for each subset. The set of measures are then used to find an optimal solution to the covering problem. The argument on which this measure is based proceeds as follows.

In the circumstances leading to Rule 1 (Algorithm II.1) of the reductions, the i^{th} object belongs to precisely one subset, and so gives complete information in the sense that this subset belongs in *every* covering solution. Similarly, if an object belongs to just two (distinct) subsets, at least one (if not both) must belong in every covering solution. This argument can be taken as forming the essential part of proving the following:

> PROPOSITION II.1. *The **more** subsets to which an object belongs, the **less** information given by that object about which subsets must belong in every covering.*

The same argument can be taken to prove the dual:

> PROPOSITION II.2. *The **more** objects a given subset contains, the **more** likely is that subset to belong to an irredundant covering.*

At this stage, the ground is prepared for introducing two non-negative measures, namely, $\mathbf{p}$, an m-element non-negative vector reflecting the importance of the subsets with respect to a covering, and $\mathbf{q}$, a n-element non-negative vector that measures the

information given by each object relevant to the subsets to which it belongs about the subsets' participation in a covering. The relationship between **p** and **q** is based on the duality with respect to **A** of **p** and **q**.

If **q** is a vector measuring the relative importance of each object, $\Sigma q_i = 1$, and **p** measures the relative importance of each subset, $\Sigma p_k = 1$, then from Proposition II.1

$$\mathbf{A}^*\mathbf{p} = \alpha\mathbf{q}$$

where $\mathbf{A}^* = \{1 - a_{ik}\}$ and α is a scalar. Similarly, since **p** measures the relative importance of each subset with respect to a covering, then from Proposition II.2

$$\mathbf{A}^T\mathbf{q} = \beta\mathbf{p}$$

where ß is a scalar. These two relationships form a proof of the following theorem, important because **p** and **q** play major roles in conditional clustering.

THEOREM II.1. (a) *The vector **p** is the column eigenvector associated with the largest eigenvalue of A^TA^*; and*
(b) *the vector **q** is the column eigenvector associated with the largest eigenvalue of A^*A^T.*

Remark. Two proofs of this theorem are given; the first does not provide any obvious interpretation of **p** and **q**; the second does.

Proof.(First proof). Substituting $\alpha^{-1}\mathbf{A}^*\mathbf{p}$ for **q** in $\mathbf{A}^T\mathbf{q} = \beta\mathbf{p}$ proves (a), i.e., $\mathbf{A}^T\mathbf{A}^*\mathbf{p} = \alpha\beta\mathbf{p} = \lambda\mathbf{p}$, so that **p** is an eigenvector of $\mathbf{A}^T\mathbf{A}^*$; similarly, substituting $\beta^{-1}\mathbf{A}^T\mathbf{q}$ for **p** in $\mathbf{A}^*\mathbf{p} = \alpha\mathbf{q}$ proves part (b), i.e.,

$$\mathbf{A}^*\mathbf{A}^T\mathbf{q} = \beta\alpha\mathbf{q} = \lambda\mathbf{q}. \qquad \text{Q.E.D.}$$

(*Second proof*; Lefkovitch 1985*a*). Define

$$X = \{\mathbf{x} : \mathbf{Ax} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m\},$$

as the set of all distinct coverings (not only irredundant) permitted by the constraints. In X, the k^{th} column of $\mathbf{A}$ (i.e., the k^{th} subset of objects) is indicated m_k times in X, where $m_k = \Sigma_{x \in X}(x_k)$; define $M = \Sigma_k m_k$. The number of ways a given set of m_k values can be realized is the multinomial coefficient

$$w = M!/\Pi_k(m_k!).$$

The greatest number of ways that the assignment can be achieved, and therefore is the least prejudiced, is if this function is maximized, subject to constraints that depend on $\mathbf{A}$. Defining $u_k = m_k/M$, using the Stirling approximation to the factorials, justified by the facts that $M = |X| = O(2^m)$, and the m_k tend to be large, after taking logarithms, gives

$$\log_e w \approx -M \Sigma u_k \log_e u_k,$$

which can be recognized as entropy. Since M is a constant, although unknown, it can be ignored. The more frequently the k^{th} column of $\mathbf{A}$ is indicated in X, the larger is m_k and hence u_k. An m_k will be large iff the objects in the k^{th} subset belong to few others, and so for such objects, the value of $v_i = \Sigma_k a_{ik} u_k$, i.e., the sum of the set of u_k to which object i belongs, will also be large. The problem, therefore, is to determine the u_k and v_i in such a way that

$$\mathbf{Au} = \mathbf{v}$$

where v_i is large if the number of subsets to which object i belongs is small and is small if this number is large, *without* forming X. If

$$Y = \{\mathbf{y} : (\mathbf{A}^*)^T\mathbf{y} \geq \mathbf{1},\ \mathbf{y} \in \{0,1\}^n\}$$

is the set of all representations of the m subsets in the *complementary* problem, and $\hat{v}_i$ the relative frequency of the number of times object i occurs in Y, then the logarithm of the corresponding multinomial coefficient is proportional to $-\Sigma_i \hat{v}_i \log_e \hat{v}_i$, and, omitting a few steps paralleling those above, $\hat{\mathbf{v}}$ needs to satisfy

$$\mathbf{A}^*\hat{\mathbf{v}} = \hat{\mathbf{u}}.$$

Since $\mathbf{A}$ and $\mathbf{A}^*$ are complementary and so represent equivalent sets of propositions, it follows that $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are the same as $\mathbf{u}$ and $\mathbf{v}$. Combining these relationships gives

$$\mathbf{A}^*\mathbf{A}^T\mathbf{u} = \mu\mathbf{u}$$

and

$$\mathbf{A}(\mathbf{A}^*)^T\mathbf{v} = \mu\mathbf{v}.$$

Because the derivation of $\mathbf{u}$ and $\mathbf{v}$ assumes the principle of indifference in its use of the maximum entropy principle and aims at a *maximum* opportunity for each subset to contribute to a covering, contrasting with the objective functions (described below) that seek to *minimize* the number of subsets in the optimal solution, the vectors $\mathbf{p}$ and $\mathbf{q}$ are obtained by reversing the roles of $\mathbf{A}$ and $\mathbf{A}^*$ in the previous arguments, so that

$$\mathbf{A}^T\mathbf{A}^*\mathbf{p} = \lambda\mathbf{p}$$

and $\mathbf{A}^*\mathbf{A}^T\mathbf{q} = \lambda\mathbf{q}$

as before. Q.E.D.

Note that there is no need to maximize entropy explicitly, since the first proof shows that the solutions for **p** and **q** are unique, i.e., these vectors of probabilities have maximum entropy. In consequence, they satisfy the requirements for information operators discussed by Shore and Johnson (1980).

The uniqueness of the solutions is also a consequence of the non-negativity of the elements of **A**, which contains only zeros and unities, so that $\mathbf{A}^T\mathbf{A}^*$ has non-negative (integer) elements, with zeros on the diagonal and perhaps elsewhere. If **A** is reducible (i.e., there exists a permutation of the rows, **R**, and columns, **C**, of **A** such that **RAC** is block diagonal, where the diagonal blocks, which are either scalars, square, or rectangular, include some unities, and all off-diagonal blocks are strictly zero), the clustering problem is simplified into subproblems each consisting of the subset of objects belonging to one such block. Note that the set of blocks forms a partition of the objects, possibly providing a sufficient clustering for these data. This condition is relatively rare, however; without loss of generality, it will be assumed that **A**, and hence $\mathbf{A}^T\mathbf{A}^*$, is irreducible. At this stage, it follows from the Perron–Frobenius theorem (see Gantmacher 1959) that $\mathbf{A}^T\mathbf{A}^*$ has one real positive eigenvalue not exceeded in modulus by any other, and that the corresponding row and columns eigenvectors are the only ones having strictly non-negative elements. Since each row and each column of **A** contains at least one unity, it is not difficult to show that the elements of **p** and **q** are strictly positive.

There is a log-linear model that leads to the same solution procedure as that for **p** and **q**. Suppose n students are given m questions, whose answers can be either right or wrong, and it is desired to assign scores to the students weighted according to the

difficulty of the questions. It is not known which questions are easy and which are hard, and so there is the need to assign scores to these as well. Clearly, a question answered correctly by all students can be regarded as easy, and a student who correctly answers all questions should rank high. If **A** is the $n \times m$ array of zeros and unities, where $a_{ik} = 1$ denotes a correct answer by the i^{th} student to the k^{th} question, then Rasch (1960) proposed determining v_i, the score for the i^{th} student, and u_k, the difficulty of the k^{th} test, from a model which asserts that

$$\Pr(a_{ik} = 1) = v_i/(v_i + u_k),$$

from which it is apparent that

$$\Pr(a_{ik} = 0) = u_k/(v_i + u_k),$$

so that the odds are

$$\Pr(a_{ik} = 1)/\Pr(a_{ik} = 0) = v_i/u_k.$$

Taking logarithms, and after normalisation of **u** and **v**,

$$\log [\Pr(a_{ik} = 1)/\Pr(a_{ik} = 0)] = \alpha + \log(v_i) - \log(u_k),$$

which is a log-linear model. Fienberg (1984) showed that the solution for **u** and **v** satisfies

$$\mathbf{Au} = \alpha\mathbf{v}$$

and

$$(\mathbf{A}^*)^{\mathrm{T}}\mathbf{v} = \beta\mathbf{u},$$

which together lead to

$$(\mathbf{A}^*)^T\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$$

i.e., the row eigenvector corresponding with the Perron–Frobenius eigenvalue of $\mathbf{A}^T\mathbf{A}^*$; this solution contrasts with the column eigenvector of the same matrix, the solution obtained above. If the focus is on the students, the solution wanted is $\mathbf{v}$ in

$$\mathbf{A}(\mathbf{A}^*)^T\mathbf{v} = \lambda\mathbf{v},$$

which again is different.

A comparison between the Rasch and the clustering models (Table II.1) shows that the scores to be assigned to the tests in the Rasch model are given by the Perron–Frobenius row eigenvector of $\mathbf{A}^*\mathbf{A}^T$, and the object scores for clustering given by the corresponding column eigenvector. Of more interest are the candidate scores, which are given by the Perron–Frobenius row eigenvector of $\mathbf{A}^T\mathbf{A}^*$; the subset scores, of particular interest in the present study, are given by the corresponding column eigenvector. Thus the solution methods for the Rasch and the present models are dual to each other. Since the Rasch model is log-linear (Fienberg 1984), it follows that there is a log-linear model from which that of clustering can be derived. There may also be a connection with the model proposed by Sutcliffe (1986), who describes a procedure in which objects are to be ordered from the more to the less differentiated, and attributes from the more to the less differentiating. However, the fact that two or more models have a common solution procedure does not necessarily imply their equivalence.

In ecology, empirical data sometimes consist either of the proportion either of some fixed number of samples, or of the total flora or fauna for each of n species at each of m sites. In taxonomy, if an object is a species, the empirical data for such an

Table II.1 Comparison between the Rasch and the clustering models

	Rasch model		Clustering model
1.	Students The more tests students get correct, the higher their score	1.	Objects The more subsets to which an object belongs, the lower its score
2.	Tests The more students who are correct in a test, the easier is the test	2.	Subsets The more objects in a subset, the more likely is it to be in a covering
3.	(a) Test scores $\mathbf{v}$ $\mathbf{A}(\mathbf{A}^*)^T\mathbf{v} = \lambda\mathbf{v}$	3.	(a) Subset scores $\mathbf{p}$ $\mathbf{A}^T\mathbf{A}^*\mathbf{p} = \lambda\mathbf{p}$
	(b) Student scores $\mathbf{u}$ $(\mathbf{A}^*)^T\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$		(b) Object scores $\mathbf{q}$ $\mathbf{A}^*\mathbf{A}^T\mathbf{q} = \lambda\mathbf{q}$

attribute may be an estimate of the proportion of individuals showing presence. One possibility for such data is to choose some threshold value, e.g., 0.5, and define the **A** matrix accordingly. This arbitrary choice can be avoided, at least for estimating the probabilities, by a simple extension of the binary data model as follows. Define the matrix **B** to consist of the probabilities of occurrence of each species in each site, and let these probabilities be estimated by the proportions; **A** can then be regarded as a special case of **B** in which the probabilities are either 0 or 1. The matrix, $\mathbf{A}^*$, is replaced by $\mathbf{B}^* = \mathbf{1}\mathbf{1}^T - \mathbf{B}$ and the probabilities for the species and sites obtained from $\mathbf{B}^T\mathbf{B}^*$. The constraint on the solution, however, must be $\mathbf{B}\mathbf{x} \geq \mathbf{b} > \mathbf{0}$; although it is not clear if **b** has a best set of values, it seems that each element should be specified to be greater than 0.5. An interpretation of the elements of $\mathbf{B}^T\mathbf{B}^*$, as well as of possibilities for the constraints, is made

later in this chapter. In Chapter X, "André's data" illustrates the direct study of an empirically observed **A**; "Plant frequencies" considers a **B**.

To treat **p** and **q** as probability measures for the subsets and objects, respectively, it only remains to normalize each to sum to unity. An iterative procedure now follows, which obtains both **p** and **q** simultaneously, which requires only additions and subtractions other than the normalization, and which avoids the need to form $\mathbf{A}^T\mathbf{A}^*$, whose elements can exceed unity. Let *eps* be a small positive quantity:

ALGORITHM II.2. *Given A, calculate p and q:*

step 0: initiate $q^0 = \{1/n\}$; $t = 0$; $p^0 = \{1/m\}$;
step 1: calculate $p^t = A^T q^{t-1}$;
step 2: $p^t = \{p_k^t / \Sigma\, p_k^t\}$;
step 3: if $|p^t - p^{t-1}| < eps$, *terminate*;
step 4: calculate $q^t = A^* p^t$;
step 5: $q^t = \{q_i^t / \Sigma\, q_i^t\}$; *and*
step 6: $t = t + 1$; *go to step 1.*

Because the elements of **A** and **A*** are either 0 or 1, the matrix multiplications in steps 1 and 4 require additions only. If there is no interest in the relative importance of the objects, the normalization of **q** in step 5 can be omitted. The rate of convergence of this algorithm depends on the ratio $|\lambda_2|/\lambda_1$, where λ_1 is the Perron–Frobenius eigenvalue, and λ_2 the second eigenvalue in modulus; the smaller the ratio, the quicker the convergence. It also depends on *eps*; the smaller the value, the more steps are required. A rough guide for *eps* is based on the value φ, which is the largest value such that a test of $(1.0 + \varphi)$ as equal to 1.0 is reported as true in fixed word length (computer) arithmetic. Then *eps* is defined as φ/m. This choice is perhaps oversensitive, because it will bring about more iterations than are

really needed to obtain a clustering, but there may be circumstances where great numerical accuracy is needed. The norm used in step 3 may also be replaced, for example, by the sum of the squared differences (replacing *eps* by its square), or the maximum of the absolute values of the differences.

Because the definition of **p** and **q** can be based on arguments depending on entropy, the elements of these two vectors (both standardized to sum to unity) can be interpreted as probability measures corresponding to each subset and object, as appropriate. In other words, this process has assigned a measure, which can be regarded as a probability, to each element of the Borel algebra of the subsets and of the objects. These probabilities are not limiting values of frequencies, even though they can be derived from counting arguments; they are unique, and, from the entropy argument, represent the least structure consistent with the requirement that the information in the rows of **A** is dual to that of the columns of **A*** (Lefkovitch 1985*a*). Examples of these two sets of probabilities are given in Chapter *X* "Plant frequencies," "Arctic grasses," "Letters," "Fescue grasses," and "Blood and language."

Since in **p** (and **q**) some elements may differ by an amount sufficiently small that it is reasonable to believe that in exact arithmetic, or with minimal changes in **A**, they would be identical, it is reasonable to assign a tolerance to **p**, and hence to any optimal solution based on it. Suppose **b** is a set of hypothetical probabilities for **p** (e.g., $\mathbf{b} = \{1/m\}$); then the minimum discrimination statistic

$$G^2 = \|\mathbf{A}\|_1 \sum p_k \log_e(p_k/b_k),$$

where $\|\mathbf{A}\|_1$ is the number of unities in **A**, can be regarded as a χ^2 with $m - n - 1$ degrees of freedom (if $n > m$, replace the test by that for the corresponding set representation, i.e., reverse the

roles of rows and columns) for examining the consistency of **p** with **b**. The set of assignments to **b** such that $\Pr(G^2) \geq \alpha$ defines an α-tolerance class. From the concentration theorem for entropy (Jaynes, 1983), the range of values for the entropy of the members of this class is given by

$$(H(\mathbf{p}) - \chi^2_{m-n-1}/\|\mathbf{A}\|_1) \leq H(\mathbf{b}) \leq H(\mathbf{p}),$$

where $H(.)$ denotes entropy. If this outer interval is small, then there is every reason to believe that the **p** is well defined; it will be smaller, for example, the larger is $\|\mathbf{A}\|_1$.

Missing values: reductions and probabilities

In performing the reductions directly on empirically obtained data, it is not unlikely that some entries in the $\{0,1\}$ array, **A**, may not be known, or may be variable. This situation leaves some uncertainty into what should be done to arrive at a solution. Although admissible actions depend on the circumstances bringing about the uncertainty, four possibilities are now examined.

(1) One reduction can be carried out without any loss of information. Suppose object i, other than having a missing value for subset k, would have been deleted by object j. If $a_{jk} = 0$, object i can be deleted; it cannot be deleted with complete confidence if a_{jk} is missing or is unity. The proof that this reduction does not lose information is trivial.

(2) Replace the missing values by 0, and continue as usual; this action includes the previous proposal but also allows other reductions.

(3) Replace the missing values by 1, and continue as usual; this action is perhaps the most dubious.

(4) If most of the missing values are for a very few of the subsets, eliminate these subsets; less satisfactory is to eliminate objects for the same reason, unless many objects appear to represent a few groups, and the deletion of objects is unlikely to have excluded a group.

There is every reason to do all three of proposals 2–4, and to compare the results.

In computing the probabilities based on the original array, the most neutral approach is to replace the missing values by 0.5 in the iterative procedure.

Irredundancy

Irredundancy, briefly mentioned earlier in this chapter, is now explored further. If $\mathbf{x} = \mathbf{1}$ and there is no null row in $\mathbf{A}$, then $\mathbf{Ax} \geq \mathbf{1}$ is satisfied; in fact, there may be very many distinct $\mathbf{x}$ satisfying this constraint. At this stage the principle of parsimony is introduced; here it is expressed by restricting $\mathbf{x}$ to be such that there is no redundancy in the choice of subsets, i.e., deleting any one of the indicated subsets violates the constraint. By contrast, the effect of the inclusion of others is to increase some of the elements of $\mathbf{Ax}$.

Amongst the set of irredundant solutions, some satisfy a stronger form of parsimony, namely, the choice of a *minimal* number of subsets. In the present context, this requirement gives rise to the **minimal set-covering problem**, which can be represented as the need to find $\mathbf{x}$ to

$$\text{minimize } \{\mathbf{x}^T\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m\}.$$

Vercellis (1984) provided a method for obtaining a very rough estimate of the value of the optimal $\mathbf{x}^T\mathbf{x}$. If the elements of $\mathbf{A}$ are independent (clearly, they are not) and are unity with constant

probability, π, (there is no reason to make this assumption), then for large numbers of subsets and constraints, as well as a number of other conditions, the ratio between the optimal $\mathbf{x}^T\mathbf{x}$ and $z = \log_e n/\log_e(1/(1 - \pi))$ tends to unity.

Because it may be true that there are several solutions for $\mathbf{x}$ such that $\mathbf{x}^T\mathbf{x}$ takes the same minimal value, it becomes of importance to find a rule for choosing from among these irredundant or minimal solutions. The proposal for determining such a rule is based on the use of the vector $\mathbf{p}$ to choose the best irredundant or minimal covering. Two possibilities are considered; the arguments for each are different, although related, and depend on the fact that $\mathbf{p}$ and $\mathbf{q}$ are probability measures. As noted above, the elements of $\mathbf{p}$ and $\mathbf{q}$ have no interpretation as frequencies, nor are they degrees of subjective belief. Consider the fact that the unreduced $\mathbf{A}$ can be used to examine a whole series of propositions such as:

- "objects i and j belong in subset k"
- "subset k contains objects i_1, i_2, etc., but does not contain objects j_1, j_2, etc."
- "subset k belongs in every covering"

and others of a similar nature. These propositions may be shown to be true or false by direct reference to $\mathbf{A}$. Some propositions are relevant to an irredundant covering, such as:

- "subset k is a member of every covering"
- "subset k is a member of an irredundant covering"

and so on. Some can be shown to be true or false (see the reduction rules of Algorithm II.1); for others, an inspection of $\mathbf{A}$ does not of itself give any absolute answer. However, the counting

arguments leading to the fact that of all non-negative probability vectors corresponding with $\{0,1\}^m$ and $\{0,1\}^n$ constrained as described above, the **p** and **q** have maximum entropy, implies that in **p** those elements having the largest values will tend to be present the most often in a covering (not necessarily minimal) of the objects; thus

> the value of p_k is a measure of the degree of support given by **A** to the proposition that subset k belongs in the optimal covering.

For example, all subsets of the n objects eliminated by the reductions, as well as those belonging to the power set of N but absent from **A**, have zero support, while the mandatory subsets have absolute support (of unity). One criterion emerging from the parsimony principle is to choose those subsets such that the support for the joint proposition that they form the optimal covering is larger than (or, at least, not exceeded by) that of the support for any other such proposition. In formal terms, this reasoning leads to an objective function, which is to maximize the joint probability. With the convention that $0^0 = 1$, the problem to be solved can be expressed as:

$$\text{determine } \mathbf{x} : \mathbf{x} \in \{0,1\}^m \text{ so that } \Pi p_k^{x_k} \text{ is a maximum.}$$

Since the p_k are probability measures, the solution obtained is the joint probability of the choice. Since all objects have to be considered, **x** is constrained so that $\mathbf{Ax} \geq \mathbf{1}$, as is always required. Before re-expressing this program in a form more suitable for computation, note that not only will redundant subsets be excluded, since the inclusion of such will necessarily reduce the value of the joint probability, but also, for the same reason, $\mathbf{x}^T\mathbf{x}$ will be a minimum at the optimal solution. Thus it is apparent that the solution, **x**, is an irredundant minimal covering, and that there is

no need as such to determine all the minimal coverings and find which is best, since the solution to the problem is necessarily confined to such coverings. However, it is useful to remember that although **p** and **q** can be determined from the original **A** (but not necessarily, as discussed below), the search for a covering by any algorithm should be based on the reduced **A** (Algorithm II.1), keeping the appropriate values of p_k derived from the original **A**, since the amount of work required will be reduced considerably.

If the elements of **p** are replaced by the negative of their logarithms, maximizing the joint probability of the choice is equivalent to

$$\begin{gathered} \text{minimizing } -\sum x_k \log_e p_k \\ \text{subject to } \mathbf{Ax} \geq \mathbf{1}, \\ x_k \in \{0,1\} \end{gathered}$$

or, more compactly, after defining $c_k = -\log_e p_k$, k = 1...m, (note that $p_k > 0$) by

$$\min \{\mathbf{c}^T\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m\},$$

where **c** is a vector of "costs" and $\mathbf{c}^T\mathbf{x}$ is called the objective function. This formulation is an ordinary linear least-cost set-covering problem, one of the more tractable integer programming problems. Three general algorithms for this problem are described in some detail later in this chapter. One is heuristic and fast and almost always has resulted in the optimal solution (where the latter is known). The second is exact but requires considerable arithmetic, perhaps unjustified considering the empirical nature of the data used in clustering. The third is stochastic and obtains the optimal solution with probability unity (depending on the number of iterations); it generalizes to more than one noncommensurate cost or if the objective functions are nonlinear.

The arguments leading to an objective function based on joint probability necessarily lead to solutions that are minimal and therefore can be considered to be too strong a requirement. The following discussion leads to a solution, which, although irredundant, is not necessarily minimal. This circumstance is not a disadvantage, because, as argued in Chapter I, the stronger are the scientist's imposed requirements, the more restricted is the choice of solutions, so converting the solution into one in which groups tend to be formed rather than revealed. It is preferable, it is now claimed, to use the weaker condition and inspect the solution to see if it is also minimal.

It has already been remarked that underlying the definition of $\mathbf{p}$ is the maximum entropy principle; as is well known (Jaynes 1983), the maximum entropy probability vector is the assignment of probabilities to the elements showing the least departure from uniformity, i.e., the least structure, consistent with the constraints on the system. (An early statement of this property is by Woodward (1953, p. 21–25), who also proved that "for a given mean squared value of x [assumed continuous and unbounded], the Gaussian distribution is the most random of all.") Thus the assumptions of information theory are implicit in these probabilities, and so remaining within its theoretical framework for any further development does not introduce any different concepts, as does joint probability.

It has already been shown that the entropy of $\mathbf{p}$, defined by $-\sum p_k \log_e p_k$ (with the usual convention that if $p_k = 0$, then $p_k \log_e p_k$ is set to zero) is a maximum subject to constraints depending on $\mathbf{q}$ (and vice versa). As a consequence, it has least structure with respect to the relationship between the rows and columns of $\mathbf{A}$. By contrast, one objective of clustering is to select subsets so that the choice has the most structure, i.e., to maximize the information (minimize the entropy; Watanabe 1981) of the choice; this objective can be expressed as choose $\mathbf{x} \in \{0,1\}^m$ to minimize $-\sum x_k p_k \log_e p_k$ subject to $\mathbf{Ax} \geq \mathbf{1}$.

Defining $c_k = -p_k \log_e p_k$, this program can also be written as

$$\min \{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m\},$$

which differs from the expression for the solution using maximum joint probability only in the definition of **c**. This formulation is still that of an ordinary, linear, least-cost set-covering problem, since the p_k, and hence the $-p_k \log_e p_k$, are known. It is apparent that there will be no redundancy in the optimal solution, since the inclusion of additional subsets will increase the objective function; but also there is no guarantee that $\mathbf{x}^T\mathbf{x}$ will be a minimum.

The solution obtained by maximizing information can be interpreted as a balance between two different aspects of the entropy principle; maximum entropy is used to determine the probabilities, maximum information (equivalent to minimum entropy) for the choice of subsets. This balance contrasts with the maximum joint probability procedure, which uses maximum entropy to obtain the probabilities, followed by maximizing the joint probability, which, although related to maximum information, is nevertheless a different principle. If the axiom systems for maximum entropy of Shore and Johnson (1980), Jones and Byrne (1990), Paris and Vencovská (1990), and Csiszár (1991), and the resulting theorems are appropriate for the clustering problem, consistency is guaranteed for the maximum information solution, although not for the maximum joint probability solution, unless they happen to coincide. If maximizing entropy is equivalent to an increasing variance principle, and maximum information to one of variance reduction, the solution obtained can be regarded as being an expression of the balance between the entropy-generating and entropy-reducing processes, which is proposed in Chapter I as a replacement for the principle of parsimony in phylogenetic inference (evolutionary reconstruction).

A further objective function of some interest can also be based on **p**. Defining again the column vector

$$\mathbf{c} = \{c_k\} = \{-\log_e p_k\},$$

and $\mathbf{D} = \{\text{diag } p_k\}$, the two objective functions, $-\sum x_k \log_e p_k$ and $-\sum x_k p_k \log_e p_k$, can be written as $\mathbf{x}^T\mathbf{I}\mathbf{c}$ and $\mathbf{x}^T\mathbf{D}\mathbf{c}$, respectively. If $-\log_e p_k$ is interpreted as the (information) measure of subset k, the value of the objective function for the solution, $\mathbf{x}$, gives the combined measure of the solution, in which the prior probability is uniform for each subset in the joint probability solution, $\mathbf{x}^T\mathbf{I}\mathbf{c}$, and is equal to $\{p_k\}$ in the maximum information solution, $\mathbf{x}^T\mathbf{D}\mathbf{c}$. This objective function suggests the possibility of a more general matrix, $\mathbf{P}$, to replace $\mathbf{D}$, in which there are nonzero off-diagonal elements. A definition of $\mathbf{P}$ arises naturally from $\mathbf{p}$ and $\mathbf{q}$:

DEFINITION II.1. *For subsets r and s, let k_{rs} denote an n-element vector in which the unities point to the objects in their intersection; note that $|r \cap s| = |k_{rs}| \geq 0$, with strict equality to zero for disjoint subsets; then*

$$\boldsymbol{P} = \{p_{rs}\} = \{q^T k_{rs}\}.$$

It follows that $p_{rr} = d_{rr}$, and if subsets r and s are disjoint, $p_{rs} = 0$. Defining $\mathbf{Q} = \{\text{diag } q_i\}$, $\mathbf{P}$ can be equivalently defined as $\mathbf{A}^T\mathbf{Q}\mathbf{A}$, i.e., the objective function is $\mathbf{x}^T\mathbf{A}^T\mathbf{Q}\mathbf{A}\mathbf{c}$, in which $\mathbf{x}$ is unknown. Since $\mathbf{P}\mathbf{c}$ is an unchanging vector for any given problem, the objective function $\mathbf{x}^T\mathbf{P}\mathbf{c}$ is still linear. With this modified objective function, choosing $\mathbf{x}$ to minimize $\mathbf{x}^T\mathbf{P}\mathbf{c}$ tends to produce solutions that minimize the number of overlapping subsets. The real issue is whether this restriction is desirable; as discussed in Chapter I, if the objective is to reveal the "true" groups within the data, because $\mathbf{P}$ imposes more structure on the solutions than $\mathbf{D}$, which in turn imposes more than $\mathbf{I}$, only $\mathbf{I}$ should be used. In contrast, if the objective is to form the objects into groups for which it is desirable that so far as possible they form a partition, there is good reason to adopt this extended $\mathbf{P}$. If the

goal is to generate hypotheses for subsequent evaluation, the use of all three may have value.

A further reduction

Although other costs, **c**, are discussed below, it is convenient at this stage to introduce a further reduction rule, which makes use of them. This rule, which extends Algorithm II.1, may allow others of those rules to become available, with the result that the further simplification may complete the vector **x** without need of any of the more complicated algorithms.

ALGORITHM II.1 (continued). *Cost-based reductions*:

RULE 4. *If* **k** *is a subset with cost* c_k, *and there exist other subsets,* k_1, k_2,... *such that* **k** *is a subset of their union and also the sum of their costs does not exceed* c_k, *then* **k** *can be deleted from* **A**.

Such a **k** is redundant because any covering requiring objects belonging to **k** can be covered by one or more of the other subsets at lower cost. From the definition of **p**, cost-based reductions are not particularly effective for the maximum joint probability objective function but may be so for costs used in maximizing information, especially if the probabilities differ greatly; they are often very helpful if the costs are independently provided. It is apparent that rules 1–3 (Algorithm II.1) should be used to achieve the maximum reduction possible before applying Rule 4, because they produce a minimal set of constraints together with a partial, if not complete, solution that is *independent* of costs.

Conversion of coverings to partitions and dendrograms

One of many objectives of clustering is to obtain a classification that satisfies the requirements of the rules of biological

nomenclature. One requirement of these hierarchical systems is that a taxon cannot belong to more than one category at any given level of the system; a species, for example, can belong to just one subgenus, a subgenus to one genus, a genus to one tribe, a tribe to one subfamily, and so on. Because the overlapping groups that may be produced by the set covering appear inconsistent with this requirement, some additional action or further reasoning is needed to achieve at least a candidate classification. One possibility is to change the constraints on the system so that it becomes the strict equality

$$\min \{\mathbf{c}^T\mathbf{x} \mid \mathbf{Ax} = \mathbf{1}, \mathbf{x} \in \{0,1\}^m\},$$

but some consequent difficulties may arise. For example, no $\mathbf{x}$ may satisfy these equality constraints for empirically obtained $\mathbf{A}$. More to the point, however, is that even though all partitions are coverings, not all coverings are partitions; as already noted, a partition is more solidly supported if an optimal covering is inspected and *found* to be a partition. Garfinkel and Nemhauser (1972) described the reductions and methods for solving linear least-cost set-partitioning problems.

Interpreting overlapping subsets

Numerical procedures, which may give rise to overlapping subsets, have been proposed several times as a solution to the clustering problem. For example, Needham (1961, 1965), Auguston and Minker (1970), Lefkovitch (1975), and, more recently, Opitz and Wiedemann (1989) all described methods based on maximal cliques (maximally connected subgraphs). As in the procedures of this book, nondisjoint subsets generate some interesting problems—to determine why the subsets overlap and also the status of the objects in an intersection. It is doubtful if notions of simplicity and parsimony alone are relevant to an answer.

In an optimal covering that is not a partition, suppose two subsets, I and J, in this solution have a nonempty intersection, $I \cap J$. What is the status of $I \setminus J$, $J \setminus I$, and $I \cap J$?

(1) The "statistical" explanation: there are "real" groups for which the range of variation is such that some of the members of a group are closer to the centroid of others than they are to their own.

Example. In two multivariate normal populations having different centroids, it follows that $I \setminus J$ and $J \setminus I$ belong in distinct subsets, and that objects in the intersection $I \cap J$ belong to "tails" of the parent distributions; additional data may be able determine which of $I \cap J$ belong to I and which to J.

(2) The "procedural" explanation: there is only one "real" group, which has been partly divided because of the interaction between the clustering method and the measure of dissimilarity.

Example. A threshold-determined clique is necessarily contained within a hypersphere in the dissimilarity space; there is no logical reason to expect that a single true group can be contained within just one subset at the chosen threshold. Thus $I \cup J$ represents one group, and the formation of I and J is an artifact of the calculations. It follows that the objects in $I \cap J$ should be considered to be central (Jardine and Sibson 1971).

(3) The "irrelevant" explanation: the groups are a consequence of external circumstances.

Example. The species associations in ecological studies may be no more than a reflection of abiotic factors. Ruling out this class of explanation is very important. If there are provenance data, such as known geographical distribution, or known abiotic (or biotic) differences, the data under study should be subjected to some formal null hypothesis

evaluations. Only if the data are consistent with the null hypotheses does clustering become a useful technique. Further, once groups have been established by clustering, it may be of value to construct a contingency table classified by the cluster-generated groups and the provenance data.

(4) The "evolutionary" explanation: $I \bigcup J$ represents a taxon in the process of subdivision, that $I \bigcap J$ represents the central (typical, ancestral?) form, while $I \setminus J$ and $J \setminus I$ are the diverging forms.

(5) The "hybridization" explanation: $I \bigcap J$ represents a hybrid between $I \setminus J$ and $J \setminus I$.

Explanations (4) and (5) are probably indistinguishable using purely numerical techniques unless these relate to additional biological information.

(6) The "continuum" explanation: $I \bigcap J$ cannot be distinguished either from $I \setminus J$ or from $J \setminus I$, which, however, can be distinguished from each other.

This explanation differs only in degree from that of (3) but is included here because it is reminiscent of Menger's (1979) characterization of the physical continuum by a failure of transitivity:

$$A = B,\ B = C, \text{ but } A \neq C,$$

i.e., A and C cannot be distinguished from B but can be distinguished from each other. In these circumstances, the three relationships imply that A, B, and C are all different.

If it can be determined that (3) does not apply, the remainder present a dilemma. Perhaps (2) should be assumed to hold from small sample considerations if either or both the number

of objects and attributes are small; (2) then represents a cautious interpretation, which is that the solution is conditional on the attribute set, and for small such sets, the degree of approximation to the "true" state of nature is perhaps poor. The first consistency assumption (Chapter I) implies that if the true classification is represented by a binary vector, $\psi \in \{0,1\}^m$, then

$$\lim \mathbf{x} = \psi \text{ as } n \to \infty.$$

For small m, therefore, what can be concluded?

There is also a pragmatic consideration; in biology, it is generally believed that distinct taxa differ morphologically, physiologically, ecologically, and so on to a greater or lesser degree, and that failure to name each distinct entity can result in considerable confusion and the loss of painfully collected and expensive information. Thus information is lost in not giving a distinct label to objects for which evidence supports some distinction in the attributes under investigation. Furthermore, while it may ultimately prove to be true that objects assigned to differently labeled subsets are subsequently found to belong to the same taxon, no information is lost, because the literature on the components can be combined; the only penalty is the nuisance of having to include another name in the formal synonymy of the taxon. The rules of nomenclature are meant to serve communication in biology and are to be used intelligently for this purpose.

Although the argument for "splitting" may be persuasive, there is also a contrary one, particularly for relatively small m. Because the choice of attributes to some extent is uninformed, in the (relatively) small subset of those possible which are chosen to represent the objects, there may be attributes distinguishing subsets having little (if anything) to do with genuine taxa. Thus the

existence of overlapping subsets in an optimal solution suggests that evidence only supports one taxon, namely, $I \bigcup J$. Thus to form a partition from a covering is no more than the formation of the union of overlapping subsets, each called a **muster** (Chapter VIII discusses musters). In some ways, this "lumping" procedure has advantages over splitting for the following reasons:

(1) If there is only one muster, to which all n objects belong, the existence of more than one entity is at best only weakly supported; if uncertainty remains, more data (attributes, objects) need to be collected to settle the issue.

(2) If there is more than one muster, each may be separated from the remaining and further studied alone, preferably with further attributes, to see if in the absence of other musters, greater resolution may be obtained.

The second of these is a reflection of a very important principle; it is that **the numerical properties within one distinct group**, once it is recognized as being distinct, **should not enter in determining the structure within any other**, because there is no reason to require the implied parallelism. There is every reason to **observe** that it may be true, and to report the fact, but to **require** it imposes a solution. Furthermore, a subset of the objects may generate probabilities, **p** and **q**, which differ appreciably from those of the corresponding subset when based on the complete set of objects. More interestingly, separating the objects into musters in this way allows a hierarchy to be established, because it yields an algorithm in which each not necessarily binary division depends only on the objects and attributes under consideration, and not on others.

With this preamble, a procedure for sequential division of the objects can be constructed.

ALGORITHM II.3. *A sequential divisive hierarchical procedure based on optimal set covering*:

step 0: the original data form a single muster;

step 1: for each remaining muster, form the A;

step 2: for each A in turn, obtain an optimal set covering and form the musters; and

step 3: repeat steps 1–3 until objects can no longer be distinguished.

If the state exhibited by an attribute is uniform within a subset, the attribute conveys no information on the internal structure of the subset; in consequence, it is quite likely that the data will be exhausted before each object forms a subset by itself.

Examples of muster formation given in Chapter X include "Letters" and "Fescue grasses."

Suboptimal solutions

In $\mathbf{x}_{opt}$, defined as

$$\mathbf{x}_{opt} = \{\mathbf{x}: \min(\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m)\},$$

some of the n original constraints in $\mathbf{A}$ are inactive, including those eliminated by the reductions; also parts of $\mathbf{c}$ are not involved in this solution. With respect to $\mathbf{A}$ and $\mathbf{c}$, there is no doubt that $\mathbf{x}_{opt}$ is the most interesting solution, yet information remains in the inactive constraints and in those parts of $\mathbf{c}$ not involved in this solution. If this unused information implies a grouping of the objects not resembling that implied by $\mathbf{x}_{opt}$, it may be concluded that those from $\mathbf{x}_{opt}$ are suspect. Two methods of obtaining some

of this information are now considered; the first involves no new computational procedures, the second is somewhat more complicated.

(1) Define $\mathbf{X}$ to be the subset of the columns of $\mathbf{A}$ corresponding with $\mathbf{x}_{opt}$, and $\mathbf{B}$ to be $\mathbf{A}$ but with these columns deleted, deleting also the corresponding elements of $\mathbf{c}$. Iff the i^{th} row of $\mathbf{B}$ is zero, add further columns to $\mathbf{B}$, which are zero except for the i^{th} element, which is set to unity; include additional elements of $\mathbf{c}$ set to zero. The solution to

$$\min\{\mathbf{c}^T\mathbf{x} \mid \mathbf{Bx} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m\}$$

is clearly suboptimal to $\mathbf{x}_{opt}$ and, except for the mandatory columns, which are the added single unity element columns, is independent of the first. This sequence can be repeated until $\mathbf{B} = \mathbf{I}$, i.e., until all columns of $\mathbf{A}$ have been used. It is remarkable how often the first few suboptimal solutions generate groupings very similar to those indicated by $\mathbf{x}_{opt}$.

(2) Suppose in place of obtaining a series of suboptimal solutions, the "quality" of $\mathbf{x}_{opt}$ is assessed by determining an upper bound. Clearly, *maximizing* $\mathbf{c}^T\mathbf{x}$ is given by $\mathbf{c}^T\mathbf{1}$,

i.e., the trivial solution $\mathbf{x}_{max} = \mathbf{1}$, in which there is no interest. However, suppose $\mathbf{x}$ is further constrained so

$$\mathbf{x}^T\mathbf{x} \leq \mathbf{x}^T_{opt}\mathbf{x}_{opt} = b,$$

which is the number of subsets in the optimal solution for cost $\mathbf{c}$, then any solution to this more constrained problem, if any exist, is certainly of interest. For example, if the joint probability of the

chosen subsets forms the objective function, then, since $\mathbf{x}_{opt}$ gives a minimum cover, the upper bound is a solution confined to *minimum* covers having *maximum* cost, i.e., *minimum* joint probability. For an objective function based on maximum information, since $\mathbf{x}_{opt}$ is not necessarily a minimum cover, it is conceivable that the upper bound could be associated with fewer than b subsets. The upper bound on cost is given by the value of $\mathbf{c}^T\mathbf{x}$ in the program

$$\max\{\mathbf{c}^T\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{1}, \mathbf{1}^T\mathbf{x} \leq b, \mathbf{x} \in \{0,1\}^m\}.$$

The additional constraint changes the problem into a more general 0–1 integer program, which may be solved either by using methods more general than those of optimal set covering, or by modifying the stochastic solution method (Appendix 2), restricting the candidate solutions to satisfy the additional constraint. Note that the reductions described above *cannot* be used. It is quite possible that the optimal solution to this program is $\mathbf{x}_{opt}$, because there may be only one covering of size b.

Hypothesis testing

A "contrived" classification is the tendency of a clustering procedure to generate a grouping where in fact none exists (Eilbert and Christensen (1982) discussed this circumstance). This danger is well recognized by those occupied in practical clustering, but the lack of formal diagnosis for this possibility has led to little beyond intuitive assessments. For example, it is "well known" that the single-linkage (nearest-neighbor) clustering procedure tends to produce dendrograms that exhibit "chaining," so that if chaining is observed, it is either asserted as being an artifact of the procedure, or the results of the computation are not reported at all. To some extent, the conversion of coverings to partitions, described above, is a first step in detecting a possible contrived

solution; for example, if the union of the nondisjoint subsets in the optimal covering is such that all objects are in one muster, such a solution could be interpreted as being contrived. Similarly, the resemblances and differences shown to the optimal solution by the suboptimal solutions defined above may also be informative. One possibility is to examine the array **A** under the hypothesis that $\Pr(a_{ij} = 1) = \Pr(a_{ij} = 0) = \frac{1}{2}$, as proposed by Buser (1983); Buser also derived a test statistic, which reflects the possibility that more values of unity occur in the j^{th} column than should occur by chance. Buser provided tables to aid examination of this hypothesis.

The following discussion is included primarily to show other possibilities for examination of hypotheses that may also lead to the recognition of a contrived solution; a full theory is yet to be developed.

In the process of going from the original **A** to the subsets in the optimal covering, information has been discarded. The first hypothesis of interest is to determine if the amount eliminated is large. One way to measure the amount is by comparing $\sum n_i \log_e p_i$ with $\sum x_i n_i \log_e (p_i x_i / \sum x_i p_i)$, which implies replacing m parameters by $\mathbf{x}^{\mathrm{T}}\mathbf{x}$. A second class of hypotheses considers different solutions to the problem, e.g., a comparison between the covering that maximizes the joint probability and that which maximizes the information; which is to be preferred?

For the second class of hypotheses, the joint probability, given that the solution is a partition, is the familiar

$$L_1 = n! \, \Pi_i (p_i^{n_i} / n_i!),$$

where n is the number of objects, n_i the number in each subset, and the $\{p_i\}$ the solution to $\mathbf{A}^{\mathrm{T}}\mathbf{A}^*\mathbf{p} = \lambda\mathbf{p}$, the $\mathbf{p}$ being standardized so that $\sum p_i = 1$. If the subsets form a covering that is not also a

partition, clearly $\sum n_i > n$ and so L_1 is not applicable. Suppose (for the moment) that the intersection of three (and more) distinct subsets is empty, let n_{ij} denote $|I \cap J|$, where I and J denote the objects in two distinct subsets, and let $p_{ij} = p_i p_j$. The problem, therefore, is to adjust L_1 for these intersections. In particular, the numerator should be reduced by the size of the weighted probability against intersection, i.e., by $(1 - p_{ij})^{n_{ij}}$, and the denominator reduced by $n_{ij}!$. This revision gives

$$L_2 = n! \, \Pi_i(p_i^{n_i}/n_i!) \, \Pi_{i,j}(n_{ij}!/(1 - p_{ij})^{n_{ij}}).$$

The generalization to the intersection of $3 \ldots m$ subsets is immediate, and with an obvious extension of the definition of n_{ij} and p_{ij}, the joint probability is

$$L_m = n! \, \Pi_i(p_i^{n_i}/n_i!) \, \Pi_{i,j}(n_{ij}!/(1 - p_{ij})^{n_{ij}}) \ldots \Pi_{i,j,\ldots,m}(n_{ij\ldots m}!/(1 - p_{ij\ldots m})^{n_{ij\ldots m}}).$$

For m of any reasonable size, this expression is rather formidable to compute, but the following observations make it less difficult than it may seem at first.

(1) Many multiple intersections are empty, and so

$$(n_{(t)}!/(1 - p_{(t)})^{n_{(t)}}) = 1.$$

(2) If for t subscripts all intersections are empty, then so will be those for $t + 1, t + 2, \ldots, m$.

(3) A good approximation to L_m by using L_2 (or perhaps L_3) follows from:

LEMMA II.1. *As* $t \to m$, $(n_{(t)}!/(1 - p_{(t)})^{n_{(t)}} \to 1$.

Proof. As $t \to m$, then
(a) $n_{(t)} \to 0$, i.e., $n_{(t)}! \to 1$
(b) $p_{(t)} \to 0$, i.e., $(1 - p_{(t)})^{n_{(t)}} \to 1$.
Combining (a) and (b) completes the proof. Q.E.D.

Computationally, the Stirling approximation to the factorials simplifies the calculations of L_i. The only serious computation, therefore, is generating the intersections of all subsets while ensuring that there are no repeats.

If the joint probabilities, L_i, are taken to define the likelihood of the covering, the logarithm of the relative likelihood of two solutions can be obtained as the difference in the natural logarithms of the joint probabilities and, in a hierarchical context, may be used to examine the various hypotheses.

Choosing among alternative clusterings

Given a set of objects described by some attributes, an array **A** is generated; **A** can either be used directly (Chapter IV), or via dissimilarity vectors (Chapter V), or via one or more scalar dissimilarity coefficients (chapters VII and VIII), with one or perhaps multiple objective functions (Chapter IX), to obtain clusterings of the objects. If the clusterings so obtained do not coincide, and assuming that there are no external criteria available, it is of interest either to determine which is to be preferred based on the data themselves, or to determine a consensus from them. This section considers how a choice may be made; a later section considers how to obtain a consensus.

The reader will have noted that while **p** and **q** have been obtained from any given **A**, only **p** is used to choose the optimal solution, and that **q** is not used explicitly except as part of a heuristic solution procedure; in what follows, both play a role. Prior to defining a statistic that can be used for choosing among different solutions, each optimal for a different objective criterion

and/or family of subsets of N, it is necessary to describe the motivation for the choice. Two assertions are made; if they are unacceptable, the proposed method of choice should be omitted. They are as follows:

(1) **Partitions are simpler solutions than coverings**; the greater the degree of overlap, the more complicated is the solution.

(2) **A solution with few subsets is simpler than one with many**; the simplest solution consists of one subset of all objects, and the more complicated consist of many (overlapping) subsets.

Notice that a partition may include many subsets, in conflict with the second assumption. The principle on which the statistic is based is that of choosing the simplest grouping consistent with the data; simplicity here is understood in terms of a compromise between the two assumptions just made. Even though these assumptions seem reasonable, and the decision criterion to be defined is based on an easy computation consistent with these, simplicity, plausibility, and consistency do not logically imply biological truth, nor does the last necessarily imply any of the others. Ultimately, more is required than a decision criterion.

Let $\mathbf{x}$ be an optimal solution based on $\mathbf{A}$; then suppose it is possible to determine $\Pr(\mathbf{x}|\mathbf{A})$, and to use this probability to compare different $\mathbf{x}$ or $\mathbf{A}$, or both. By Bayes theorem,

$$\Pr(\mathbf{x}|\mathbf{A}) = \Pr(\mathbf{x})\Pr(\mathbf{A}|\mathbf{x})/\Pr(\mathbf{A}).$$

To exploit this relationship, it is necessary to define the terms on the right-hand side of this expression. Let $\mathbf{z} = \mathbf{A}\mathbf{x}$ represent the

multiplicity of the objects in the solution **x**. The three terms are considered in turn:

$$\Pr(\mathbf{x}) = \Pi_{i=1\ldots n} q_i^{z_i}, \tag{1}$$

which is clearly smaller the more the subsets overlap.

$$\Pr(\mathbf{A}\,|\,\mathbf{x}) = \Pi_{k=1\ldots m}(1 - p_k)^{(1-x_k)}; \tag{2}$$

this representation of the probability is preferred over $\Pi_k p_k^{x_k}$, because once a solution has been found, the (posterior) probability of the chosen subset is unity. This definition is such that the probability is larger the fewer the subsets of high individual probability that are excluded, i.e., it satisfies the requirement of being larger the fewer subsets in the solution.

For fixed **A**, it is possible to compare two or more solutions without needing to consider Pr(**A**); for example, if **x** and **y** are two solutions, and $\mathbf{w} = \mathbf{Ay}$, then **x** is preferred to **y** if

$$\Pi_{i=1\ldots n} q_i^{z_i} \Pi_{k=1\ldots m}(1 - p_k)^{(1-x_k)} > \Pi_{i=1\ldots n} q_i^{w_i} \Pi_{k=1\ldots m}(1 - p_k)^{(1-y_k)}$$

or, defining $\alpha = \{\log_e(1 - p_k)\}$, $\beta = \{\log_e q_i\}$, in matrix notation if,

$$\beta^{\mathrm{T}}\mathbf{Ax} + \alpha^{\mathrm{T}}(1 - \mathbf{x}) > \beta^{\mathrm{T}}\mathbf{Ay} + \alpha^{\mathrm{T}}(1 - \mathbf{y}).$$

Interestingly, a cost vector, **c**, defined by

$$-\mathbf{c} = \alpha^{\mathrm{T}}\mathbf{1} + (\beta^{\mathrm{T}}\mathbf{A} - \alpha)^{\mathrm{T}}\mathbf{x}$$

(the first term is a constant and can be ignored) can be used in the objective function for linear least-cost set-covering. If so, it will

find solutions that tend to have few overlapping subsets, i.e., it will be close to a partition.

(3) The remaining problem is to define Pr(**A**) so that different families of subsets may be compared in terms of the solutions they give. The two sets of probabilities, **p** and **q**, are equivalent to **A** in the sense that together they summarize the essential marginal information in **A** relevant to set covering. Thus any definition must use both. The arguments used for defining Pr(**A**|**x**) above show that **p** should be used in the form $\{1 - p_k\}$, since before a solution is found $\mathbf{x} = \mathbf{0}$, which implies

$$(1 - p_k)^{(1-x_k)} = (1 - p_k),$$

it follows that

$$\Pr(\mathbf{A}) = \Pi_i q_i \Pi_k (1 - p_k).$$

It further follows that

$$\Pr(\mathbf{x}|\mathbf{A}) = [\Pi_i q_i^{z_i} \Pi_k (1 - p_k)^{(1-x_k)})] \; / \; \Pi_i q_i \Pi_k (1 - p_k)$$

$$= \Pi_i q_i^{(z_i - 1)} \; / \; \Pi_k (1 - p_k)^{x_k},$$

which is more easily interpreted in logarithmic form as

$$\log_e(\Pr(\mathbf{x}|\mathbf{A})) = \sum (z_i - 1)\log_e q_i - \sum x_k \log_e (1 - p_k).$$

Similarly, $\log_e(\Pr(\mathbf{y}|\mathbf{B}))$ can be calculated in exactly the same way, and so would appear at first to provide a decision criterion. This conclusion is not yet true, because it is easy to see that if the number of subsets in **A** is less than that in **B**, **x** tends to be

preferred to y almost independently of these vectors. The solution is to standardize by n and m, i.e., define the criterion to be

$$n^{-1}\sum(z_i - 1)\log_e q_i - m^{-1}\sum x_k \log_e(1 - p_k).$$

This decision criterion is analogous to Akaike's information criterion, in which the principle is that a model is preferred if the price paid is minimal in terms of the number of parameters when balanced against the goodness of fit (Akaike 1973).

Consensus among coverings

The reader may be surprised that in this book I do not indicate a "best" method even within the framework of conditional clustering. I do not even claim that the conditional component of subset formation, which forms the basis of much of chapters V–X, is superior to the unconditional principle used in most other methods of clustering. I do not believe there to be a best method because "any given set of data may admit of many different but meaningful classifications" (Anderberg 1973). The same author also commented that "cluster analysis is a device for suggesting hypotheses" and that "a set of clusters is not itself a finished result but only a possible outline." He also emphasized that "cluster analysis methods involve a mixture of *imposing* a structure on the data and revealing that structure which actually exists" [italics added], so that the use of different clustering procedures, each perhaps imposing themselves differently on the data, may in total reveal the "true" structure. It is for this reason that the elucidation of a consensus among classifications is so important; one can hope that the consensus will tend to minimize the influence of the various impositions. Certainly the results of a single clustering rarely satisfy the cautious practitioner. To quote another distinguished author:

> We cannot expect to have as a result of a couple of hours computer time something which will straightforwardly replace the insight and effort of the scholar (which will have been needed anyway in the selection of properties, and will be needed in the interpretation of the results). What we can hope is that groups will be thrown up for consideration which have not previously been noticed, and which may be useful; that groups well accepted in the profession will perhaps not turn up - because they are obsolete and not supported by the data. (Needham 1965)

These opinions, as well as those held by the writer, make clear the need for a variety of clustering procedures; perhaps the best sources for many in English (and Fortran!) are to be found in the books by Anderberg (1973) and Hartigan (1975).

It is not uncommon that different clusterings are obtained for the same set of objects either using different sets of attributes (an example having important classification consequences for the organisms—mosses—is given by Rohrer 1988), or from different clustering methods. Two problems need to be solved to obtain a consensus among two or more clusterings; first, how to represent the clusterings; and, second, dependent on this representation, how to define a consensus from them, which can be recognized as being that of the clusterings.

For simplicity, consider the consensus between two clusterings. An obvious way to represent a clustering is by the columns of **A** corresponding with the unities in the solution **x**, i.e., deleting the columns corresponding with $x_k = 0$; suppose such an array is called **X**. Similarly, for a solution **y** for the same n objects in the same order, but based on **B** (which could also be **A**), form **Y** in the same way. Several observations are immediate; first, perfect agreement exists between **X** and **Y** if a permutation of the columns of **Y** makes it identical with **X**; second, partial agreement

exists if some columns of **Y** are identical with some of **X**; third, imperfect agreement exists if the union of some (not all) columns of **Y** equals the union of some (not all) columns of **X**. Otherwise, there is disagreement.

Without loss of generality, after deleting those columns of **X** and **Y** that are identical and do not intersect with other columns, put on one side the parts of the solutions that completely agree but intersect with other subsets. In the remaining part of each of the two arrays, the rows of **X** and **Y** corresponding with the objects in these subsets become empty; these are deleted. Here assume that n refers to the number of *remaining* objects, with m_x and m_y the number of remaining columns; **X** and **Y** refer to these reduced arrays.

A direct procedure for forming the consensus of two or more subsets is based on arguments similar to those for forming musters. Two possibilities are now considered:

(1) Form the union of all intersecting subsets no matter from which solution they are derived.

The justification for this action is that if one data set gives one assignment and another gives a different one but with some individuals common to both, and because there is only one(?) "true" evolutionary classification, the inconsistency is due to the particular subsets of attributes used in obtaining **X** and **Y**. The best decision, therefore, is to join such subsets. Such a subset subsequently can be studied alone.

(2) For two intersecting subsets, form three groups, namely, those in common and those belonging to each alone.

This action, however, tends to produce a multiplicity of single-element subsets when there are many solutions and so is not recommended.

Two somewhat more elaborate procedures are now discussed. The first remains within the framework of the subsets, while the second embeds these into a Euclidean space, and then does things in that space before reversing the process.

For the first of these proposals, **X** and **Y** refer to different subsets. Two cases are distinguished depending upon whether **X** and **Y** are obtained from a common **A** or from different ones.

Common A Let **Z** be formed from **X** and **Y** by adjoining, i.e.,

$$\mathbf{Z} = [\mathbf{X} : \mathbf{Y}].$$

It is apparent that each object is included at least twice in **Z**, and so there may be some redundancy. Assuming that now a **z** is sought so that

$$\mathbf{Z}\mathbf{z} \geq \mathbf{1},$$

there is no difficulty in treating **Z** as if it were an **A**, obtaining a new set of probabilities and a covering solution based on it. This solution can be called a consensus. Because it is easy to adjoin as many solutions as are available, this procedure is easy to implement.

Different A For most purposes, there is every reason to use the same procedures as those for a common **A**. However, **Z** can also be formed from the original subsets, say **A** and **B**, i.e.,

$$\mathbf{Z} = [\mathbf{A} : \mathbf{B}],$$

and the process completed by obtaining the probabilities based on **Z** (after removing duplicate subsets). Again, it is easy to adjoin as many families of subsets as are available.

The embedding procedure for obtaining a consensus is as follows: **X** (resp. **Y**) is embedded in a m_x (resp. m_y)-dimensional Euclidean space, and then **X** (resp. **Y**) is re-expressed by referring the arrays to a common basis, which is $\mathbf{I}_n$. In fact, this re-expression leads to the polar decomposition of **X** (resp. **Y**), which is obtained by an orthogonal rotation into two parts, namely, a symmetric and an orthogonal matrix. The symmetric part of each is standardized to unit Euclidean norm and then averaged (i.e., in Euclidean space, including those of any other solution). This average, which represents the consensus in Euclidean space, then needs to be converted back to a grouping. This inverse process is performed by converting the symmetric consensus into a dissimilarity array, and performing a conditional clustering (Chapter VIII). Algebraically, the steps are as follows: replace **X** by $\mathbf{X}/\|\mathbf{X}\|$, using a Euclidean norm, and let a singular decomposition of **X** be

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}},$$

where $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{VV}^{\mathrm{T}} = \mathbf{I}$, **S** diagonal, non-negative. The polar form of **X** is given by

$$\mathbf{W} = \mathbf{USU}^{\mathrm{T}} = \mathbf{XVU}^{\mathrm{T}} = \mathbf{XH}.$$

Note that for fixed rows, **W** is unique for any matrix **X**, unlike the singular decomposition, which is unique only up to permutations of the columns of **S** (and hence the columns of **U** and rows of **V**). Let $\mathbf{W}_i$ denote the i^{th} such matrix; then the consensus is given by

$$\mathbf{C} = k^{-1}\sum_{i=1\ldots k}\mathbf{W}_i.$$

If a singular decomposition of **C** is $\mathbf{PDP}^{\mathrm{T}}$, a set of principal coordinates is given by **PD**. The Euclidean distance between the

rows of **PD** give the dissimilarities to be used in a conditional clustering (Chapter VIII), so yielding the consensus.

The weakness of using an embedding in a Euclidean space, followed by taking advantage of the continuity of that space, is apparent but has many precedents. For example, even if it is known that the solutions to a set of equations must be integers, an embedding into the real (or complex) space is often adopted, and the solutions considered acceptable if they happen to be integers; if not, they are converted to integers at the end. It is hoped that some better method may be identified involving fewer assumptions. An example of consensus formation is given in Chapter X, "Cabbages."

Other mathematical programs for clustering

M.M. Rao (1971) considered a more restricted class of mathematical programs for use in clustering than those described here, by adding the further constraint that the number of groups, G, is specified. This program he described by

$$\min\{f(\mathbf{c},\mathbf{x}) \mid \mathbf{Bx} = \mathbf{b}, x_s \in \{0,1\}\},$$

where the first n rows of **B** are identical to the matrix **A** and the $(n+1)^{st}$ (last) row consists of unities. The first n elements of the vector **b** also consist of unities, but the $(n+1)^{st}$ element is set to G. The solutions are constrained to be partitions. A more flexible formulation is to remove the last row from Rao's matrix **B** and

$$\begin{gathered}\text{minimize } f(\mathbf{c},\mathbf{x}) \\ \text{subject to } \mathbf{Ax} = \mathbf{1} \\ \text{and } \mathbf{t}^T\mathbf{x} = G \text{ for precisely } G \text{ groups} \\ \text{or } \mathbf{t}^T\mathbf{x} \geq G \text{ for at least } G \text{ groups} \\ \text{or } \mathbf{t}^T\mathbf{x} \leq G \text{ for no more than } G \text{ groups} \\ \text{and } x_s \in \{0,1\},\end{gathered}$$

where $\mathbf{t}$ is a column vector of m unities corresponding to the last row of Rao's matrix $\mathbf{B}$. For the strict equality, namely, $\mathbf{t}^T\mathbf{x} = G$, Fréhel (1975) showed that for a linear cost function, there is a smallest scalar, v, which can be found by trial and error, and which will

$$\text{minimize } (\mathbf{c} + v\mathbf{t})^T\mathbf{x}$$

subject to the usual constraints. If v is very large, so that $c + vt \approx vt$, the solution is a minimum partition or covering. If G is less than the number of columns in this solution, no feasible solution exists. By extension of Fréhel's technique, the inequality conditions can also be satisfied. White and Gillenson (1975) provided a graph-covering algorithm, which can be modified for the present purposes and which finds v automatically.

Mulvey and Crowder (1979) described a special case of the uncapacitated facility location and m-median problems as mathematical programs and showed how it may be solved by use of Lagrange multipliers.

Other objective functions

This section draws attention to several other objective functions but does not describe them in detail. The costs that M.M. Rao (1971) proposed for each subset generated by various methods (Chapter VIII) all involve some measure of within-subset heterogeneity; in consequence, single-object subsets cannot be considered. Rao also showed that several such measures cannot be expressed as a linear function $\mathbf{c}^T\mathbf{x}$, and that, for others that can be, sometimes only suboptimal solutions are found. The latter deficiency is serious, but only if the subsets in the solution are identified unreservedly with the underlying groups. For practical reasons, it is sensible to choose as a measure of separation one that not only is simple to compute but that allows a linear formulation. An extension of

W.D. Fisher's (1958) principles along the lines that Rao also considered was given by Gordon and Henderson (1977).

This book is based essentially on the concept of the probability of a subset belonging to a covering, and so it is inappropriate to review concepts of heterogeneity or homogeneity. Nevertheless, I want to draw attention to the existence of the generalization of heterogeneity, called a **scattering measure** by Emptoz and Fages (1980), and which they claim encompasses many others, including that which Watanabe (1969) called cohesion. Although it has been argued that either information or joint probability should provide the objective function, some less obvious possibilities should be discussed.

Kruskal's measure of homogeneity

One interesting possibility for the assessment of homogeneity was proposed by Kruskal (1972). His argument proceeds approximately as follows. Consider a set of points clustered tightly around a twisted curve; choose any point, and construct a sphere around it of radius r big enough to contain the scatter around the curve, but small enough so that the curve does not bend much, so that locally the curve is nearly equivalent to a straight line. In these circumstances, the number of points is proportional to r, and the *internal dimension* is unity. If the points are diffusely distributed in d-dimensional space, the number of points in a sphere of radius r is proportional to r^d, and the internal dimensionality is d. If the points are clustered tightly around a few centres, the number is independent of r, and the internal dimensionality is zero. Kruskal proposed the coefficient of variation of the interpoint distances as the basis for the estimate of internal dimensionality. If $\mathbf{X}$ is a set of coordinates for a set of points, and s_i is the i^{th} largest singular value, $i = 1 \ldots r$, of $\mathbf{X}$, the *shape of a set of points* is defined as follows. Let

$$\sigma_i = (s_i/\Sigma s_i)^2, \text{ and } g_i = \sigma_i/(1 + \sigma_i);$$

then the **global dimension of the shape** is

$$(\Sigma g_i)^2/\Sigma g_i^2.$$

These interesting concepts appear not to have been applied in clustering, because their roles, other than describing the properties of a solution, are not clear. However, a low internal dimensionality clearly implies homogeneity and so is of general application (the distance need not be Euclidean), while the global dimension of shape certainly requires Euclidean conditions for the rotations in obtaining the singular decomposition, if a meaningful interpretation is to be made. Another measure of the shape of a finite set of points is proposed in Chapter IX.

Separation

Because within-subset homogeneity is obtained if all objects form subsets of unit cardinality, the concept of separation, the second of the two desirable attributes of a cluster (Cormack 1971) has been proposed. Separation is now interpreted to imply that the distance between a subset and its nonmembers should be as large as possible. A simple and natural measure of separation is the smallest distance between a member and nonmember; denote this distance for the s^{th} subset by z_s. This measure of separation ignores some relevant factors, such as the size, shape, and number of objects in the subset, as well as the position of the nearest neighbor in relation to the shape characteristics. To understand something about this, I performed an experiment (Lefkovitch 1978, appendix) to obtain an empirical function, to measure separation by using an "artificial intelligence" approach. I formed into a measure of separation a mathematical expression summarizing human responses to specified questions. Because shape perception, generalization, and inference depend on an individual's previous experience (Luria 1976), I interviewed separately 16 subjects,

including taxonomists (who are professionally sensitive to differences), statisticians (who tend not to reject "outliers" unless very different), and others uninvolved in these matters. I asked each to give their opinions of the membership of a specified point on a plane to a group of others arranged in a loose ellipse. These opinions were investigated in terms of the distance along the directional lines at which it appeared that the scorer was neutral about membership or nonmembership of an object to a planar set, where y is the break-even distance (i.e., the neutral point), $|s|$ denotes the number of objects in the ellipse, ϵ the eccentricity of the underlying ellipse, a and b its major and minor semi-axes, and Θ the angle between the major axis and the line from the nearest neighbor of the candidate point to the centre of the ellipse. After a series of linear regression analyses, an empirical relationship was defined as

$$\begin{aligned} \text{constant} &= y\,[1 + (\log_e|s| + \sin \epsilon\Theta)/e^{(a+b)/2}] \\ &= y\,(1 + x) \end{aligned}$$

in which the constant appears to depend on the prior experience of the individual. The term $\log_e|s|/e^{(a+b)/2}$ is an empirical measure of density that is zero only for null subsets; the term $(\sin \epsilon\Theta)/e^{(a+b)/2}$ is a shape-angle-size characteristic. This expression is such that the more objects present, or the greater the eccentricity, or the greater the angle with the major axis, or any combination of these, the more effectively is the candidate considered to be isolated from the group. (For further details of this experiment, see Lefkovitch [1978].)

A quadratic measure of separation

Another measure of separation can be based on a $m \times m$ matrix $\mathbf{T}$ of distances among the m subsets; choosing subsets to maximize

some function of these (e.g., a norm) also focuses on separation. The problem, therefore, is that of quadratic set covering, namely,

$$\text{maximize } \{\mathbf{x}^T\mathbf{T}\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{1}, \mathbf{x} \in \{0,1\}^m\},$$

where $\mathbf{T}$ is a non-negative, positive semidefinite matrix. If $\mathbf{t}$ is defined as a column vector of m elements all equal to unity since the elements of $\mathbf{T}$ are all non-negative, it follows that maximizing $\mathbf{t}^T\mathbf{T}\mathbf{x}$ is equivalent to maximizing $\mathbf{x}^T\mathbf{T}\mathbf{x}$; because $\mathbf{t}^T\mathbf{T}$ is a constant vector, this quadratic problem is equivalent to a linear one in which each subset is characterized by some aggregate of the measure of separation that it has with the remaining subsets. Computing $\mathbf{T}$ involves $O(n^4/4)$ arithmetic operations, which may not be practical even for the modest n.

If the aggregated measure of separation over all subsets, indicated by the elements of the vector $\mathbf{x}$ of this chapter, is incorporated into the function $f(\mathbf{x},\mathbf{z})$, then choosing an $\mathbf{x}$ so that this function is maximized will result in an optimal solution. For example, maximizing $\mathbf{z}^T\mathbf{x}$ for the nearest-neighbor measure of separation results in an optimal linear solution.

Multiple objective functions

While there is no obvious way to choose a "best" measure of separation, it may be possible that using more than one simultaneously may be preferable. Let there be g measures of separation, of which z_t is the t^{th} $(t = 1...g)$; the combined set-covering problem becomes the search for efficient solutions, i.e., the Pareto problem (described later in this chapter), for which a simulated annealing algorithm is given.

Bitran (1977) described a different approach, in which each set of the $g - 1$ separation vectors generates a further set of constraints for the remaining one. Hansen and Delattre (1978) described a partitioning algorithm that seeks to satisfy two criteria

simultaneously, namely, to maximize homogeneity and to maximize separation, which they referred to as a bottleneck type of problem.

Because these measures of separation are monotonically increasing functions of the distance between the subsets, and because the solution maximizes some combination of these measures, the number of subsets to be considered can be reduced further by deleting any column of **A** if the separation of the subset, however defined, is less than some specified value, unless this action results in row i becoming a zero vector. By this heuristic, large problems can be further reduced to manageable proportions with only minor impact on the probability that the optimal solution is contained within the reduced number of subsets.

Algorithms for the solution of set-covering problems

This section describes three algorithms for the solution of linear least-cost set-covering problems; such problems are **L**inear **O**bjective **CO**mbinatorial, called LOCO by Edmonds (1971). It is important to distinguish heuristic from approximate solutions. An approximate solution is defined as being one that guarantees obtaining a solution bounded by some known ratio to the optimal solution. A heuristic solution procedure consists of rules that do *not* guarantee that the solution is correct. For a heuristic solution, the expected time, effort, and computational complexity are appreciably less than those either of an approximation or of an exact solution, and in fact their use may be the only way to obtain any solution. The failures and errors using heuristics are not random but systematic, and, once it is understood how the heuristic works, classes of problems can be constructed that produce the wrong solution or fail to produce a solution at all.

The first algorithm is a version of the greedy algorithm; it gives a solution never more than $2\log_e n$ times the true optimum (Johnson 1974) and can simultaneously yield a lower bound

(Hey 1981). It consists of a set of plausible and computationally effective rules that obtain a solution satisfying the constraints, and for which numerical experience has shown that, with either $\{-\log_e p_k\}$ or $\{-p_k \log_e p_k\}$ as costs, the optimal solution has always been achieved. Even if not, the value of the objective function at termination provides a good upper bound for other methods, and so this algorithm is always worth using. For example, any single subset whose cost exceeds the total cost of the approximate solution can be eliminated, perhaps permitting the reduction rules of Algorithm II.1 to be used again. Alternatively, if the reduced array has only a few columns, little computational effort is needed to examine each irredundant covering.

ALGORITHM II.4. *Approximate solution for the linear least-cost set-covering problem*:

input: A (preferably fully reduced), c, and q;

step 1: for each remaining subset, calculate

$$h_k = c_k/(\sum q_i \mid i \in k)$$

where the summation is over all uncovered objects in the subset;

step 2: find that subset for which h_k is a minimum, say k_{min};

step 3: set $x_{k_{min}} = 1$, delete from all subsets all objects newly covered by k_{min};

step 4: if objects are still uncovered, go to step 1; and

step 5: calculate $\mathbf{c}^T\mathbf{x}$, the value of the objective function.

This algorithm is modified from that of Chvátal (1979) by incorporating $\mathbf{q}$, the relative importance of the objects; it chooses

the next subset to be included as that for which the most remaining objects are included for the least cost per additional object; each object is weighted by its relative importance. Chvátal's original proposal weights the objects not by **q** but equally; others have proposed using the number of objects in the subset, or its logarithm. Other weighting schemes may be more appropriate, for example, $\{-q_i \log_e q_i\}$, but the use of **q** seems to be the best of several tried. For independently provided costs (discussed in Chapter VIII), it seems that Chvátal's original proposal may be best if the costs are linearly uncorrelated with $\{-\log_e p_k\}$. Because of the numerical experience and the close relationship of **p** and **q** with **A**:

> CONJECTURE II.1 *(a) If the costs are $\{-\log_e p_k\}$, Algorithm II.4 always obtains a minimal covering;*
> *(b) if the costs are either $\{-\log_e p_k\}$ or $\{-p_k \log_e p_k\}$, Algorithm II.4 obtains a least-cost covering.*

A related heuristic solution procedure, described by Fishburn and Gehrlein (1988), is a modification apparently appropriate for large number of objects or subsets, or both, but has not been studied in the present context. It differs by including more than one subset at each step; with an informed choice of which of the remaining to include, it can be seen that the execution time of the procedure can be reduced even further than adoption of some heuristics. Vasko and Wilson (1986) discussed other hybrid heuristics for set-covering problems.

If the array **A** is totally unimodular (i.e., every square nonsingular submatrix of **A** has determinant equal to either +1 or −1), then the optimal solution to the linear relaxation of the program is also the solution to the set-covering problem (Padberg 1975). However, it is sometimes difficult to show that any given

A is totally unimodular, but the possibility of taking advantage of the linear relaxation remains. An exact solution for the linear least-cost set-covering problem, described by Garfinkel and Nemhauser (1972), is based on the least-cost linear programming relaxation of the problem, namely,

$$\min\{\mathbf{c}^T\mathbf{z} \mid \mathbf{Az} \geq \mathbf{b}, 0 \leq z_k \leq 1, k = 1...m\},$$

which is most conveniently solved in the dual form, to obtain a lower bound, from

$$\max\{\mathbf{b}^T\mathbf{w} \mid \mathbf{A}^T\mathbf{w} \leq \mathbf{c}, w_i \geq 0, i = 1...n\}.$$

For $\mathbf{b} = \mathbf{1}$ and z_k not confined to being either zero or unity, $\mathbf{c}^T\mathbf{z}$ at the optimum can be less than $\mathbf{c}^T\mathbf{x}$, the optimal set-covering solution, and so provides a lower bound. If this lower bound is equal to the upper bound given by Algorithm II.4, the (or an) optimal solution has been obtained. If not, it becomes possible to determine some of the elements of the unknown binary **x**, and so to reduce the problem further. Thus by alternating the approximate solution procedure and the linear relaxation formulation, a solution can often be found quite quickly, since a reduced **A** may in fact prove to be totally unimodular. The solution procedure requires two components additionally to Algorithm II.4, namely, how to solve the linear program, and how to obtain further elements of **x** from **z**. Because procedures are easily available for solving linear programs (for example, Fortran code is given by Kuenzi et al. 1971), I do not describe this aspect here. Consider the vector **z**; any element equal to zero implies that the corresponding element of **x** is zero, so that the corresponding column of **A** can be deleted. Any z_k equal to unity implies that x_k is unity; objects covered by these subsets are then deleted, so that these subsets become empty and hence deleted. Each noninteger element of the solution vector

z generates two new (smaller) sets of constraints, each of which generates a new linear program to be solved and followed in turn, until either the branch terminates, or it becomes apparent that any further solutions on that branch of the search tree are necessarily worse than the best current upper bound. After any further consequent reductions (usually none are possible), a new **A** remains, which together with the remaining elements of **c** are used in a further relaxation to a linear program. This process is repeated until all objects are covered, or equivalently, until **A** becomes null. To summarize:

> ALGORITHM II.5. *Exact solution of the linear least-cost set-covering problem:*
>
> *step 1: perform Algorithm II.4 and record the (new) upper bound;*
>
> *step 2: solve the corresponding linear relaxation program;*
>
> *step 3: if the objective function equals the upper bound, terminate and report the solution of step 1 as optimal; otherwise delete subsets if the indicator variable is zero, set* $x_k = 1$ *if* $z_k = 1$, *delete the covered objects and null subsets from A; and*
>
> *step 4: generate the new sets of constraints, and for each, go to step 2.*

Algorithm II.4 for an approximate solution gives an upper bound for the optimal solution, which is never more than $2\log_e n$ times the true optimum (Johnson 1974) and can simultaneously yield a lower bound (Hey 1981). The upper bound can be used to eliminate single subsets of greater cost from further consideration. Garfinkel and Nemhauser (1972) also gave an additional cost-based reduction

that is sometimes very effective in reducing the size of the problem, because it may be combined with those described in the previous section. A review of further heuristics for this problem was given by Hey (1981) and by Moret and Shapiro (1985); a general strategy for the solution of set-covering problems was given by Hey (1981). Procedures for assessing how well an approximate solution may represent the data were described by Zemel (1981).

The next algorithm is essentially very simple. It is a version of the simulated annealing algorithm (Aarts and van Laarhoven [1987] provided a useful description of the general procedure), but it may also be regarded as a stochastic programming solution. For single linear least-cost set-covering problems, it is less efficient than the heuristic method, but it appears to have no real competitors either for general nonlinear objective functions (although for special cases, e.g., fractional objective functions, there are effective algorithms) or for multiple objectives for which Pareto (efficient, undominated) solutions are required. Here it is described for a single objective function, $f(\mathbf{c},\mathbf{x})$, which is not necessarily linear, which is to be minimized (maximization problems are easily converted). The general strategy of this procedure is as follows: given the current candidate solution, $\mathbf{x}_b$, generate at random a new feasible candidate, $\mathbf{x}_t$, in the neighborhood of $\mathbf{x}_b$; if the new candidate is better than the so-far globally best $\mathbf{x}_g$, retain it as the global best; else replace $\mathbf{x}_b$ with $\mathbf{x}_t$ with probability which depends on

$$\exp[f(\mathbf{c},\mathbf{x}_b) - f(\mathbf{c},\mathbf{x}_t)]$$

(note that the term in square parentheses is negative); after k no further improvements, generate a new feasible solution, $\mathbf{x}_t$, *not* constrained to be in the neighborhood of either $\mathbf{x}_b$ or $\mathbf{x}_g$, and repeat the procedure k times, where k is a preset number.

ALGORITHM II.6. *Determining a random feasible solution:*

step 0: initially consider all subsets;

step 1: choose a subset with uniform probability from those remaining and include it in the cover if it is irredundant; if it is redundant, ignore it; and

step 2: if all objects not covered, go to step 1.

ALGORITHM II.7. *Determine a feasible solution in the neighborhood of the current best:*

step 0: let x_b *be the current solution; define a vector* $w = 2 + x_b$;

step 1: choose subset k uniformly at random from the subsets remaining (initially all);

step 2: choose a uniform random number, u; if $(u \leq 1/w_k$ *and* $x_{kb} = 1)$ *or* $(u > 1/w_k$ *and* $x_{kb} = 0)$ *and the* k^{th} *subset is irredundant, then set* $x_{kt} = 1$;

step 3: if objects are still uncovered, go to step 1; and

step 4: if $f(c,x_t) < f(c,x_b)$, *w is replaced by* $w + x_t$.

This algorithm can be regarded as being an approximate method for determining **p**, since as the number of trials increases, the vector $\{w_k/\Sigma w_k\}$ approaches **p**. In fact, step 2 can be replaced by the following:

step 2': choose a uniform random number, u; if ($u \leq mp_k$ and $x_{kb} = 1$) or ($u > mp_k$ and $x_{kb} = 0$) and the k^{th} subset is irredundant, set $x_{kt} = 1$.

Clearly, this focuses on those subsets having high probability of being members, and which are not currently in the best solution.

The remaining component, namely, to determine when to accept a poorer solution than the current best, is in fact a key element of this algorithm. A function having the following properties is desirable:

- as more and more random starts have been adopted, it should be more "difficult" to accept a poorer solution
- it should be proportional to how much poorer is the current solution.

Let v count the number of random starts; accept a worse solution if a random number, u, is such that

$$u < \exp[\ (f(\mathbf{c},\mathbf{x}_t) - f(\mathbf{c},\mathbf{x}_i))/v\ ].$$

Liepins et al. (1987) described a related solution procedure, based on the concepts of genetic algorithms, which "mutates" the current best solution rather than being completely random. Another genetic algorithm (Goldberg 1989) is a procedure used to solve optimization problems by generating candidate solutions and using concepts borrowed from genetics; it is based on (a tautology) "survival of the fittest" in simulated evolution. Candidate solutions are represented by "organisms" having a single "chromosome" whose loci correspond with the parameters; in combinatorial problems, a chromosome is often a bit string. The "fitness" of each organism is determined by its chromosome, from which the fitness value is calculated, and from which the probability of

reproduction and survival into the next generation is obtained. Usually, the number of organisms per generation remains fixed. A set of "genetic" operators is applied to the chromosomes:

- mutation, defined as changing some of the parameter values (e.g., in binary problems, replacing a bit by its complement)

- recombination, defined as single and multiple crossing-over (exchanging a substring in one chromosome with the corresponding region of another)

- inversion (reversing a part of a chromosome).

The effects of these operators is either

(1) that offspring are generated whose fitness is not greater than that of their parents, in which case the offspring "die" and the parents survive into the next generation; or

(2) that offspring are produced having a fitness greater than that of its parent(s), in which case the parents "die," and the offspring replace them in the next generation and reproduce with increased probability.

Because the number of organisms is fixed, the effect is to replace inferior solutions with ever-increasing improvements. This algorithm can be seen to be a form of multiple hill-climbing (in the maximization context) and clearly lends itself to parallel processing.

In application to the set-covering maximization problem, the following observations can be made:

- there is no reason to distinguish an organism from its chromosome

- each chromosome contains m loci, where m is the number of subsets in **A**, and consists of a bit string

- for n objects, it is unlikely that a solution containing more than $k < n/2$ subsets is of any interest, and so it can be assumed that no more than k organisms need be considered

- fitness is determined as

$$g_j = \begin{cases} f(\mathbf{c},1), & \mathbf{Ax} \not\geq 1 \\ f(\mathbf{c},\mathbf{x}) & \text{otherwise} \end{cases}$$

where **x** is a candidate solution, and $f(.,.)$ is the function to be maximized

- the probability of reproduction for an organism is defined either as

$$h = (f(\mathbf{c},1) - f(\mathbf{c},\mathbf{x}))/(f(\mathbf{c},1)$$

or as

$$h = \exp(f(\mathbf{c},\mathbf{x}) - f(\mathbf{c},1)).$$

The proposed procedure is as follows:

ALGORITHM II.6. *A genetic algorithm for solving set-covering problems:*

initialization: assign k; generate at random k binary m-element vectors, $\mathbf{x}_j$, $j = 1...k$, *which satisfy* $\mathbf{Ax} \geq \mathbf{1}$, *and calculate their fitness and reproduction probability;*

repeat until termination condition satisfied:
choose one of {mutation, recombination, inversion} at random;
(a) mutation: choose parent j at random, and with probability 1 – h_j, randomly replace each bit in x_j by its complement, and determine the fitness of the new organism; if fitter, then replace x_j by the changed organism;
(b) recombination: choose parents j and j' with probability h_j and $h_{j'}$, respectively; within the chromosome of each, choose at random the same position and exchange the left segments; calculate the fitness of the two new arrangements, and replace the two parents by the fittest pair of the four (which may be the parents); and
(c) inversion: choose parent j at random, and a position at random within the chromosome; with probability h_j, reverse the left half, and determine the fitness, and then the right half; if either is fitter than the parent, it replaces the parent.

Note that the fittest organisms tend to reproduce in each generation. The process may be terminated if one of two different conditions is satisfied:

- either that a prescribed number of generations passes in which the maximum fitness does not change

- or that the majority of the organisms become alike.

For single linear, objective functions, this algorithm is not competitive with the heuristic and exact method; but for multiple linear, objective functions, preliminary trials suggest that it is no worse and perhaps better that the simulated annealing procedure, with which it has some similarity. For nonlinear objective functions and also for additional constraints, such as specifying the

number of unit elements in the optimal $\mathbf{x}$, it seems to have some merit.

With the exception of how to modify the annealing algorithm to solve multi-objective problems, I have now described the main general algorithms required. In other chapters I introduce further specialized algorithms as required.

Claus (1973) proposed a more general method for solving set-covering problems; the program was written as:

$$\text{maximize } \{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{b}, x_i \in \{0,1\}\},$$

where $\mathbf{A}$ and $\mathbf{b}$ are arrays whose elements need not be restricted to zeros and unities. Claus proved two theorems for reducing the number of rows in $\mathbf{A}$; these theorems are used to find the hypersphere consisting of the intersection of the n constraints; suboptimal solutions are obtained during the process until the best is found. The Claus algorithm, a predecessor of the ellipsoid algorithm for linear programming, is more general than that of Garfinkel and Nemhauser (1972), since both $\mathbf{A}$ and $\mathbf{b}$ are general arrays, but does not seem to be as efficient as that of Fiala (1973). Babaev (1978) described a modification of the Claus approach. Etcheberry (1977) described an enumeration solution procedure together with some computational experience.

III Numerical representation of attributes

In taxonomic studies, the attributes chosen to represent the objects clearly should avoid those known to show considerable variation associated with environmental differences. This possible association is most obvious in plant species able to live in a wide variety of conditions and exhibit form that reflects this ability; it can also occur in animals. Those investigating genotype–environment interaction face this problem regularly. To avoid creating classifications no better than that of the environments, one experimental procedure is by reciprocal transplantation (if possible raising the offspring of all individuals in all-natural environments), or, failing that possibility, raising all candidate individuals in a common environment (preferably several such), so that, after considering carry-over effects, differences within an environment can be assigned to genetic causes. Different environments may elicit different aspects of the genome as well as environmentally different responses. Because this opportunity tends to be available only rarely (although it is common in agronomy), investigators must use their judgement, perhaps guided by previous studies in the group, to select attributes least subject to environmental influences. Both for plants and animals, such attributes are usually associated with the primary reproductive organs—flowers, genitalia, birth mechanisms, and so on—because any major sensitivity of these organs to environmental influences is likely to reduce reproductive success. Perhaps vegetative reproduction and parthenogenesis further reflect this sensitivity, and so should be considered as aspects of the genotype relevant to revealing the groups that exist.

The subject matter of this chapter is relatively familiar, and is therefore kept brief. As already noted in Chapter I, the empirically defined and observed attributes need to be converted

into numerical forms appropriate for the current purpose. The objective now is to indicate how this conversion may be done so that the numerical values represent the objects adequately *not* for their description, but rather on how best they may be compared. The comparison of two objects almost always implies the existence of a third, in particular,

> for any three distinct objects, $\{i, j, k\}$, decide if objects i and j are more alike with respect to the attribute than each is to object k.

Not all requirements for the numerical representation of attributes are described in this chapter, because some further conditions need to be satisfied for the scalar dissimilarity coefficients described in Chapter VII. Nevertheless, the theme behind the variations, each of which corresponds with the numerical representation of a particular class of attribute, is that an appropriate measure of resemblance between two objects can be computed as the inner product between two vectors.

Careful consideration of *what* is being measured is always necessary. Some attributes are lengths, areas, volumes, weights, and so on, all of which are components of size, so that if these are to be combined in some way, care has to be given to the units of measurement. If shapes are to be described, a set of lengths and angles capturing differences is to be preferred rather than a single number to represent the whole, because their correlation is often as informative as the separate values. For variables such as angles, if averages are to be calculated, transformations to cosines are almost standard. Numerous arguments exist against the use of ratios, including the loss of information when two variables are replaced by one, the unpredictable nature of the variance of the ratio, and the spurious correlations with other ratios which can be generated. It is important, however, to avoid using what can be called "over-descriptions;" for example, a structure approximately

triangular can be represented either by two sides and the included angle, or by one side and the angles at its ends, or by three sides; introducing extra angles or sides does not define the shape more closely. Similar minimal sets can be found for approximate quadrilaterals, circles, ellipses, and other simple geometric shapes, and the principles involved extended to represent irregularly shaped structures. A decomposition of these irregular shapes into simple components sometimes gains sufficient resolution. Advantage can also be taken of the use of the coefficients of a Fourier decomposition of a closed shape, especially if there are no reliable "landmarks" on the objects.

Not all measured attributes need be used for clustering. For example, if an attribute shows the same state for all objects, it conveys no information on the existence of subsets within them; it conveys information that it makes sense to consider the objects together. If the object by attribute array is reducible, nothing is lost by separating the objects into irreducible blocks and processing each alone. These decisions are both obvious and innocuous, which may ease the computational load considerably as well as increase resolution. A further simplification, which requires some thought and which cannot be applied routinely, is to retain only one of each attribute state that is duplicated over all objects by another. For example, if a state of one attribute is perfectly correlated with a state of another, should both be retained? The principal argument for retention is that a classification receives support by the concordance of many attribute states across the objects. Among several arguments for deletion are the following:

- many gene complexes have pleiotropic effects

- attributes may vary together because of environmental influences

- the logical dependence among attributes is often poorly understood (a well-understood case is surface area and weight; many other such dependencies must be as yet unrecognized).

A solution to this problem is open; perhaps the best strategy is to proceed both with and without such deletions, to try to understand any differences among the classifications that may emerge, and to use the classifications arising as competing hypotheses.

As is apparent, each attribute needs to be considered carefully, because no automatic procedure is advocated. Furthermore, because an understanding of the attributes is based partly on the objects to be studied (as well as on the accumulated knowledge, experience, and intuition of the observer), the circumstance represents an example of a hermeneutical circle; the attributes are understandable by virtue of the classification of their owners, and the classification of the owners is understandable (or possible) only by the properties of the attributes. The key concept to which these remarks are leading is that of attribute homology; without knowing the phylogeny, it is impossible to be sure of the homology. For this reason, Sneath (1983) proposed a less evocative term, **isology**, which can be interpreted as capturing some elements of the concept of "sameness."

The first step, therefore, is to describe the attribute classification system adopted here and the resulting numerical representation. The scheme used is based essentially on that of Gower (1971*a*) as simplified by me (Lefkovitch 1976, appendix). In the simplification, two types of attribute are recognized, namely, unordered and ordered, each of which is further subdivided by the number of possible states that an attribute can exhibit.

Unordered attributes

An unordered attribute shows no natural ordering (statisticians call these nominal); an object can exhibit only one of several mutually exclusive states of such an attribute. Suppose the number of states of an unordered attribute is $s > 0$; then, for this attribute, a particular object is represented by an s-element binary vector, **v**, defined by

$$v_k = \begin{cases} 1, \text{ if the object shows state } k, \\ 0 \text{ otherwise,} \end{cases}$$

for $k = 1 \ldots s$[1]. The major empirical problem is to distinguish between the situations $s = 1$ and $s = 2$. For the first, called a one-state attribute, the attribute is either *present* or *absent* (for example, a particular sex-linked structure); absence is *not* considered to be a state. To emphasize further and to anticipate later, *no information is assumed to be given* about the resemblance between two objects by this attribute *if it is absent in both*, i.e., just three of the four possibilities give information on the pairwise resemblance. An example is the absence of feathers both in mammals and annelids, which scientists would agree gives no information about their resemblance. In fact, there is an infinity of attributes which, if jointly absent, give no information on the resemblance of a pair of objects. By contrast, if joint absence is considered to give information on resemblance, the attribute is two-state; a clearer example of a two-state attribute is one in which there can be just two colors, one or other of which is *necessarily*

[1] Although values for v_k here are considered to be elements of {1,0}, in Chapter VII, {1,−1} is also considered.

present in an object. If two objects show the same state for such an attribute, they are alike with respect to it; otherwise they are unalike. All four possibilities, therefore, give information on their pairwise resemblance. Gower (1971*a*) calls one-state attributes "dichotomies," two-state attributes "alternatives," and ($s > 2$)-state attributes "multistate unordered."

Ordered attributes

Ordered attributes, which include those that are either discrete (e.g., counts) or continuous (e.g., lengths, areas, and weights), are perhaps more difficult to deal with than the unordered. It is almost always true that although there is an ordering, the direction is not an intrinsic component of the attribute. Some numerical values are larger than others, and so a direction is implied by the size of the integers or real numbers empirically observed, but this sequence may have nothing to do with any biological direction. For this reason, the degree of resemblance between a pair of objects given by ordered attributes usually depends on the absolute value of the *difference* in the numerical scores; in consequence it is necessary to ensure that a unit difference in the numerical representation of the attribute measures the same amount of dissimilarity throughout the whole range. It is easy to see that this condition is true for unordered attributes, when any difference in state can be assigned a unit dissimilarity. The key problem, especially for scalar dissimilarity coefficients (Chapter VII), is to determine a uniform scale for ordered attributes; does a unit difference (e.g., from 1 to 2, from 11 to 12, or from 101 to 102) reflect the same degree of biological difference? Probably not, but what should be the proper scale? In a sense, this question is impossible to answer, because the data under study represent a mixture of within- and between-group differences, which cannot be distinguished until a hypothesis on group membership has been suggested. This dilemma implies that several passes of the data may be needed

when ordered attributes are involved. Special problems also exist if zero is one of the possible values; for a non-negative ordered attribute, does zero signify logical absence? If so, it may be sensible to replace such an attribute by two, the one indicating presence or absence, the other, conditional on the first, being the positive value (or its transform).

Discrete ordered attributes

Almost without exception, the values empirically recorded here consist of the non-negative integers. Dealing with an easy case first; if there can be only two values, y_1 and y_2, this circumstance is equivalent to a two-state unordered attribute, and so can be represented by a two-element vector, because the direction does not really play a role. This conclusion immediately suggests a pragmatic recommendation for multistate continuous and ordered attributes; if the values fall into two disjoint subsets (i.e., there is an intermediate range of values for which there are no observations), the attribute can be replaced by a two-state unordered attribute. For more than two distinct values, the more usual situation, it is necessary to record the values exactly.

Discrete ordered attributes can be of two main types: ordinal discrete and interval discrete.

Ordinal discrete attributes

For an ordinal discrete attribute, it does not make sense to talk of the spacing or distance between the categories. An ordering does not determine a unique metric and can be isometrically embedded into an infinitude of different geometries, each of which can represent the ordering. To obtain a set of numbers that can be used for the purposes of clustering, the naïve procedure is to construct an intuitive transformation of the scores, e.g., to percentages, and to use these as if they are interval values, i.e., as real numbers. This questionable procedure is often seen in the statistical analysis

of rating data in plant science (Little 1985), but, as McCullagh (1980) and Agresti (1984) pointed out, the analysis of rating data does not need to follow that procedure. Although much of the statistical development is irrelevant for the present purposes, underlying it is a useful scenario, which I now briefly review before proposing a numerical protocol.

Assume that there exists a probability space, S, on which a metric, which is a member of the class of metrics associated with the scores under consideration, is defined. For any two elements, p and $q \in S$, let $G_{pq}(x)$ denote the probability that the distance between scores p and q is less than x, i.e.,

$$0 \leq G_{pq}(x) \leq 1, \forall x \geq 0;$$

i.e., G is an integral transform of the scores. If $x < y$ implies that $G_{pq}(x) < G_{pq}(y)$, it follows that G_{pq} is a probability function (a distance distribution function; Menger 1942). Suppose q now corresponds with one of the two extremes of the ordering, which without loss of generality can be taken to be the lower, the proposed measure of distance between p and p' is

$$d_{pp'} = |G_{pq} - G_{p'q}|.$$

To estimate the values of G corresponding with all empirical scores, form the empirical cumulative frequencies of them scaled so that the total frequency is unity, and determine the corresponding cumulative proportion. These numbers can now replace the scores, and the distances can be computed remembering that they are estimates of a probability, and so should be treated as such. A numerical example is shown in Table III.1. Inspection of first differences in the scaled values suggests that there may be too many categories, that too fine a discrimination exists, and that only little is lost if categories 1–6 are combined.

Table III.1 Example of 10 ordered score categories

Score	Observed frequency	Cumulative frequency	Scaled to unity
0	10	10	0.127
1	2	12	0.152
2	3	15	0.190
3	5	20	0.253
4	2	22	0.279
5	4	26	0.329
6	3	29	0.367
7	16	45	0.570
8	17	62	0.785
9	17	79	1.000

Although the number of objects (79) in this example is somewhat larger than often seen in clustering, the procedure works quite well even with as few as 15 objects; the modest amount of computation is worth the insight that is given in understanding the differences among the scores.

Interval discrete attributes

For an interval discrete attribute, the values are treated as category averages, medians, or midpoints, so that differences between the scores are interpreted as a measure of separation between the categories. These are of two types, bounded and unbounded; for example, the transformation of ordinal attributes just described constructs a bounded interval variable, which because only a finite number of ordinal classes exist, the values should still be regarded as discrete.

Bounded discrete ordered attributes For simplicity, and without loss of generality, it is assumed that the empirical data have been rescaled so that the lower bound is zero, and the upper is unity;

the bounds here are usually *not* determined by the observed range of the data but by theoretical considerations. This kind of attribute can be considered as being of binomial-type and has special properties that need to be considered if, in addition to systematic differences among the unknown groups, random variation is also a possibility. The problems needing attention can be illustrated by considering some extreme cases; if the mean of a single group is 0.01, the possible within-group variation is positively skewed, if the mean is 0.99 it is negatively skewed; only if the mean is 0.5 is it potentially symmetric. Thus the within-group variation changes shape as well as having the potentiality to be of different size. Furthermore, even though a difference from 0.01 to 0.02 may be considered to be equivalent to that between 0.98 and 0.99, that between 0.495 and 0.505 indicates much less of a biological difference, even though the numerical differences are 0.01 in all three. From this reasoning, it follows that stretching the scales as the values approach zero or unity, leaving the values in the vicinity of 0.5 essentially unchanged, is appropriate. One transformation satisfying these conditions is the logit; assume y is the proportion, i.e., $0 \leq y \leq 1$; then the logit transform is

$$z = \log_e(y/(1 - y)),$$

so that $y = 0.5$ implies that $z = 0$. Since $y = 0$ results in z equal to $-\infty$, and $y = 1$ to z equal to ∞, precautions have to be taken in a computer program. These may include adding a small value to the numerator and denominator which map the infinite range of z to a smaller, e.g.,

$$z' = \log_e((y + 0.00375)/(1.00375 - y))$$

so that $-5.59 \leq z' \leq 5.59$, using natural logarithms. This range can be rescaled to lie between 0 and 1. A little investigation shows

that this added constant hardly makes a difference in the numerical values of the differences, except for the empirically determined bounds. Other possible transformations, and other constants, can achieve a stretching of the scales as y approaches zero or unity, but they hardly differ from the logit in the interval (0.05, 0.95).

Consider again the y prior to transformation; clearly, y and $1 - y$ are equivalent representations of the attribute, since the first shows the proportion of the scale of the attribute *shown* by the object, and $1 - y$ shows the proportion of the scale *not shown*. It follows that two-state ordered attributes, which have been argued above as being equivalent to two-state unordered attributes, can be considered to be extreme cases of binomial-type attributes. For consistency, a binomial-type attribute should be represented by a two-element vector whose numerical values are the logit transforms of y and $1 - y$, which differ only in sign.

Unbounded discrete ordered attributes There are two cases to consider, attributes with a lower but no upper bound, and attributes that are discrete and unbounded in both directions.

For the first case, it is almost always true that the attribute is a count, and that the lower bound is zero. Suppose it is assumed (a critical assumption) that the distribution of counts *within* each true group is Poisson. Since there is potentially an unknown mixture of populations present in N, each presumably with a distinct mean, it is almost certain that the sample variance for the N will exceed the sample mean, perhaps making a compound Poisson distribution more appropriate. One possibility is to assume that the means follow a gamma distribution, so that the compound can be represented by a negative binomial. Then a transformation of y, which generates an approximately normally distributed variable, is

$$z = y^{-1/3},$$

which can further be transformed assuming a normal distribution to obtain an approximate uniform scale. Clearly, this chain of reasoning is a very tortuous, so that other transformations may investigated (for example, $\log(y + c)$, $y^{1/2}$, and so on) until one is found that appears to be satisfactory, even though this criterion is intuitive. Once some groups have been established, the transformation can be reconsidered and perhaps changed, but there is always the danger that the original choice will start things off in the wrong direction, as it were, so that the chances of revealing the true groups are reduced.

For the second case, it is difficult to imagine a truly discrete attribute unbounded in both directions, but because there does not seem to be any obvious suggestions differing from those of continuous attributes, they are considered implicitly in the next section.

Continuous ordered attributes

For many empirical, continuous measurements, such as lengths, areas, and weights, there is necessarily a lower limit of zero (or above), but there may be no theoretical upper limit, even though there is an ill-defined value for which higher observed values are extremely unlikely. Although this value together with the lower bound may be used in determining the range, it is likely that the within-group variation is positively skewed for all the (unknown) groups and, because the lower limit is zero, is also a monotonically increasing function of the means. If this dependence seems to hold (and it is now proposed that each measurement be examined from this point of view) then both for counts and measurements, it may be true that equal intervals on a geometric scale, e.g., counts 1, 2, 4, 8, 16, and so on represent equal steps on a scale of differences. This possibility suggests the transformation

$$z = \log_e(y + o),$$

where o is an offset, sometimes zero, but usually unity or greater for counts, and at least *eps*, defined as the maximum (absolute) measurement error for continuous measurements. If the logarithmic transformation appears to be too severe, in the sense that many originally large values become too close, then

$$z = (y + o)^{1/2}$$

will compact them somewhat less; the Box-Cox family of transformations (described by Atkinson 1985) offer a wider range of possibilities than those described here. The choice of a transformation, however, is an empirical decision, because no mathematical theory seems to be helpful in the absence of the objects being assigned to distinct classes.

An empirical transformation based on an integral transform of an empirical estimate of the probability density (Izenman [1991] gave a review) is also possible and has proved to be of use in a few data sets. Let x_i be the measurement of variable X for the i^{th} object, and let $K(u)$ be any unimodal function for which

$$\int K(u)du = 1, \text{ and } \int K^2(u) < \infty.$$

(The Epanechnikov [1969] function, which is essentially

$$K(u) = \begin{cases} 0.75(1 - u^2/5), & u^2 < 5 \\ 0, & \text{otherwise} \end{cases}$$

is said to be optimal with respect to the mean integrated squared error for any underlying distribution, although according to West [1991]

$$K(u) = \tfrac{1}{2}e^{-|u|}, \quad (-\infty < u < \infty)$$

is the only function satisfying marginalization consistency.) Then the estimate of the empirical probability density estimate at any point x is

$$p(x) = n^{-1} \sum_i (h_i^{-1} K(x - x_i)),$$

where h_i, the "window," is the distance between x_i and its nearest neighbor. The value of

$$y_i = \int_0^{x_i} p(x)dx$$

can be used as a replacement for x_i. This integral, which estimates the empirical cumulative distribution function for each x, can be obtained by numerical integration. A relationship exists between this transformation and that proposed above for ordinal attributes.

For bounded continuous variables, e.g., that proportion of the area exhibiting a particular state, two comments can be made. If the variable is a *computed* ratio, first, it has replaced two numbers by one and so discarded information; second, it has generated a variable whose variance properties are somewhat unpredictable as a result of measurement error. For most purposes, it is preferable to retain both observations in the ratio as distinct attributes; however, if there are good biological grounds for retaining ratios, their logit transformation appears to be an appropriate numerical representation for them, assuming that no ratio exceeds unity.

Other considerations may aid in determining numerical values for ordered attributes, which become of particular use if the number of objects is large. If a graph of y plotted against its empirical cumulative frequency distribution falls into a line or a curve without any plateaux, it can be inferred that the attribute

shows no evidence of the existence of subsets, and so perhaps can be discarded. The existence of one or more plateaux gives good grounds for discretizing the measurement using the points of inflection. It can then be asked if the attribute under consideration, although recorded on a continuous scale, is in fact such that each part of the range represents the expression of some *discrete* biological phenomenon.

Categorizing continuous attributes

In several different contexts, it seems that the processing of continuous variables presents difficulties, which are eased if the variable can be categorized. Archie (1985) proposed one such method in the context of cladistics; later, Goldman (1988) suggested several others. References cited in these publications refer to other methods.

Because categorizing a continuous variable is equivalent to clustering on the real line, with the distinction depending on whether the focus is on the objects or the variable, this problem is considered again in Chapter VI in the context of unidimensional clustering.

IV Clustering without pairwise resemblances

This chapter describes a method for the direct application of set-covering methods to an object–attribute incidence array, which avoids the need to choose from the multiplicity of similarity coefficients (see Chapter VII for a selection of these), their transformations to distances (Gower and Legendre 1986), and the even larger number of heuristic clustering methods (Sneath and Sokal 1973, Hartigan 1975). Together, they offer a sometimes bewildering array of choices to the taxonomist wishing to use numerical procedures, which in turn leads to the eclecticism characteristic of publications in this field.

Suppose a set of n objects is described by the states shown by m one-state (presence–absence) attribute data (Chapter III); these data can be assembled into a $n \times m$ incidence table in which each object is represented by a distinct row, each attribute by a distinct column, presence by unity, and absence by zero. After clustering, the grouping of the objects can also be represented as a binary incidence table with n rows and as many columns as there are groups. The group incidence table, not surprisingly, can often be recognized as being a subset of the columns of the original attribute incidence table, but now the columns represent clusters and not attributes. This chapter describes some direct methods by which an attribute table may be converted into a group incidence table. For convenience, the arguments are presented in terms of the clustering of objects, but they can apply equally to obtaining associations among the attributes.

Direct clustering of incidence arrays

Let the **A** of Chapter II be defined by

$$\mathbf{A} = \{a_{ik},\ i = 1...n,\ k = 1...m\},$$

where n is the number of objects and m is the number of one-state attributes (Chapter III). **A** is therefore the incidence matrix corresponding to attribute presence (for examples, see Chapter X, "Butterflies," Table A1a; "Beetles," Table L1). One solution to the clustering problem consists of equating this **A** with that defined in Chapter II, and then determining the vector **x**, which, because each object must be located in at least one of the associations (Chapter II), must satisfy the constraint $\mathbf{Ax} \geq \mathbf{1}$. The objective, therefore, is to decide if x_k should be set either to unity, which implies the choice of column k of **A**, corresponding with the k^{th} attribute, as an object association, or to zero, implying that it is not chosen.

Although this set of circumstances hardly requires more comment than that given in Chapter II, any clustering procedure depends on the choice of empirical data. Thus some preliminary thought and decisions are needed prior to any attempt at grouping. For example, an attribute uniformly present (or absent) in all objects gives no information on the object associations; it can therefore be eliminated. Some consideration should be given to attributes that are either present or absent infrequently (e.g., in one object); a strong case exists for eliminating such attributes, because they give only limited information about associations. Similarly, those objects showing presence just for one attribute can be put on one side. These decisions, however, should not be made for computational reasons but should be based on the experience and intuition of the ecologist or taxonomist.

An open question is whether the probabilities should be obtained from the original **A** or from this array after duplicate attributes (i.e., identical columns in **A**) have been eliminated. Even after allowing for the different standardization, the numerical values can differ appreciably if the duplication is considerable for

some attributes. A decision to retain duplicate columns clearly depends on the original sampling procedure for their choice; if it was random, it seems preferable to use the original **A**. It is not difficult to obtain the probabilities and groupings from both arrays; clustering, after all, is a hypothesis-generating procedure and not an evaluation.

Empirically observed incidence arrays are quite likely to include missing data, namely, elements which, for some inadvertent reason, no data are available. Some proposals on how to proceed are made in Chapter II.

In a sense, this chapter could be terminated here, were it not for some somewhat unexpected applications. Two of these are now considered.

Phylogenetic hypothesis generation

As discussed in Chapter I, the problem of obtaining a phylogenetic hypothesis is fraught with logical difficulties yet often takes as its starting point 0–1 data arrays resembling the matrix **A** of Chapter II. There is some interest, therefore, in describing how such arrays may be used, and to consider the assumptions opening this class of problem to numerical procedures. A series of definitions allow this.

DEFINITION IV.1. *A rooted tree, $T = T(A)$, for A is such that each object is attached to exactly one leaf (terminal vertex other than the root) of T, that each of the states of the m attributes (columns of A) is associated with exactly one edge of T (more than one attribute can be associated with an edge), and that for any leaf of the tree, the states of the attributes associated with the edges along the unique path from the root to it exactly specify the attribute vector of the objects at the leaf.*

The key feature of $T(\mathbf{A})$, if it exists, is that each attribute is associated with exactly one edge of the tree. The properties of this mathematical object are in themselves perhaps of some interest, but with three assumptions:

- that zero represents the ancestral state of each attribute
- the root of T represents an ancestral object, i.e., one for which all elements of the attribute vector are zero
- each attribute changes from the zero state to the unity state no more than once in any path between the root to a leaf, and never from the unity state to the zero state,

then T is called a phylogenetic tree. The problem of numerical cladistics can be expressed as:

given **A**, determine if a $T(\mathbf{A})$ exists, and if so, exhibit it.

Esterbrook et al. (1975) proved a lemma establishing the existence or otherwise of $T(\mathbf{A})$ given an **A**, and Gusfield (1991) described an algorithm, which is essentially no more than the second step of the procedure for permuting **A** to block-diagonal form (Chapter II), which will test if $T(\mathbf{A})$ exists using arithmetic of $O(nm)$. Such a tree exists for Example II.1 illustrating the block-diagonal form in Chapter II, but not for Example II.2. An efficient algorithm for constructing a phylogenetic tree from **A** is also described by Gusfield (1991), assuming that one exists. If $T(\mathbf{A})$ does not exist, or it is not known which of the states of each attribute is ancestral, it becomes necessary to consider each of the 2^m possible choices for the root, leading to a major computing problem for more than a modest number of attributes. If despite considering each of (some selection of) the 2^m choices no $T(\mathbf{A})$ exists, then minimizing

reversals (i.e., contradictions to the definition of the tree) is often combined with the choice of root to yield a solution that is regarded as a best approximation to a phylogenetic tree, but which does not satisfy the definition of $T(\mathbf{A})$; discussion of algorithms for these problems is beyond the present scope. Note that this computation is in addition to a search among the trees having n labeled vertices of unit degree, of which there are $(2n - 5)!! = 1*3*5*...*(2n - 5)$ (Cavalli-Sforza and Edwards 1967).

Diagnostic sets of attributes

A quite different context for the direct application of set covering is the determination of diagnostic attributes, both for descriptions and identification keys. Numerical methods for the generation of identification keys, e.g., for a set of species belonging to a genus, have tended to be based on the rule:

> choose the attribute dividing the species as nearly as possible into equally sized subsets.

This method gives an identification protocol for which the number of steps in obtaining an identification is approximately $\log_2 n$. Because it is rare for the set of species to fall even approximately into equal halves on an attribute, the half criterion is sometimes replaced by that of maximum entropy, i.e., choose that attribute for which the entropy of the division is a maximum, or by various extrinsic criteria, for example, based on the work required to assess the attribute (minimizing work is clearly desirable), or by its reliability (which should be a maximum). Although the general objective is the minimization of the average amount of work required to obtain a correct identification, these procedures are usually no more than locally optimal; there is no guarantee that global optimality is obtained (Moret 1982). Surveys of methods for

key generation (Payne and Preece 1980), decision trees (Moret 1982) and criteria (Brown 1977, Payne 1981) make a detailed review of previous work unnecessary, although they do not appear to be widely known (Leuschner and Sviridov 1986).

The aim of this section is to describe a method to obtain an irredundant set of attributes, which can be used to distinguish the species (Chittineni 1980, Payne 1981, Pankhurst 1983, Roberts 1984, Moret and Shapiro 1985). This discrete discrimination problem has been described in terms of a vector inequality (Lefkovitch 1987*c*). For species i, form the rectangular matrix, $\mathbf{D}_i$, which for n species has $n - 1$ rows, and m columns for m attributes, with

$$d_{jk} = \begin{cases} 1 \text{ if it is known that species } i \text{ can be distinguished from species } j \text{ using attribute } k \\ 0 \text{ otherwise (i.e., if they cannot be distinguished or if it is not known that they can be distinguished)} \end{cases}$$

as a typical element. It follows that

LEMMA IV.1. *Species i is distinguishable from all other species under consideration using the m attributes iff* $\mathbf{D}_i\mathbf{1} > 0$.

It is always valuable to verify that an individual identified as a particular species shows the appropriate pattern of differences from all others, but it is also desirable to reduce work and to increase the possibility of memorizing the diagnostic features by using the minimum number of attributes needed for complete accuracy.

The first of three problems discussed here is the selection of a **minimal diagnostic set** (MDS) of attributes to be used to

distinguish species i from the remaining $n - 1$ (Pankhurst 1983). The solution to this problem can be written as the need to determine a binary vector, $\mathbf{x}$, to

$$\text{minimize } \{\mathbf{x}^T\mathbf{x} \mid \mathbf{D}_i\mathbf{x} \geq \mathbf{1}\},$$

where $x_k = 1$ here indicates the k^{th} attribute; the minimization of $\mathbf{x}^T\mathbf{x}$ ensures the irredundancy.

The second problem, a generalization of MDS, is to choose a **minimal identification set** (MIS) of attributes to distinguish all species (Ledley 1973, Pankhurst 1983, Moret and Shapiro 1985). MIS relates to MDS as follows. Since the j^{th} row of $\mathbf{D}_i$ is identical with the i^{th} row of $\mathbf{D}_j$, the $\mathbf{D}_i$, $i = 1...n$, after removing duplicate rows, can be arranged in a rectangular array having $n(n - 1)/2$ rows and m columns

$$\mathbf{D} = \{d_{ijk} : j = 2...n;\ i = 1...j - 1;\ i \neq j;\ k = 1...m\}.$$

LEMMA IV.2. *The n species are distinguishable from each other using the m attributes iff* $\mathbf{D1} > \mathbf{0}$.

The proof of this is immediate from Lemma IV.1. It follows, therefore, that solutions to MIS are given by any binary $\mathbf{x}$, which

$$\text{minimizes } \{\mathbf{x}^T\mathbf{x} \mid \mathbf{Dx} \geq \mathbf{1}\},$$

where $\mathbf{1}$ is now a column vector of $(n(n - 1)/2)$ unities.

Clearly, $\mathbf{D}_i$ and $\mathbf{D}$, which represent the existence of pairwise differences among the species for the attributes under consideration, and the optimal choice for $\mathbf{x}$, together concisely represent the procedures described by Kautz (1968), Pankhurst (1978), and Moret and Shapiro (1985). Pankhurst (1983) described verbally the essence of lemmas IV.1 and IV.2. The present

notation, however, makes it apparent that both MDS and MIS are examples of determining minimal set coverings (Moret and Shapiro 1985), and also the obtaining of a family of distinct representatives from the representative graph (Chapter IX) obtained from **A** (Chapter II).

As pointed out in Chapter II, the set-covering problem may have many solutions, but although having several solutions is an advantage for verification, it may not be for identification. To choose from among these minimal coverings, either further (intrinsic) information, extracted from the $\mathbf{D}_i$ or **D**, or extrinsically obtained information is needed to measure the cost (i.e., the work required) in assessing each attribute. This further information allows the finding of a globally minimal-cost set (or sets) of attributes for MDS and for MIS.

The third problem arises if there is more than one measure of the cost of an attribute; it is to choose an **optimal compromise set** (OCS) for an identification scheme. If the costs can be combined into a single index, the problem is no different from that in the previous paragraph; but without a reasonable index, an optimal compromise solution is needed. Determining this solution is an example of multi-objective set covering; the solution proposed is such that any other choice degrades the solution for at least one measure of cost.

Since MIS and MDS are essentially the same problem, no distinction is made between them unless necessary; **A** is used to refer both to $\mathbf{D}_i$ and **D**. Ignoring any inequality in the work required to observe the state of the attributes they represent, it is not uncommon for two or more columns of **A** to be identical, i.e., to be equally effective in distinguishing the taxa corresponding to the unities. In forming **A** from $\mathbf{D}_i$ and **D**, therefore, duplicate columns are represented by just one, with a record kept of the fact that such columns represent more than one attribute or attribute state. This information may be of value in determining an optimal solution to MIS and MDS.

It is apparent that the minimal set-covering problem can be written as follows:

$$\text{find any (or all) } \mathbf{x} \text{ which}$$
$$\text{minimize } \{\mathbf{1}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{1}\}.$$

Because **A** is often a large matrix, the constraint set is somewhat complicated. It is often possible, however, to simplify **A** without changing the solution (or solutions). These simplifications, the reduction rules in set covering described in Chapter II (see also Garfinkel and Nemhauser 1972), sometimes give a unique solution to the problem and may make redundant the remaining parts of the procedures described below. Thus if **A** is emptied by the reductions, the attributes indicated by those elements of **x** that are unity form a unique minimal covering, and so are a minimal set of attributes.

Because a unique minimal covering is rare, and because of the preprocessing, each unit element of **x** may point to a column of **A** that represents more than one attribute, so that it is sometimes necessary to introduce further criteria to choose a solution. These are incorporated into the problem in the following manner.

Let **c** be a m-element vector of given costs corresponding with each column of **A** (based on the difficulty in observing the attribute, the reliability of the observation, the number of attributes corresponding with a column), and $f(\mathbf{c},\mathbf{x})$ the combined cost of the solution **x**; then a least-cost set-covering is given by those **x** that

$$\text{optimize } f(\mathbf{c},\mathbf{x}) \mid \mathbf{A}\mathbf{x} \geq \mathbf{1}\}.$$

In the minimal covering situation, $f(\mathbf{c},\mathbf{x}) = \mathbf{x}^T\mathbf{x} = \mathbf{1}^T\mathbf{x} = \mathbf{c}^T\mathbf{x}$, where **c** consists of unities.

Even though the costs in **c** can be obtained extrinsically, an intrinsic set can be based on the separation number (Rypka et al. 1967), which is the number of nonzero elements in each column of **A**. Because not all objects are equally informative about which attributes are useful for identification (consider, for example, those objects eliminated by the reductions), a better measure can be obtained from the interrelationships in **A** among the species and attributes, in particular, on the logical probability, p_k, that column k is a member of the optimal covering. The background to these probabilities is given in Chapter II.

The procedure to obtain **p** can be generalized to allow a measure of uncertainty in the distinguishability of each species by each of the m attributes. This generalization makes use of the matrix **B** described in Chapter II, the elements of which are defined as

> the probability that object i can be distinguished accurately from object j using attribute k.

The empirical problem of the estimation of $b_{ij,k}$ (the probability that individuals of species i can be distinguished accurately from those of species j using attribute k), from which the elements of **A** are obtained, can be considered as follows. For simplicity, assume attribute k has two states; a two-way table for species i and j can be constructed as

	state	
species	1	2
i	a	b
j	c	d

where $\{a,b,c,d\}$ represent the frequency of the specified state in the specified species, and with the columns of the table sequenced so that $b + c$ is a minimum. It is apparent that if $a/(a + b)$ and simultaneously $d/(c + d)$ both tend to unity as $(a + b)$ and $(c + d)$ both tend to infinity, then attribute k may be useful for distinguishing the pair of species. If so, it seems reasonable to estimate $b_{ij,k}$ by

$$ad/((a + b)(c + d)).$$

These values can be tested for exceeding a critical level, say 0.5, as follows. The rows of the table represent the frequency of the two states in a sample of size $a + b$ for species i, and of size $c + d$ for species j, and so can be considered as samples from two independent binomials. This interpretation allows a series of tests based on the cross-entropy with the critical value. These problems, and their extension to more states, attributes, and objects, have not been investigated empirically.

Define an indicator variable, $I_{ij,k}$, which is unity for success and zero for failure. The sum of each row of $\mathbf{B}$ is $m\mathrm{E}(I_{ij})$, and interpreted as $m\mathrm{E}(b_{ij})$, where $\mathrm{E}(b_{ij})$ measures the probability of success that an attribute chosen uniformly at random from the m will distinguish species i and j. The corresponding diagonal element of $\mathbf{B}^{\mathrm{T}}\mathbf{B}^{*}$, where $b^{*}_{ij,k} = 1 - b_{ij,k}$, namely $\Sigma_k b_{ij,k}(1 - b_{ij,k})$, is proportional to the variance of success, and is

$$m\mathrm{E}(b_{ij})(1 - \mathrm{E}(b_{ij})) - \mathrm{var}(b_{ij}),$$

where $\mathrm{var}(b_{ij})$ is the "variance" among the $b_{ij,k}$ (Johnson and Kotz 1970, p. 80), and each off-diagonal element, $\Sigma_k b_{ij,k}(1 - b_{i'j',k})$, is proportional to the covariance between success in distinguishing i and j, and failure to distinguish i' and j'. This interpretation of $\mathbf{B}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{B}^{*}$ allows an equivalent one to be made of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}^{*}$.

The reasoning used to obtain **p** from **A** can be extended to **B**, so that the probabilities to be assigned to an attribute to obtain an optimal covering are given by the Perron–Frobenius column eigenvector of $\mathbf{B}^T\mathbf{B}^*$. Although it is possible to find an optimal covering by finding those **x** that

$$\text{minimize } \{-\Sigma_k x_k \log_e p_k \mid \mathbf{Bx} > \mathbf{0}\},$$

this problem is no longer one of traditional set-covering because the elements of **B** are not {0,1} but in the interval [0,1]. However, solving this new problem does not seem worthwhile, because it depends on deciding how much greater than zero is the probability of identification acceptable. The solutions obtained by using the **p** obtained from **B**, with the constraints $\mathbf{Ax} \geq \mathbf{1}$, where $a_{ij,k}$ is unity if $b_{ij,k}$ exceeds a specified threshold (0.5 would seem to be an appropriate choice), and is zero otherwise, appear to be adequate. Having obtained the reduced **A** and also the corresponding elements of **p**, the undetermined elements of **x** can be found as described in Chapter II.

If, besides the (intrinsic) **p**, there are extrinsic criteria (e.g., it takes t_k units of time, or costs g_k units of money, or has relative difficulty for a novice of r_k, or has an identical pattern of zeros and unities with s_k others), it is unlikely that a choice based on **p** alone would be optimal for all. If there is just one extrinsic criterion, or a meaningful index can be defined for the combined cost, **c**, further solutions are possible. The *least* interesting is to choose **x** to minimize $\mathbf{c}^T\mathbf{x}$, which generates solutions independently of **p**; somewhat more interesting is to choose **x** to minimize the expected cost:

$$\text{minimize } \{\Sigma_k p_k c_k x_k \mid \mathbf{Ax} \geq \mathbf{1}\}$$

(Lefkovitch 1985*a*). Because both **p** and **c** are known, these problems remain as linear least-cost set-covering programs needing no new solution procedures. Clearly, different attributes may be chosen for different **c**.

A nonlinear objective function arising in the present context is the information in the chosen subsets after normalizing their probabilities, namely,

$$-\Sigma_k[(x_k\, p_k/z)\, \log_e(p_k/z)],$$

where $z = \Sigma_k x_k p_k$. Other nonlinear objective functions, including several that are quadratic (Roberts 1984), are beyond the scope of this study.

The combining of noncomparable measures, such as time, money, difficulty, and probability, into an index is somewhat artificial; an alternative is to keep them distinct and find a solution to a multi-objective set-covering problem. Let $f_\alpha(x)$ denote the value of the α^{th} objective function evaluated at **x** for the α^{th} criterion; the objective functions are assumed to be optimal when minimized. An efficient (Pareto-optimal, nondominated, noninferior) solution, $\mathbf{x}_{\text{opt}}$, to a multi-objective least-cost set-covering problem is

$$\mathbf{x}_{\text{opt}} = \{\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{1}, f(\mathbf{x}_{\text{opt}}) \leq f_\alpha(\mathbf{x}),\ \alpha = 1 \ldots \zeta,\ \zeta \geq 1\}.$$

In words, for any solution other than $\mathbf{x}_{\text{opt}}$, at least one of the objective functions is degraded. Clearly, for $\zeta = 1$ and f a linear function, the problem is identical with that of ordinary least-cost set covering. It is unlikely that **x** is unique (i.e., there exists a so-called utopia solution) in the sense that it is also the optimal solution for each objective function alone. Thus the attributes chosen by $\mathbf{x}_{\text{opt}}$ form an OCS.

Algorithms for the exact solution of the multi-objective set-covering problems (note that there is no known polynomial algorithm for it) were described by Bitran (1979) and by Kiziltan and Yucaoglu (1983), but, because the computational effort for finding an exact solution even for small problems does not seem to be justified in the present context, a heuristic solution, such as that described by Gabbani and Magazine (1986), is probably acceptable. A multistart stochastic programming solution to the problem, algorithms II.6 and II.7, appears to show promise both for multi-objective and nonlinear set-covering problems. Chapter X, "Arctic grasses," illustrates the methods described for the choice of attributes for diagnosis.

The formation of an identification key and a diagnostic system has a number of distinct steps, not the least of which is to define attributes and to demarcate their states. These are empirical matters, demanding considerable taxonomic knowledge and experience. Once the data have been assembled, however, much of the work that traditionally has occupied taxonomists, namely, preparing descriptions and constructing diagnostic keys, can now be automated, even to the extent of preparing multilingual versions (e.g., Watson et al. 1986). It is now easy to obtain special-purpose solutions if some attributes are seen only at special times in the life cycle, or if a specimen is incomplete.

This section has attempted to address two soft areas in the key construction procedures. First, the sequential choice of the best attribute for the next decision in the key, although locally optimal for minimizing the number of steps in obtaining an identification, is not necessarily globally (or locally) so either for minimizing the number of attributes in the key or for finding the set having the best compromise among a number of different criteria. By contrast, the optimal solution to the set-covering problem based on **A** and **p** provides a globally minimal set of attributes, which can then be used both as input to any key generating program and to provide a set for a minimal diagnosis.

As the "Arctic grasses" example in Chapter X illustrates, a key based on the minimal set may hardly differ in length from one based on a wider choice.

The other area considered is the definition of a minimal set of attributes to distinguish a species from all others under consideration. This subset is never larger than that for all species and is often appreciably smaller. As shown in an example (Lefkovitch 1987*c*) of 25 attributes, 6 were sufficient to identify all 10 species (there were nine possible combinations), but, for example, 2 were sufficient to distinguish one of the species from the remaining (there were 28 possible combinations for the pair).

A computer program can easily form the **D** array(s) from the species by attribute data. Because $d_{ij,k} \in \{0,1\}$ can be stored explicitly in bit form, **D** requires only modest amounts of computer space even for large problems.

Some other direct clustering methods

Some other direct clustering methods described for binary data ought to be compared with that of the present chapter. As noted by Lefkovitch (1985*a*), there is a resemblance between the probabilities as obtained here and the values of the object ordination given by correspondence analysis. One representation of the latter procedure is essentially as follows: with **A** as in Chapter II (without missing values and ignoring various normalizations), the reciprocal averaging solution is to find **v** and **w** so that

$$\mathbf{Av} = \alpha\mathbf{w}$$

and

$$\mathbf{A}^{\mathrm{T}}\mathbf{w} = \beta\mathbf{v}.$$

If the interest is in $\mathbf{v}$, the solution is given by the Perron-Frobenius eigenvector of

$$\mathbf{A}^T\mathbf{A}\mathbf{v} = \zeta\mathbf{v},$$

which clearly differs from

$$\mathbf{A}^T\mathbf{A}^*\mathbf{p} = (\mathbf{A}^T\mathbf{1}\mathbf{1}^T - \mathbf{A}^T\mathbf{A})\mathbf{p} = \lambda\mathbf{p}$$

of Chapter II.

In the present model, the rows (= objects), columns (= attributes), and elements of $\mathbf{A}$ are *not* regarded as being random. Chapter II has already described the superficially similar set of circumstances arising from item analysis (Rasch 1960, Andersen 1980, Tjur 1982), which by contrast assumes that the a_{ik} are independent Bernoulli random variables, with

$$p_{ik} = \Pr(a_{ik} = 1) = \alpha_i/(\alpha_i + \beta_k),$$

where the row parameter α_i increases with the increasing "ability" of object i to show the suite of attributes under consideration. The column parameter β_k decreases with the increasing "difficulty" of attribute k to be shown by the objects under consideration. The objective of the analysis, which is to estimate α_i, differs from that of the present study, which is to identify recurrent sets of individuals. The Rasch model leads to determining the set representation probabilities of $\mathbf{A}$, given that they balance the covering probabilities of $\mathbf{A}^*$; so α_i is equivalent to q_i of Chapter II. It should be emphasized, however, that in the set-covering model, there is no probabilistic interpretation of the elements of $\mathbf{A}$, and that $\mathbf{p}$ and $\mathbf{q}$ have meaning only with respect to providing evidence relevant to propositions about the grouping of objects. Were either the rows, columns, or elements of $\mathbf{A}$ to be regarded

as random samples from populations of rows or columns, then the Rasch model would be of interest, and advantage could be taken of any relevant hypothesis tests. Note that the array **B** (see Chapter II) can be used for the Rasch model if, rather than the scores being either 0 or 1, some proportional measure of correctness is used.

Another direct procedure for clustering binary data was described by Buser and Baroni-Urbani (1982). In their model, for a single 0–1 attribute in which 1 denotes presence, and S, the frequency of presences over all objects, then the sign test gives the probability density function of S as

$$p(S) = \binom{n}{S} p_1^S p_0^{n-S},$$

where p_1 is the probability of presence of the attribute, p_0 the probability of absence, and $p_1 + p_0 = 1$. It is then assumed that $p_1 = p_0 = \frac{1}{2}$ "as there are no concrete reasons for assuming other values"; however, there is no reason to accept $p_1 = p_0 = \frac{1}{2}$ in any real taxonomic situation, since not only are the states unlikely to be equiprobable, but also the unequal sampling representation of the groups (see Chapter I) is likely to make this assumption dubious at best. Buser and Baroni-Urbani then defined two distances (the first for one-state attributes, the second for two-state) between the data under study and the "most probable configuration" and considered the division of the data into k subsets, for which they obtain the entropy, defined by them as

$$\Sigma_i \log_e p_i(S).$$

Their clustering criterion is: "the lower the entropy, the better the order of the corresponding cluster, since the entropy is a measure of disorder." After considering more than one attribute, with the attending modifications of their theoretical development, they then "require the best separation of the data set into two nodes," i.e.,

introduce a sequential procedure based on the greedy algorithm. Because this criterion is local, and because the whole procedure, as currently developed, depends on the assumption that $p = \frac{1}{2}$, I discuss their procedures no further.

For data that are not necessarily binary, but for which the entries in the array are all comparable, i.e., measured in the same units, Hartigan (1972, 1975) and Eckes and Orlick (1991) described a direct clustering procedure based on computations familiar in statistics. The procedure is based on the marginal means of the object by attribute array; the first step is to permute the rows and columns so that the marginal means are monotonically increasing. Within any previous division, the columns (rows) are sequentially separated into two subsets and then subjected to a one-way analysis of variance; the decision rule is to divide the columns or rows under consideration into two subsets based on the maximum F-ratio. The assumption of homogeneous variance implicit in the analysis of variance, together with the requirement that the entries in the array must be comparable (Hartigan illustrated the method with percentage data), restrict the applications of this method. However, applications to the problem of genotype-by-environment interaction seem possible but do not appear to have been published.

V Boolean dissimilarity

The procedures described in Chapter IV are based solely on the logical relationships exhibited by the incidence array of the various objects. Although measures of pairwise relationship are not an integral component of that proposal, they are intuitively appealing and of potential value in representing the relationships among the objects. This chapter, an expanded version of the account given by Lefkovitch (1991*a*), investigates a nonscalar measure of pairwise relationship retaining the information given by their individual attributes and describes how to generate a family of subsets based on this measure.

Vector dissimilarity

For simplicity, consider a set of m one-state attributes as defined in Chapter III; extension to other types of attributes is considered in a later chapter. For the i^{th} object, let $\mathbf{i}$ be the (Boolean) vector in which the p^{th} element is unity if object i shows the state "presence" for attribute p and is zero otherwise; it is assumed that at least one element of $\mathbf{i}$ is unity. Denote by $\mathbf{B}$ the $m \times n$ array for which the columns are formed by the n vectors, $\mathbf{i}$, each corresponding with one of a set, N, of n objects. Let $\mathbf{i}$ and $\mathbf{j}$ be any two such vectors, and define

$$g_p(\mathbf{i},\mathbf{j}) = \begin{cases} 1 \text{ if } i_p \neq j_p \\ 0 \text{ if } i_p = j_p, \end{cases}$$

where i_p denotes the state of the p^{th} attribute shown by object i. Thus there are m values, $g_p(\mathbf{i},\mathbf{j})$, $p = 1...m$, for the two objects.

DEFINITION V.1. *Vector dissimilarity, $\mathbf{g(i,j)}$, is defined by $[g_1(i,j),\ldots,g_m(i,j)]$ taken in the same sequence as the attributes in* $\mathbf{i}$ *and* $\mathbf{j}$.

COROLLARY V.1. *The object-attribute and object-vector dissimilarity spaces are the same, namely $\{0,1\}^m$, and the function $\mathbf{g(.,.)}$ maps the object attributes onto the vector dissimilarities.*

By inspecting the elements of $\mathbf{g(i,j)}$, those attributes in which objects i and j differ can be identified immediately. Ellis (1951) used the norm of $\mathbf{g(i,j)}$ as a measure of the distance between $\mathbf{i}$ and $\mathbf{j}$; it is not difficult to show that most common scalar coefficients are functions of various norms of this vector. In what follows, it is assumed that

- the Boolean sum of vectors is given by the component sums $1 + 1 = 1 + 0 = 0 + 1 = 1$
- the Boolean product of vectors is the vector of the Boolean product of corresponding elements
- the inequality between Boolean vectors is defined by the component ordering, $0 \leq 0 < 1 \leq 1$
- $\mathbf{0}$ denotes a vector of zeros, $\mathbf{1}$ a vector of unities, and $\bar{\mathbf{i}}$ the complement of $\mathbf{i}$.

It is easy to verify that the function $\mathbf{g(.,.)}$ defines an abelian group in $\{0,1\}^m$, so that commutative as well as associative mathematical operations can be performed.

It is necessary to develop a number of further properties of $\mathbf{g(.,.)}$ before describing its application to clustering. One

consequence of defining vector dissimilarity is given by the fact that **g(.,.)**, although a vector retaining the identity of the attributes, satisfies the three following conditions:

- **g(i,j) = 0** implies and is implied by **i = j**
- **g(i,j) = g(j,i)**
- **g(i,j) ≤ g(i,k) + g(k,j).**

The proof of these, which is omitted, assumes the component Boolean vector sums and inequalities described above; it was given by Blumenthal (1952). These three conditions can be restated as

THEOREM V.1. (Blumenthal 1952): *g(.,.) is a metric in* $\{0,1\}^m$.

The following are simple consequences of the definition of **g(.,.)** or of Theorem V.1.

COROLLARY V.2. $g(i,0) = i, \forall i \in \{0,1\}^m$.

This corollary shows that each description vector, **i**, can be regarded as a vector dissimilarity with **0**.

COROLLARY V.3. $g(i, \bar{i}) = 1, \forall i \in \{0,1\}^m$.

Corollaries V.2 and V.3 show that **g(.,.)** defines a lattice for which **0** and **1** are the universal lower and upper bounds.

COROLLARY V.4. $g(\bar{i}, \bar{j}) = g(i, j)$.

Corollary V.4 implies that vector dissimilarity depends only on whether the objects show the same or different states, and not on which is coded unity or zero.

In forming $\mathbf{g(i, j)}$ from the original data, the m-element Boolean vector descriptions of pairs of the n objects are replaced by a single m-element Boolean vector. In consequence, $n(n-1)/2$ such vectors exist; it is these on which clustering procedures can be based. Since $\mathbf{g(i, j)} = \mathbf{0}$ implies that each of the m attributes of object i exhibits the same state as the corresponding attribute of object j, these objects are immediate candidates for grouping; in the context of other objects, either can represent the pair. It is unlikely that the objects will fall into a few groups within which vector dissimilarity is uniformly $\mathbf{0}$, so that further tools need to be developed. It is remarkable that they are a consequence of the next definition.

DEFINITION V.2. *The neighborhood of any vector $\mathbf{i}$ in the space $\{0,1\}^m$ consists of $\mathbf{i}$ and those vectors differing from it in a single element.*

Denote the neighborhood of $\mathbf{i}$, which consists of $\mathbf{i}$ and the adjacent vertices on the unit cube, by V_i. Define $\mathbf{e}_p$ by

$$e_{pq} = \begin{cases} 0, & q = 1 \ldots m,\ q \neq p \\ 1, & \text{otherwise,} \end{cases}$$

so that the m different vectors $\mathbf{e}_p$, which form the set $\mathbf{E} = \{\mathbf{e}_p\}$, are a basis for $\{0,1\}^m$; $\mathbf{E}$ can be represented by an identity matrix. It is easy to see that

$$\mathbf{g(i,(j} \in V_i)) \in \{\mathbf{e}_p\},\ \mathbf{i} \neq \mathbf{j}$$

is one of this set.

Since V_i may not contain any points corresponding with real objects (other than i itself), and because clustering needs to deal

with all n objects, it is necessary to establish a relationship between any $\mathbf{i}$ and any $\mathbf{j}$ in $\{0,1\}^m$. Such a relationship may be established by forming a chain

$$[\mathbf{i}...\mathbf{j}] = [\mathbf{i}, \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{q-1}, \mathbf{j}],$$

where each $\mathbf{c}_s$, $s = 1...q-1$, is in the neighborhood of the two points adjacent to it; the $\mathbf{c}_s$ need not correspond with a real object.

LEMMA V.1. $g(c_s, c_{s+1}) \in E$.

Proof. Immediate from the definition of a neighborhood. Q.E.D.

There are many chains which may connect $\mathbf{i}$ and $\mathbf{j}$, and although it is tempting to consider one chain to be shorter than another if there are fewer steps, using the number of steps introduces a metric (path length) equivalent to using a traditional (scalar) measure of dissimilarity. Lemma 1 has already established that $\mathbf{g}(.,.) \in \mathbf{E}$ for all adjacent vectors on a chain, and since $\mathbf{E}$ is a basis, this fact can be exploited to eliminate duplication.

DEFINITION V.3. *A minimal chain is a chain for which all vectors* $g(i,c_1)$, $g(c_1, c_2),\ldots,g(c_s, c_{s+1})$, ..., $g(c_{q-1}, j)$ *are different.*

Thus the set of vector dissimilarities between the adjacent members of a chain form a basis for the subspace defined by the Boolean vector sum of $\mathbf{i}$ and $\mathbf{j}$.

Although more than one minimal chain may exist between any pair of objects, the number of steps, q, is the same for all. Avoiding the use of q, the length of a chain can be defined as the Boolean sum

$$g[i...j] = g(i,c_1) + g(c_1, c_2) + \ldots + g(c_{q-1},j),$$

which is also an element of $\{0,1\}^m$. From the triangle inequality (Theorem V.1), it follows that

$$g(i,j) \leq g[i...j],$$

with strict equality not infrequently observed in practice.

Following Blumenthal (1952), the vectors on a straight line defined by **i** and **j** are any (and all) pairs of vectors, **u** and **t**, such that

$$\mathbf{u} = \mathbf{it} + \mathbf{j}\bar{\mathbf{t}},$$

where $\bar{\mathbf{t}}$ is the complement of **t**. If there are straight lines, there is also a concept of betweenness; it is this concept on which the clustering procedures described below are based.

DEFINITION V.4. (Blumenthal 1952): *if*

$$g(i,j) = g(i,k) + g(k,j),$$

then ***k*** *is between* ***i*** *and* ***j***.

This concept has a natural relationship with that of betweenness for the real line. In fact, objects are between **i** and **j** if they lie on the sublattice having the universal bounds **i** + **j** and **ij**. Furthermore, if **k** is between **i** and **j**, then **k** is part of a minimal chain connecting **i** and **j**.

The practical determination of the betweenness relationship among *i*, *j* and *k* takes advantage of

COROLLARY V.5. *If* $k = k + ij$, *then* ***k*** *is between* ***i*** *and* ***j***.

The proof that Corollary V.5 is equivalent to Definition V.4 can be verified by considering all eight possible cases.

Note that if a row of **B** either consists entirely of unities or of zeros, or is identical with another row, the betweenness relationships are unchanged; this situation implies that the corresponding attributes need not be considered.

Vector dissimilarity and subset generation

The procedures described in Chapter II require a family of subsets of the objects to be obtained from the empirically observed data. If a subset, S, of the n objects is defined by a n-element Boolean vector, **a**, with elements

$$a_i = \begin{cases} 1, \text{ object } i \in S \\ 0, \text{ otherwise,} \end{cases}$$

obtaining these subsets can be regarded as a function that maps vectors belonging to $\{0,1\}^m$ onto vectors belonging to $\{0,1\}^n$. The formulation of such a function arises from an appropriate answer to the following

> *Question.* If an arbitrary subset, S, of the N objects is formed, which others of $N \setminus S$ should also be included?

An answer in the case of the real line is informative. Let

$$z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}$$

be the order statistics corresponding with a measurement of some continuous variable for the n objects; if the objects corresponding with $z_{(i)}$ and $z_{(i+2)}$ are included in S, then all objects corresponding with any points between them, e.g., $z_{(i+1)}$, should also be included

(W.D. Fisher 1958). This heuristic principle has the consequence that if the objects corresponding with $z_{(1)}$ and $z_{(n)}$ are included in S, then so should all of N. Fisher's answer to the question for the real line, therefore, leads to a rule which can be summarized as

> HEURISTIC V.1. *Include in S those members of* $N \setminus S$ *between any (pair of) members of S.*

For m-dimensional continuous spaces, $m > 1$, arising from a scalar measure of the distance between objects, betweenness is not defined, and so this principle has to be modified (Lefkovitch 1982; see Chapter VIII). For vector dissimilarity, however, no modification is needed for the concept of Boolean betweenness given above, allowing Heuristic V.1 to be restated for practical implementation as:

> HEURISTIC V.1′. *Include k in S if k is between i and j, where i and j are a pair of (not necessarily distinct) members of S.*

Betweenness is easily determined using Corollary V.5.

It is unnecessary to consider each of the nonempty subsets of N in turn to initiate an S for several reasons. If $\mathbf{g(i,j)} = \mathbf{i} + \mathbf{j}$, equivalent to $\mathbf{ij} = \mathbf{0}$, where i and j are part of the initial subset, then objects i and j have no attribute exhibiting the identical state (excluding absence and attributes not included in the study) so there is no interest in the subset of objects formed by them and those between (the Jaccard scalar similarity (see Chapter VII) between such objects is zero). Furthermore, the subset initialized by three (or more) containing such a pair is represented by the union of two or more of the subsets initiated by other pairs chosen from the three (or more); eliminating the threes, fours, and so on, however, requires that the optimal selection of subsets by the

programming methods described in Chapter II must not be constrained to be a partition but allowed to be a covering. Those pairs for which $\mathbf{g(i,j) = i + j = 1}$ can also be excluded from the initial set, because every other object is between them. Similar remarks follow for all initial subsets of three, four, and so on. These rules collapse into eliminating i and j as an initial pair if either $\mathbf{i + j = 1}$, or $\mathbf{ij = 0}$, or both. Thus only the $\binom{n}{2}$ initial pairs of objects need be considered to initiate the family of subsets for subsequent study. This upper limit on the number of distinct subsets is rarely reached, not only because of the disqualifications described above, but also because different initial pairs may generate the same final subset. Thus a distinct subset can be denoted by

$$\mathbf{a = \{k : k = k + ij; i + j \neq 1 \text{ and } ij \neq 0\}},$$

which is one of the candidate subsets for choosing the optimal covering. If H distinct subsets are generated by betweenness, they are assembled into a $n \times H$ matrix, $\mathbf{A}$, for which the procedures described in Chapter II can be applied.

Extending vector dissimilarity to s-state unordered attributes is straightforward, using standard procedures for converting these to s one-state characters; for ordered attributes (discussed in Chapter VI), betweenness is well understood. In this more general setting, the inclusion of object k in the subset defined by objects i and j depends not only on betweenness for the s-state unordered attributes but also on whether the measurement for each ordered attribute of k is not outside the range determined by i and j. A case for extending the range, especially if the attribute is a random variable, is made in Chapter VI.

The fact that a clustering procedure can be based on vector dissimilarity illustrates the potential of this measure of pairwise relationships, especially as *betweenness* has a very natural

interpretation as a heuristic clustering principle. A numerical example (see Chapter X, "André's data") shows, not surprisingly, that the solution obtained need not coincide exactly with those obtained using other methods (one without any dissmilarities, the other with the scalar Jaccard similarity), although considerable resemblance is seen. Because the objective of clustering is to generate hypotheses about group existence and membership, differences such as these are more likely to be helpful in deciding about relationships than in adopting a single method asserted to be the best. Because it is not known which grouping is correct, the combined results are suggestive about those objects that appear to belong together without doubt, and also about those whose positions are perhaps uncertain.

Interior objects

Consider objects i, j, and k, represented by **i**, **j**, and **k**, the attribute vectors. Let $I_{ij}(k)$ take the value of unity if **k** is between **i** and **j**, and zero if not; $k \neq i, j$; $i \neq j$. Define also

$$I(k) = \Sigma_{i \in N, j \in N, j \neq i} I_{ij}(k),$$

which is the number of occasions that object k is between others. Let $I_{max} = \max(I(k), \forall k \in N)$; then as defined in Appendix 1,

> DEFINITION V.5. *$K = \{k : I(k) = I_{max}\}$ are a set of focal points.*

Thus K is the **kernel** of the set, N. The objects belonging to K can be regarded as being central in the set N and so may be useful in providing the candidates for the selection of a type specimen; those of $N \setminus K$ form the **frontier** of N and represent the range of possible variation.

Boolean similarity

Although $\mathbf{g(i,0) = 0}$ implies that the origin of the description vectors is **0**, interpretable as a vector describing "absence," other origins can be used to represent pairwise relationships. If the origin is **1** (i.e., "presence"), **i** as originally defined is replaced by $\bar{\mathbf{i}}$, but $\mathbf{g(.,.)}$ is unchanged, as shown by Corollary V.4; using **1** as the origin is a mapping that preserves distance and is therefore a motion (Ellis 1951). With **1** as the origin, however, the concept of similarity may be more natural, and, although vector similarity can be represented by the complement of $\mathbf{g(.,.)}$, neighborhoods, minimal chains, bases, and clustering rules are sometimes more awkward to define.

Boolean (discrete) derivatives

Scalar dissimilarities map the relationship between i and j onto a point of a space usually considered to be continuous (often treated as isomorphic to a Euclidean space); thus they allow derivatives to be defined. To develop the idea of a discrete derivative, let **i** be any vector belonging to $\{0,1\}^m$ and $\mathbf{i}(p)$ be the same vector with the p^{th} element replaced by its complement.

DEFINITION V.6. *The discrete derivative of* $g(i,j)$ *with respect to* j, *denoted by* $g'(i,j)$ *is an* $m \times m$ *matrix with elements*

$$g'(i,j) = \begin{cases} 1 \text{ if } g(i,j) \neq g(i,j(p)),\ p = 1 \ldots m \\ 0 \text{ otherwise.} \end{cases}$$

Clearly, $\mathbf{g'(i,j)}$ consists of the set of values $\mathbf{g(i, j \in V_j)}$.

The discrete derivative may be useful if compound objects are constructed, i.e., a group of objects not necessarily identical in the states of each attribute, but considered to be a single entity,

e.g., a species. If for two compound objects (including single objects) $\mathbf{g}'(\mathbf{i},\mathbf{j})$ is a zero matrix, good grounds exist for replacing i and j by a compound formed from them. However, the most promising application of $\mathbf{g}'(\mathbf{i},\mathbf{j})$ is in dealing with missing data. If the p^{th} value for object i is missing, $g'(\mathbf{i},\mathbf{j}) = g'(\mathbf{j},\mathbf{i}) = 0$; the more zeros in $\mathbf{g}'(\mathbf{i},\mathbf{j})$, the less well founded is the measure of dissimilarity.

Vector dissimilarity and phylogenetic reconstruction

The clustering procedure based on vector dissimilarity described in this chapter is appropriate in any context requiring a classification based on a set of attributes. Vector dissimilarity is immediately applicable to phenetics in biology, especially if the objective is to assemble *individuals* into groups. It also has a role in cladistics, where the objective is to model the phylogeny of the assembled groups (such as species).

One criticism of phenetics made by advocates of numerical cladistics is that the use of overall measures of (dis)similarity is not helpful for phylogenetic reconstruction, and that each attribute should be considered separately. Although vector dissimilarity does not include a sense of direction (Corollary V.4) in each attribute, it does retain the separateness of the attributes required by numerical cladistics but without the biologically false assumption of independence implicit in much cladistic practice. Furthermore, if the zero state is *defined* as being ancestral to that represented by unity, it becomes possible to interpret the vectors $\mathbf{ij}$ and $\mathbf{i} + \mathbf{j}$ phylogenetically. For example, it is almost (but not absolutely) a truism that the states shown by an ancestral form include those that are uniform in the group of taxa supposedly descended from it; thus for objects i and j, since $\mathbf{ij}$ is between $\mathbf{i}$ and $\mathbf{j}$, a hypothesis for the states of their ancestor is $\mathbf{ij}$, and the diversity of attributes in the combined taxa is $\mathbf{i} + \mathbf{j}$. The attributes that have changed in

i correspond with the unities in **g(i,ij)**; the unities in **g(i+j,ij)** indicate those attributes changed in both. If **ij** is null, it must be concluded that the description of the objects is insufficient to determine the states shown by a common ancestor. There is a role for vector dissimilarity in cladistics as well as in phenetics.

VI Clustering on the real line

In Chapter III, the need to consider carefully the units of measurement for ordered or continuous attributes was discussed in some detail. Grouping a set of objects into subsets by any single attribute yields solutions that may also be considered as defining an ordered categorization of the variable; as mentioned in Chapter III, the categorized attribute may then replace the original attribute with only modest loss of information. This replacement is of lesser importance for group formation than it is in at least two other contexts:

- in constructing formal identification keys, deciding where to divide an ordered variable for diagnosis

- for categorizing ordered variables in current numerical cladistics (Almeida and Bisby 1984; Archie 1985; Goldman 1988).

Methods for the grouping of objects using continuous (and discrete ordered) measurements are the subject of this and the next two chapters. This chapter considers unidimensional data, and the simultaneous consideration of more than one but without combining them into a single index as discussed in Chapter VII. The use of such indices for clustering is postponed until Chapter VIII.

The starting point for unidimensional clustering is a vector, $\mathbf{z}$, of n elements for which z_i is the measurement on object i. It is assumed (see Chapter III) that a unit difference in two measurements indicates the same degree of difference throughout the whole range of values, and that each z_i is an observation from

a random variable, Z, for which the probability distribution or density is unknown. A second assumption of lesser importance is that the whole real line is potentially available for a value of z_i, or perhaps that there is a sufficiently large interval, extending beyond the empirical observational bounds, for which any value for z_i is possible but unlikely.

Betweenness

As noted in Chapter V, W.D. Fisher (1958) claimed if the $z_{(i)}$ are the (ascending) order statistics, to include $z_{(i)}$ and $z_{(i+2)}$ in a subset but to exclude $z_{(i+1)}$ is unreasonable. This heuristic principle receives substantial support from lemmas 2.1 and 2.2 of Boros and Hammer (1989). For continuous variables, forming subsets by use of a criterion equivalent to that of betweenness suffers from an obvious weakness; it is that $z_{(i-1)}$ may be only trivially less than $z_{(i)}$, or $z_{(i+3)}$ trivially greater than $z_{(i+2)}$, or both, so that a group formed only of $z_{(i)}$, $z_{(i+1)}$, and $z_{(i+2)}$ has no more than a passing interest. The next section of this chapter describes how the boundaries may be extended in such a way as to avoid forming subsets unlikely to be of any lasting interest.

Outliers

The remarks about $z_{(i-1)}$ and $z_{(i+3)}$ have been leading towards the notion that making a decision about the membership of an object in a subset is analogous to considering the probability of its being an outlier. If the probability is sufficiently low, the object should be included in the subset. To consider this notion as a possible model for subset generation, two assumptions, virtually the same as those of Scott (1965), have to be made. These assumptions are that a subset is a candidate for inclusion in an optimal covering if:

(1) For any subset of objects, there exists a probability density for an arbitrary point in the dissimilarity space which

depends only upon the distance between this point and the subset.

(2) Given a subset of objects, the remaining objects are mutually independent, at least locally.

These remarks and assumptions lead naturally to considering the probability of membership as being one of extreme values. Using the exponential extreme-value distribution, for many (all?) distributions belonging to the exponential family

$$\Pr(Z \leq z) = \exp(-\exp[-(z - \rho)/\Theta]),$$

where ρ and $\Theta > 0$ are parameters to be determined. Their moment estimators, which are probably satisfactory in the context of clustering, depend on $\hat{\sigma}$, the sample standard deviation, which can be obtained from the average absolute difference among all pairs of elements of $\mathbf{z}$. If this average is δ, which has an expected value of $2\sigma/\pi^{1/2}$ for normal distributions,

$$\hat{\Theta} = \hat{\sigma}\, 6^{1/2}/\pi = \hat{\delta}\, (3/2\pi)^{1/2}$$

and

$$\hat{\rho} = \mu - \gamma = -\gamma(3/2\pi)^{1/2},$$

where μ is the centroid (mean) of the population, here assumed to be zero, and γ is the Euler number 0.57722... . To a reasonable degree of approximation, suppose S is a nonempty subset of the objects, then

$$\Pr(i \in S) = \exp(-\exp[-([w_i + 0.4\delta]/0.7\delta)]),$$

where w_i is the average (absolute) difference between object i and the members of S. If $V(S)$ denotes the neighborhood (e.g., betweenness) of subset S, this reasoning leads to the following heuristic:

HEURISTIC VI.1. *Given a subset S_ν at stage ν of a generating process, $S_{\nu+1}$ consists of those objects*

$$S_{\nu+1} = \{i : Pr(i \in V(S_\nu)) \leq \alpha_\nu\},$$

where $\alpha_\nu = \alpha(S_\nu)$ is some function satisfying the two conditions

$$\lim \alpha_\nu = \begin{cases} 1, & \delta_\nu \to \infty \\ 0, & \delta_\nu \to 0. \end{cases}$$

Thus for a very large average distance among the members of S_ν, all n objects should belong, and for smaller averages, then essentially fewer (or even no) other objects should belong. These properties imply and are implied by the requirement that the number of objects in the subset should tend to increase in relationship to the increase in variability. From these considerations, a definition of α_ν can be obtained that is data dependent. Let ξ_ν be the characteristic largest value, defined here as the absolute maximum difference in the subset, i.e., the range of values of z_i for $i \in S_\nu$. Then

$$\lim_{(\xi\nu - \delta\nu) \to \infty} \begin{cases} |S_\nu| = n \\ \alpha_\nu = 1 \end{cases}$$

and

$$\lim_{(\xi\nu - \delta\nu) \to 0} \begin{cases} |S_\nu| = 1 \\ \alpha_\nu = 0 \end{cases}$$

are a re-expression of the same limits set for α_ν but with respect to δ_ν. Using the probability corresponding with ξ_ν to define $\alpha_{\nu+1}$ gives

$$\alpha_{\nu+1} = \exp(-\exp[-([\xi_i + 0.4\delta]/0.7\delta)]).$$

For practical application, this expression decides on membership of object i in a subset by the following very simple rule, which is the key element not only for single univariate clustering, but also for sets of attributes considered simultaneously and for dissimilarity indices.

> HEURISTIC VI.2. *If the average distance of object i to the members of S_ν does not exceed the maximum among them, include object i in subset $S_{\nu+1}$.*

If $d_k(S_\nu)$ denotes the average distance of object k to the members of S_ν, this heuristic can be written alternatively as

$$S_{\nu+1} = \{k : d_k(S_\nu) \leq \max(d_{ij}) \,|\, i, j \in S_\nu\}.$$

Each of the $\binom{n}{2}$ pairs of objects defines a subset formed from them, from those between, and perhaps some just external to them. The process is repeated, with $S_{\nu+1}$ replacing S_ν until $S_{\nu+1} = S_\nu$.

Although this recursion rule is derived from reasoning based on extreme value theory, with several approximations and assertions, its simplicity speaks in its favor. Furthermore, because the extreme value distribution is largely independent of the underlying probability distribution (or density) of the variable under discussion, the rule has elements of robustness. Perhaps the most telling argument in support is that it is extremely plausible and would not be considered unusual or unbelievable even if it were to be proposed as a clustering principle *de novo*.

Any procedure satisfying Heuristic VI.1 is an example of conditional clustering and provides an answer to the question asked in Chapter V, and now repeated:

Question. If an arbitrary subset, S, of the N objects is formed, which others of $N \setminus S$ should also be included?

Outlier tests

Standard statistical tests for outliers, described in detail by Barnett and Lewis (1984) for normal, gamma, exponential, Poisson, binomial, extreme value, and other distributions, can be imagined to provide an alternative to the procedure described above.

HEURISTIC VI.3. *Examine each nonmember of S_v in turn as if it were an outlier from the appropriate distribution, and include it in S_{v+1} if the probability of its being an outlier is sufficiently low.*

The problem with these standard statistical tests is that they are functions of the number of objects belonging to S_v; but this number is unlikely to represent anything other than the behavior of the collectors of the objects or the rarity of the unknown groups (see Chapter I). The primary (possibly, the only) objective of clustering is to determine the existence of a group, *not* the probability of an extreme member of it being found in a sample of size $|S_v| + 1$.

Nevertheless, one standard test can be modified for this purpose. Let $d_{(ks)}$ be the smallest and $d_{(km)}$ the largest distance between the candidate, k, and any current member of the subset. According to Likeš (1966), the exclusive excess range statistic (Dixon 1950, Barnett and Lewis 1984) for an upper outlier from an underlying exponential distribution is

$$c = d_{(ks)}/d_{(km)}$$

with test statistic

$$F_n(c) = 1 - (n-1)(n-2)\, B([2-c]/[1-c],\, n-2),$$

where $$B(\alpha,ß) = \Gamma(\alpha)\Gamma(ß)/\Gamma(\alpha+ß).$$

Tables for this test are presented in Barnett and Lewis (1984). As just noted, the fact that n, the number of objects under study, is in this expression is the main deficiency, given the usual collecting procedures for biological objects. To modify this test for determining subset membership, however, it seems reasonable to consider that the objects number just three, namely, the candidate and the closest and furthest subset members. After a little algebra, the test criterion becomes

$$F_3(c) = c/(2 - c)$$

from which it is easy to show that $c = ⅔$ implies $F_3(c) = 0.5$ for reference to the tables. If the average absolute distance to the members of the set is approximated by $0.5d_{(ks)} + 0.5d_{(km)}$, the criterion of Heuristic V.2 is equivalent to a probability of 0.2. Although this value may appear to be somewhat low, there is a fundamental difference between statistical testing, which tends to need strong evidence to support a decision that something is different, and the requirements of taxonomy, which needs to keep things separated for relatively weak reasons until they are shown to be unsubstantiated.

More than one ordered variable

For more than one ordered variable, traditionally the procedure has been to combine them into a scalar pairwise measure of dissimilarity (see Chapter VII), and to use a clustering procedure that operates in a multidimensional space. However, such indices

are usually compounds of incomparables (shapes, colors, lengths, areas, and so on) and can be considered as being too artificial. An early alternative to such indices was sequential (Williams and Lambert 1959, 1960); an ordering of the attributes was determined and the objects divided into subsets, each of which has the property of being relatively homogeneous for a state of the selected attribute. This monothetic, sequential, divisive procedure has the weakness that all divisions beyond the first are conditional on those previously made, and so there is a nontrivial probability of failure to construct a meaningful hierarchy of more than one level, although the procedure is certain to produce groups that are essentially uniform in the states of the chosen attributes. If that is the objective, as in the maximal predictive classification procedure of Gower (1974), there is much to recommend it. A possible sequence for introducing the ordered attributes can be determined by use of the following criterion. For a non-negative ordered attribute in which there is at least one nonzero value, the Gini (hierarchy) index is defined by

$$g(z) = (\sum_{j=2\ldots n} \sum_{i=1\ldots j-1} |z_i - z_j|)/(n(n-1)\bar{z}),$$

can be shown to satisfy $0 \leq g(z) \leq 1$ (Stuart and Ord 1987, §2.25). The greater this index, the more structured are the data. A set of attributes can now be ordered from the greatest to the least Gini index. For two (or more) attributes, z_k, the value of $g(\pi z_k)$ can be compared with $\pi g(z_k)$.

Rather than a sequential introduction of the attributes, their simultaneous use is now considered. This and the previous chapters have led to a common notion; that a subset defined by objects i and j consists of them and any objects located between them (Chapter V) or, for ordered variables, between them and just outside their range (this chapter). An alternative procedure is now proposed, which also keeps each variable separate; it extends the

procedures of Chapter V to apply both to ordered and to any combination of ordered and unordered variables.

Let a subset generated by i and j be denoted by $V(i, j|\mathbf{z})$, where $\mathbf{z}$ is the set of attributes used; there may be just one attribute (as so far in this chapter) or several (see Chapter V). In Chapter V, it is apparent that $V(i, j|\mathbf{z})$ is formed by the simultaneous consideration of the one-state attributes, and that it requires only one of the attributes of a candidate object to fail to satisfy the criterion of betweenness to disqualify the object from membership in $V(i, j|\mathbf{z})$. Extending this idea to ordered variables is straightforward and leads to a generalization of Heuristic VI.2:

HEURISTIC VI.4. *Multicriteria subset generation.*

If $\{d_{ij}|u\}$ *is the set of distances between objects i and j based upon the set of attributes* $\{u\}$;
$\{\delta_v|u\}$ *is the set of maximum distances for each attribute among the members of the subset at stage v in the procedure for the set of attributes* $\{u\}$; *and*
$\{\bar{d}_{kv}|u\}$ *is the set of average distances between object k and the members of the subset for the set of attributes* $\{u\}$;
then include object k in the subset at stage $v + 1$

$$\text{iff } \{(\bar{d}_{kv}|u)/(\delta_v|u)\} \leq 1,$$

i.e., if for all variables, the average distance for each object has to the members of the subset does not exceed the maximum for that variable. This simultaneous consideration of many sets of distances is an example of multicriteria decision making discussed by Zeleny (1982). If $\mathbf{u}$ consists of many perfectly correlated variables, the whole set is equivalent to just one of them, and so the effect of duplication of information is removed (this lack of

effect is not necessarily an advantage, but it does tend to eliminate inadvertent duplication). If there is no association among the variables, then the final family of subsets will consist just of the $\binom{n}{2}$ pairs of objects, indicating that perhaps some other method may be preferable. An example where this combined algorithm may be applied is given in "GE interaction" (see Chapter X).

It is not difficult to combine the Heuristic VI.4 for multiple continuous attributes with that of vector dissimilarity betweenness by extending the criterion to include k in S_{v+1}

$$\text{iff } \{(\bar{d}_{kv}|\mathbf{u})/(\delta_v|\mathbf{u})\} \leq 1$$

and

$$\mathbf{g(i,j) = g(i,k) + g(k,j).}$$

Note, however, that because the dissimilarity vector criterion for subset membership is not iterative, it suffices to consider it first, and then to retain or reject the subsets based on the criterion for ordered variables.

Chapter IX continues the discussion of multiple decision criteria using concepts based on the relative neighborhood graph; the latter is described in some detail in Chapter VIII. The concept of betweenness for multivariate data, via a generalization of the relative neighborhood graph by Ichino and Sklansky (1985), is also discussed in Chapter IX, section 6.

Grouping means

To complete this chapter, I consider a recurring problem in statistics, namely, the grouping of means, which can be recognized as clustering on the real line. Some solutions for this problem have been based on assumed distributions, others on Bayesian considerations, and yet others on the application of heuristic clustering procedures (e.g., O'Neill and Wetherill 1971, Binder 1978, Basford and McLachlan 1985, McLachlan and Basford 1988).

Calinski and Corsten (1985), dissatisfied with the fact that many solutions consist of overlapping subsets, proposed three principles for forming a partition, namely, that

- the number of groups is as small as possible
- homogeneity within groups is a maximum
- there is good separation among the groups.

These principles contain the elements of a mathematical program, in particular, of a least-cost set-partitioning problem. The components of this program are conceptually simple, namely, given a family of subsets of the objects, together with a measure of heterogeneity for each subset, choose from among these to minimize the pooled within-group heterogeneity of the choice, subject to the constraint that each object be included precisely once. How the number of groups is to be chosen, or how to pool heterogeneity, or how to decide what is an acceptable level of within-group pooled heterogeneity all need investigation. The remaining part of this section describes how Heuristic VI.2 can be used to provide a solution to the problem of grouping means.

Adopting W.D. Fisher's principle, described prior to Heuristic V.1 as re-expressed by Heuristic VI.2, each of the $\binom{n}{2}$ pairs of class means defines a subset. If the subset is acceptable (e.g., it contains fewer than n means; other conditions may also be imposed), it is retained; the process is then repeated with another of the pairs. With the decision rule redescribed below, not all $\binom{n}{2}$ pairs need be used; Chapter VIII shows that in fact only those pairs adjacent on the relative neighborhood graph (Toussaint 1980) need be used; for unidimensional data this graph is identical not only with the minimum spanning tree but also with the (unbranched) shortest path among the objects, so that there is a maximum of $n - 1$ subsets. If an object proves not to be a member of any of the accepted subsets, it forms a subset by itself.

Defining the distances used in Heuristic VI.2, therefore, is the key component, because it determines which objects are to be included in a subset and which are not. As shown above, betweenness, coupled with some blurring at the ends, leads to the following recursive form of Heuristic VI.2:

> HEURISTIC VI.2 (recursive form). *If d_{ij} is the distance between objects i and j, δ_v the maximum distance among the members of the subset at stage v in the procedure, and $\bar{d}_{kv}$ the average distance between object k and the members of the subset, then include object k in the subset at stage $v + 1$*
>
> $$\text{iff } \bar{d}_{kv}/\delta_v \leq 1,$$

i.e., if the average distance an object has to the members does not exceed the maximum among them. For unidimensional data, δ_v is the range of the variable in the subset. If $z_{(i)}$ is the i^{th} ordered value of a continuous variate in a sample of size n, $(i) = 1 \ldots n$, the range is related to the estimate of the standard deviation, s, by

$$\delta_v = |z_{(n)} - z_{(1)}| \leq 2(n-1)^{1/2}s.$$

Because two estimated means will be distinct with probability 1 the range will be positive. Since for $v = 1$, the number of distinct means is necessarily 2, the recursive Heuristic VI.2 can be specialized as

> HEURISTIC VI.2′. *Include object k in the subset at stage $v = 2$ if at stage $v = 1$, $\bar{d}_{k1}/s \leq 2$,*

i.e., if the ratio of a mean to its appropriate standard deviation (its standard error) does not exceed 2. This ratio can be regarded as an

internally studentized range, which, in normally distributed samples, has a distribution free of the unknown true mean and standard deviation (David 1981, p. 89).

As the subset size increases, the expected range also increases approximately in proportion to the square root of the number of objects and so includes more of the objects under consideration. It follows that at least three possibilities exist for the decision rule for $v > 1$.

HEURISTIC VI.2′.1. *Let the decision criterion be* $2(n_v - 1)^{\frac{1}{2}}$, *where* n_v *is the cardinality of the current subset.*

Thus s (the standard deviation) remains constant throughout and is the value appropriate to the generating pair, i.e., δ_v remains constant at δ_1. The argument against this assumption is that it depends on normality, which, in a clustering context, can hardly be justified, and so this possibility can be rejected.

HEURISTIC VI.2′.2. *Ignore the fact that the cardinality increases, and let the empirical range be determined for the current subset; leave the decision criterion at 2.*

The argument against this possibility is that it is largely intuitive. The principal support comes from the assertion that there lack grounds for believing that the objects under study are a fair sample of those in the population (Chapter I) also the fact that the rule has worked well in numerous real data sets.

HEURISTIC VI.2′.3. *Let the decision rule depend both on the cardinality and the estimated s, i.e., regard the criterion as a test of significance analogous to a t-test by incorporating* n_v *and the value of Student's t tabulated for* n_v *degrees of freedom in the rule.*

The principal argument against this possibility is that since

$$s \geq \delta/[2(n_\nu - 1)^{1/2}],$$

the decision rule is

$$d_{k\nu}/s_\nu = 2(n_\nu - 1)^{1/2} d_{k\nu}/\delta_\nu \leq 2.$$

If $\delta_\nu = 1$ and $d_{k\nu} = 0.5$, object k is almost certainly contained within the bounds determining the range. It is easy to see that at stage $\nu + 1$, object k would be included if $n \leq 5$, but not if n is greater. This consequence is unacceptable, and so Heuristic VI.2′.3 is rejected.

The arguments against heuristics VI.2′.1 and 2 seem weaker. It is interesting that, in simulated combined samples from normally distributed populations having different means but the same variance, the first two rules gave very much the same groups as each other; however, they gave different groupings for skewed distributions, with the second rule almost always giving groups corresponding with the known generating parameters. The rule based on Heuristic VI.2′.2, therefore, appears to be the best of the three. Two further points to note in connection with this heuristic are as follows:

- the value of s_ν changes as the subset changes

- when the subset is finally accepted, the value of s_ν is neither assumed nor required to be the same as that of any other accepted subset.

It follows that the accepted subsets are allowed to differ not only in their mean values (i.e., if z_0 is some origin for the unidimensional data, the means are then defined by the average

distance of the members of each subset from this origin) but also in their permitted internal variation. A further question is to decide if the within-class variance (or range) should play a role in subset generation, because here the objects under consideration are sample means. One role for these estimated variances is in relation to the units in which the grouping is to take place. It is often true that a relationship between the mean and variance is ignored, even though this contradicts the assumption of constant second-order conditions in an ANOVA. For clustering, such a relationship implies a changing metric, and so computing δ and $\hat{d}$ needs some prior preparation; in particular, it is desirable to transform the data so that the within-class variances are effectively constant and independent of the means, either by theoretical considerations or by an appropriate choice from the Box–Cox family (Atkinson 1985). Let σ_e^2 be the estimate of this constant variance, and let d_{ij} denote the pairwise distance between classes, measured by the absolute value of the difference in the class means, in the same units, i.e., d_{ij} is replaced by d_{ij}/σ_e. Clearly, such a replacement does not alter the value of the decision criterion, d_{kv}/δ_v; it follows that, other than in determining the metric, σ_e^2 plays no role.

If there is heterogeneity in the variances apparently unrelated to the means, or if the intrinsic variability is genetically controlled separately from the mean (such as may be true of different varieties or species), a transformation may be neither identifiable nor desirable. A proposal for these circumstances is now described. For each of the $\binom{n}{2}$ pairs of means, compute

$$d_{ij} = |\bar{z}_i - \bar{z}_j| / \text{ s.e.}(i, j),$$

where s.e.(i, j) is the estimated standard error of the difference between $\bar{z}_i$ and $\bar{z}_j$, and form the array

$$\mathbf{D} = \{d_{ij}\}.$$

D will be unidimensional if all s.e.(i, j) are equal but otherwise will almost certainly be multidimensional, for which the simple concept of betweenness has to be replaced. This increase in dimensionality leads naturally to the subjects discussed in Chapter VIII, where it is considered in detail.

An alternative, which may also lead to an increase in dimensionality, is to measure the dissimilarity, d_{ij}, between two means using the estimated first two moments, $\bar{z}_i$ and s_i^2 by

$$d_{ij}^2 = (\bar{z}_i - \bar{z}_j)^2 + (s_i - s_j)^2.$$

This Fréchet (neighborhood) distance (see Appendix 3) is zero iff both the means and variances are equal; it can be extended to include the differences in the cube roots of the third moments, and so on, should it appear that useful information is conveyed thereby. If the distributions of the variable for the two objects are known, then the squared Hellinger distance

$$d_{ij}^2 = \int (p_i^{1/2}(z) - p_j^{1/2}(z))\, dz,$$

where the integral, which is over the admissible domain of z, can be replaced by a summation for discrete ordered variables.

At this point, a number of subsets of the n objects have been found. Some subsets may contain few objects while others may contain larger numbers. There is no assumption of a common within-subset variance. Each accepted subset has been *found* independently of any other but is related to others by the extent that they are nondisjoint. The accepted family of subsets can be represented by the matrix **A** as defined in Chapter II. It is unlikely that the subsets either form a partition or even an irredundant covering. The problem of how to choose from among the generated subsets to obtain a partition of interest from an irredundant covering is discussed at length in Chapter II. To

summarize, a probability measure based on the relationships among the objects and subsets is given by the Perron–Frobenius column eigenvector, $\mathbf{p}$, of $\mathbf{A}^T\mathbf{A}^*$. The solution to a least-cost set-covering is found based on these probabilities. A covering solution is chosen because

- the covering may in fact also be a partition
- no partition may exist, and so requiring one will make the problem appear to have no solution
- the overlapping of two or more groups in the optimal covering indicates a sufficiently close relationship that they ought to be combined.

The third of these is the key, because acting on it converts the optimal covering into a partition. In particular, if the partition consists of just one group, there is a strong indication that there may be no distinct subgroups of any interest. Furthermore, this process of forming **musters** (Lefkovitch 1982) remedies some, if not all, the inadequacies of the subset generation phase, because in it, a subset that would have been initiated by members of now distinct components would probably have generated a subset of all n objects and have been rejected.

In addition to this solution, another programming procedure is now described, based on Calinski and Corsten's (1985) second principle, so that some of its properties in the present context can be recognized. Let the heterogeneity of the k^{th} subset be measured by its variance, σ_k^2 (or perhaps by δ_k^2, which is related to the variance), and define $\mathbf{A} = \{a_{ik}\}$ to be the incidence matrix as before. Then consider the linear least-cost set-*partitioning* problem, which is to find a binary vector, $\mathbf{x}$, that

$$\text{minimizes } \{\Sigma_k x_k \sigma_k^2 \mid \mathbf{Ax} = \mathbf{1}, x_k \in \{0,1\}\}$$

(M.M. Rao 1971). This program may not have a solution because the subsets in **A** may exhibit a pattern of relationships not allowing a partition. This possibility can be avoided by replacing the equality by $\mathbf{Ax} \geq \mathbf{1}$, i.e., to obtain a covering solution. Of more importance, however, is that the original data will have been used not only to generate the **A**, but also to obtain the σ_k^2 for the objective function, $\Sigma_k x_k \sigma_k^2$. The consequence of this double usage is that subsets of low cardinality are chosen, since the variances corresponding with them tend to be small (e.g., a group consisting of one object has $\sigma_k^2 = 0$ and necessarily is included in the optimal solution).

Yet another solution procedure, based on Calinski and Corsten's (1985) third principle, proves to have a similar consequence in a programming formulation; suppose each subset is characterized by the distance to the nearest member of the other subsets (Lefkovitch 1978). Then a solution that minimizes the number of groups and maximizes the sum of the distances among them seems of interest. Leaving aside the problem of how to measure the distance between nondisjoint subsets, it is clear that single-object subsets tend to be chosen, and that furthermore, the original data again will have been used twice.

The case study "ANOVA means" (Chapter X) considers a number of data sets used by Calinski and Corsten (1985). Conditional clustering as applied to the grouping of means gave much the same results as some other methods in four out of five examples. In these circumstances, what advantages are there in using it? It appears that they are largely theoretical, because the amount of computation involved is not prohibitive in any of them. A summary of the major differences between the procedure described here and that described by Calinski and Corsten (1985) are given in Table VI.1. That differences exist is apparent, and so the reasons for a choice must be based on theoretical considerations; in particular, I claim that

it is preferable to choose that class of procedures which appears to assume less and which exploit the consequences of these minimal assumptions as fully as possible.

Table VI.1 Comparison between the proposed procedure and those given by Calinski and Corsten (1985)

Conditional clustering	Calinski and Corsten
Subsets	
Generated independently	Generated simultaneously
Local significance test, with (implied) fixed α	Global significance test required, with user α
Together form a covering	Always form a partition
Largely distribution free	Assumed asymptotic normality
Independent of within-sample variance	Significance test depends on within-sample variance
Uses arithmetic of $O(n^2)$	Uses arithmetic of $O(n^3)$
Optimal solution	
Uses only the relationships among among generated subsets (i.e., data used once)	Given at subset phase, uses global significance test (i.e., data used twice)
Trivial solutions avoided unless necessary	Trivial solutions avoided automatically only by specifying a cutoff
Approximate procedure uses arithmetic of $O(n)$	Uses arithmetic independent of n

Conditional clustering uses the data to obtain subsets based Heuristic VI.2; using the maximum entropy principle, it obtains set of probabilities and a solution from the subsets without further recourse to the data. The other methods considered here use the means to generate the subsets but require them again, together with a significance test (and its assumed distributions) at a chosen level, to obtain the solution. Conditional clustering may produce a solution that is not a partition, but which can be formed into one, whereas the other methods are constrained to form a partition and thus impose one even if it is inappropriate. A covering solution obtained by the programming phase of conditional clustering, if not a partition, indicates that some groups are not well separated; but if they are not well separated, they represent a group somewhat more elongate (for the unidimensional problem) than the decision rule allows. Thus forming musters remedies some of any deficiencies in the subset-generating phase. A further property of the conditional clustering method is that it can easily be extended to the case of nonhomogeneous variances, and even further by using Fréchet (neighborhood) distances (Lefkovitch 1985*b*) to incorporate differences in distribution beyond the first two moments. However, it is in the subset-generating phases that the two classes of procedures are essentially different. In the Calinski and Corsten method using complete linkage, subsets are generated using a greedy algorithm; this algorithm has considerable computational power, being based on local optimality, but has the weakness that once a decision is made, it can never be undone within the framework of the algorithm, (which may explain the difference in the location of Q in data set 5 of "ANOVA means" in Chapter X). By contrast, in the subset-generating phase of conditional clustering, each subset is generated independently of any other, and so decisions made about the membership of an object to a subset, which are conditional on the current members,

may differ from that to another subset, even if the two subsets ultimately are found to overlap.

Stanfel (1982) described another procedure for partitioning unidimensional data, which can be written as an integer-programming problem with linear constraints and a nonlinear objective function. The objective function, the difference between the average within-group and average between-group distances, is of some interest; but because it apparently does not seem to have a natural extension for overlapping subsets, and also because the program itself appears unsuitable for finding coverings that need not be partitions, Stanfel's proposal is not considered further in the present study.

VII Scalar dissimilarity coefficients

Although scalar dissimilarity between a pair of objects is a reasonably well-discussed parameter, it is usually based on attributes (see Chapter III), i.e., concepts more primitive than a direct measure. Curiously, in spite of many publications discussing dissimilarity (or similarity) in numerical taxonomy,[1] its definition seems to have no more content than it is "the number obtained as the result of some specified sequence of calculation." Because this situation is somewhat unsatisfactory, a definition based on these primitive concepts is given after some preliminary remarks.

An object, *i*, is represented by the state it exhibits of each of a number of attributes. As in Chapter III, an attribute is a subdivision of an object exhibiting one of a number of states; sometimes the number of states is finite, and sometimes the possibilities map to some continuous subset of the real line. In any particular context, the delimitations of an attribute, and the range of conditions subsumed by a single state are defined empirically. Each state is demarcated according to the degree of resolution of the measuring instruments (such as eyes, rulers). There may be some pooling of the states, if the resolution is thought to be too fine, or if a computer program imposes restrictions. The number of attributes theoretically definable is infinite, but in practice, a small (finite) sample is chosen because that set happens either to

[1] However, in a mathematical context, similarity has received some discussion; I recommend Höhle (1988) as an important recent contribution. Unfortunately for the present context, one of his axioms is a generalized transitivity, which I am loath to accept.

be accessible, or is of interest to the taxonomist (and therefore represents a bias). Arguably, there is much subjectivity in the choice of attributes and definition of states.

Similarity as a probability

It is convenient to begin this discussion with the estimation of similarity coefficients, for which it is necessary to make explicit what is being estimated. Consider the following geometric model. Imagine two congruent, nonisosceles triangles each of which is completely colored by some unknown, nonoverlapping regions of blue and red; there is also a template identical in outline to the triangles, but with holes of varying shapes, sizes, and position. The template is placed on each triangle in turn, and the revealed pattern of colors is recorded. Some holes may correspond entirely with red in both triangles, some entirely with blue, some with red in the one and blue in the other; others may be partly red and partly blue. The objective is to determine the probability that in a randomly chosen perforation, the two color patterns on the triangles are the same. If the perforations in the template are chosen at random, it seems reasonable to estimate this probability by the ratio of the area exposed by the perforations in the template in which both triangles show the same color to the total area of the perforations.

Note the following correspondence between this triangle model and the definition of attributes:

(1) The number of points on any triangle (and the number of points in each exposed region) is nondenumerably infinite, as is the number of attributes of an object.

(2) The number of perforations in any template is finite, as is the number of empirically chosen attributes, even though

the number of possible choices and delimitations of perforations in the templates is denumerably infinite.

If the two triangles correspond with two objects in which homologies can be identified with (reasonable) certainty (e.g., the vertices of the triangles, and the positions of the perforations), a single attribute corresponds with a single perforation, and the colors with its possible states. If the perforations of the template are chosen randomly, this situation corresponds with a random choice of attributes; if in some biased way, e.g., along the edges, corresponding with choosing attributes on objects known to be diagnostic or are known not to be diagnostic, the estimated probability will be biased.

Leaving the triangular model, which implies the existence of some unambiguously identifiable landmarks, suppose there are m' attributes, where m' is denumerable; then the similarity of two objects as used in this book is as follows:

DEFINITION VII.1. *The similarity, s_{ij}, between objects i and j is the probability that an attribute chosen at random in object i shows the same state as the corresponding attribute in object j.*

This definition assumes not only that attributes can be unambiguously identified in the two objects, but also that each and their corresponding states are homologous. If all m' attributes can be examined, measuring s_{ij} presents no difficulty; but because only a very small number, m, are ever examined, chosen for reasons of accessibility, then some estimation procedure is needed. Presumably, consistency should obtain, i.e., that for each set of attributes unbiasedly chosen, the estimated similarity would be about the same.

Let the set of attributes be denoted by Z, where z_{ik} is the state of the k^{th} attribute of object i; $|Z|$ may be finite for practical purposes, but in fact is infinite. Suppose each member of N is described by the states shown by each of $|Z| \geq 1$ attributes, and let a characteristic function be defined as

$$s_{ijk} = \begin{cases} 1, z_{ik} = z_{jk} \\ \qquad\qquad , k = 1 \ldots |Z|. \\ 0, z_{ik} \neq z_{jk} \end{cases}$$

Then

DEFINITION VII.2. *The similarity between objects i and j, denoted by s_{ij}, is given by*

$$s_{ij} = E(s_{ijk}),$$

where E denotes the expectation operator over Z.

The whole class of similarity coefficients can be regarded as being based on differently interpreted characteristic functions.

Similarity is the joint probability that an attribute chosen at random in two objects shows the same state. Let $\Pr(i, j)$ be such a probability; it is estimated by s_{ij}. Assuming (local) independence of the two objects, it is now postulated that

$$\Pr(i, j) = \Pr(i)\Pr(j),$$

where $\Pr(i)$ represents some property of object i now to be interpreted. Suppose another object, i', belonging to the same "very local subgroup" (e.g., a sibling of object i) had been chosen in place of i; $\Pr(i)$ is the similarity between object i and i', i.e., $\Pr(i) = \Pr(i,i') \leq 1$.

The values of s_{ij} can be substituted for $\Pr(i, j)$, but s_i is unknown; what we have is s_{ii}, which is unity for all objects. Replacing $\Pr(i, j)$ by s_{ij} gives

$$s_{ij} = s_i\, s_j,$$

or in logarithmic form,

$$\log_e s_i + \log_e s_j = \log_e s_{ij}.$$

The problem, therefore, is to estimate s_i, for which the only data are the $\{s_{ij},\ i = 1...n, j = 1...n, j \neq i\}$. For $n = 2$, there can be an infinite number of solutions for the two unknowns, although this set may be limited somewhat by requiring that $1 \geq \{s_i, s_j\} \geq s_{ij}$. For $n = 3$, there are three equations in three unknowns, but the constraints may not allow an exact solution; $n > 3$ leads to the (overdetermined) set of $n(n - 1)/2$ equations in n unknowns

$$\begin{bmatrix} 11... & \text{etc} & .. \\ 1.1.. & \text{etc} & .. \\ & \text{etc} & \\ ..... & \text{etc} & 11 \end{bmatrix} \quad \begin{bmatrix} \log_e s_1 \\ \log_e s_2 \\ ..... \\ \log_e s_n \end{bmatrix} \quad = \quad \begin{bmatrix} \log_e s_{12} \\ \log_e s_{13} \\ ..... \\ \log_e s_{n-1,n} \end{bmatrix}$$

which can be solved by minimizing some norm, subject to the constraints that $1 \geq s_i \geq s_{ij}$, $\forall j{=}1...n$ (i.e., the array is diagonally dominant). There may be no solution unless some rank for the final similarity array is specified; because there is a resemblance to determining the communalities in factor analysis, methods for obtaining estimates of communalities in that procedure may be used to determine a solution. Regarding s_i as a binomial parameter, the variance within the "very local subgroup" containing object i can be approximated by $ms_i(1 - s_i)$.

Because similarity is here regarded as a probability, it can be generalized to define the similarity between sets of objects. Let I be a subset containing $|I| \geq 1$ objects, and J of $|J| \geq 1$ objects; there is no assumption that I and J are disjoint, and in fact they can be identical. Then the definition of the similarity between I and J is

$$s_{IJ} = \mathrm{E}(s_{ij} \mid i \in I, j \in J),$$

the expectation being over Z. This generalized definition collapses to the simple one if I and J consist of one object; it also defines a measure of the similarity of an object to a group of which it can be a member, and for nondisjoint groups in general.

This definition can be expressed as follows:

DEFINITION VII.3. *The similarity between two nonempty subsets is the probability that the state of an attribute chosen at random in an individual chosen at random from the one subset is identical with that of the same attribute in an individual chosen at random from the other subset, i.e.,*

$$Pr(z_{ik} = z_{jk} \mid k \in Z; i \in I, j \in J).$$

It follows that even if $I = J$, this probability need not be unity unless the members of I are identical with respect to Z, e.g., if I consists of one object.

The probability of identity of the states of attributes of subset J conditional on the states shown by subset I, dropping reference to Z for convenience, is now defined by

$$\Pr(z_{jk} = z_{ik}) \,/\, \Pr(z_{ik}),$$

while conditional on the states shown by both i and j, it is

$$s_{ij} = \Pr(z_{jk} = z_{ik}) \,/\, (\Pr(z_{ik})\ \Pr(z_{jk}))^{1/2}.$$

Since similarity is considered to be a probability, then so is dissimilarity, which is the probability that the objects do not show the same state for an attribute chosen at random, i.e.,

$$d_{ij} = 1 - s_{ij}.$$

The mean dissimilarity between an object and the members of a set, S, now satisfies the relationships

$$\begin{aligned} w_i &= 1 - \mathrm{E}(s_{ij}) \\ &= 1 - s_i(S) \end{aligned}$$

where $s_i(S)$ is the similarity of object i to S. The mean dissimilarity among members of a set becomes

$$\delta = 1 - s(S)$$

where

$$s(S) = \mathrm{E}(s_{ij}).$$

If similarity is no more than the value resulting from some calculation, and if all attributes are discrete (nominal or ordinal), the similarity space (and hence the dissimilarity space) cannot be continuous, as is implicit in many of the algebraic operations seen in cluster analysis. Considered as a probability, however, continuity is assured, even though for a given set of objects and attributes, there can be only a finite set of possible values.

Even though similarity is often the framework used to describe clustering algorithms, it is in terms of distance that they are understood; Appendix 3 describes some necessary properties of a distance space. Two objects for which $s_{ij} = 1$ have a distance

d_{ij} of zero; and two for which $s_{ij} = 0$ are separated by the maximum distance possible. Thus similarity, which is a value in the (open) interval (0,1), corresponds to a distance in the open interval $(\infty,0)$, and there is a one-to-one mapping between these measures. Thus to obtain the distance between sets of objects in the context of clustering, a transformation, $d_{ij} = f(s_{ij})$, is needed to satisfy the functional relationships:

$$\begin{gathered} 0 \leq d_{ij} \leq \infty \text{ for } 0 \leq s_{ij} \leq 1, \\ \infty = f(0), \\ 0 = f(1), \\ s_{ij} > s_{ij'} \Leftrightarrow d_{ij} < d_{ij'}. \end{gathered}$$

Because dissimilarity is a probability, these conditions show that dissimilarity, as defined above, is not the same as a distance. Assuming independence among the objects, it follows that f should satisfy

$$f(s_{ij}s_{ij'}) = f(s_{ij}) + f(s_{ij'}),$$

which is one form of Cauchy's equation (Saaty 1981). (Note that this functional relationship also underlies the entropy of mutually exclusive events.) With these conditions, a solution is

$$f(s_{ij}) = -\alpha \log_e(s_{ij}),$$

where α is an arbitrary positive constant (conveniently taken to be unity). This distance can be recognized as the Hartley information relevant to the states of the attributes. It is easy to show that $d_{ij} = -\log_e(s_{ij})$ is only a semimetric, since the triangle inequality need not be satisfied. In fact, what is defined is a preordering, $\leq$, on pairs of objects for which $d_{ii} = 0$, $\forall i$, and for which $d_{ij} = d_{ji}$, $\forall i, j$. In the next chapter, this definition of distance allows an

interpretation of subset generation to be that objects are included in a subset until too much chaos is created, i.e., until the probability of identity of attribute states in two objects chosen at random becomes too small in relation to some initial condition.

The metric and other properties of dissimilarity coefficients were discussed by Gower and Legendre (1986); Baulieu (1989) proposed an axiomatic system for presence–absence dissimilarity coefficients and examined 21 of them to see if they satisfy the axioms. However, neither publication appears to be clear about what a dissimilarity is supposed to be measuring, although they both offer interesting comments on the properties of some well-known coefficients. Bacelar-Nicolau (1987), by means of an integral transform of similarity

$$\text{Prob}(s_{ij} \leq s^*),$$

showed that many of the binary coefficients described below are distributionally equivalent, and that for arrangements of the empirical data, such as in Table VII.1 under an assumption of fixed margins and independence, a standard normal distribution obtains for the transformed estimate. Even if there is disagreement with interpreting similarity as a probability, the probability, which is the integral transform above, is also a measure of similarity.

Table VII.1 Two-way frequency table for m binary (i.e., one- or two-state) attributes for two objects

		Object i state		
		1	0	sum
Object j state	1	a	b	$a + b$
	0	c	d	$c + d$
	sum	$a + c$	$b + d$	m

Attribute-based dissimilarity

With this understanding of distance, dissimilarity, and their functionally related similarity (it is assumed that a one-to-one relationship exists between them), a more-detailed study can be undertaken. The rest of this chapter discusses three main subjects. The first is the estimation of similarity from attributes and examines a number of well-known coefficients in the light of its definition. Different types of attributes require different treatment, so there is a need to consider not only how the attributes were chosen but also how to combine them. The second is directly measured pairwise relationships, and how they can be converted into dissimilarities and distances; because several types of pairwise relationships need to be considered, each requires its appropriate treatment. The third subject is the role of scalar distances in the context of clustering and describes a nonmetric transformation, which is likely to emphasize the separation of distinct subsets, but which does not change their internal structure.

In trying to define a similarity as a probability, several technical problems arise because of the subjective nature of the attributes and their states. As a result of this subjectivity, the attribute population, called Z, is ill-defined, even more so than their states. Moreover, it is often true in practice that what is noticed about the objects are local differences, which then are combined to define an attribute. Leaving aside these technical problems, the *intent* of the definition is clear and allows the examination of similarity coefficients as estimates of this probability. If similarity is not a probability but a separate concept, the following evaluation of the various coefficients will need a different emphasis.

Similarities for binary data

A variety of empirical similarity coefficients have been proposed for what are loosely called binary data; a survey of the literature

using these in real applications rarely discusses the appropriateness of the one chosen, or if more than one is used, why, except to make comments along the lines "the results obtained are consistent with the classification obtained by ...." It has already been noted (Chapter III) that there is often a lack of appreciation that attributes called "binary" may be one of two distinct categories, namely, dichotomies and alternatives (Gower 1971*a*), (one-state and two-state attributes, respectively, in the terminology of Chapter III), and that a set of binary attributes may contain both. The distinction between the two kinds of binary attributes is clear in definition but is not always so in practice; a one-state attribute can be present or absent but, in particular, is characterized by the fact that information is given on the similarity of a pair of objects iff it is present in at least one of the pair, and that no information is given by it on the similarity if it is absent simultaneously from them. By contrast, a two-state attribute is characterized by the fact that two objects are equally similar if they agree in the state shown by the attribute; if they disagree, they are dissimilar. With this distinction in mind, a number of well-known empirical similarity coefficients (for some others, see Wolda 1981) proposed for binary data can be classified as being appropriate for one-state, two-state, or are ambiguous.

Let there be m "binary" attributes, each of which may be scored 1, which are taken to indicate presence of a particular state of the attribute, or 0, which indicates either absence of the attribute, or presence of the alternative state. Assuming no unassessable (missing) values, a contingency table can be constructed for two objects, i and j (Table VII.1). In that table, each letter represents the frequency of (binary) attributes in the available data exhibiting the marginally indicated states simultaneously in objects i and j. The symbol, $\hat{s}_{ij}$ is used to denote the estimate of similarity between objects i and j; if more than one proposed estimate of similarity is discussed, the notation $\hat{s}_{ij}(.)$ is used, where (.) gives the originator's name.

Comments on ten similarity coefficients for binary data

For one-state attributes

(1) Jaccard (1908):

$$\hat{s}_{ij} = a/(a + b + c).$$

Since neither d nor m is involved in this coefficient, it follows that it is appropriate for one-state attributes. If the attributes are chosen at random, it seems reasonable to assume that this coefficient is an unbiased estimator of the probability that object i and j show the state for a randomly chosen one-state attribute.

(2) Kulczynski (1927), first coefficient:

$$\hat{s}_{ij} = a((a + b)^{-1} + (a + c)^{-1})/2$$

This coefficient is appropriate for one-state attributes, since neither d nor m is involved; its numerical value is in the interval [0,1]; it is not clear if it should be considered as a probability. If the attributes are chosen at random, it is biased upwards.

(3) Czekanowski (1932):

$$\hat{s}_{ij} = 2a/(2a + b + c)$$

For one-state attributes, since neither d nor m is involved, and if the attributes are chosen at random, it is clearly biased upwards.

(4) Ochai (1957):

$$\hat{s}_{ij} = a/((a + b)(a + c))^{1/2}$$

For one-state attributes, by including a twice in the denominator, while b and c are included once, the estimate is biased upwards.

(5) Sokal and Sneath (1963):

$$\hat{s}_{ij} = a/(a + 2(b + c))$$

For one-state attributes, since the number of mismatches is included twice in the denominator, this estimator of the probability is biased downward.

For two-state attributes

(6) Jaccard (1912), Dice (1945), Sorensen (1948):

$$\hat{s}_{ij} = 2(a + d)/(2a + 2d + b + c)$$

This coefficient is a generalization of that of Czekanowski (see coefficient 3); if the attributes are chosen at random, this estimate is biased upwards.

(7) Kulczynski (1927), second coefficient:

$$\begin{aligned}\hat{s}_{ij} &= (a + d - b - c)/(a + d) \\ &= (m - 2(b + c))/(m - (b + c)) \\ &= 1 - (b + c)/(a + d).\end{aligned}$$

Although mismatches are ignored in the denominator, the numerator is also reduced by double the amount, and so for randomly chosen attributes, this estimate is biased. Because it can be negative infinite, it cannot be a probability, but since its maximum value is unity, it can be transformed to lie in the interval [0,1] as

$$\hat{s}'_{ij} = \exp(\hat{s}_{ij} - 1).$$

(8) Russell and Rao (1940):

$$\hat{s}_{ij} = a/m$$

This estimate involves m; but by not using d in the numerator, the properties of this estimate are somewhat ambiguous. If m is relevant, it seems curious that d is not used in the numerator; assuming random sampling, it is a downwards-biased estimate.

(9) Zubin (1938), Dumas (1955), Sokal and Michener (1958):

$$\hat{s}_{ij} = (a + d)/m$$

Assuming a random choice of attributes, this coefficient is unbiased. Sometimes called the simple matching coefficient, it can be recognized as the probability that objects i and j both show the same state for a randomly chosen two-state attribute.

(10) Rogers and Tanimoto (1960):

$$\hat{s}_{ij} = (a + d)/(a + d + 2(b + c))$$

This coefficient parallels that described by Sokal and Sneath (see coefficient 5) and is a downwards-biased estimate of the probability, because the number of mismatches is included twice in the denominator.

After examining these ten coefficients, it seems that only coefficient 1, for one-state attributes, and coefficient 9 (possibly also 7), for two-state attributes, provide unbiased estimates of similarity, as here interpreted as a probability, given that the attributes are uniformly and randomly chosen. If they are not chosen in this way, then others may provide appropriate estimates.

For example, if the attributes are chosen randomly from a subset *known* to be diagnostic, then it may be desirable to give a reduced weight to the agreements and to emphasize the disagreements (such as in coefficients 5 and 10), so to reduce the bias. By contrast, if the choice of attributes has avoided those known to be diagnostic, then emphasizing agreements in comparison with the mismatches (such as in coefficients 3 and 6) also tends to reduce the bias. However, each coefficient may be appropriate given the particular choice procedure for the attributes. Thus in estimating the similarity among a set of objects, the first problem is to determine how the attributes were chosen, and then (if all are binary) to group them into one-state and two-state, and to subdivide them further into subsets "known to be diagnostic," "known not to be diagnostic," and "random" (or noncharacterizable). This classification allows a two-way table to be formed (Table VII.2), which gives a coefficient in each category using the list above. Because an overall measure of similarity is the objective, the six possible values may need to be combined in some way. Let $\hat{s}_{ij}(p,q)$ represent the $(p,q)^{\text{th}}$ cell of Table VII.2, with missing combinations

Table VII.2 Appropriateness of various estimates of similarities for binary attributes; ξ_{ij} denotes any similarity coefficient for an i-state attribute belonging to the j^{th} diagnostic/nondiagnostic/random category

	One-state coefficient	Two-state coefficient
Known to be diagnostic	5* ξ_{11}	10 ξ_{12}
Known to be nondiagnostic	3(2?) ξ_{21}	6(7?) ξ_{22}
Random (not characterized)	1 ξ_{31}	9 ξ_{32}

* The numerical values refer to the list of 10 binary coefficients in the text.

set to zero; because similarities are assumed to be probabilities, it seems that the best combined estimate is

$$\hat{s}_{ij} = \exp\left(\xi^{-1} \Sigma_{p=1\ldots3}\Sigma_{q=1,2}\, \xi_{pq}\log_e \hat{s}_{ij}(p,q)\right) \mid \hat{s}_{ij}(p,q) > 0,$$

where $\xi = \Sigma_p\Sigma_q\xi_{pq}$.

The definition of any one of the similarity coefficients in Table VII.2 assigns equal weight to each considered attribute; those attributes not considered are given zero weight. However, the attributes used may also be given unequal nonzero weights to reflect their *known* prior importance. For example, in Zubin's coefficient (9), if w_k is the weight assigned to the k^{th} attribute, and if $\delta_{ijk} = 1$ (if objects i and j both show the same state for attribute k), and is zero otherwise, then

$$\hat{s}_{ij} = \Sigma_k \delta_{ijk} w_k / \Sigma w_k$$

is a weighted estimate of the similarity. Analogous definitions for the other coefficients are easily formed. However, other than by subjective decisions, it is not always clear from where these weights are to be obtained. Because the empirical nature of attributes does not ensure their mutual independence, the weights may be chosen to minimize the effects of duplication of information.

Since the Jaccard (1) and Zubin (9) coefficients are those most likely to be used, it is useful to show that they differ only in what constitutes the same state. Suppose the marginal classification sets of Table VII.1 are replaced by "same state" and "different state," as in Table VII.3, and $\{\alpha,\beta,\gamma,\delta\}$ are the number of different cases observed. Clearly, $\delta \neq 0$ is impossible, and, by symmetry, the assignment to β and γ is arbitrary, and so both can be assigned to β; thus the table collapses to two entries, namely, $\alpha' = \alpha$ and $\beta = \beta' + \gamma'$. It follows that the Jaccard and Zubin

Table VII.3 Alternative tabulation for binary attributes*

		Object i	
		same state	different state
Object j	same	α'	$ß'$
	different	γ'	δ'

* For explanation of how this arrangement is derived from Table VII.1, see text.

similarity are both $\alpha/(\alpha + ß)$, with the distinction between them as to what constitutes "same state"; for the Jaccard coefficient, joint absence is ignored. In terms of Table VII.1, for the Jaccard coefficient,

$$\alpha = a \text{ and } \alpha + ß = a + b + c,$$

while for the Zubin coefficient,

$$\alpha = a + d \text{ and } ß = a + b + c + d = m.$$

An application of the Jaccard coefficient in a context outside of clustering is given in Chapter X, case study P.

Similarities for multistate unordered attributes

Suppose an attribute has $s > 2$ states; the traditional procedure is to continue as for two-state attributes, i.e., to construct Table VII.1, and then to compute the similarity as before (Gower 1971, Lefkovitch 1976). However, there may be a defect in this practice.

Consider one such attribute in object i; there is only one way object j can be alike object i with respect to this attribute, but s - 1 ways it can differ. Assuming a uniform random distribution, it can be argued that since the probability that j is alike object i by chance is $1/s$, and of being different is $(s - 1)/s$, then it is reasonable to weight these possibilities so that the "score" assigned to resemblance equals the combined scores of all possibilities of being different. In terms of α and ß of Table VII.3, the computation can be expressed in the pseudocode given in Table VII.4. For example, let $s = 3$ for all attributes, where one of the states of each is absence, but the attribute, when present, can be in one of two conditions. Then the Jaccard coefficient becomes $2a/(2a + b + c)$, which is the same formula as that for the

Table VII.4 A general algorithm for computing similarities

```
input:   s is the number of states for the k^th attribute;
         x is the entry for object i, attribute k;
output: similarity coefficients.
for i:= 2...n
   for j:= 1...i-1
      α:=ß:= s(i,j):=0;
      for k:= 1...m
         if (x_ik not comparable with x_jk) go to next k;
         if (a_t is unordered) then if (x_ik = x_jk) α:= α+0.5
                                    else ß:= ß+1/2(α -1)
                               go to next k;
                               else y:= abs(x_ik-x_jk);
                                    y:= sin^-1(sqrt(y));
                                    α:= α+y; ß:= ß+1-y;
         end k;
         if((α+ß)>0) then s(i,j):= α/(α+ß)
      end j;
   end i;
end
```

Czekanowski coefficient (3), which, for one-state attributes, is appropriate for attributes chosen with the belief that they are nondiagnostic (Table VII.2). The Zubin coefficient (9) generalized to three-state attributes is clearly the same as the Jaccard–Dice–Sorensen coefficient (6).

Similarity based on randomness

If it can be assumed that the attributes have been chosen without any deliberate bias, then another way to estimate similarity for unordered attributes becomes of interest.

Case 1: one-state attributes For object i, those attributes present represent a description of it; the zero elements represent absent attributes. The number of the latter is finite for the attributes under consideration but is clearly infinite for all possible attributes. Thus the information available on presence can be regarded as conditional on the subset of attributes, of which there are m in Table VII.1. Suppose that presence (=1) and absence (=0) are regarded as Bernoulli random variables, and that a + b and a + c in Table VII.1, the number of attributes equal to unity in objects i and j, respectively, are fixed quantities. Given the assumptions, the probability of obtaining the observed number of attributes simultaneously equal to unity in both objects, given the total m, can be regarded as an estimate of similarity and obtained using combinatorial arguments as

$$\Pr(a) = \binom{a-b}{a}\binom{m-a}{c}/\binom{m}{a+c}$$

$$= \binom{a+c}{a}\binom{m-a}{b}/\binom{m}{a+b}$$

$$= (m-a)!(a+b)!(a+c)!/(m!a!b!c!).$$

This quantity can be calculated (for large x) using the (modified) Stirling approximation (McCullagh and Nelder 1989),

$$x! \approx (2\pi(x + 1/6))^{1/2} x^x e^{-x},$$

which is an accurate approximation and is defined even for $x = 0$.

Example 1: One-state attributes (1 = presence, . = absence).
Object i: {11111.....}; Object j: {1111.11...};
a = 4, b = 1, c = 2, d = 3, m = 10
Pr = 5/14 = 0.3571; Jaccard similarity = 4/7 = 0.5714

Case 2: s-state attributes Using the reasoning for case 1, if each of the s-state attributes is replaced by s one-state attributes, the probability can be calculated as above. Because the value of m changes from the number of attributes to the sum of the number of states, the calculated value must be multiplied by the number of attributes and divided by the number of states (this factor is unity for one-state attributes).

Example 2: Same data as example 1, transformed for two-state attributes.
Object i: {1.1.1.1.1..1.1.1.1.1}; Object j: {1.1.1.1.1..1..1..1.1.1};
a = 7, b = 3, c = 3, d = 7, m = 20
Pr = 10*143/(20*323) = 0.2214; Zubin similarity = 7/10 = 0.7

Having computed the probability of a random association under a Bernoulli assumption, it becomes possible to estimate the variance, and also to compare it with hypothesized values.

A correlation between individuals for qualitative attributes

Consider two objects described by m unordered attributes each having at least one state. Let A be the number of attributes in which both objects agree in their state, and D the number of attributes in which they disagree, i.e., $A + D = m$.

DEFINITION VII.4. *The binary correlation between individuals i and j is*

$$r_{ij} = (A - D)/(A + D).$$

The justification for regarding this expression as a correlation follows from the fact that $-1 \leq r_{ij} \leq 1$; the bounds are achieved as follows:

(1) If $D = 0$, then $A = m$, so that $r = 1$, implying complete *positive association* between the states of the attributes shown by the two objects.

(2) If $A = 0$, then $D = m$, so that $r = -1$, implying complete *negative association* between the states of the attributes of the two objects.

(3) If $A = D$, then $r = 0$, and there is *no association* between the states of the attributes shown by the two objects.

A vector form of the definition is easily obtained. Assume the m attributes are two-state, and no values are missing; let state 1 of each attribute be scored as 1, and state 2 as -1. Denote by $\mathbf{x} = \{x_i\}$ the column vector of these scores for one object, and $\mathbf{y} = \{y_i\}$ for another. Then since the quadratic norms of each vector are $m^{1/2}$, it follows that

$$r_{xy} = \mathbf{x}^T\mathbf{y}/(\|\mathbf{x}\| \, \|\mathbf{y}\|) = m^{-1} \sum x_i y_i = m^{-1}(A - D).$$

Since r_{xy} is the inner product of two vectors of unit norm, it can be regarded as giving the cosine of the angle between them, i.e.,

$$r_{xy} = \cos \Theta_{xy}.$$

Thus distance can be considered to be given by Θ_{xy}, which is in the interval $(0,\pi)$. Missing values in either **x** or **y** or both are replaced by zero, and the computing procedure represented by $r_{xy} = \mathbf{x}^T\mathbf{y}/(\|\mathbf{x}\|\,\|\mathbf{y}\|)$ is used.

Comparison with some qualitative similarity coefficients described above is of interest. Using the notation in Table VII.1, if the attributes are all one-state (Chapter III), d is by definition zero, and so $A = a$, $D = b + c$ resulting in

$$r_{ij} = (a - b - c)/(a + b + c).$$

The difference from the Jaccard coefficient (1) of similarity, $a/(a+b+c)$, is

$$s_{ij} - r_{ij} = (b + c)/(a + b + c).$$

If the attributes are all two-state, $A = a + d$, and $D = b + c$, so that

$$r_{ij} = (a + d - b - c)/(a + b + c + d).$$

The Zubin coefficient (10) is $(a + d)/(a + b + c + d)$, so that their difference is

$$s_{ij} - r_{ij} = (b + c)/(a + b + c + d).$$

Perhaps the similarity coefficient having the closest resemblance to r_{ij} is the second coefficient of Kulczynski (7), which is defined as $(a + d - b - c)/(a + d)$; thus their ratio is

$$r_{ij}/s_{ij} = (a + d)/m,$$

i.e., Zubin's coefficient (9). Note that the definition using A and D is equally applicable to attributes having more than two states.

Similarities for ordered attributes

An important class of ordered attributes consists of a string of symbols, e.g., DNA sequences. The basic measure of distance between two strings in the Levenshtein; it consists of the minimum number of substitutions, deletions, and insertions that converts one string into the other. A dynamic programming algorithm for this is given by Kohonen (1989).

In the following discussion of a purely attribute-based estimate of similarity and dissimilarity for a set of objects based on a set of continuous variables, it is assumed that each numerical variable has been transformed so that a unit difference represents the same amount of biological difference throughout the whole range. The unit may differ from one variable to another (indeed, they may be quite different kinds of attribute, e.g., lengths measured in millimetres, weights measured in kilograms), yet to obtain a meaningful estimate of similarity, they need to be combined in some way so that none dominates the remaining purely by reason of the units of measurement. Thus it is necessary to find a scaling for each variable so that this domination does not occur, i.e., an estimate of the similarity has to be invariant under admissible transformations of the variables. One way to achieve this scaling is to arrange that each variable has a mean of zero and a variance of unity; another is to re-scale each of the variables so that the minimum observed value is mapped to zero, and the maximum to unity; Milligan and Cooper (1988) show by simulation that the latter rescaling tends to reveal the true clusters. This "uniforming transformation" (the terms "standardized" and "normalized" have special interpretations and are here avoided) is generally applicable to all variables, including multistate ordered, even though extreme *aberrant* values tend to "squeeze" up the remaining.

In more concrete terms, the preliminary step ensures that a unit difference in the score has the same meaning with respect to similarity throughout the whole range, i.e., the difference between two measurements, $|y_i - y_j|$, denotes equal amounts of dissimilarity throughout the range of the attribute (see Chapter III). If this equality is not true, then it is necessary to rescale the difference to be

$$|y_i - y_j|/c_{ij}$$

where c_{ij} is a scale (units) factor which can depend on i and j. To ensure that the estimate of the maximum dissimilarity cannot exceed unity, so that the estimate of the minimum similarity is zero, each of the $\binom{n}{2}$ different values is divided by the largest among them and then subtracted from unity. For constant c_{ij}, this maximum is the observed range of y. The set of values for all n objects for y can now be assembled into a $n \times n$ array, which has unities on the diagonal, from which principal coordinates can be computed, if desired. For a single attribute, there is just one principal coordinate of positive length if c_{ij} is constant, but the rank may be greater if c_{ij} depends on i and j. For rank unity, it is easy to see that the array consists of the values defined by Gower (1971*a*) for similarity, namely,

$$\hat{s}_{ij}(y) = 1 - |y_i - y_j|/\text{range}(y),$$

which is easily shown to be very much the same as

$$\begin{aligned} \hat{s}_{ij}(\theta) &= \sin\theta_i \sin\theta_j + \cos\theta_i \cos\theta_j \\ &= \cos(\theta_i - \theta_j), \end{aligned}$$

where $$\theta_i = \tfrac{1}{2}\pi(y_i - \min(y))/\text{range}(y)$$

(Lefkovitch 1976). For $i = j$, the estimate, $\hat{s}_{ij}(\theta)$, collapses into the familiar $\sin^2 + \cos^2 = 1$, which for $|y_i - y_j| = 1$ takes the value of $\cos(\pi/2)$, i.e., zero. A geometric model comparing these two is easily constructed; the Gower value is the complement with unity of the linear distance between the ends of two unit vectors radiating from a common origin, while the trigonometric value is the cosine of the arc distance. In terms of distances, these two hardly differ for estimates of similarity approaching zero or unity. They differ most in the vicinity of the midpoint, with a maximum deviation of

$$|2(1 - \sin(\pi/4)) - \surd(0.5)| \approx 0.121.$$

Thus objects showing a Gower similarity of 0.5 for a variable show one of about 0.621 for the trigonometric definition. The *prima facie* reason to prefer the trigonometric definition over that of Gower is the computational convenience of allowing the use of linear algebraic procedures, that it can be easily generalized for combining all classes of attributes, and thereby for incorporating unequal weights. To represent $\hat{s}_{ij}(\theta)$ in vector notation, define the two-element row vector,

$$\mathbf{z}_i = [\sin(\theta_i) \quad \cos(\theta_i)],$$

from which it follows that

$$s_{ij}(\theta) = \mathbf{z}_i\mathbf{z}_j^{\mathrm{T}}.$$

If all m ordered variables are each replaced by the two elements $\mathbf{z}$, the $n \times m$ matrix $\mathbf{Y}$ is replaced by the $n \times 2m$ matrix $\mathbf{Z}$, from which similarity, computed as the average

$$\hat{s}_{ij}(\theta) = m^{-1}\Sigma_k s_{ijk},$$

is given by the $n \times n$ array

$$\mathbf{S} = m^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}.$$

If $\mathbf{W}$ is a diagonal matrix of weights, e.g., $\mathbf{W} = \mathrm{diag}\{\mathbf{Z}\mathbf{Z}^{\mathrm{T}}\}$, the definition can be further generalized to

$$\mathbf{S} = \mathbf{Z}\mathbf{W}^{-1}\mathbf{Z}^{\mathrm{T}};$$

$\mathbf{W}$ may be permitted to be a more general matrix.

If an object under study is made up of several not necessarily identical units, the corresponding subvector in $\mathbf{Z}$ can easily be modified to accommodate the possibility that more than one (mutually exclusive) state is represented by the different units. Consider a two-state unordered attribute, for which there is a proportion α for the first state, and $1 - \alpha$ for the second. In terms of the inner product, a two-element vector is sought such that the inner product with itself is unity but with another such vector is zero iff they are orthogonal (orthogonality can occur iff $\alpha_i = 1$ and $\alpha_j = 0$). A two-element subvector, with elements $\alpha^{1/2}$ and $(1 - \alpha)^{1/2}$, satisfies these conditions and can also be extended to $(s > 2)$-state attributes. It is interesting to note that if α and β are two such row vectors, then

$$(\Sigma(\alpha^{1/2} - \beta^{1/2})^2)^{1/2}$$

is the discrete Hellinger distance between two multinomial vectors, which in turn is a special case of the continuous Hellinger distance

$$\sqrt{(\int (p_i^{1/2}(x) - p_j^{1/2}(x))^2\, dx)}.$$

Thus if an object under study is made up of several not necessarily identical units, an appropriate subvector can be incorporated in $\mathbf{Z}$.

Note that the value of m remains unchanged in computing S; it is the number of distinct attributes, *not* the number of columns in Z.

From this definition of similarity, the distance between objects i and j, as described earlier in this chapter, is defined by

$$d_{ij} = -\log_e \hat{s}_{ij}$$

which is not the same as dissimilarity, which is $(1 - s^2_{ij})^{1/2}$.

Having defined similarity for ordered attributes, it needs to be combined with that for unordered attributes. This combination can be made by incrementing α by either $\hat{s}_{ij}(y)$ or $\hat{s}_{ij}(\theta)$ and incrementing ß by either $1 - \hat{s}_{ij}(y)$ or $1 - \hat{s}_{ij}(\theta)$ in Table VII.4 to achieve a combined measure of similarity.

Combined similarities for unordered and ordered attributes

It is easy to extend the vectors $\mathbf{z}_i$ for ordered attributes to include the s-state unordered attributes. Each of the latter can be represented by an s-element row subvector whose inner product with itself must be unity, and with the corresponding subvector from another object is either zero or unity. A trigonometric interpretation for these inner products is immediate. With this representation for the attributes so far considered, each unordered s-state attribute contributes s elements, either zero or unity, to a row vector $\mathbf{z}_i$ for object i, while an ordered attribute is represented by two elements (a sine and a cosine). Let the total number of elements for an object be K. Thus a numerical representation for a set of m attributes for n objects is $n \times K$ matrix, Z, $K \geq 1$, which is such that $m^{-1}\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$ has unities on the diagonal, and that any element is the average of the similarities computed from the separate attributes.

Correlation and attribute weights

A matrix, **W**, has been incorporated above in a definition of similarity, and, other than proposing that it may be a particular diagonal array and also suggesting that nothing prohibits it from being more general, no discussion has been offered for the latter. Attributes assembled for cluster analysis are usually correlated, and therefore include an element of redundancy. The computation of (linear) correlation among continuous variates, of ordinal association among ordered variates, and also of association among categorical variates are standard problems of statistics; they require no further discussion here; **W** may be formed from these values.

There are circumstances, however, when paired observations are unavailable but for which it is desired to estimate some measure of association. Let A and B be two distinct such attributes; there are n_a observations available for A, with calculated mean of $\bar{x}_a$, and correspondingly for B. Three possibilities for this estimate, considered here for ordinal (continuous or discrete) attributes, have been extracted from the statistical literature; none appears to be well known.

(1) The **coefficient of weak monotonicity** is defined as

$$(\bar{x}_a - \bar{x}_b)/[(\Sigma_{i\in A}\Sigma_{j\in B}|x_{ia} - x_{jb}|)/n_a n_b].$$

(2) The **correlation ratio** defined as

$$|\bar{x}_a - \bar{x}_b|/[s^2{}_a + s^2{}_b - (\bar{x}_a - \bar{x}_b)^2]^{1/2}.$$

(3) A **coefficient of ordinal association**, defined as

$$(\Sigma_{i\in A}\Sigma_{j\in B}\text{sign}(x_{ia} - x_{jb})/n_a n_b.$$

These three may also be used for paired observations and are in increasing order of generality. The correlation ratio may be computed if only the means and variances are known.

These coefficients may be assembled into an array analogous with a correlation matrix, which need not be positive semidefinite (p.s.d.). It may be converted to be p.s.d. by the nonmetric transformation described later in this chapter. Its eigenvalues can then be examined, as in principal components analysis, to determine the number of "independent" factors implied.

Direct measures of pairwise relationship

Numerous possibilities exist for assembling data measuring some aspects of the pairwise resemblance among objects. Each of these requires its own attention, and although it is not possible to discuss many in detail, some general classes of data are usefully distinguished.

Probability estimates

In several branches of genetics, probability estimates are often computed; for example, the coefficient of common parentage, which is the probability that a gene chosen at random is identical by descent with the corresponding gene in another individual, conditional on some ancestral origin. Such values are based purely on the breeding and crossing history of the ancestry of the two individuals. Its definition as a probability leads at once to considering it as a similarity and to transforming it to a distance as described above.

Another probability is based not on ancestry, but on the observed or estimated gene frequencies in a set of experimental populations. These give rise to a set of multinomial frequencies, from which genetic similarity is computed, whose logarithm, known as Nei's genetic distance, satisfies all the requirements of

a distance considered here. (Note that a Hellinger-based distance using the square root of the probability is also possible.)

One component of multiple discriminant analysis is the computation of generalized squared distances among the (centroids of) populations. The generalized distances, when negated and exponentiated, estimate the probability of identity of the populations; it is apparent that they are also appropriate in the terms of this study. The advantages of generalized distances for dendrogram formation have been explored by Dagnelie and Merckz (1991); other discussions of distance measures for populations of objects have been made above and are considered again below.

Confusion arrays

In some psychological experimentation, subjects are asked to compare two stimuli and to record whether they are distinguishable or not. These data, after appropriate replication, can be regarded as a similarity if expressed as the proportion said to be indistinguishable. Note that although not all supposedly indistinguishable stimuli need be recorded as being identical, the degree to which the diagonal of the similarity matrix departs from unity gives some measure of variability. Perhaps that variability should be used to define the amount of difference to be regarded as trivial. However, sometimes ambiguities can exist in interpreting this class of data. Consider the well-known Rothkopf (1957) Morse-code data, which in Kruskal (1971) consist of the proportion of times that each of the $2 \times \binom{36}{2}$ ordered pairs of symbols sent to unpracticed listeners was recorded as being identical. The rows represent the first symbol of the pair, the columns the second. These values, usually termed similarities, are not symmetric, and although the array of them is diagonally dominant, it does not have unities on the diagonal. Clearly,

asymmetry is present not only in the values but also in what the rows and columns represent. The average proportion is usually used as the estimate of the similarity; but also investigated is the skew symmetric array, which, when added to the average, reconstructs the original table. In fact, the resemblances among the symbols may even have been captured perhaps from just a few of the pairs of symbols, e.g., the first k columns or the first k rows of the array (without the listeners' knowledge); this interpretation makes it clear that these data should be regarded as a set of 36 objects on which 36 variables have been measured, these variables being proportions that estimate

$$\Pr(i \text{ not confused with } j),$$

where i is the first and j the second symbol, and *not* as similarities. With this interpretation, the square array can be regarded as being analogous to the frequency data, but with the added complication that it is necessary to decide if it is the columns or rows for which the probabilities **p** of Chapter II are to be obtained. This decision can be made in the following way. Since the row symbols are the first in the pair, it becomes a reference for the second, in the sense that the listener has to answer the question:

> Is the second heard symbol of the pair the same as the first or is it different?

If it is this question that consciously or unconsciously is in the minds of the subjects, it is the probabilities associated with the columns that are of primary interest for grouping the symbols based on the confusion.

In the experiment described by Nosofsky (1989), subjects were first exposed to a series of 16 shapes in which two factors

(size, and angle of presentation with respect to the horizontal, each at four levels) were different; they were then presented with a shape randomly selected from the set and were asked to identify which level of each of the two factors the shape exhibited. The data so collected gave rise to an asymmetric array (the number of correct responses for each shape) resembling that of the Morse-code data, from which Nosofsky obtained similarities between the shapes, ultimately converted to distances. Following the reasoning described above, the subjects can be considered as being asked to consider if the shape is the same as their *memory* of the training experience, and so it seems reasonable to treat Nosofsky's Table 1 as a set of vectors describing the shapes, which have a two-factor structure.

Other direct measures

These measures are legion: immunological incompatibilities, diallel crosses, DNA hybridization, correlations, and so on, each of which has different properties. An interesting discussion of similarity and intensity in a psychological context is given by Sjöberg (1975). An important class arises from measuring within-location diversity in ecological applications (Magurran 1988), which sometimes permits the measurement of the degree of similarity in diversity among locations. Spearman rank correlations are also asserted as being particularly appropriate for measuring the pairwise relationships among ecological variables (Bölter and Meyer 1986). The key responsibility for the scientist is to decide what needs to be done so that the data provide a fair measure of distance. It is especially important for the distance to be unbiased and consistent for the smaller values, because it is these that are critical for clustering. This fact will become apparent later in this chapter, where a nonmetric transformation of distances having some merit for clustering is described.

Permuting similarity matrices

In Chapter II, noninvasive procedures were described that attempt to convert a 0–1 array into block-diagonal form by means of permutation matrices. The more this conversion can be achieved for similarity (resp. dissimilarity) arrays, the less will the apparent clustering depend on heuristic principles. An invasive possibility is to choose a threshold value for the similarities (dissimilarities) and form an array in which unity replaces higher similarities (lower dissimilarities) and is zero elsewhere, to allow the methods described in Chapter II to be applied; note that if $\mathbf{Z}$ is such a transformed array, the permuted matrix has to be $\mathbf{PZP}^{\mathrm{T}}$, where $\mathbf{P}$ is a permutation matrix. However, if $\mathbf{P}$ is sought so that

$$\sum_{ij} s_{ij}(i-j)^2$$

is minimized, then the permuted similarity array tends to have its highest values near the diagonal (Gourlay 1979). Other schemes are possible.

As originally defined (Toussaint 1988), the sphere of influence graph is appropriate for two-dimensional "skeletons" of planar objects. Its definition, modified slightly in notation, follows:

> DEFINITION VII.5. *Let $N = \{x_i\}$, $i = 1...n$, be a finite set of points in a plane. For each point $x_i \in N$, let r(i) be the smallest distance to any other point in the set. Let C(i) be the circle of radius r(i) centred at x_i. The sphere of influence graph is a graph on N with an edge between points x_i and x_j iff C(i) and C(j) intersect in at least two places.*

Toussaint illustrated this graph with 54 figures, each consisting of a set of points and the corresponding sphere (better, circle?) of influence graph. These illustrations show that the graph need not

be connected, that one connected component can be completely enclosed within another, and that the graph can be complete. Let $\mathbf{E} = \{e_{ij}\}$ denote the adjacency matrix of the graph; since two circles intersect if $f(i,j) = C(i) + C(j) - d_{ij}$ is greater than zero, the elements of **E** can be defined by

$$e_{ij} = \begin{cases} 1, f(i,j) > 0 \\ \\ 0, f(i,j) \leq 0, \end{cases}$$

from which it is apparent that the set of points need not be planar. Thus **E** defines a set of edges among a set of n vertices in any dissimilarity space.

Using **E** and any method to determine a spanning forest, the vertices forming disconnected subgraphs can be identified. At least to a first approximation, these subgraphs can be processed separately. Note, however, that the smallest isolated subgraph contains at least two vertices. Further, permuting **D**, based on a block-diagonal permutation of **E**, sometimes reveals the structure effectively and is an alternative procedure to those based on global thresholds.

Euclidean and non-Euclidean dissimilarities

For most, if not all procedures based on linear algebraic methods applied in biology, sociology, and so on, there is an assumed underlying Euclidean metric. However, pairwise distances calculated from attribute data (as in taxonomic studies), or from presence and absence data (as in ecological surveys), or from subjective assessments (as in psychometric measurements) rarely satisfy the Euclidean conditions; although they may do so approximately, they usually fail to satisfy the triangle inequality and so are no more than a semimetric. If they are Euclidean, standard linear algebraic procedures may be used directly without

discarding information from the array. Another advantage for having data satisfying the Euclidean conditions is that if n points are multivariate normally distributed in n-dimensional space, the squared distances are jointly exponentially distributed subject only to the constraint that a valid n-point configuration is prescribed (Clifford and Green 1985). Coupled with a Euclidean representation of the data, this result may lead to a comparison of the empirical squared distances with an exponential distribution, and hence to a test for the existence of too many small values, i.e., clusters.

Recognizing non-Euclidean data

There are easier ways for determining if the data are consistent with a Euclidean metric than counting the number of failures to satisfy the triangle inequality out of the $\binom{n}{3}$ possibilities. The degree of departure from Euclidean conditions can be measured in several ways, of which only one based on the eigenvalues is considered. Since the diagonal of dissimilarity arrays is uniformly zero, some eigenvalues are necessarily positive and others negative (some may also be zero; all will be uniformly zero if all elements are zero). Thus, if the dissimilarities are consistent with a Euclidean metric and the dissimilarities are all greater than zero, there is just one negative eigenvalue. Inconsistency with the Euclidean metric is indicated by the number of negative eigenvalues together with their absolute size compared with those that are positive. Assuming again that all dissimilarities are greater than zero, the inconsistency with the Euclidean conditions can be estimated independently of the sizes of the eigenvalues by the matrix sign function, which identifies the positive, negative, and null subspaces of a matrix; this function is obtained from the Newton iteration

$$\mathbf{X}_{t+1} = \tfrac{1}{2}(\mathbf{X}_t + \mathbf{X}^{-1}_t), \quad \mathbf{X}_0 = \mathbf{D},$$
$$\operatorname{sgn}(\mathbf{D}) = \lim_{t\to\infty} \mathbf{X}_t$$

using a Moore–Penrose inverse. The matrix sign function, sgn(.), consists of the (diagonal) elements of the limit, each of which is an element of {−1,0,1}. One less than the number of elements equal to −1 is the dimensionality of the negative space. The inertia of a (square) matrix is a three-element vector consisting of the number of positive, negative, and zero eigenvalues. If $\|\mathbf{I} - \mathbf{D}^2\| < 1$ for any matrix norm, the iteration

$$\mathbf{X}_{t+1} = \tfrac{1}{2}\mathbf{X}_t(3\mathbf{I} - \mathbf{X}_t^2), \quad \mathbf{X}_0 = \mathbf{D}$$

also converges to sgn(**D**) and avoids the need for matrix inversion (Kenney and Laub 1991). Because dissimilarities are relative quantities, $\mathbf{X}_0$ can be rescaled to satisfy the norm requirement.

Determining rank

Besides numerical matrix rank, which requires determining how small an eigenvalue must be for it to be regarded as zero, there are several other methods for determining a value for the rank of a matrix. The *intrinsic dimensionality* (Pettis et al. 1979) is based on the assumption that the objects are independently distributed in d-dimensional space; if so, an equation can be derived to estimate d based on the number of neighbors within prescribed distances of each (occupied) point. The applicability of the assumption of independence to the clustering problem in biology, however, is questionable, because it cannot be assumed that there is only one population, so that there is support for little more than (very) local independence.

The *effective dimension* of a matrix is more interesting, in that fewer assumptions are made. This parameter also presents a measure suitable as an index for the effects of any transformations. Let s_i be the strictly positive singular values of any rectangular

matrix, arranged so that $s_1 \geq s_2 \geq \ldots \geq s_n$; then the effective dimension is defined as

$$\rho(\alpha) = s_1^{-\alpha} \Sigma s_i^{\alpha}, \ \alpha \geq 0$$

(Oldford 1987). For $\alpha = 0$, the effective dimension is equivalent to ordinary rank; in the present context, three further values are considered to be informative, namely $\alpha = 0.5$ (since distances can be regarded as quadratic forms), $\alpha = 1$ (Thisted 1982), and $\alpha = 2$ to illustrate the effect of different α.

Dimensionality reduction

A dissimilarity matrix of order n may be of rank $n - 1$; it is often believed that some of these dimensions represent "noise," and that only some subspace, of relatively low dimensionality, contains the information of interest. In identifying this subspace, usually called dimensionality or rank reduction, two of many possibilities are considered here. The first of these attempts to recognize which directions in the principal coordinate space (Gower 1966) can be regarded as random, discarding them as "noise"; the second attempts to obtain a representation of low dimensionality, perhaps using a nonmetric transformation of the data. The second, therefore, *imposes* a solution on the data and so reflects the scientist's opinions about the data in question. The first, by contrast, seeks both to reveal what can be reasonably considered as being nonrandom, and to distinguish it from what can be regarded as being of no interest. Only the first case is described here; some nonmetric transformations and their possible consequences are discussed subsequently.

Rank reduction using principal components

Denote the principal coordinates (Gower 1966) obtained from the possibly transformed distances by $\mathbf{US}$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{U}\mathbf{U}^T$ is

idempotent and S is diagonal, consisting of the square roots of the eigenvalues. Since the centroid of the principal coordinates is the null vector, it follows that S is the sum of squares, i.e., the eigenvalues of the principal *components* corresponding with the nonzero coordinates. With this assumption, a standard test used in principal components analysis may be adopted (Anderson 1958). This test is sequential: given that k components have been accepted, $k = 0,1,\ldots,\text{rank}-1$, the null hypothesis is that no direction in the remaining space has variance significantly greater than any other. If the null hypothesis is not disproved, the conclusion is that the remaining data are approximately spherical, therefore representing random variation, and so can be discarded. While the rejection level for the hypothesis can be adjusted to achieve the desired rank, accepting too few directions can result in loss of useful information, while accepting too many may confuse any subsequent computations with artifactual patterns.

The dimensionality of the negative space is sometimes greater than zero, i.e., the dissimilarities are non-Euclidean. In consequence, there have been many proposals for converting them into a form that satisfies the Euclidean conditions. The easiest procedure is simply to replace the negative eigenvalues by zero, i.e., to join the negative with the null space; if the dimensionality of the negative space is large, the effect of this reduction in rank is quite unpredictable.

Rank reduction by smoothing

Another possibility, little used, is to smooth the dissimilarities. It is apparent that each empirically obtained dissimilarity is a random quantity and so may be considered to consist of true and random parts. In the absence of information about the distribution of the random components, or knowledge of the true value, to identify these components seems to be an impossible task. However, using empirical probability density estimation can result

in some smoothing of the empirical values; the assumption is that the parts smoothed away represent a segment of the random component.

Let $V(i)$ be the set of objects for which the dissimilarities between them and object i have been obtained empirically. Consider the dissimilarity d_{ij}; the average dissimilarity object k has to i and j is $(d_{ik} + d_{jk})/2$; the smoothed estimate of $d_{ij}(0) = d_{ij}$ is now proposed to be

$$d_{ij}(t+1) = (2d_{ij}(t) + \tfrac{1}{2}\Sigma_k (d_{ik} + d_{jk}))/|V(i) \bigcup V(j)|,$$

where $k \neq i, j$. For a complete set of dissimilarities, the denominator is n. If the process is repeated, eventually the dissimilarities will become equal. There seems little reason to consider $t > 1$, unless there are special circumstances.

The following numerical example illustrates the impact of this transformation. Consider the three objects i, j, and k for which five two-state attributes have been observed, some of which are randomly missing:

i	j	k
1	*	1
*	2	1
*	1	1
2	2	1
2	2	2

Then $d_{ij} = 0$, $d_{ik} = \frac{1}{3}$, and $d_{jk} = \frac{1}{2}$. Because of the missing values, the comparison between i and j is based on two attributes, that between i and k on three, and between j and k on four. The first-order smoothed values are $d_{ij} = 5/36$, $d_{ik} = 11/36$, and $d_{jk} = 7/18$.

If dissimilarities can be smoothed, then so can the empirical values of the attributes. For a set of continuous attributes,

Winsberg and Ramsey (1983) proposed a monotone transformation of the values for each variable based on integrating what are called basis splines; they provided a computer program to achieve this smoothing.

Other methods of rank reduction

The problem discussed here is to determine how many of a set of directions are to be accepted, *not* to determine (further) transformations of the dissimilarities to achieve lower dimensionality. Three possibilities are now summarized.

Method 1 Choose a cutoff value, C^*, somewhere between 70 and 90%, and determine the rank to be the smallest number of eigenvalues accounting for at least C^* of the observed variation.

Method 2(a) Let $\lambda_1 \geq \lambda_2 \geq \ldots$ and plot the eigenvalue, λ_i, against i; by inspection, determine the rank as the largest value of i for which the slope on the left remains steep (Cattell 1966).
(b) This method is identical with 2(a) but uses the logarithms of the eigenvalues (Craddock and Floud 1969).
Methods 2(a) and 2(b) tend to discard those eigenvalues that on the right, fall into a straight line.

Method 3 Let $\lambda_1 \geq \lambda_2 \geq \ldots$; if the value of $\lambda_i - \lambda_{i+1}$ is sufficiently small, small changes in the empirical data are likely to bring about changes in the directions of the subspace spanned by the corresponding eigenvectors. The proposal is that rank is determined so that there are large differences among the adjacent λ_i and λ_{i+1}, i.e., only eigenvalues $\lambda_{1\ldots i-1}$ should be retained (assuming a large value for $\lambda_{i-1} - \lambda_i$). Determining what is sufficiently large for Method 3 is still an empirical problem, but clearly, some statistic based on

$$2(\lambda_i\lambda_{i+1})^{1/2}/(\lambda_i + \lambda_{i+1}),$$

i.e., the ratio of the geometric to the arithmetic means, is perhaps suitable. If this ratio approaches unity, the differences are small, and the accepted rank is at most i (or $i - 1$).

Nonmetric conversion of non-Euclidean to Euclidean arrays

Some of the problems of obtaining acceptable Euclidean representations of non-Euclidean data have been discussed by Williams et al. (1971) and Gower (1984). In an interesting proposal to solve the problem, Thu (1978) identified a set of parameters, γ, such that d^γ satisfies the triangle inequality for all triples, where d here is an element of the dissimilarity array. Because this transformation tends to enlarge the smallest dissimilarities and reduce the larger, it brings together groups well separated with respect to the untransformed dissimilarities. It was also conjectured that the largest γ yields the space of smallest dimensionality. Many other published proposals also include the same objective, which is to obtain a low-dimensional solution, often of no more than a specified rank. The combination of these two steps has spawned several variations on this theme since Kruskal's (1964) early statement of it (see Schiffman et al. 1981, Davison 1983). In this section, these steps have been separated, and the main focus is converting empirical distances to be Euclidean.

The motivation for the nonmetric transformation redescribed from Lefkovitch (1984) arises from a consideration of distances based on attribute data. However, if a directly obtained distance is considered to be an integrated comparison of a set of unknown multistate attributes, the same arguments apply. The conversion of similarities into distances, which depends on the coefficient, is discussed above. The following remarks, adapted from Lefkovitch (1984), form the basis of the model.

Consider objects i and j described by m unordered, equally weighted and independent attributes, each having s mutually

exclusive states. Then for fixed i, it can be shown that the number of ways object j can differ from object i in $k = 1...m$ attributes, for each of which dissimilarity, d_{ij}, using almost any attribute-based definition, will take the same value,

$$w_k = \binom{m}{k}(s - 1)^k.$$

Since

$$k < (s - 1)(m + 1)/s \text{ implies } w_k/w_{k-1} > 1,$$

and since dissimilarity increases monotonically with k, the states shown by the attributes of object j can be predicted by those of (fixed) object i with fewer errors if d_{ij} is small than if it is large. (These formulae are easily extended to unequal numbers of states and to ordered and continuous attributes.) Furthermore, the pattern shown by w_k remains essentially the same if the attributes are correlated. The conclusion is that whole arrays of dissimilarities may be less valuable in determining relationships than is an informed selection of their smaller elements. This conclusion was also reached by Chen and Andrews (1974), Williamson (1978), Clymo (1980), Minchin (1987), and Bradfield and Kenkel (1987) by simulation and other noncombinatorial considerations.

One example of the problems created by ignoring this inequality of information can be seen in the Euclidean case in the context of principal coordinates (Gower 1966). Since the first (most important?) principal coordinate computed from the $\mathbf{D} = \{d_{ij}\}$ maximizes the squared distance among the objects, the major contribution must come from the **largest distances**; but it is precisely these that **inform least on close relationships**. In clustering, especially the sequential, agglomerative, hierarchical and nonhierarchical procedures, it is the smallest dissimilarities that are used to form groups; they determine the internal properties

of each subset, while the largest are hardly ever used for these purposes.

It is not difficult to make plausible selections of the smallest dissimilarities; for example, those remaining after discarding the maximum dissimilarities (Williamson 1978, Clymo 1980, Minchin 1987), the local condition of the k nearest neighbors of each object, with fixed k (this restraint is related to the proposal made by Bradfield and Kenkel 1987), or the neighbors of each object within a specified (i.e., global) dissimilarity (Faith et al. 1987), each provide subsets of interest. Because a particular value of k, or a critical (global) dissimilarity, has to be chosen and thus is externally imposed, these procedures cannot be completely satisfactory. To avoid these problems, a selection of the smallest dissimilarities can be based on a definition of neighbors that is local in its operation and does not depend on externally specified parameters. Toussaint (1980) gave such a definition; two objects are relative neighbors if they are at least as close to each other as they are to any other. More formally, the **relative neighborhood graph** (RNG) defines objects i and j as relative neighbors if

$$e_{ij} = \begin{cases} 1, & d_{ij} \leq \min\{\max(d_{ik}, d_{jk})\} \; \forall \; k \neq i,j \; ; \; i \neq j; \\ 0, & \text{otherwise.} \end{cases}$$

E, considered as a matrix, gives the adjacencies of the RNG. Toussaint (1980) showed that the RNG is a supergraph of the **minimum spanning tree** (MST), which is the shortest connected graph having no cycles; it follows that the RNG is connected. In common with the MST, the RNG has neither metric nor dimensionality implications and can exist in any kind of space equipped with a concept of relative closeness. Toussaint (1980) gave an algorithm to obtain this graph for objects located in spaces of any dimensionality; Urquhart (1982) and Supowit (1983) gave detailed studies of its properties.

At this stage, the nonlinear transformation follows almost naturally. The proposal, similar to that of Bradfield and Kenkel (1987) is applied to the RNG rather than the k nearest neighbors as follows:

DEFINITION VII.6. *The RNG-path distance transformation of* **D** *replaces the empirical distance between two objects by the length of the shortest path in the* **D** *array corresponding to edges in* **E**, *i.e., in the RNG.*

Let $\mathbf{Z} = \{z_{ij}\}$ be these shortest-path distances computed from **D** and **E**; it follows from the construction of the arrays **E** and **Z** that the dissimilarities among the objects that are most alike, including those objects adjacent on the MST, are identical in **Z** and **D**.

THEOREM VII.1. *The metric of* **Z** *is Euclidean.*

Proof: *a.* Since the RNG is a subgraph of the Gabriel graph, which is Euclidean, the assertion is true for objects adjacent on the RNG.

b. For objects i and j not adjacent on the RNG but separated by object k, since z_{ij} is the shortest path distance and so is equal to $z_{ik} + z_{jk}$, then

$$z^2_{ij} \geq z^2_{ik} + z^2_{jk}.$$

c. For objects i and j separated by more than one object, the proof follows by mathematical induction from *b*. Q.E.D.

It can be seen that this transformation, which converts an arbitrary semimetric into a Euclidean distance, changes only the larger

distances, corresponding to the poorest empirical mutual information. It can be used for any semimetric, including, for example, the Euclidean case.

To describe the consequences of the RNG transformation, the following observations are of interest. Let d_{jk} and d_{ik} be given; then the range of values for which the metric is Euclidean for d_{ij} is that

$$|d_{ik} - d_{jk}| \leq d_{ij} \leq (d_{ik} + d_{jk})$$

be true for all i, j, and k. In the RNG, if $e_{ij} = 0$ and $e_{ik} = e_{jk} = 1$, then $z_{ij} = d_{ik} + d_{jk}$. It follows that if the original metric is Euclidean, so that $d_{ij} \leq d_{ik} + d_{jk}$, then $z_{ij} \geq d_{ij}$. Thus the largest distances are increased, and the smallest are unchanged, so that compared with the original values, the variance of the interpoint distances is increased. As a result, dimensionality tends to be reduced (Kendall and Moran 1963), and the squared lengths of the largest principal coordinates increase disproportionately to the smallest.

From numerical experiments, it seems that the direction of the principal coordinates is hardly changed. The effect of this transformation, for example in a clustering context, is further to separate distinct groups, while leaving closely related objects virtually unchanged in their mutual proximity. Thus many sequential agglomerative procedures, which are based largely on the smallest dissimilarities, are little affected by this transformation.

The non-Euclidean case is a much larger set than the Euclidean, but it is still possible to make some general remarks. It is convenient to consider the two reasons for the failure of the Euclidean metric. Assume $e_{ij} = 0$ and $e_{ik} = e_{jk} = 1$ as before:

(1) Suppose $d_{ij} > d_{ik} + d_{jk}$; it follows that $z_{ij} < d_{ij}$.

(2) Suppose $d_{ij} < |d_{ik} - d_{jk}|$; this case cannot satisfy the assumptions about e_{ij}, e_{ik}, and e_{jk} but, with a suitable relabeling of the vertices, becomes identical with (1).

It follows, therefore, that in the non-Euclidean case, at least some of the largest distances are reduced, and, as in the Euclidean case, the smallest are unchanged. With rank being the number of nonzero eigenvalues in **D**, it can be shown that the rank in general is not changed, that *all* the eigenvalues are likely to change, not only the negative ones. (Note that **D** is symmetric, so that there are no complex eigenvalues.) There is also empirical evidence that the directions of the eigenvectors are changed in some way dependent on the number and relative sizes of the negative eigenvalues, but no predictable pattern has yet emerged.

An additional step may be incorporated in this transformation. Because the edge distances in the RNG are the smallest, they are about equal, so that the transformed edges in the graph complementary to the RNG are approximately integer multiples of the average of those in the RNG. This fact suggests a further transformation of the distances:

> DEFINITION VII.7. *The RNG-edge distance between two objects is the number of edges in the shortest path between them in* **E**.

Since the metric properties of the path distances carry across to the edge distances, and since all triangles whose sides are the edge distances are isosceles, it follows that they are ultrametric. Furthermore, because these distances are integers, the minimum dimensionality of the Euclidean space that can reproduce them exactly tends to be less than that of the path distances. If the RNG is also a MST, it follows that both the path and edge distances are ultrametric.

In the context of clustering, the following proposition illustrates some of the effects of the RNG-path distance transformation.

> PROPOSITION. *The MST derived from the RNG-path distance transformation is identical with that of the original distances.*
>
> *Proof.* The distances in the RNG are unchanged by the RNG-path distance transformation; those in the complementary graph consist of at least the sum of two edges in the RNG. Since the MST is a subgraph of the RNG, the assertion follows. Q.E.D.

Beyond the present context, because the single-linkage dendrogram is equivalent to the MST (Gower and Ross 1969), there are no effects of this transformation for this clustering method. Other dendrogram-generating procedures, which recompute distances after grouping, are likely to give somewhat different dendrograms after the RNG-path distance transformation. After the RNG-edge distance transformation, however, even the single-linkage algorithm may produce different results, because at least $n - 1$ edges have identical lengths, so that the MST is not unique. Chapter X, "Caste skulls," "Fescue grasses," and "Blood and language" give examples of the RNG-path transformation and its subsequent effect on the groupings.

An alternative transformation

The arguments above for retaining the smallest distances are based on the conclusion that these give the most useful mutual information. However, if closely related objects differ only randomly, while distantly related ones differ systematically, it is

the *larger* values that should be retained, and the smaller replaced. In these circumstances, the edges corresponding with the **relative external graph** (REG), defined as

$$e_{ij} = \begin{cases} 1, & d_{ij} \geq \min(d_{ik}, d_{jk}),\ \forall\ k \neq i,j;\ i \neq j \\ 0, & \text{otherwise,} \end{cases}$$

plays the same role as the RNG. The transformed matrix, **Z**, is formed from **D**, by the values corresponding to the unities in **E**, and, for example, for two nonadjacent objects, i and j, separated by just object k, by $z_{ij} = |z_{ik} - z_{jk}|$. Several conjectures based on this transformation seem to follow, supported by numerical examples.

CONJECTURE VII.1. *In the REG-path distances transformation, the smallest distances tend to be increased.*

In consequence, the variance of the interpoint distances is reduced, which results in the eigenvalues tending to become more uniform, i.e., a tendency to sphericity. If this conjecture is true, it follows that, in the less than full rank situation (i.e., if the rank is less than $n - 1$), the rank of the transformed array may be increased. There are no obvious consequences referring to the directions of the principal coordinates themselves, except those following from the near-sphericity, i.e., their computed directions may differ considerably from those of the untransformed data. Furthermore, since the REG transformation increases the smallest distances, it reduces the (relative) separation among distinct groups.

For several data sets, including some for which the rank, $\rho(0)$, does not change (this circumstance is rare for RNG-edge distances), the following is believed to be true. Let **D** be a matrix

of distances obtained from dissimilarities, **M** the REG-path distances, **Z** the RNG-path distances, and **G** the RNG-edge distances; then

CONJECTURE VII.2. *For* $\alpha > 0$,

$$\rho(\alpha;\mathbf{M}) \geq \rho(\alpha;\mathbf{D}) \geq \rho(\alpha;\mathbf{Z}) \geq \rho(\alpha;\mathbf{G}).$$

Support for this conjecture is from numerical observation; thus the transformation that appears to reduce dimensionality the most is likely to be based on the RNG-edge distances.

The case for retaining of the smallest dissimilarities and replacing the largest by the path distances on the RNG seems stronger than that based on retaining the largest and replacing the smallest by the path distances on the REG. It is supported by the numerical examples. For most of the numerical examples (e.g., Chapter X, "Caste skulls"), the REG-path distances made the data more nearly spherical and also produced a disposition of the objects largely disagreeing with the complete set of principal coordinates and with prior knowledge; those depending on the RNG were more elliptical and agreed very well. However, both possibilities are available; the RNG-based transformation may be more appropriate in taxonomy, the REG-based method may be preferable in psychometry.

Although the RNG-path distance transformation is not directly comparable with others in common use, it is useful to contrast it with three others. The first model, essentially that of Gabriel (1978), can be described as follows: let **D** be a distance matrix; it is desired to obtain a matrix **Y** so that

- $\mathbf{D} + \mathbf{Y}$ is Euclidean
- $\|\mathbf{Y}\|$ is a minimum

the rank of $\mathbf{D} + \mathbf{Y}$ should not exceed some particular value.

The motivation behind the third condition is clear, and, because larger values need to be changed the most to satisfy conditions 1 and 2, the numerical results of the Gabriel procedure and those described for the RNG-path distance transformation tend to agree quite closely. The advantage of the RNG-path distance procedure over that of Gabriel is therefore largely computational; obtaining $\mathbf{Y}$ requires an iterative minimization phase, while RNG-path distance computation is direct and of $O(n^3)$. The rank-reduction component of the Gabriel procedure, however, is not an integral part of RNG-path distances, and so, if the desire is to obtain a low-dimensional representation that happens to be Euclidean, rather than the reverse, there may be an advantage is using it. The integers closest to $\rho(0.5)$ or $\rho(1)$ provide an empirical estimate of an "adequate" dimensionality.

In the second model, essentially that of Kruskal (1964), the distances are arranged in ascending order and then monotonically smoothed to achieve a low Euclidean rank. Missing values, as in the RNG-path distance transformation but unlike the Gabriel procedure, present no problems. Because all distances present are treated equally, the small ones are as likely to be changed as the larger, and so, it is now claimed, the details of the close relationships tend to be changed; however, Chen and Andrews (1974) considered cost functions, which penalize changes in the small values more than the larger. Because the Kruskal procedure is also iterative, the RNG-path distance transformation offers a computational advantage.

A third model also changes all values in the dissimilarity array. As noted above, the transformation proposed by Thu (1978) to obtain data satisfying the triangle inequality changes the smaller dissimilarities more than the larger; it tends to bring together

groups that are well separated in the original dissimilarity measure. However, it is easy to change Thu's proposal so that the reverse happens; rather than transforming the dissimilarities, first convert them to similarities, i.e., $s = f(d)$, and find the minimum γ such that the inverse transformation, $d = f^{-1}(s)^{\gamma}$, satisfies the triangle inequality for all triples. The larger similarities, corresponding with the smaller dissimilarities, are changed the least by this procedure, so leaving proximal objects relatively unchanged in position. The Fortran program given by Thu can be easily modified to include whatever similarity transformation is appropriate.

The dual to the object space of principal coordinates analysis is the space of the variables in principal components analysis. For simplicity, suppose all m variables are standardized so that the estimated covariance matrix, $\mathbf{R}$, has unities on the diagonal. Assuming a linear relationship, more information about variable j is given by variable i if $|r_{ij}| \to 1$, than if $r_{ij} \to 0$; there are many more ways in which variable j can differ from variable i if $r_{ij} = 0$ than if $|r_{ij}| = 1$. Since the first principal component is the direction in the space maximizing the variance, and since variance is here equivalent to (squared) distance, the major contribution to the first principal component is based on those covariances that approach zero in absolute value. There seems to be a case for a nonmetric transformation of covariance matrices, analogous to that for principal coordinates. The proposed procedure is as follows:

(1) Obtain the principal components of $\mathbf{R}$ as the complete solutions to $\mathbf{AV} = \mathbf{\Lambda V}$.

(2) Obtain the array $\mathbf{\Lambda}^{1/2}\mathbf{V}$ and obtain the Euclidean distances between the rows (corresponding with the variables).

(3) Transform these distances using the same procedure as that proposed for dissimilarities.

(4) Obtain the eigenvalues and eigenvectors of the array of transformed distances.

This final set are the adjusted principal components.

Are the dissimilarities consistent with a clustering?

The belief that the small dissimilarities are consistent with objects belonging together, and that large dissimilarities indicate objects that do not, is perhaps more than intuitive. Exploiting this belief to produce a test for the existence of clusters has been the subject of several studies, including Fillenbaum and Rapoport (1971), Hubert (1974), Ling and Killough (1976), Hawkins et al. (1982), and McArdle (1991). Such a test is now considered. Let t be a threshold value, and define a graph, $G(t)$, having n vertices and an edge between vertices i and j iff $d_{ij} \leq t$. Let k_t be the number of edges in $G(t)$, and t_c be the smallest threshold value, such that $G(t_c)$ is connected and contains k_c edges. Then $k_c \geq n - 1$ may sometimes be seen, but the upper limit of $\binom{n-1}{2} + 1$ is unlikely. Referring the value of k_c to the tables provided by Ling and Killough (1976), and perhaps making use of the macro given by McArdle (1991), provides some guidance on the existence of clusters.

An alternative procedure for testing for the existence of clusters is via the RNG. Lefkovitch (1984, 1985*c*) proposed examining the number of edges in the RNG in excess of $n - 1$, i.e., the number in the MST. Because the MST is a subgraph of the RNG, if there are many more edges in the RNG than $n - 1$, small differences in the dissimilarities may well give a different

MST. The number of edges in the RNG more than $n - 1$, if considered as a random quantity bounded by zero and $(n - 1)(n - 2)/2$, can be used for testing. Let the data consist of two lists: the edges in the MST, and the edges in the RNG. Form the following:

		MST	
		In	Not in
RNG	In	a	b
	Not in	c	d

in which it is apparent that

$$a = n - 1,\ c = 0, \text{ and } a + b + c + d = n(n - 1)/2.$$

If b is zero, I claim that any clustering is likely to be stable; to the extent that b exceeds zero, the less stable is any clustering, which leads to the hypothesis of $b = 0$ versus the alternative that $b > 0$. Since the expected number of edges in common to two random spanning trees is 2 for large n, and assuming the binomial distribution, $B(\, n(n - 1)/2,\ 2/n\,)$, leads to comparing

$$X^2 = b^2 n/((n - 1)(n - 2))$$

with the chi-squared distribution with 1 degree of freedom in a one-tailed test. If the evidence against the null hypothesis is high, there will be many spanning trees in the RNG of about the same (dissimilarity) length, so that small changes in the empirical dissimilarities may give rise to a MST having a different topology from that under study and so may lead to a different clustering. It follows that a clustering based on the current data should be regarded as unstable.

A related test (Lefkovitch 1985*c*), based on a displaced binomial distribution and its approximation for small b by a displaced Poisson distribution, is perhaps not needed. Other tests based on extracting more of the structure in the MST were given by Lefkovitch (1985*c*).

Missing distances

In many sets of directly obtained distances, some pairwise values are missing either randomly or systematically. Both cases can be considered together, and, other than the degree of approximation resulting from incomplete data, the results obtained by replacing missing values by a large value in the distance array should be acceptable. However, it is useful to distinguish these two cases to substantiate this remark.

Case 1: randomly missing data The randomness here is meant to indicate that there was *no* deliberate decision to obtain the distances between objects belonging to distinct subsets and not to obtain distances among the objects within each subset.

Assuming that the number of missing values is not too large, so that the RNG exists as a connected structure, it is conjectured that there is a high probability that the RNG obtained from the incomplete data will be virtually identical with that given by the complete data. The following model provides heuristic support for this assertion. An urn contains $n(n-1)/2$ balls, of which m are black (corresponding with edges in the RNG) and the remainder are white. Since, in a sample of size k, the number of black balls follows the hypergeometric distribution, with the expected value being $2km/(n(n-1))$, then for at least $m/2$ balls to be black in the sample, $E(k) \leq n(n-1)/4$ need to be sampled. Thus if the number of missing values is appreciably less than half the number of edges in the complete graph, the RNG is likely to be a good approximation to the true one. In fact, the situation is

probably even more satisfactory, because the missing values are often not random, in the sense implied by the sampling procedure implicit in the urn model, but tend to be those impossible to collect because of some biological incompatibility (i.e., implying a high distance) or because "everyone knows" that the objects are no more than distantly related. Note also that, since the number of edges in a RNG tends to be less than about $3n$ (Lefkovitch 1984), only if the number of missing values exceeds $n/6$ and these coincide largely with the distances that would have been represented on the RNG is there any serious risk of distortion.

Case 2: systematically missing data An example of this class is given by those problems normally studied by the methods called "unfolding." Suppose the objects are divided into two disjoint subsets, I and J, and the only distances available are $\{d_{ij}, i \in I, j \in J\}$. If these distances are already Euclidean, they can be unfolded by the singular decomposition of the block of complete data (see below), and a complete coordinate system obtained. If the data are not Euclidean distances, the missing data can be replaced by a sufficiently large value, and the RNG obtained as before. The path distances only in the nonmissing block now replace the original values and the modified data can then be unfolded as before.

Suppose a set of objects is divided into two disjoint subsets I and J, $n_1 = |I|$, $n_2 = |J|$, $n_2 \geq n_1$, and only the pairwise similarities (distances should be transformed appropriately) between objects belonging to the different subsets obtained. It is desired to obtain a set of coordinates for the combined set of objects in whatever dimensional Euclidean space will reproduce the known similarities.

Consider the following model; suppose $\mathbf{X}$ consists of a set of coordinates for the $n = n_1 + n_2$ objects whose origin is their centroid; then

$$\mathbf{X}^T\mathbf{X} = \left[\begin{array}{c|c} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \hline \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{array}\right]$$

Assume that $\mathbf{X}_2^T\mathbf{X}_1$ are the only data available. Let the singular value decomposition be

$$\text{svd}(\mathbf{X}_2^T\mathbf{X}_1) = \mathbf{U}\mathbf{S}^2\mathbf{V}^T,$$

where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$, and $\mathbf{S}^2$ is diagonal and non-negative. It follows that the principal coordinates of $\mathbf{X}_1$ are $\mathbf{V}\mathbf{S}$, those of $\mathbf{X}_2$ are $\mathbf{U}\mathbf{S}$, and that $\mathbf{X}^T\mathbf{X}$ can be completed as

$$\mathbf{X}^T\mathbf{X} = \left[\begin{array}{c|c} \mathbf{V}\mathbf{S}^2\mathbf{V}^T & \mathbf{V}\mathbf{S}^2\mathbf{U}^T \\ \hline \mathbf{U}\mathbf{S}^2\mathbf{V}^T & \mathbf{U}\mathbf{S}^2\mathbf{U}^T \end{array}\right]$$

The principal coordinates for the n objects are obtained from the completed matrix in the usual way.

Several other methods for multidimensional unfolding have been published; perhaps the most easy to implement is given by Greenacre and Browne (1986).

VIII Subset generation using scalar dissimilarities

Three main circumstances lead to the procedures discussed in this chapter. First, at the end of Chapter VI, it was shown that if, for univariate data, the metric is not constant throughout the range but depends on the pairs, the complete array of standardized differences may be multidimensional. Second, attributes are subjective divisions (Chapter I) of an object, which are used to estimate a measure of dissimilarity, the array of which will almost certainly be multidimensional (Chapter VII). Third, many sets of data consist of relationships (similarities, dissimilarities, and so on) measured directly on the pairs of objects, which again may lead to an array of dissimilarities that is multidimensional. For multidimensional arrays, there is no natural ordering (although Barankin and Takahasi 1978*a*, 1978*b*, and Goodman and Pollack 1983 discussed possibilities for these circumstances), so that a simple definition of betweenness, as used in Chapters V and VI, is not immediate. In this chapter, an alternative concept is developed and used to generate subsets; the procedures discussed in Chapter II can then become applicable.

The assumptions

For clustering in a multidimensional context, two assumptions are common to almost all clustering methods and can hardly be avoided; they are

ASSUMPTION VIII.1. *The objects can be represented as points in space.*

ASSUMPTION VIII.2. *The resemblance between all pairs of points can be measured.*

These relationships are here referred to as distances, so that similarity, dissimilarity, and related concepts have been transformed to that class of measures. Even though a dissimilarity may be a distance (Chapter VII), it is not necessary to assume that the space is Euclidean, linear, or even continuous. For the purposes of subset generation, two regularity conditions generalize the principles discussed in Chapter VI when considering the real line; they are

> ASSUMPTION VIII.3. (first regularity condition). *As the maximum distance among members of a subset approaches infinity, so the cardinality of the subset approaches n.*

> ASSUMPTION VIII.4. (second regularity condition). *As the maximum distance among members of a subset approaches zero, so the cardinality of the subset approaches unity.*

Assumptions VIII.3 and 4 enable the subset-generating process described below to avoid much redundant arithmetic. Their role emerges in what follows, which provides the details of a subset-generating procedure fundamental to the present circumstances.

A general subset-generating procedure

In Chapters V and VI, some essential components of a subset-generating procedure were described but were specialized to particular classes of data. Those procedures can be considered as particular cases of a more general class, outlined in Table VIII.1. In the pseudocode of Table VIII.1, notice that because the acceptability of a group is defined by a pair of objects, this algorithm can be implemented using parallel processing, with processors discarded either when subsets become duplicated, or if there is no change in the k-loop leading to storage of the group. It will be seen below that the number of processors is unlikely to exceed $3n$.

Table VIII.1 An outline of the subset-generating procedure

```
for i:= 2...n
    for j:= 1...i-1
        group:= {(object i) ⋃ (object j)};
        if acceptable(group) = .true. then
        label:   for k:= 1...n
                     if newmember(group, (object k)) = .true.
                       then group:= {group ⋃ (object k)};
                 end k;
        if the group has been changed on the last pass
          through the for k-loop, go to label;
        else store(group);
    end j;
end i.
```

The key elements to be completed in the algorithm are to define acceptability, and how new members of a group can be recognized. In addition, the concept of **join**, i.e., the process by which some subsets are joined to others, is interpreted as a mathematical operation in Appendix 1.

It is assumed that the $n(n - 1)/2$ pairwise relationships among n objects are empirically observed or can be computed from one or more of their attributes. Without loss of generality, it is also assumed that the distances are scaled so that the maximum value is unity.

Preprocessing

Let N be a finite set of n objects, too large for any reasonable clustering procedure. As usual, $\{d_{ij}\}$ is some measure of distance between all pairs of the objects, which can either be stored in computer memory (or be easily computed) or is available in

backing store. The objective is to find a quick way to divide the objects into sufficiently small-sized components such that each is amenable to some efficient clustering procedure. The model now described represents the objects by a graph on n vertices, with a path between vertices i and j iff the objects they represent belong to the same component.

A procedure for this division (already described in Chapter VII) is Toussaint's (1988) sphere of influence graph (SOIG). In summary, let r_i be the distance between object i and its nearest neighbor:

> DEFINITION *VIII.1. The generalized sphere of influence graph is a graph, G, on n vertices with edges* g_{ij} *defined by*

$$g_{ij} = \begin{cases} 1, & r_i + r_j > d_{ij} \\ 0, & \text{otherwise.} \end{cases}$$

This graph can be formed for any measure of distance (it need not be a metric) between objects i and j in any dimensional space. Note that

- the graph need not be connected
- one connected component can be completely enclosed within another
- the graph can be complete
- the subset of the $\{d_{ij}\}$ corresponding with $g_{ij} = 1$ is locally a metric.

If G is not connected, subpopulations are equated with the connected components; their existence and membership may

be determined quite easily by use of algorithms for spanning trees operating on G. The components first can be processed alone and then reassembled; in this way, a large problem can be broken into smaller subsets.

Generalizing betweenness to neighborhoods

The genesis of the ideas developed further here can be seen in Chapters V and VI. W.D. Fisher (1958) argued that if the objects are colinear, a subset containing two objects but excluding those between them is not to be considered as a candidate for a cluster. Fisher's heuristic can be regarded as defining the neighborhood of a pair of points. The challenge is to extend the concept of betweenness for unidimensional data to neighborhoods for multidimensional circumstances. It is useful first to review some preliminary attempts at this extension before giving what is currently considered to be best.

Some early neighborhood definitions[1]

Neighborhoods are here into two categories, asymmetric and symmetric, depending on the status of the individuals contained within them.

Asymmetric neighborhoods

If the distances among objects are such that the points corresponding to them are coplanar, Dagnelie (1966) proposed that a decision to include three of them in a subset but to exclude any (occupied) point contained within the triangle formed by the three is unreasonable. This model can easily be extended to more dimensions by replacing the triangles by simplexes, as follows. Let

[1] The reader uninterested in the discarded definitions may continue with "C-neighborhoods" (p. 231).

X be a $n \times m$ set of principal coordinates for n objects in m dimensions; choose any subset of $m + 1$ distinct rows, $\mathbf{X}_{m+1}$, assumed to define a nondegenerate simplex, and any other distinct row, x. Define

$$\mathbf{Y} = [\mathbf{X}_{m+1}:1]^{\mathrm{T}} \text{ and } \mathbf{y} = [\mathbf{x}:1]^{\mathrm{T}};$$

then

LEMMA VIII.1. *x is contained within the simplex defined by X_{m+1} iff the elements of α satisfying*

$$Y\alpha = y$$

are strictly positive.

The proof is omitted. This generalization of Dagnelie's proposal is here called a D-neighborhood.

Vinod (1969) extended Fisher's notion in the spirit of Dagnelie's into what he called a **generalized string**, and which here is called a V-neighborhood:

> any object k for which $d_{ik} \leq d_{ij}$ belongs to the V-neighborhood of the ordered pair (i, j).

In two-dimensional Euclidean space, the V-neighborhood is a circular region of radius d_{ij}, centred on i and including all points corresponding to objects k located within the circle. There is nothing in Vinod's proposal restricting the dimensionality. If all pairs of objects are considered, there are n^2 subsets (including the cases $i = j$). Although the generating pair differs, several of these subsets may be identical in the sense that the union of the objects included equals their intersection, so that the number of distinct subsets is often less than n^2.

Another definition of a neighborhood is that of **α-talons**, which are:

> for a specified α, the subset generated by object i and all objects k for which $d_{ik} \leq \alpha$.

This definition appears to achieve a modest number of subsets (no more than n) at the price of having to select a value for α.

However, the possibilities so far considered have a defect; it is that there are two classes of objects, namely, the generators, and those subsequently admitted. One reasonable additional requirement is that of *symmetry*:

> *all* objects belonging to a subset should be equivalent.

V-neighborhoods and α-talons are asymmetric in that the subset centred on object i may differ from that centred on j. The asymmetry arises from the fact that object i bears a different relationship to the remaining objects in the subset than the latter do among themselves; in an α-talon, one object is no further than α from all others, which may be separated by as much as 2α distance units from each other.

Nevertheless, consideration of these proposals leads to a general neighborhood principle:

> if certain specified objects, the generating objects, belong to a subset, any other object satisfying some proximity criterion to them also belongs to the subset.

Symmetric neighbourhoods

A symmetric neighborhood rule is that of **α-cliques**, or α-maximally complete subgraphs. In these subsets, no individual

is further than α from each other; the number of these subsets is unpredictable. Both for α-talons and α-cliques, the choice of α is critical; there may be a "best" value, but finding it can involve much computation (Lefkovitch 1975). However, if for some value of α these subsets are disjoint (i.e., no member is closer than α to any object in another clique), the partition formed is of direct interest. These disjoint subsets, here called α-**partitions**, are not common in real examples, except when α is chosen so that many of the objects form subsets of one member.

A defect in this symmetric rule, which it shares with the asymmetric, is that they all imply hyperspheroidal neighborhoods in the space of the distances; if these neighborhoods are conjectured to coincide with the underlying (unknown) true groups, the implication is that they also are hyperspheroidal, albeit differing in position and size. There is no justification for believing that a true group can be contained within a ball.

An improvement on this situation is an elliptical neighborhood in which neither size, shape, nor orientation are constant. Such neighborhoods can be derived by combining V-neighborhoods with D-neighborhoods as follows:

> for objects i, j, and k, any object l satisfying
>
> $$d_{il} + d_{jl} \leq d_{ik} + d_{jk}$$
>
> belongs to the Dagnelie–Vinod (DV) neighborhood defined by the ordered trio (i, j, k).

In two-dimensional Euclidean space, these neighborhoods are ellipses in which objects i and j are situated at the foci, and k on the boundary; the eccentricity is

$$d_{ij}/(d_{ik} + d_{jk}).$$

Denoting the subset generated by $\{i, j,k\}$ as $S(i, j,k)$, then:

(1) $S(i, j,k) = S(j, i,k)$.

(2) $S(i,i,k)$ is equivalent to a V-neighborhood.

(3) $S(i,i,i)$ consists of the object i and all those of identical position.

(4) $S(i, j,i)$ and $S(i, j, j)$ consist of objects i and j and those on the straight line between them. Since the probability of there being any objects satisfying (4) is virtually zero, $S(i, j,i)$ can be replaced by $S(i,i, j)$, and $S(i, j, j)$ by $S(j, j, i)$, i.e., V-neighborhoods.

(5) The maximum number of distinct subsets for this combined procedure is

$$(n^2 + 2)(n - 1)/2.$$

A problem with using DV-neighborhoods is that generating the subsets requires arithmetic of $O(n^4)$. However, if the maximum distance between two objects is unity, and global and near-global subsets are unlikely to be of interest, there is no need to consider subsets if $d_{ik} + d_{jk} \geq 1.0$ or if $d_{ij} \geq 0.5$; these lower limits may be further reduced on heuristic grounds. Setting the interfocal distance to $\max(d_{ij}, d_{ik}, d_{jk})$ and placing the remaining object on the boundary reduces the arithmetic by a factor of 3; subsets having an eccentricity near unity also can be excluded, because they are approximately equivalent to a straight line.

Unfortunately, in widening the class of subset-containing shapes, asymmetry is still present; further, the reductions still leave the arithmetic to be $O(n^4)$. To reduce the arithmetic to $O(n^3)$

or less, subsets should be initiated by pairs of objects. F-neighborhoods (**flexible strings**; Lefkovitch 1978) are an attempt at this. Suppose from assumptions VIII.3 and 4 the shape of the ellipses is altered by arranging that the eccentricity is smaller the greater the distance between the pair of generating objects. To be present in the neighborhood, an object needs to be close to the straight line between them if the generating pair are close, but can be relatively more remote if the generating pair are further separated. Defining a parameter $\gamma \geq 1$,

any object k for which

$$d_{ik} + d_{jk} \leq \gamma(e^{d_{ij}} - 1)$$

belongs to the F_γ-string defined by the pair $\{i, j\}$.

It follows that the eccentricity is $d_{ij}/\{\gamma(e^{d_{ij}} - 1)\}$. Because the maximum distance is unity and hence the sum of any two distances cannot exceed 2, the effect of γ is to eliminate from consideration those subsets for which the generating distance is greater than $\log_e(1 + 2/\gamma)$. There is an obvious criticism of a criterion that compares the sum of two distances with an exponential of a third; such a comparison is akin to comparing a length with an area, which is perhaps meaningless. However, this proposal results in reducing the amount of arithmetic while having an enlarged family of shapes, even though it is not symmetric in that the objects in the final set fall into two classes—the initiators, and those included subsequently. The following neighborhood rule weakens this distinction, effectively removing it from having any serious effect.

C-neighborhoods

Even though the neighborhoods described above begin to satisfy some of the requirements, they still imply a restricted family of

shapes in the distance space, namely, ellipsoids (including spheroids), even though they can be of different shapes, sizes, and positions. The neighborhoods described here permit a larger family of shapes, and the asymmetry is essentially removed.

The reasoning presented in Chapter VI for clustering unidimensional data defined a sequential procedure, which is

> subset S_{t+1} consists of all objects whose average distance to the members of S_t does not exceed the maximum among objects in S_t,

i.e.,

$$S_{t+1} = \{k : \text{ave}(d_{ik} |\ i \in S_t) \leq \max(d_{ij} |\ i, j \in S_t)\}.$$

Nothing in this representation requires unidimensionality; the argument depends only upon distance. The definition of the C-neighborhood is the region satisfying

$$\forall w : \text{ave}(d_{iw} |\ i \in S_t) \leq \max(d_{ij} |\ i, j \in S_t),$$

where w is a point in the distance space. Thus for an initial pair of objects, i.e., at stage $t = 0$, the criterion assigns all objects in the region defined by the rule as members of the subset. At stage $t = 1$, and in two dimensions, this region is an ellipse; in more than two dimensions, it is an ellipsoid, hyperellipsoid, and so on. Notice the deceptively simple but quite fundamental aspect of this definition: if the neighborhood at stage $t + 1$ contains objects other than those at stage t, the process is repeated by determining the average distance of each nonmember to *all* members, which is then compared with the maximum among the members. Notice also that because

$$\max(d_{ij} |\ i, j \in S_t) \geq \text{ave}(d_{ij} |\ i, j \in S_t),$$

the generating procedure can be written as

$$S_{t+1} = S_t \cup \{k:\text{ave}(d_{ik} |\ k \notin S_t,\ i \in S_t) \leq \max(d_{ij} |\ i, j \in S_t)\}.$$

Appendix 5 shows that for any initiating S_0, there is a stage such that $S_{t+1} = S_t$; this fact allows the following:

DEFINITION VIII.1. *A C-neighborhood subset satisfies* $S_{t+1} = S_t$.

Thus a C-neighborhood subset is conditional on the initial members; this dependence is the source of what I call **conditional clustering**. Some properties of C-neighborhoods as a subset-generating procedure are as follows:

(1) Since all members of S_t are used in obtaining S_{t+1}, the process is symmetric.

(2) Many subsets initiated by more than two proximal objects are also considered during the sequences, which implies that subset generation is $O(n^3)$ and can be restricted to the $\binom{n}{2}$ pairs as starting points.

(3) The range of shapes of the neighborhoods in the distance space consists of multifocal "ellipses," their higher-dimensional counterparts, and includes standard bifocal ellipses.

A possible defect, which does not appear to be major, is that an average distance is being compared with a maximum. An alternative possibility, which does not use averages, is of the Dixon-type used in testing for statistical outliers; it is

include object k in the subset S_{t+1} if the minimum distance it has to a member of S_t does not exceed the maximum among S_t.

A neighborhood based on this principle of inclusion seems plausible, but empirically it has been found to generate few and apparently heterogeneous subsets, consistent with the conservative rejection rules characteristic of statistical decisions. Another related possibility is to determine

include object k in subset S_{t+1} if
$$\min\{d_{kk'} \mid k' \in S_t\} \leq \max\{d_{jk'} \mid j \in S_t\},$$

but this decision criterion also has been found to generate very few and heterogeneous subsets.

A modification that changes the average distance criterion by a parameter $0 < \gamma < 1$ is conceivable, but if γ depends on an externally chosen value, it is not recommended. Attempts to define $\gamma_t = \gamma(S_t)$ as some function of S_t, such as

$$\gamma_t = \gamma(\min\{d_{ij}\}, \max\{d_{ij}\}), \forall i, j \in S_t,$$

failed to be satisfactory; either many small subsets (two or three objects only) or few large subsets were produced. Furthermore, $\min\{d_{ij}: i, j \in S_t\}$ depends on the sample size, which in turn depends partly on the collectors' behavior. The consequences of biased collecting have already been discussed and need to be avoided (Chapter I).

Experiments based on these neighborhood definitions, coupled with the supposed desirable properties for a subset, have led me to conclude that C-neighborhoods represent the widest range of possibilities coupled with a simple definition and modest

amounts of computation. Some aspects of the geometry of multifocal ellipsoids are discussed in Appendix 4; the convergence of the subset-generating procedure based on C-neighborhoods is investigated in Appendix 5.

Reducing the amount of arithmetic

It is not wrong to use each of the $\binom{n}{2}$ pairs of objects to initiate a subset to obtain C-neighborhoods. Some pairs will generate the improper subset (i.e., containing all n objects), which is hardly of interest. The same proper subset may be generated by different initial pairs. The purpose of this section is to show how to avoid this duplication without eliminating the subsets likely to participate in the optimal covering and so reduce the quantity of arithmetic. The general strategy is to eliminate from $\{S_0\}$ superfluous initial pairs.

Determining S_0 using a four-point condition

Fitch (1981) defined a **neighbors relation** for the pair of objects $\{i, j\}$ relative to the pair of objects $\{p,q\}$, which can be written as follows:

> assuming $i \neq j$, $p \neq q$, $\{i, j\} \neq \{p,q\}$ and **D** is a metric, then i and j are neighbors iff

$$d_{ij} + d_{pq} < \min(d_{ip} + d_{jq}, d_{iq} + d_{jp}).$$

This definition can be used to select the S_0 in the following way. If $\{p,q\}$ ranges over all pairs chosen from N satisfying the definition, the total number of instances where $\{i, j\}$ are neighbors provides a score for $\{i, j\}$. Pairs with many neighbors can be regarded as being more central than those having a lower score and so are likely to be good candidates for subset generation. The pairs can be used in descending order of the scores, continuing

until the generated subsets consist of N. Because the distances are assumed to be a metric, and also because determining the scores requires arithmetic of $O(n^4)$, more than using all $\binom{n}{2}$ initial pairs, this proposal is rejected for the present purposes.

Determining the S_0 by graph theory

Suppose three objects in the distance space form the vertices of a nondegenerate equilateral triangle. Then for each of the three pairs, the first iteration in the C-neighborhood procedure includes the third vertex, because the average distance the third has to the initial pair is equal to the distance between them. Thus the three initial pairs generate the same subset, and so only one of the pairs is required. For nonequilateral triangles, using the pair adjacent to the shortest side generates a subset just of the pair (in the absence of other objects in the vicinity), while the subset initialized by the longest side generates a subset of all three. However, the intermediate length side almost certainly also generates the subset of three unless the "remote" vertex is far removed from the nearer of the pair, when the subset formed just by those three is of little interest. It is concluded that the pairs separated by the larger distances are likely to generate subsets that either are also generated by a less separated pair, or are unlikely to be a part of the optimal covering.

Thus the problem can be reformulated as determining a graph on n vertices, such that the adjacent pairs of objects generate the subsets of interest. The possibilities for this graph considered here are based on different definitions of neighbors.

The nearest neighbor graph (NNG) This graph is defined by joining each object to its nearest neighbor in the distance space. With no requirement for the triangle inequality to be satisfied in recognizing nearest neighbors, this graph requires no particular

metric. This graph need not be connected, is cycle free, and consists largely of isolated pairs of vertices, and so the number of edges may be as few as $n/2$; any large structures in the data tend not to be generated. This graph can be formed with $O(n \log_e n)$ arithmetic.

The minimum spanning tree (MST) This well-known graph is connected, has $n - 1$ edges (therefore is cycle-free), and is of minimum length in the distance space. It is unique if all distances differ, and it includes the NNG as a subgraph. Using just the adjacent objects as the initial pairs, some groups of well-separated objects may be formed. As for the NNG, there is no requirement for the distances to satisfy the triangle inequality. This graph can be constructed with $O(n \log_e n)$ arithmetic for spaces of any dimensionality.

(1) *The square of the MST* The square of *any* simple graph consists of the graph together with further edges joining vertices separated by two edges of the original graph. The square of a spanning tree is a trigraph (Bondy 1989).

(2) *The principal weighted spanning tree decomposition* In a MST, $n - 1$ elements of the distance matrix, **D**, are selected so that $(n - 1)(n - 2)/2$ remain unused. Suppose small changes in the empirical data result in some of the elements of the new **D** to change in such a way that edges between different pairs may form part of the MST. These new pairs are almost certain to be those for which the distances are small in the original **D**; hence they are near neighbors but are excluded from the MST because of the latter's acyclic connected structure. These pairs of objects are almost certainly adjacent in the shortest spanning tree formed from the original **D** excluding edges in the MST.

That such a second tree exists is assured for $n > 3$, unless the MST is also a star graph. Together with those pairs adjacent on the MST, these provide an another initial set to form **A**. Thus the procedure is to determine a spanning tree from **D** that satisfies two conditions:

- it has minimal length
- pairs adjacent in any previous tree cannot be used.

The identification of such spanning trees can be repeated until the graph becomes disconnected; there are at most k such trees, where k is the largest integer not exceeding $n/2$. For n even, all edges may have been used; for n odd, the remaining edges do not form a connected graph.

Let $\mathbf{E}_i$ be the adjacency matrix corresponding with the edges of the i^{th} weighted spanning tree, $i = 1...k$; let $\boldsymbol{E}_{k+1}$ be the adjacency matrix of the remaining edges, with the possibility that $\mathbf{E}_{k+1}$ may be null. By definition, these adjacency matrices are additively orthogonal; further:

THEOREM VIII.1. $\boldsymbol{E}_i\boldsymbol{E}_j = \boldsymbol{0}$, $i \neq j$, *i.e., the adjacency matrices are pairwise orthogonal.*

Proof. The assertion follows from the symmetry of the adjacency matrices, from the positions of the zeros, and from the definition of the decomposition. Q.E.D.

Because the sum of the distances corresponding to the edges of these spanning trees are of minimal length conditional on those previously extracted, and the set of adjacency matrices are additively and multiplicatively orthogonal, the set of spanning trees

is here called a **principal weighted spanning tree decomposition** (PWSTD) of the distance matrix. A numerical example is given in Table VIII.2, in which, if the initial pairs for subset generation are based on $\mathbf{E}_1$ and $\mathbf{E}_2$, only the two pairs in $\mathbf{E}_3$ are excluded. Although for larger n, three (or more) spanning trees may be used, it seems that for data in which there are distinct groups, two trees may be adequate.

Table VIII.2 An example of the principal weighted spanning tree decomposition

$$\mathbf{D} = \begin{bmatrix} . & & & & \\ 0.5 & . & & & \\ 0.4 & 0.3 & . & & \\ 0.6 & 0.8 & 0.7 & . & \\ 0.2 & 0.9 & 0.3 & 0.8 & . \end{bmatrix} \qquad \mathbf{E}_1 = \begin{bmatrix} . & & & & \\ . & . & & & \\ . & 1 & . & & \\ 1 & . & . & . & \\ 1 & . & 1 & . & . \end{bmatrix}$$

$$\mathbf{E}_2 = \begin{bmatrix} . & & & & \\ 1 & . & & & \\ 1 & . & . & & \\ . & . & 1 & . & \\ . & . & . & 1 & . \end{bmatrix} \qquad \mathbf{E}_3 = \begin{bmatrix} . & & & & \\ . & . & & & \\ . & . & . & & \\ . & 1 & . & . & \\ . & 1 & . & . & . \end{bmatrix}$$

Computing the PWSTD using any algorithm for the MST presents no difficulty; after identifying the edges in $\mathbf{E}_1$, set the corresponding elements in $\mathbf{D}$ to a value greater than the sum of the elements of $\mathbf{D}$, and use the MST algorithm; this sequence can be repeated for as many trees as are thought to be needed.

The relative neighborhood graph (RNG) The adjacency matrix of this graph is defined by

$$\mathbf{E} = \{e_{ij}\} = \begin{cases} 1, \; d_{ij} \leq \min\{\max(d_{ik}, d_{jk})\}, \; \forall k \neq i, j; \; i \neq j \\ 0, \text{ otherwise.} \end{cases}$$

If $e_{ij} = 1$, then objects i and j are relative neighbors (Toussaint 1980). This graph is connected; the MST is a subgraph (Toussaint 1980) and may contain cycles. In special cases, the RNG may be identical with the MST. In consequence, the number of edges may be as few as $n - 1$, or, if all triangles are equilateral, as many as $\binom{n}{2}$; empirical investigation (Lefkovitch 1984) indicates that no more than about $3n$ edges are found in random distances, and fewer in structured data. In planar graphs, Urquhart (1983) has shown that an upper bound for the number of edges is $3n - 10$ for $n > 7$. This graph can be formed in $O(n^3)$ arithmetic for spaces of any dimensionality; the distances need not be metric.

(1) *The square of the RNG* The number of edges in the square of this graph is unpredictable. For example, the square of a triangle is identical with the triangle, and so it is conceptually possible that no new edges are inserted.

The Gabriel graph (GG) Originally defined for the plane (Gabriel and Sokal 1969, Matula and Sokal 1980), two vertices are adjacent in this graph iff the unique hypersphere of radius $d_{ij}/2$ passing through them (called the **hypersphere of influence**) is empty of occupied points. This graph, which can be formed in $O(n^3)$ arithmetic for spaces of any dimensionality, assumes that the distances are Euclidean.

Voronoi neighbors graph (VNG) Voronoi neighbors are defined by

$$\mathbf{V} = \{v_{ij}\} = \begin{cases} 1, \; d_{iw} = d_{jw} = \min\{k,w \mid i, j,k \in N\} \\ 0, \text{ otherwise} \end{cases}$$

where w is *any* point in the space, and i, j, and k correspond with real objects. Toussaint (1980) has shown that the VNG is a

supergraph of the RNG and noted that its dual is the Delaunay triangulation. The conditions on the points $\{w\}$ require a metric. This graph is easy to form in the plane but requires arithmetic that increases exponentially with the dimensionality of the space.

Sphere of influence graph (SOIG) From Definition VII.1, objects i and j are SOIG neighbors iff $r_i + r_j > d_{ij}$, where r_i is the distance to the nearest neighbor of object i. It follows that the NNG is a subgraph of the SOIG. Because the number of edges in such a graph may exceed the number of Voronoi neighbors, this method for selecting S_0 is ignored. However, if this graph has been formed to determine connected components, there may be merit in excluding edges from the RNG that connect components in the SOIG.

The traveling salesman graph (TSG) One further possibility is to select those initial pairs of objects that are adjacent on the graph computed from a shortest length "traveling salesman's tour" of the distance array. Frieze (1987) described an $O(n^3 \log_e n)$ randomized algorithm for this problem, which obtains the optimal solution in (integer) edge-weighted graphs with limiting probability of unity for $n \to \infty$. Lau (1986) provided a Fortran program to find a solution guaranteed to be no worse than 1.5 times the length of the optimum. Neither procedure has been studied for the present purpose, because they involve more computational steps than using all distinct pairs of objects for subset initiation.

Because the NNG is likely to provide few initial pairs, it is not considered further. From among the remaining graphs, the adjacent vertices on the MST offer a useful set of starting pairs, but the simultaneous requirement of connectivity and lack of cycles may link pairs of vertices that are not close neighbors in the distance space; on regular grids, moreover, there are no useful treelike descriptions that are simultaneously useful for the present

purposes. Because the RNG does not disqualify cycles, this defect of the MST is compensated by additional edges, at least potentially. This compensation also exists in the square of the MST, by using two (or more) trees from the PWSTD, and the VNG. The VNG requires the definition of new points, which in turn depend on the metric of the space; this requirement contrasts with the NNG, the MST, the RNG, and their squares and with the combined spanning trees, which depend only on an ordering of the numerical values of the distances separating the real objects. The SOIG need not be connected but is too "richly" connected for the present purposes. For similar reasons, the geographic neighbor graph (Yao 1982) and the GG, both of which are also supergraphs of the MST, are rejected for the present purposes. Because the RNG does not require that the distances satisfy the triangle inequality, it is the richest of the neighborhood graphs, neither requiring nor implying a metric in the distance space nor having special requirements beyond being connected, and is the preferred of those considered. Other than the RNG or the graph formed from combining two or more spanning trees, there may be a simply-formed graph not requiring a metric, which is not necessarily connected and is also not as rich as the SOIG, but such a graph is not known to the author. Nevertheless, Kirkpatrick and Radke (1985) discussed a mathematical framework for the definition of neighbors in two-dimensional space, which may provide a starting point for defining such a graph.

It may be argued that confining the initial pairs of objects to those adjacent on the RNG may result in the division of large groups (large in the sense of distance, not cardinality); this possibility is readily admitted. The worst of any resulting problems (practical experience suggests all) can be remedied by forming musters (Chapter II and Appendix 6). Musters are defined as the union of intersecting subsets, formed because the membership of an object in two or more subsets implies either inadequate data, or

that the boundary of a true group in the subset space does not coincide with a member of the family of multifocal ellipses. Expressed in another way: why should a true group represented by some dissimilarity measure necessarily be containable within a single member of the family? Muster formation is an integral part of the procedure, whether or not the RNG supplies the only initial pairs.

In summary, recognizing which initial pairs of objects satisfy "acceptable(group)" (Table VIII.1) is given by those pairs adjacent in the RNG. Thus in Table VIII.1, the outer two loops can be replaced by this condition. A further heuristic can be added; eliminate any RNG pair if they are not adjacent on the SOIG, because the subset generated is likely to be contained within a large proportion of the distance space. This additional heuristic may be replaced by retaining the group if

$$\max(d_{ij} | i, j \in S_t) < \tfrac{1}{2}\max(d_{ij} | i, j \in N),$$

or, somewhat more questionably, if

$$|S_t|/n < 0.5,$$

assuming the maximum distance is 1.0.

The optimal solution

The subsets generated by the recursive C-neighborhoods described above are assembled into the matrix **A** (Chapter II), and the optimal solution is found by the procedures described there. Having found a minimum-cost solution, it still remains to decide what it means. To identify each subset in the solution with an underlying group is to assume that the space in which the subsets were generated has a uniform metric throughout. This assumption in turn implies that the measure of distance is appropriate, and that

the cost vector is well chosen. However, the subsets in a minimum-cost solution obtained by the methods described here cannot be assumed to coincide with one of the underlying groups, even though the C-neighborhoods encompass a wide variety of sizes, shapes, and orientations in the distance space. The doubt can be seen from the following reasoning. If even one of the underlying groups, as represented by the points in the sample, is elongate or branched (using these words to describe the subjective boundaries) in the distance space with respect to the C-neighborhoods, more than one subset in the optimal covering may correspond with that group. For example, a minimum-cost solution may consist of several ellipses approximating the boundary of a circle, surrounding another ellipse contained within, so that many overlapping or contiguous regions may correspond to one underlying group. Further treatment of the optimal solutions is discussed under the name of musters in Appendix 6.

In examples of real data, several of the subsets in the solution consisted of single objects. If each singleton is thought to correspond with a distinct underlying group, it can be removed and the remaining subsets processed further. Subsequently, it is advantageous to remove all but one member of each multiple-object subset and repeat the whole process with them together with the removed singletons, thus obtaining a grouping at a higher level of taxonomic relationships. Each multiple object disjoint group or nondisjoint set of groups can also be processed separately and its internal structure investigated at a lower level of classification.

In Chapter X, the fifth data set of "ANOVA means" illustrates the application of C-neighborhoods to the grouping of means based on t-values, which, because the variances are apparently not homogeneous, is a multidimensional problem. Also in Chapter X, "Letters" gives a brief example from psychometry.

Consistency

As indicated in Chapter I, if clustering is to be anything other than referring to a given set of objects as described by a given set of attributes, some conditions for consistency need to be satisfied. An attempt is made here to expand on consistency. The arguments can be followed most easily if all attributes are considered as being one-state, but they carry through, albeit with complications, for multistate attributes (including nominal, ordinal, or continuous) as well as for dissimilarities. Two assumptions are made:

(1) The objects studied are a random sample of the members of a finite but unknown number of "true" groups.

(2) The true groups differ amongst each other with respect to the attributes from which the family of subsets has been generated.

Let $S(\mathbf{A})$ be the fully reduced $\mathbf{A}$, i.e., after the completing all possible reductions described in Chapter II; let $r(N|M)$ be the number of rows in $S(\mathbf{A})$, where N is the set of objects under study, and M is the set of attributes used to describe them.

PROPOSITION VIII.1. *For a sufficiently large randomly chosen set of attributes, M,*

$$lim_{|N|\to\infty} r(N|M) \to R < \infty,$$

i.e., the number of rows in S(A) has a finite limiting value independent of $|N|$.

Proof. Suppose that not each "true" group is represented in N; for any set N, no more than $|N|$ groups can be

represented. Assume that g groups are represented, i.e., $g \leq |N|$; if further objects are included one at a time, then either g is unchanged at each inclusion, i, in which case $r(N \cup i|M) = r(N|M)$, or it is increased by unity. If it is increased by including object i, then $r(N \cup i|M) = r(N|M) + 1$. Eventually, all groups will be represented, i.e., each of the "true" groups has at least one representative included in N; the inclusion of an additional set of objects N', identical with some of those either currently in or covered by those currently in, or which cover some currently in **A**, will not change the number of rows, i.e., $r(N \cup N'|M) = r(N|M)$. Q.E.D.

Suppose a family of m subsets is generated from M attributes, and let $c(M|N)$ be the number of columns in $S(\mathbf{A})$.

PROPOSITION VIII.2. *For a fixed set of objects, N,*

$$lim_{|M| \to \infty} c(M|N) \to C < \infty,$$

i.e., the number of labeled subsets has a finite limiting value independent of M.

Proof. The proof is similar to that of Proposition VIII.1. Suppose that not each of the "true" groups represented in N can be distinguished by the M attributes. For any set of M, no more groups than the number of distinct combinations of their states can be distinguished. Assume that this number is g; if further attributes are included one at a time, then either g is unchanged by including a single attribute, j, in which case $c(M \cup j|N) = c(M|N)$, or it is increased by no more than the additional number of

combinations. Let this additional number be J; then $c(M \cup j \mid N) = c(M \mid N) + J$. As further attributes are included, eventually all groups included in N become distinguishable. At this stage, the M are a set of attributes that correctly distinguish objects belonging to the "true" groups included in N and that give rise to a family of subsets, **A**, of N, which reduce to $c(M \mid N)$ in $S(\mathbf{A})$. Including M' additional attributes whose states coincide with some of those currently in M, or which form a partition of some of those currently in M, or for which the states of those currently in M form a partition of some of those in M', does not change the number of columns, i.e., $c(M \cup M' \mid N') = c(M \mid N)$. Q.E.D.

Remark VIII.1. For this proof, when applied to dissimilarity coefficients estimated from attribute data, the estimates are also assumed to be consistent, i.e., as more attributes are included, the value of the dissimilarity converges to a (nontrivial) limit.

COROLLARY VIII.1. *The membership of the labeled subsets becomes stable, i.e.,*

$$lim_{|N| \to \infty, |M| \to \infty} S(A) \to \underline{A},$$

a matrix of R rows and C columns.

Remark VIII.2. No matter which objective function is used, should one be needed to obtain an optimal covering, the fact that the S(**A**) is a consistent estimator of $\underline{A}$ means the grouping obtained is also consistent.

Remark VIII.3. It seems likely that the rate of approach to $\underline{A}$ is slower for increasing $|M|$ than for increasing $|N|$ because of the lack of independence among attributes of individuals.

These propositions suggest that the probabilities should be determined from $S(\mathbf{A})$ rather than $\mathbf{A}$. By doing so, the problem of the lack of consistency of the probabilities, which has been pointed out for the Rasch model by Baker (1992), is not an issue. However, the theory described by Neyman and Scott (1948) can be used to obtain consistency.

IX Special applications and additional topics

1 A multiple-entry identification protocol

An optimal covering for a set of n objects can be represented by the $n \times q$ array $\mathbf{Q}$ (where $q = \mathbf{x}^T\mathbf{x}$) consisting of the columns of $\mathbf{A}$ for which $x_k = 1$. Suppose the empirical data from which $\mathbf{A}$ was derived were p one-state attributes (extension to other types of attribute are considered below) for the n objects, denoted by the $n \times p$ array $\mathbf{P}$; sometimes, $\mathbf{A}$ is identical with $\mathbf{P}$ (Chapter III). Thus the empirical data for i^{th} object is represented by the i^{th} row of $\mathbf{P}$, here denoted by $\mathbf{p}_i$, and its membership of the various subsets in the optimal covering by $\mathbf{q}_i$, the i^{th} row of $\mathbf{Q}$; just one element of $\mathbf{q}_i$ may be unity. $\mathbf{Q}_k$ denotes the k^{th} column of $\mathbf{Q}$.

This section defines a procedure to obtain a correct prediction of the (possibly unknown) $\mathbf{q}_i$ from a $\mathbf{p}_i$, i.e., the membership of object i in the optimal covering. The **predictor vector** for the i^{th} object therefore is $\mathbf{p}_i$, and the **dependent vector** is $\mathbf{q}_i$. A $q \times p$ **matrix of templates**, $\mathbf{V}$, with elements equal to zero or unity, is formed by this procedure.

One-state attributes

Two steps are required: first, to establish predictor templates based on known data from the set of objects, and second to determine an assignment rule. It is convenient to define two algebraic operations, $\oplus$ and $\ominus$, at this stage:

(1) (a) $(1 \oplus 1) = (0 \oplus 0) = 1$;
(b) $(1 \oplus 0) = (0 \oplus 1) = 0$.

(2) The $\oplus$-inner product between two column $\{0,1\}^p$ vectors is
$$\mathbf{a}^T \oplus \mathbf{b} = \sum (a_i \oplus b_i).$$

The product between two matrices is defined as an extension of these operations.

Establishing initial templates

As will become apparent, at least three templates are desirable. Although a random choice may be used, q of these may be obtained from $\mathbf{P}$ and $\mathbf{Q}$. Let $\mathbf{E}$ be a $n \times p$ array of unities. An initial estimate of $\mathbf{V}$ has elements equal to unity iff the corresponding elements of $\mathbf{Q}^T\mathbf{P}$ exceed the corresponding elements of $\mathbf{Q}^T\mathbf{E}/2$ and is zero otherwise. The justification for this estimate is that there is a high probability that the templates will differ for each group. In the unlikely event that any rows are duplicated, all but one can be removed and replaced by vectors chosen at random from $\{0,1\}^p$. A value for the initial number is indicated below.

Refining the templates; establishing the identification rule

The next steps both modify the initial templates and define the decision rule.

step 1: define
(a) n q-element vectors $w_k = 0$;
(b) an n-element vector, $z = 0$.

step 2: for $i = 1 \ldots n$,
(a) calculate $V \oplus p^T_i$ *(i.e., the number of elements in which* p_i *agrees with each template.*

(b) if this number is not less than p/2, flag these templates; if the maximum is no greater than p/2, flag the corresponding template(s); let F_i *denote the set of*

indices corresponding with the flagged templates for object i.

(c) for $\forall k \in F_i$,

$$w_{kj} = \begin{cases} w_{kj} + 1, & q_{ij} = 1 \\ w_{kj} - 1, & q_{ij} = 0; \end{cases}$$

(d) define the q-element predicted vector, y, as

$$y_i = \begin{cases} 1, & \sum_{k \in F_i} w_{ki} > 0 \\ 0, & \sum_{k \in F_i} w_{ki} \leq 0; \end{cases}$$

(e) if $y = q_i$, *then* $z_k = z_k + 1$, $\forall k \in F_i$ *(i.e.,* q_i *is correctly predicted by y);*

(f) next i.

step 3:

(a) delete all templates for which $z_k = 0$ *(they are not involved in any correct prediction);*

(b) replace a remaining template for which the score is lowest by a new one formed as follows; the z^{th} *element takes the same value as in the templates with the two highest scores where the elements are identical and otherwise are randomly chosen. If this new template is identical with any of the remaining, choose a different random assignment; if this is not possible, use the first and third best, etc.; if all fail, choose at random.*

step 4: repeat steps 1–4 either until convergence or until the proportion of incorrect predictions falls below some prespecified level.

The final set of templates, coupled with the prediction rule, provides the identification protocol.

Multistate and ordered attributes

As described in Chapter III, an attribute having $s > 1$ mutually exclusive states is represented by s one-state variables (absence is not considered to be a state). It follows that the only change needed in the procedure described for one-state attributes is in step 2(b), where, rather than determining if the number of elements in which $\mathbf{p}_i$ agrees with a template exceeds $p/2$, the value be replaced by half the number of attributes.

Ordered, including continuous attributes

Within any of the subsets of objects defined by $\mathbf{Q}$, ordered attributes are likely to exhibit a range of values, so requiring two modifications to the basic procedure.

First, in step 2(a), if the attribute is not outside the range of values given by the template, it will be said to agree. This agreement can be represented by an extension of the definition of $\oplus$. Let α denote either a *single* value associated with an object or a *range* of values associated with a group of objects, and β the *range* of values defined by the template; then

$$\alpha \oplus \beta = \begin{cases} 1 \text{ if } \alpha \subseteq \beta \\ 0 \text{ if } \alpha \not\subseteq \beta \end{cases}$$

allows the operator $\oplus$ to be used without further definition.

Second, in step 3(b), even though the new range may be chosen at random from one of the pair of templates from which the replacement is to be formed, other possibilities include the union

of the ranges, the intersection of the ranges, and others perhaps based on more formal statistical methods.

It is apparent that the identification of object j need not depend on a single template, but on several, and that there need not be an exact match with any one. Since there are $2^m - 1$ nonempty subsets of the templates, up to this number of different predictions are conceptually possible. Thus if the m initial templates are distinct, this number should be more than sufficient to provide an identification protocol. This procedure tends to be most effective if the numbers both of objects and attributes are large, the number of subsets is relatively small, and the variability among the members of the subsets is relatively low.

The whole of the second phase, especially step 3b, may be regarded as analogous to the biological theory of natural selection, because it eliminates the weak predictors, generates new ones from the most successful, and fills gaps by immigration.

Example IX.1.1 In this numerical example, it is assumed that templates have been obtained.

I. To establish the survival of the templates, suppose for the i^{th} object:

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array}\begin{array}{c} \mathbf{V} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{W} \\ \begin{bmatrix} 6 & -8 \\ -9 & 5 \\ 3 & 2 \\ 5 & -5 \end{bmatrix} \end{array} \qquad \begin{array}{c} \mathbf{z} \\ \begin{bmatrix} 8 \\ 9 \\ 2 \\ 6 \end{bmatrix} \end{array}$$

Let the next $\mathbf{p} = [1\ 0\ 0\ 0\ 1\ 1]$, $\mathbf{q} = [1\ 0]$.

The number of agreements of $\mathbf{p}$ with $\{a,b,c,d\}$ is $\{2,2,5,4\}$.
Thus the F_i templates are $\{c,d\}$.
The sum of W corresponding with F_i is $\{8,-3\}$.
This implies $\mathbf{y} = [1\ 0]$.

Since $y = q$, update W and z:

$$\mathbf{W} = \begin{bmatrix} 6 & -8 \\ -9 & 5 \\ 4 & 1 \\ 6 & -6 \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} 8 \\ 9 \\ 3 \\ 7 \end{bmatrix}$$

Continue with the next known object.

II. Suppose object *i* had been the last in the current pass:

Delete row *c* from V since the value of z_c is the minimum.
Form *c'* from V(*a*) and V(*b*), in which the elements are chosen equiprobably, e.g., [1 0 1 1 0 0].

Restart the process with:

$$\begin{array}{c} \\ a \\ b \\ c' \\ d \end{array} \mathbf{V} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{z} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Example IX.1.2. This larger numerical example shows two iterations.

Iteration 1:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \\ I \\ J \end{array} \mathbf{P} = \begin{bmatrix} 111111..... \\ .111111..... \\ ..11111..... \\ ..111111.... \\ ..1111111... \\ ......1111.. \\ .....111111. \\ ....11111111 \\ ......111111 \\ ...1...11... \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} 1.. \\ 1.1 \\ 1.. \\ 1.. \\ 1.1 \\ 11. \\ .11 \\ .1. \\ .1. \\ 11. \end{bmatrix}$$

$$\mathbf{Q}^T\mathbf{P} = \begin{bmatrix} 125655543200 \\ 000110455432 \\ 002223322210 \end{bmatrix} \qquad \mathbf{Q}^T\mathbf{E}/2 = \begin{bmatrix} 3.5 \\ 2.5 \\ 1.5 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} 001111110000 \\ 000000111110 \\ 001111111100 \\ 111100001110 \end{bmatrix}$$

Iteration 2:

$$\mathbf{Z} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} 000 \\ 000 \\ 000 \\ 000 \end{bmatrix}$$

$$\mathbf{V}\Theta\mathbf{P}^T = \begin{bmatrix} 8 & 1 & 6 & 7 \\ 10 & 2 & 8 & 5 \\ 11 & 4 & 9 & 4 \\ 12 & 5 & 10 & 3 \\ 10 & 7 & 12 & 5 \\ 6 & 11 & 8 & 5 \\ 6 & 11 & 8 & 2 \\ 6 & 9 & 8 & 3 \\ 4 & 11 & 6 & 3 \\ 7 & 8 & 7 & 7 \end{bmatrix}$$

$$\mathbf{F} = \begin{bmatrix} 1 \\ 1,3 \\ 1,3 \\ 1,3 \\ 1,2,3 \\ 2,3 \\ 2,3 \\ 2,3 \\ 2 \\ 1,2,3 \end{bmatrix} \qquad \mathbf{W} = \begin{bmatrix} 6 & -4 & 0 \\ 0 & 4 & -2 \\ 4 & 0 & -2 \\ 2 & 0 & -2 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ 100 \\ 110 \\ 110 \\ 110 \\ 010 \\ 100 \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 3 \\ 2 \\ 3 \\ 2 \end{bmatrix}$$

Delete V_2 and replace:

$$\mathbf{V} = \begin{bmatrix} 001111110000 \\ 001111110100 \\ 001111111100 \\ 111100001110 \end{bmatrix}$$

There were five accurate predictions. Because complete accuracy was not attained, further iterations are needed. Of the three groups in the example, the first two (= columns of **Q**) are homogeneous, but the third is not, which makes successful identification unlikely.

2 Converting set systems to graphs

This book uses integer programming, in particular, finite set covering of N, to achieve a solution of the clustering problem for a given or generated family of subsets represented by **A**. However, **A** can also be considered in terms of graph theory, leading to possible alternative solution procedures. Several different graphs are now defined corresponding with **A**, each illustrating different aspects of the relationships among the objects and the various subsets.

Some preliminary definitions are in order. Let $\mathbf{A}_j$ denote the j^{th} column of **A**, and a_j the membership of the j^{th} subset. Let T be a subset of N such that $T \cap a_j \neq \varnothing$, $\forall j$, $|T|$ a minimum.

DEFINITION IX.2.1. *$|T|$ is the transversal number of N for **A**.*

DEFINITION IX.2.2. *A family of distinct representatives is defined as*

$$\{i: i \in a_j,\ i \notin a_{j'},\ j \neq j',\ \forall j, j'\}.$$

The first graph to be defined is based on the elements of N.

DEFINITION IX.2.3. *A simple graph, **A**, corresponding with **A** has n vertices corresponding with the objects, with an edge between two vertices iff the corresponding objects occur together in at least one subset.*

Since each a_j forms a clique, if $a_k \subseteq a_j$, only the clique corresponding with a_j is identifiable. Because each connected component of this graph is equivalent to a muster (Chapter II and Appendix 6), graph-theoretic algorithms for determining the connected components are useful for identifying musters. An efficient procedure to determine the connected components and their membership is by algorithms for obtaining a minimal spanning forest; each disjoint subtree corresponds with a component.

The next graph to be defined focuses on the subsets of the objects:

DEFINITION IX.2.4. *The representative graph, R, of A is a simple graph of order m in which each vertex corresponds with a subset, a_j, and a single edge exists between two vertices, j and k, iff $a_j \cap a_k \neq \varnothing$.*

The representative graph, here the intersection graph of the set system, is perhaps more useful for clustering than the simple graph A, because it allows many important properties of the set system to be determined. For example, to determine if there is just one muster in **A** is equivalent to determining if R contains a spanning tree. If R is not connected, identifying the connected vertices in each spanning subtree in it by algorithms for (minimum) spanning trees, is computationally very efficient. Note that the representative graph R is the dual to the simple graph A, but that dual to the dual is *not* an identity operation.

Let $G = G(V,E)$ be any simple (i.e., loopless, without parallel edges) graph with vertex set V, and edge set E consisting of $|E|$ unordered pairs of distinct vertices. Then

DEFINITION IX.2.5. *The graph dual to G is a graph in which each vertex corresponds with an edge of G, i.e., there are $|E|$ vertices, and an edge between two vertices iff distinct edges in G have an element of V in common.*

Occasionally, recognizing the number of objects involved in defining a graph is of value:

DEFINITION IX.2.6. *The intersection multigraph of a set system has m vertices, with $|a_j \cap a_k|$ edges between vertices j and k; a weight, $|a_j|$ can be assigned to vertex j.*

DEFINITION IX.2.7. *If $|a_j|$ is uniform (i.e., if the diagonal elements of the adjacency matrix of the multigraph are all equal), the set system can be regarded as a uniform hypergraph.*

An even more uniform set system can be recognized if the off-diagonal elements of the adjacency matrix of the multigraph are all equal.

Example IX.2.1 In illustrating these graphs it is convenient to represent a graph by its adjacency matrix; this symmetric square array has as many rows as there are vertices, with edges between vertices indicated by unity, and no edge by zero.

(1) The simple graph, the representative graph, and its dual are derived from a subset system **A**.

	Subsets			
	1	2	3	4
N				
a	1	.	.	1
b	1	1	.	.
c	.	.	1	1
d	.	.	.	1

Subset system **A**

Adjacency matrices
(lower triangle)

N		subsets of **A**		edges of R	
a	.	1	.	I	.
b	1 .	2	1 .	II	1 .
c	1 . .	3	. .	III	. 1 .
d	1 . 1 .	4	1 . 1 .		

	Simple graph A	Representative graph R	Graph dual to R
The graphs:	b–a–d |/ c	2–1–4–3	I–II–III

(2) The adjacency matrix of the multigraph corresponding with the representative graph, i.e., with the unit elements replaced by the number of edges, is as follows:

subsets

1	2			
2	1	1		
3	.	.	1	
4	1	.	1	3

Many properties of simple graphs of interest in the study of set systems are discussed by Buckley and Harary (1990). However, the properties of simple random graphs may prove to be of most use in the clustering context. Unlike the random graphs usually studied, both the edges and vertices of the representative graph are random. This property can be seen as follows. A basis set N of n elements gives rise to the power set $P(N)$; choose subsets at random from $P(N) \setminus \{\}$, where $\{\}$ denotes the empty set, subject to the constraint that the union of the subsets is N. The number of different ways m subsets can form a disordered (i.e., no

specified sequence) covering of an unlabeled n-element set is quite large (Clarke 1990) but is ignored here. Let the number of subsets in the choice be m. Thus the number of subsets, their composition, and hence the edge set are all random. Assume not only that the vertices of the graph are mutually independent, but also that other than the end vertices of an edge, edges are independent of vertices. One way to assign values for the probabilities of occurrence of the vertices and edges in the covering is as follows:

DEFINITION IX.2.8. *The probability of vertex j is*

$$p_j = |a_j| / \sum_j |a_j|.$$

Note that $\sum p_j = 1$, and that $p_j = 0$ iff $|a_j| = 0$. If the entropy based on $\{p_j\}$, $-\sum_j p_j \log_e p_j$, is equal to $\log_e m$, the covering consists of subsets of the same size.

DEFINITION IX.2.9. *The conditional probability of edge $\{i, j\}$ joining vertices i and j is*

$$(p_{ij} | a_i \cup a_j) = |a_i \cap a_j| / |a_i \cup a_j|.$$

Definition IX.2.9 can be recognized as the Jaccard coefficient of similarity (Chapter III) between the vertices. From these two definitions, the probability of edge $\{i, j\}$, which may also be regarded as being conditional on the underlying set, N, can be obtained as

$$p_{ij} = p_i \, p_j (p_{ij} | a_i \cup a_j).$$

If the entropy based on $\{p_{ij}\}$ is zero, the cover is also a partition, and thus the nearer this entropy is to zero, the solution can be regarded as being "simpler."

A possible application of this model is in determining whether to regard as random the representative graph corresponding with the set of objects under study, because to do so may suggest that clustering is perhaps premature for these objects. Another possibility for examining consistency with randomness, based on results of Juhász (1981), is as follows. For *fixed m*, let **X** be the adjacency matrix of a simple graph, and let

$$\Pr(x_{ij} = 1) = p,\ 0 < p < 1;\ i \neq j.$$

Then for the first two eigenvalues of **X**,

(1) $\lim_{n\to\infty}(\lambda_1/n) = p$ with probability 1.

(2) $\lim_{n\to\infty} \Pr(\lambda_2 > n^{1/2+\epsilon}) = 0$, for each $\epsilon > 0$, i.e., $\lambda_2 = o(n^{1/2+\epsilon})$.

Thus p can be estimated from the first eigenvalue of **X**, and if the second is appreciably larger than $n^{1/2}$, the graph is unlikely to be random with common value of p. For *any* graph with n vertices and m edges, the adjacency matrix has n eigenvalues, and

(a) $\sum \lambda_i = 0$;

(b) $\sum \lambda_i^2 = 2m$; and

(c) $\lambda_1 \leq (2m(1 - n^{-1}))^{1/2}$.

For fixed t, let $G_n(t)$ be a graph on n vertices constructed from a complete graph by deleting edges randomly and independently with probability

$$p = 1 - e^{-t}.$$

If $P_n(t)$ denotes the probability that $G_n(t)$ is connected (i.e., there is just one muster), and if for some x

$$t = (\log_e n + x + o(1))/n$$

then

$$\lim_{n\to\infty} P_n(t) = \exp(-e^{-x})$$

(Stepanov 1970*a*). Stepanov (1970*b*) also obtained an expression for the probability that $G_n(t)$ has exactly k components (i.e., there are exactly k musters). At first examination, using these results to decide if the optimal solutions to the clustering problem differ appreciably from randomness seems promising but has yet to be exploited.

Further properties of random graphs, which may add insight into the clustering problem, were described by Bollobas (1985) and Palmer (1985).

3 The kernel of a subset

The rules of biological nomenclature require that an individual be designated as the "type" (in zoology, it is the "name bearer"); no matter the details of the written description, identification is ultimately by reference to this individual. Suppose a group of individuals are newly recognized as being distinct from previously named groups and a type is required, some mechanism to select one or more candidates is desirable. This section proposes that the **kernel** (Appendix 1) be obtained in a way that is consistent with the concepts of being central. For vector dissimilarity, this selection has been discussed in Chapter V, where those objects having the maximum frequency of betweenness are defined as forming this subset. Here, the procedure is extended first to graph-theoretic concepts, and second to measures of dissimilarity that can be embedded into a continuous space.

Graph-theoretic kernels

Based on graph theory, a number of possibilities exist for defining the vertices that form the kernel.

DEFINITION IX.3.1a. *The eccentricity of vertex v is the distance to a vertex furthest from v.*

DEFINITION IX.3.1b. *The radius of a graph is the minimum eccentricity, and the diameter is its maximum.*

DEFINITION IX.3.2. *The central-kernel consists of all vertices for which the eccentricity is equal to the radius.*

DEFINITION IX.3.3a. *The status of vertex v is the sum of the distances it has to all other vertices.*

DEFINITION IX.3.3b. *The median-kernel consists of all vertices for which the status is a minimum.*

Because the central- and median-kernels often coincide in part, it seems reasonable to choose a type (= typical specimen) from their intersection, assuming it is not empty. For further discussion of graph-theoretic centres, see Buckley and Harary (1990). An algorithm for measuring the centrality properties of trees is described by Rosenthal and Pino (1989).

Continuous spaces

For dissimilarity spaces, it is convenient to set the scene by reference to a single continuous variable. If $F(x)$ is the cumulative distribution function (CDF) of a univariate random variable X (F is assumed to be continuous), then any point x maximizing

$$m(x) = F(x)[1 - F(x)]$$

is a population median. In a very natural sense, the medians can be regarded as forming the kernel of x and are central in that sense. Suppose there is a random set of n observations, $\mathbf{x}$, from X, and it is desired to estimate the median in the following way:

> for $\forall\ i, j, k,\ i \neq j \neq k$, determine if x_k is between x_i and x_j; those k for which the frequency of betweenness is a maximum are estimates of the median (Liu 1990).

This concept forms the motivation for recognizing one or more objects, here represented by their principal coordinates, which can be regarded as forming the kernel.

Liu (1990) first considered the bivariate case and proposed determining the frequency that object i' is contained within the triangles defined by $\{i, j, k\}$, $\forall\{i, j, k\} \subseteq N$, $i' \notin \{i, j, k\}$, which he called the "simplicial depth" of i' in the set N. He then defined all i' for which the frequency is a maximum as the "bivariate simplicial medians." He then extended this notion to p dimensions (Chapter VIII shows how) by considering the frequency that each object is contained within the complete set of simplices on $p + 1$ vertices and so defines the "multivariate simplicial medians."

The computational determination of the multivariate medians depends strongly on the dimensionality, d. Since the number of simplices is $\binom{n}{p+1}$, if n is large and d is approximately $n/2$, determining the frequencies represents a major computational effort. Advantage may be taken to reduce the rank of a dissimilarity matrix (Chapter VII), but if that still fails to result in a manageable number of simplices, an approximation is to make a sufficiently large random choice of $d + 1$ objects, and to determine if each of the remaining $n - d - 1$ objects is within the simplex. There seems little reason to use multivariate medians rather than either the median- or central-kernel.

4 Species associations

If the number of individuals of a species in each of n randomly chosen quadrats of unit area is x_i, $i = 1...n$, the mean density is estimated as

$$\sum x_i/n$$

with variance

$$(\sum x_i^2 - (\sum x_i)^2/n)/(n - 1).$$

Several different and identifiable sources of variation make up the variance.

(1) A species has properties that may be reflected in many ways, e.g., how proximal two or more individuals may be.

This source of variation is not necessarily constant, and may depend on random events (who lands where), reproductive patterns, and so on, which will affect the means and variances or other statistics estimated from a set of units. This source of variation, which for convenience can be called **ecological variance**, is unit-size free in the sense that it depends (largely) on the species being sampled.

(2) The choice of quadrat size.

Generally, the larger the quadrat size, the smaller is the (relative) variance. Because we are dealing with counts, even if the ecological variance were to generate occupied points having a Poisson distribution, a sample of small plots (small in relation to the consequences of the ecological behavior pattern) generally has a variance in excess of the mean, while large plots tend to have a

variance appreciably less. For convenience, this source of variation can be called the **quadrat variance**. The effect of this source of variation is often masked, or lost, when the results of the survey are standardized into the number of individuals per unit area. Thus if the area of a quadrat is v, the estimated mean per unit area is

$$\sum x_i/vn$$

and the variance is

$$(\sum x_i^2/v^2 - (\sum x_i)^2/v^2n)/(n - 1),$$

which differ only by a scale factor from the original definitions iff the variance is independent of the area of the quadrat. Notice that the species "knows" nothing of the size of a sample unit, only the size of its local universe, and behaves accordingly.

(3) Sampling variation.

Suppose a second, third, ... independent set of sample units is taken; each set allows a mean to be estimated, and although the expectation of the mean and variance may be the same from set to set, there is no reason to expect the calculated means and variances to be identical. The variance shown among the sets measures the sampling variability and leads to an estimate of the **sampling variance.**

If the clustering problems are in circumstances in which one species is being studied, the complexity of the problem is increased if there is more than one source of variation. The difficulty can be seen by considering the following practice.

> A common procedure used in plant ecology consists of choosing a quadrat size (and shape—usually

square) and placing it either systematically on a grid or randomly in the area under study. A census is then made of all individuals of the species of interest, and the results assembled in a two-way table of species-by-quadrats; often a further dimension gives the locations of the quadrats with respect to each other.

The problems arising from these procedures are reasonably well known; for example, if a species is clumped, its apparent probability distribution is a function of the area of the quadrat (the quadrat variance) and the behavior of the species (the ecological variance), but, because the areas are chosen by the ecologist, this combination creates a difficulty that needs to be resolved before the main objective of the study, namely, determining species associations, can be investigated. To make this clearer, suppose there are two species and n quadrats, and that n_{ij} is the number of quadrats in which species i and j both occur; then the contingency table (Table IX.4.1) appears to capture the essentials of what is required to determine if species are associated. In particular, a test for marginal independence, such as the likelihood ratio, or Pearson's χ^2, might be taken to indicate that there is a species association if significantly large. But this logic is dangerous because these tests fail to distinguish lack of independence from departure from a Poisson distribution for the species. A departure from a Poisson distribution may be a function of the size of the quadrat, the behavior of each species in the presence of the other, as well as of any species present but not included in the study. There is no escape from this confounding unless the correct ecological distributions are known and can be incorporated in an appropriately chosen test statistic. In summary, the results of the study of spatial data of this kind are not independent of the scale and aggregation effects implicit in the choice of quadrat size and shape.

Table IX.4.1 Contingency table for pairwise species association in ecological studies

n_{11}	n_{12}	$n_{1.}$
n_{21}	n_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

These problems have been avoided by replacing the quadrat sampling procedure by mapping each plant's position, and then determining the distributional properties of the species from the physical distances to the nearest neighbors of the same and other species. Ripley (1981), Panayirci and Dubes (1987), and others have discussed these procedures, but they are outside the scope of the present study. I now intend to show how information on more than that of the nearest neighbor distances may be used to determine (perhaps disjoint) regions in which a species tends to occur, followed by an extension to two or more species.

The proposed strategy follows from the procedures described in Chapter VIII but requires some measure of biogeographical distance. Individual plants are not points (consider a tree) and two species may differ considerably in size (consider a tree and a moss); furthermore, what is the distance between a tree and a moss growing under its canopy? More generally, the objects of study may be a species, encompassing the whole of its geographical range; for convenience, the term **entity** is used here to stand for the objects under study.

Because entities may live in regions of different sizes and shapes, a general measure of biogeographical distance has to include as special cases those for which there are generally accepted definitions. Let I and J be regions occupied by species i and j (these regions are not assumed to be convex, can consist of several disconnected regions, and may also be points); locate that point in I which is closest to the closest point in J, and denote it by the *ordered* pair (I,J); then locate the point in J furthest from

(I, J) and measure the (great circle) geographical distance between this point and (I,J), calling it $d(I,J)$. If there is more than one nearest point in I to J, such as in overlapping regions, then $d(I,J)$ is defined as the minimum among all possible. The biogeographical distance between I and J is now defined as

$$\beta_{IJ} = \max\{d(I,J), d(J,I)\}$$

which is a Hausdorff measure and therefore is a (not necessarily Euclidean) metric.

The biogeographical distance, β_{IJ}, can be interpreted as a measure of the potential "contact" between the members of one entity with those of another, where contact could result in gene flow given the appropriate degree of consanguinity. Five special cases illustrate the generality of the definition.

(1) If each of two entities occur in one point locality each, β_{IJ} is proportional to the length of the shortest path between the two points.

(2) If the regions occupied by two entities are disjoint but are of the same size, shape, and orientation, β_{IJ} is proportional to the distance between their centres (medians, or any two corresponding points).

Cases 1 and 2 must be rare for extant species. For nondisjoint regions, a biogeographical distance ought to depend partly on the sizes of the regions the taxa occupy.

(3) If the region occupied by one entity is completely contained within that of the other, then β_{IJ} depends partly on the (square root of the) difference in their areas.

(4) If the region occupied by one entity is not completely contained within that of the other and is of a different

shape, then β_U is proportional to a function of their closest proximity and (the square root of) the difference in their areas.

Case 4 includes both disjoint and nondisjoint regions, as well as different shapes.

(5) If the region occupied by one or both entities consists of several apparently unconnected regions, then β_U is interpretable largely in terms of potential contact among all members of both entities.

However, most unconnected distributions tend to reflect the absence of adequate collection, or inhospitable intermediate regions without necessarily implying absolute biological barriers, and so case 5 is probably of little importance.

Given this, or any other measure of biogeographical distance, the subset-generating procedure of Chapter VIII can be used to obtain subsets for each species alone; it can also be used to generate subsets *ignoring* the species distinction. This set of subsets may offer a solution to the problem. Consider the following:

> if there is no association between the species, then in any subset, the number of individuals of species i expressed as a ratio to the total numbers in the subset should be the same as in any other; furthermore, if the subsets are disjoint, this ratio is given by the proportion of species i in the total sample.

If p_{ik} is the observed proportion of species i in the k^{th} subset, n_k is the number of objects in the k^{th} subset, and π_i the expected

proportion (estimated by the ratio of the numbers of individuals of the species to all individuals of all species), then the statistic

$$G^2 = 2\Sigma n_k\, p_{ik} \log_e (p_{ik}/\pi_i),$$

which has an asymptotic χ^2 distribution with $k - 1$ df, is appropriate for the binomial case (i.e., if there are only two species under consideration), is easily generalizable to the multinomial case, and is a test of the null hypothesis of no association.

Another problem in ecology is to determine in a set of localities grouped together by one (or more) species, if another one (or more) tend to occur with it. Without loss of generality, consider two species, I and J, and suppose $\{I_p\}$ represents the p^{th} subset of I in the optimal covering, i.e., it includes all individuals of the species whose average distance to the members is less than the maximum among them. There is also a (possibly empty) subset of the individuals of J included in the neighborhood determined by I_p. This set of possibilities can be described by the subset

$$\{J|I_p\} = \{\, j|\text{ave}(d_{ij}) \leq \max(d_{ik});\ i,k \in I_p;\ j \in J \,\}.$$

Quite separately, the optimal covering of J includes subsets that can be denoted by $\{J_q\}$. A comparison between the $\{J|I_p\}$, $p = 1 \ldots P$, and the $\{J_q\}$, $q = 1 \ldots Q$, where P is the number of subsets in the optimal covering of I, etc., provides a table for investigating the conditional association of J with I.

For $P = 3$ and $Q = 4$, the subsets can be assembled in a two-way table (Table IX.4.2) where the entries are the number of individuals cross-classified. A similar table can be constructed reversing the roles (Table IX.4.3). There is no reason to expect that the second table should be the transpose of the first.

Table IX.4.2 Cross-classification for community association

$\{I \mid J_1\}$ $\{I \mid J_2\}$ $\{I \mid J_3\}$	$\{J_1\}$	$\{J_2\}$	$\{J_3\}$	$\{J_4\}$

Table IX.4.3 As Table IX.4.2, but reversing the roles of the classification set

$\{I \mid J_1\}$ $\{I \mid J_2\}$ $\{I \mid J_3\}$ $\{I \mid J_4\}$	$\{I_1\}$	$\{I_2\}$	$\{I_3\}$

Consider the following four possibilities for such tables.

(1) All entries are zero. This result can occur, for example, if the species occupy disjoint regions, i.e., the species do not coexist in the same regions.

(2) Only one nonzero value occurs in each column, which suggests that an association exists among the subsets of the two species. This result is particularly useful if it is noted that I and J need not represent just one species each but a community. If the numbers in the table approach the numbers of individuals under study (there could be just unity in a column, yet the number of individuals which could have been assigned may be quite large), there is very strong evidence of association.

(3) Suppose nonzero entries occur everywhere in the table; such a table suggests not only that the species can coexist, but also if the values are large, that there may be no special association.

(4) Suppose one table is nonzero (as in (2) and (3)) but the dual table is zero (as in (1)). This result shows that one species can live in regions from which the other is absent.

Although there are further possibilities, two general principles emerge; suppose that there are n_i individuals of species I, n_j of species J; note that

$$\Sigma|\{I_p\}| \geq n_i \text{ and } \Sigma|\{J_q\}| \geq n_j.$$

(1) A comparison between $\Sigma|\{I|J_q\}|$ and n_i indicates something of how species I tends to occur with species J; similarly comparing $\Sigma|\{J|I_p\}|$ and n_j indicates how J tends to occur with species I.

(2) A test of marginal independence of each table, if not significant, indicates that the species can coexist, but that there is no evidence of any particular association. If significant, there is evidence of an association, i.e., patches in which both species tend to occur together.

Chesser and Van Den Bussche (1988) described a related procedure; another, which is a differently motivated solution to the multispecies ecological problem, can be derived from the simplification of Boolean-valued data located on a plane (Wang et al. 1977).

5 Bootstrapping and clustering

Other than the discussion of suboptimal solutions, very little in the previous chapters enables some measure of confidence to be placed on any solution. One reason is that not only are the distributions and densities of the attributes unknown, but also that the set of objects is potentially a mixture of samples from several populations in unknown proportions.

However, almost any numerical estimate from data for which no distributional properties are known can have the distribution estimated by means of pseudosamples. These samples are of the same size as the original data selected randomly with replacement from the data (Efron 1979, Hinkley 1988). Sampling with replacement results in some units being represented more than once, and others not at all. Using the pseudosample, the computation is repeated, and another estimate is obtained. The pseudosampling is repeated many times, and the distribution of the "pseudoestimates" is used as an estimate of the distribution of the statistic for the source data. The discussion considers how bootstrapping, the name used to describe this process, may be of value in clustering.

In the present context, a decision has to be made; should the pseudosample be of the objects, the attributes, or both? In traditional taxonomy, I propose that it be the attributes, because these are subjectively defined; furthermore, the reductions of Chapter II removes duplicate individuals and unless all individuals belonging to a true group are not chosen, there should be no major impact on the resulting covering. In numerical ecology, the situation is reversed; because the sites are chosen subjectively, they may be pseudosampled rather than the species present.

Another decision to be made is whether to keep multiple copies of the same attribute in the sample; as I argued (1991*b*), these provide no more information than a single representation, because the multiplicity is an artifact of the resampling procedure;

furthermore, a more efficient estimate of the distribution (in terms of variance) is obtained from the distinct units in the pseudosample than from estimates based on the complete pseudosample.

It is also necessary to determine some criteria to compare solutions; Moreau and Jain (1987) suggested that a solution is stable if the cluster membership remains the same with moderate variations in the data, and that any other cluster solution is not stable. Moreau and Jain developed their ideas in terms of partitions of Euclidean data, which is inappropriate for the present circumstances.

6 Subset generation with more than one criterion

I have alluded to using multiple criteria for subset formation in previous chapters; indeed, the use of betweenness in Chapter V is based on sets of one-state attributes, while the simultaneous use of several univariate criteria is discussed in Chapter VI. These earlier proposals have partly anticipated the more general discussion presented here. The arguments developed here assume that one or more dissimilarity arrays are to be used, possibly with other criteria, to generate the subsets to be included in **A**. If all dissimilarity arrays are unidimensional, the procedures will collapse to those discussed in earlier chapters.

The discussion is divided into two parts: initially using one dissimilarity array combined with externally imposed criteria, followed by the consideration of two (or more) arrays.

One dissimilarity array

The procedures in Chapter VIII can be used to generate **A**, but examining the optimal solution may show that the subsets overlap to a large extent, and that many of the subsets contain virtually all the objects. In these circumstances, there may be no real groups supported by the data, giving good grounds for terminating the study until more data have been collected. Nevertheless, there may

be reason for continuing further, such as by introducing some heuristic conditions on the subsets. Two such heuristics follow.

The first heuristic is based on edge weights; it is

HEURISTIC IX.6.1. *Reject from inclusion in A those subsets occupying a large proportion of the dissimilarity space.*

This heuristic is somewhat appealing, because it is largely independent of n. It can be considered to be a relaxation of Assumption VIII.3 that

> if a subset contains the pair of objects least alike, it should contain all objects and so is of no interest,

to become that

> if the neighborhood of a subset occupies too much of the total dissimilarity space, it probably is heterogeneous and so is of little interest.

Confining the initial pairs to be those adjacent on the relative neighborhood graph (Chapter VIII) implicitly makes this assumption, because pairs capable of generating subsets occupying large regions of the space are almost certainly not adjacent. This heuristic can be implemented in two ways: either

> exclude from the initial pairs those in the relative neighborhood graph corresponding with the longest edges in the dissimilarity array,

or

exclude the subset if the longest edge in it exceeds a predetermined threshold,

or both. The use of a threshold introduces subjectivity into the procedure.

One consequence of this and similar heuristics is that one or more objects may not be included in any of the generated family of subsets (which can occur without these heuristics). The conclusion is that such objects are isolated with respect to those remaining, and that each should be considered as forming a one-element group. Such objects can be put on one side, and calculation can continue with the multi-object subsets.

The second heuristic, also subjective, is based on the cardinality of the generated subsets and is

HEURISTIC IX.6.2. *Exclude subsets containing more than a specified number of objects.*

If the objects have been independently collected globally, there may be some merit in this proposal; but, for example, if two very distinct groups are represented, the one by two individuals and the other by (say) 100, a cutoff of less than 100 may not find the larger as a single group. Eliminating the pair may expose some distinctions among the 100 in a further clustering. Even though this heuristic is somewhat questionable, it often produces the same musters, if not the same subsets, as those using the first heuristic.

A different procedure using two criteria based on a single dissimilarity array was proposed by O'Callaghan (1976). Let d_{ij} be the distance between objects i and j, $d_{i(k)}$ the distance between i and its k^{th} nearest neighbor, and Θ_{ijk} the angle at j between the lines joining i to j, and j to k. Then object j is in the neighborhood of object i if two conditions are satisfied, namely,

$$d_{ij} < cd_{i(1)},$$

and $$\Theta_{ijk} < \phi,$$

where k, c, and ϕ are specified constants. The fact that constants need to be specified is a weakness of the original proposal. However, the first condition can be replaced by the comparison between average distances and maxima, as in Chapter VIII, and it is not difficult to define ϕ to be some function of the subset to which i belongs.

A third heuristic is based on the principal weighted spanning tree decomposition (Chapter VIII). Using the example in Table VIII.2, the MST, $\mathbf{E}_1$, is equivalent to the single-linkage clustering procedure (Gower and Ross 1969), but $\mathbf{E}_2$, even though orthogonal to the first, is also based on small distances, some of which may be less than those in $\mathbf{E}_1$. Together, both trees may give a better picture of the groupings than the MST alone. One way to combine them is to construct a two-way table from the dendrograms formed from the two trees "cut" at some level (Fig. IX.6.1).

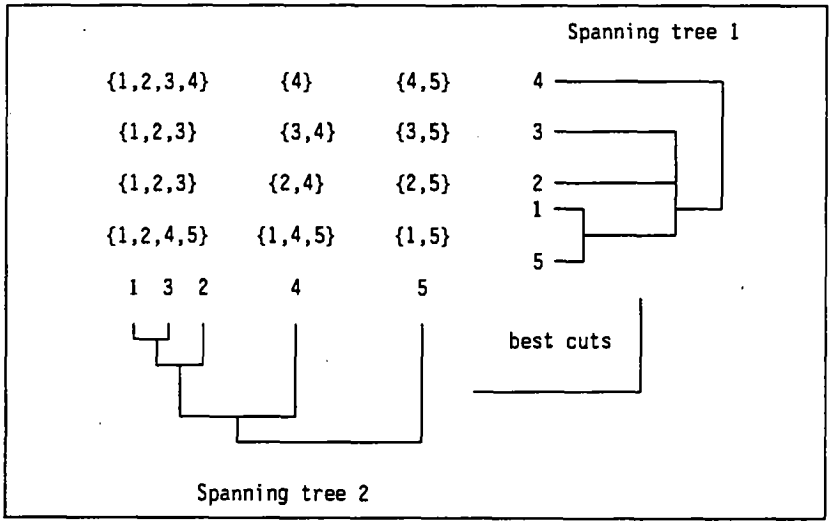


Fig. IX.6.1 The spanning trees corresponding to Table VIII.2.

The subsets formed by the intersection of the cuts can be arranged as the matrix **A**, which for the above example is given in Table IX.6.1. None of the reductions described in Chapter II are possible; there are numerous coverings, including a partition (the second and eighth columns). Computing the probabilities and determining the optimal covering as described in Chapter II is left to the reader.

Table IX.6.1 The matrix **A** corresponding with the cuts in Fig. IX.6.1

Objects	Subsets
1	1 1 1 . . . 1 . . . 1
2	1 1 1 . . 1 . . . 1 .
3	1 1 . . 1 . . . 1 . .
4	1 . 1 1 1 1 1 1 . . .
5	. . 1 . . . 1 1 1 1 1

More than one dissimilarity array

The context here is that there are two or more sets of pairwise relationships; one is used as the primary generator of subsets, the other to take some further action. For example, suppose the first array consists of morphological dissimilarities and a second array consists of the geographical adjacencies of the objects (e.g., either the Voronoi graph of the collection sites, or the Gabriel graph, or the relative neighborhood graph, or the minimum spanning tree); each of the subsets generated by morphology can be examined to determine if it forms a connected subgraph geographically. If it does, the subset is retained unchanged, but if it does not, there are two possible actions, either to modify the membership so as to satisfy the added constraints, or to reject the subset entirely. If the objects to be clustered have known geographical positions, it seems reasonable to require that geographically noncontiguous units

should not be clustered, i.e., a cluster should be compact or connected, etc.; this requirement leads to the following:

> Let **D** denote the array of dissimilarities, **E** the (geographical) adjacency matrix ($e_{ij} = 1$ iff objects i and j are adjacent, zero otherwise). Let the generating pair of objects for subset k be i and j, and let the cardinality of the set be $|k|$. Determine if the subgraph of **E** corresponding to the subset generated from **D** is connected, and if not, reject it entirely if i and j belong in different components, otherwise retain just the component containing i and j.

Unless the length of the path between two vertices is of importance, this approach is extremely efficient in computational time. The extended numerical example using the aphid genus *Pemphigus* given by Lefkovitch (1980) illustrates this.

A second possibility is to require each subset generated from **D** to be a clique (maximally connected subgraph) in **E**; given that the generating pair are adjacent in **E**, other vertices representing objects can be deleted until each retained member is adjacent to all the others. The cardinality of the largest subset estimates the clique number of the graph, and the use of the set covering and partition algorithms with subsets restricted to be cliques yields a minimum number of cliques that form a covering or a partition.

Multiple unidimensional data

Circumstances in which the dissimilarity data are unidimensional have already been discussed in Chapter VI; if some ordering of the objects is known with respect to another attribute (e.g., time or depth), several proposals for clustering have been published

(Lefkovitch 1978, gave a brief review; Gordon 1973, Hawkins and Merriam 1973, and Legendre 1987, made some interesting proposals). Here, the concepts of the relative neighborhood graph generalized by Ichino and Sklansky (1985) are used to generalize the procedures in Chapter VIII for univariate data.

Let $\mathbf{x}_i$ and $\mathbf{x}_j$ be m-element vectors corresponding with the measurements made on m attributes.

DEFINITION IX.6.1. *The rectangle of influence of $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined by the region of space $\mathbf{x}$ for which*

$$min(x_{ik},x_{jk}) \leq x_k \leq max(x_{ik},x_{jk}), \quad k = 1...m.$$

DEFINITION IX.6.2. *The rectangular influence graph, RIG, has n vertices corresponding with the objects, with an edge between vertices i and j iff there is no object within the rectangle of influence of $\mathbf{x}_i$ and $\mathbf{x}_j$.*

As Ichino and Sklansky (1985) pointed out, no special distance measure is required by Definition IX.6.1. They also showed if all attributes are quantitative, that the RIG is a supergraph of the Gabriel graph, but that there is no ordering in graph-theoretic terms with the Delaunay triangulation; it is not invariant under the rotation of the coordinate space but is invariant under any (nonlinear) scale changes of the attributes.

It is not difficult to see that the RIG can be extended to include nominal variables, e.g., the Boolean variables considered in Chapter V, and also that the blurring of the boundaries of the rectangle of influence, as described in Chapter VI, leads to a subset-generating procedure that can be summarized as follows:

step 1: construct the RIG;
step 2: for each edge in turn in the RIG, initiate a subset;

step 3: join all objects with the subset which are not outside the "blurred" rectangle of influence;
step 4: determine the enlarged rectangle of influence;
step 5: repeat steps 3 and 4 until no further objects are included;
step 6: store the subset, and go to step 2.

This procedure has not been investigated in the present study, primarily because of the subjective nature of the demarcation of the variables used to describe organisms for taxonomic purposes.

7 Applying multicriteria clustering to G by E interaction

Because conditional clustering is a general procedure, there is some value in showing how it may be applied in a specific set of circumstances.

Two aspects of the data structure in the context of genotype (G) by environment (E) experimentation (also known as the variety by location interaction problem) are of special importance, namely, the "level" aspect, represented by the marginal means, and the "shape" aspect, represented by the differential responses of individuals to one factor at different levels of the other (Lin 1982). He and others (see references cited by Lin) argued that the genotypes (or environments) should be grouped so that an analysis of variance (ANOVA) does not suggest a significant GE interaction within groups. Rarely, however, are the reasons for the presence of such an interaction considered; they can include different ranges of values that different genotypes may show among the environments, and, separately, the pattern of highs and lows, which may be shown even if the ranges are the same. This distinction can be illustrated by Table IX.7.1, in which the (hypothetical) mean yields of a replicated trial are given for four varieties grown in five environments. Considering all five, if only varieties I and IV had been grown, an ANOVA would not suggest

Table IX.7.1 Hypothetical mean yields of four varieties grown in five environments.

Varieties	Environments					Mean	Among-environments variance
	a	b	c	d	e		
I	1	3	5	7	9	5	10
II	3	4	5	6	7	5	2½
III	9	5	1	7	3	5	10
IV	11	13	15	17	19	15	10
Mean	6	6¼	6½	9¼	9½	7½	-
Among-varieties variance	68/3	251/12	107/3	323/12	139/3	-	26.58

a significant interaction (for any residual error), but it would do so for each of the other pairs (for sufficiently small residual error). For {I, II} the interaction can be explained by the different among-environments variances, for {I, III} by the different ordering of the environments with respect to the mean yields, while for {II, III} it is by both reasons. This example is meant to draw attention to the fact that the among-environments variance for each genotype conveys information relevant to their grouping that is separate from the pattern of highs and lows. This distinction is used to show that clustering is possible when there are several independent measures of relationship without the need to combine them. Other pertinent literature on the problems of GE grouping is cited by Lin (1982).

Let x_{ik} be the observed mean response of the i^{th} genotype in the k^{th} environment: because all attributes are identically defined and measured in the same units (e.g., kg ha^{-1}), a Euclidean distance is a reasonable measure of dissimilarity. The squared distance between two genotypes i and j is given by

$$\Sigma_k(x_{ik} - x_{jk})^2,$$

where the summation is over the environments. Using this distance, any clustering method, in which the distances have no (theoretical) upper bound, can be performed, including that described in Chapter VIII. Because the dissimilarities are Euclidean distances, the resemblance among genotypes can be decomposed into three independent components, which allows groupings of the genotypes to be based not only on the complete value, but also on each component separately.

Let m_i and s_i^2 denote the mean response and among-environments variance for the i^{th} genotype (the m_i and s_i^2 need not be related). If the $\{m_i\}$ are neither significantly different, nor fall into a number of subsets within which they are not significantly different, it seems reasonable to group the genotypes belonging to each of these subsets on the s_i^2; for example, the yield of the varieties belonging to a subset having the smallest among-environment variance tends to be independent of the differences in the environments. Assume for the moment that the genotypes have been grouped into subsets for which the means and among-environments variances are homogeneous; if any such group contains more than one genotype and exhibits a significant GE interaction, the explanation can only be in the patterns of ups and downs, i.e., "shape." Unlike the m_i and s_i^2, shape is multivariate and has no natural or partial ordering. Nevertheless, if the pattern is the same for some subset of the genotypes of interest, it is reasonable to assume that the genotypes in this subset are equivalent. Thus the problem becomes that of finding subsets of the genotypes for which the patterns within a subset are much more alike than they are to the patterns shown by the members of other subsets. The first problem, therefore, is to describe the difference in the patterns in some way that is independent of the m_i and s_i^2.

Assume that the m_i in the subset of the genotypes under study are not significantly different, and also that they have a common s_i^2. In these circumstances, without loss of generality, the x_{ik} can be translated so that the genotype mean becomes zero and the among-environments variance normalized to unity; this affine and scale change is achieved if x_{ik} is replaced by v_{ik} defined as

$$v_{ik} = (x_{ik} - m_i) / \|x_{ik} - m_i\|,$$

where $\|x_{ik} - m_i\|$ is the Euclidean norm. If $\mathbf{v}_i$ denotes the vector $\{v_{ik}\}$, the Euclidean distance, d_{ij}, between $\mathbf{v}_i$ and $\mathbf{v}_j$, is given by

$$\begin{aligned} d^2_{ij} &= (\mathbf{v}_i - \mathbf{v}_j)^{\mathrm{T}}(\mathbf{v}_i - \mathbf{v}_j) \\ &= 2(1 - \mathbf{v}^{\mathrm{T}}_i\mathbf{v}_j) \\ &= 2(1 - \cos\Theta_{ij}), \end{aligned}$$

where Θ_{ij} is the angle between the vectors $\mathbf{v}_i$ and $\mathbf{v}_j$. Thus d_{ij}, the linear distance between the ends of unit vectors (or the functionally related Θ_{ij}), is independent not only of the means but also of the among-environments variance. Even if the assumptions that genotypes have equal variances or means (or both) is false, this distance (or angle) focuses just on the patterns and excludes components arising from the means and variances. If the squared distance among genotypes had been computed without normalization to unit variance, the value obtained is

$$s_i^2 + s_j^2 - 2s_i\, s_j \cos\Theta_{ij},$$

which is twice the value obtained by Lin (1982) to measure differences in pattern. This value has the form of the variance of a difference, with expectation twice the error variance when the null model for GE interaction is true.

For an ANOVA to give an appropriate estimate of error mean square, it is assumed that the variance of *all* genotypes *within all* environments is constant and so is independent of the means, x_{ik}. If this assumption is false, the consequence may be an apparently significant interaction (Snee 1982) without there being any biological basis. If the means and error variances are related, a grouping based on the among-environments variance after one using the means may not result in any further subdivisions. If they are unrelated, the genotype means and among-environments variance can be combined into a Fréchet neighborhood distance (Dowson and Landau 1982; see also Appendix 3), δ_{ij} defined as

$$\delta_{ij}^2 = (m_i - m_j)^2 + (s_i - s_j)^2;$$

this distance will be zero iff the genotypes have identical means and among-environments variances. Using the Fréchet distances in a clustering context then forms groups having similar means and among-environments variances.

At this stage, therefore, the data have been replaced either by three sets of values (namely, the means, variances, and pattern distances), or by two sets (namely, the Fréchet and pattern distances). In essence, therefore, the complete grouping procedure has three steps:

(1) Forming of subsets of genotypes in each of which the means are not significantly different.

(2) Dividing each of these subsets satisfying the first criterion, if possible, into further subsets in each of which the among-environments variances are homogeneous.

(3) Dividing each subset (homogeneous for the first two criteria) into subsets in each of which the pattern of highs and lows are the same.

Because at least one set of distances is not unidimensional, a clustering procedure, such as that described in Chapter VIII, is needed if groups are to be formed empirically. Each measure of distance may be processed by *any* clustering method to form groups alike simultaneously for *all* distances. One possibility for achieving this objective is to find some compromise measure of distance (Lefkovitch 1978) by recombining the components, which, however, introduces its own problems. Here I consider a variant of the methods described in Chapter VII and section 6 of this chapter. In general terms, the procedure is that of conditional clustering given in Chapter VIII, where it is shown that the only initial pairs of objects needing to be considered are those adjacent on the relative neighborhood graph (Toussaint 1980) for *each* measure of distance under consideration. Thus given an acceptable initial pair, $S_0 = \{i, j\}$,

> include object k in S_{t+1} if it is sufficiently alike the members of S_t on *all* criteria,

which is a Pareto-type decision process (Zeleny 1982).

There are no difficulties in modifying these procedures not only to form groups of environments, but also to achieve groups homogeneous *simultaneously* for both genotypes and environments. For the moment, consider environments for which it is not possible to specify the mutual proximities; using the environments as the objects of the basic algorithm, the mean yield at the environment, the among-genotypes variances, and the pattern of genotypes at an environment are easily computed. If the mutual proximity of the environments is known and can be represented by a Gabriel or other graph, only those pairs corresponding with environments adjacent on the proximity graph need be used for subset initiation, whereas candidates for admission must also correspond

with vertices that form a connected subgraph with the current members (see this chapter, section 6).

Use of the term "environments" may have provoked the assumption that it refers to differing geographical locations. In fact, this assumption is false, because s_i^2 may equally have been defined as the among-years variance for the i^{th} genotype. Some complications but no new principles arise, however, if "environments" encompasses both spatial and temporal differences. For example, there will be an among-locations variance and pattern, an among-years variance and pattern, and also among locations*years terms. If y_i^2 represents the among-years variance for the i^{th} genotype, then

$$\delta_{ij}^2 = (m_i - m_j)^2 + (s_i - s_j)^2 + (y_i - y_j)^2$$

is also a Fréchet distance; the definition can be extended by including further terms in comparable units. Furthermore, the definition of the pattern vectors can also be extended to include further elements to represent the additional data. A numerical example is given in Chapter X, "GE interaction."

8 Local and global optimization

Published strategies for optimizing general functions that use clustering algorithms have been reviewed both theoretically and numerically by Törn and Žilinskas (1989). They arrived at a general algorithm consisting of six steps, as follows:

step 1: sample points in the region of interest;
step 2: concentrate the sample to obtain groups around local minima;
step 3: recognize these groups by a clustering method;
step 4: if a stopping condition is met, go to step 6;

step 5: transform, resample for the next iteration, go to step 2;
step 6: perform final computations and stop.

This section shows how to use conditional clustering in step 3 not only to find local optima, but also the global solution with a high probability. This demonstration is illustrated by considering each step in turn, using as an example the search for least-cost covering solutions, although the same procedure can be used for continuous problems.

Assume that m, the number of subsets in the fully reduced **A**, is such that 2^m is very large, so that it is not feasible to examine each to see if it points to an irredundant covering, and compute the value of the objective function; if that were possible, a global solution would be guaranteed. Step 1 consists of finding a set of n_1 irredundant solutions, each of which will be an **x** satisfying $\mathbf{Ax} \geq \mathbf{1}$, and evaluating the corresponding cost function (or functions; the multi-objective problem may also be included). Step 2 consists of eliminating $(1 - \alpha)n_1$ of the sample consisting of the coverings having the highest function values. From the remaining αn_1, step 3 forms subsets satisfying two criteria simultaneously, namely,

- they are sufficiently alike with respect to **x**
- they are sufficiently alike with respect to the value(s) of the objective function(s).

Each subset so obtained can be considered to be in the vicinity of a local optimum. After step 4, for each set of **x** satisfying the two criteria, form their union, and from each of them, generate a further sample of irredundant coverings. After eliminating any duplicate **x**, increase the sample to size n_1 by including further

samples from **A**, and go to step 2. The stopping condition of step 4 is the recognition that after several cycles, the same set of local optima are obtained.

This solution procedure may also be combined with simulated annealing by including the possibility of jumping to a new (randomly chosen) region in the space for further search for an optimum.

9 The shape of a set of points

In one sense, the title of this section is an oxymoron; how can a finite set of points have a shape? Various attempts to answer this question have usually been by some approximation to the hull of the points, the hull being determined on the assumption of some metric. The hull being determined, it is embedded into a continuous space and either some continuous regular shape (e.g., an ellipsoid), and so on, is used to approximate it, or the asphericity is estimated (the ratio of the radius of the largest sphere that can be contained within the hull to the smallest that can contain it). The parameters of the continuous shape (curvature, diameter, volume, and so on) or the asphericity are subsequently used to describe the set of points.

Determining the hull of a set of points in multidimensional space presents considerable computational difficulty. Even though there are efficient algorithms for obtaining the *convex* hull of a set of planar points, in higher dimensions the work is still excessive. Coupled with the subsequent need to fit one or more regular shapes to the hull makes this class of solution unappealing. Consequently, I propose a different solution; there will be almost no approximations, and the amount of computation, although still major for large n, is reduced.

In section 3 of this chapter, the diameter of a graph is defined as the maximum eccentricity among the vertices, where the eccentricity of a vertex is the maximum distance from it to any

other (definitions IX.3.1a and 1b). The concept of diameter is now generalized. The diameter, which is based on pairs of vertices, can be extended to the "k-ameter", for $k = 3,4...n$ in the following way.

DEFINITION IX.9.1. *For $k = 2,3...n$, the k-ameter of the graph is the maximum value of*

$$W(k) = max\{\sum d(i,j)/\binom{k}{2}\}$$

where the summation is over all pairs of distances among k vertices from the $\binom{n}{k}$ choices.

It is easy to see that

$$\begin{aligned} \text{diameter} &= W(2) \geq W(3) \geq \ldots W(n) \\ &= \text{ave}(d(i,j)) \geq \text{radius} \\ &= W(n+1) \end{aligned}$$

and in consequence, the set of n values can be standardized so that

$$w(k) = (W(k) - \text{radius})/(\text{diameter} - \text{radius}),\ k = 2,3...n+1,$$

for which the maximum value is unity and the minimum is zero.

Different arrangements of points exhibit different patterns in the sequence. At the present time, the patterns are unknown for any but a few standard arrangements; the main application is in forming these sequences for the different subsets in an optimal solution, and in exhibiting them for comparative purposes.

The major task in computing $W(k)$ is the need to determine its value for large k. For selecting all $\binom{n}{k}$ subsets of objects, the procedure described by Gentleman (1975) seems hard to improve. Having selected a subset, the mean distance among the individuals and also among the members of the complementary subset can be

computed, which reduces the work by half. For large n, two modifications may be adopted: first, for $k >$ ca. 15, compute the value for steps of k greater than 1 (e.g., 5); second, for very large n and k, choose a reasonable number of subsets of size k at random from the n (an efficient procedure for this selection was described by Ernvall and Nevalainen 1982), and set $W(k)$ to the maximum average distance among them. The second modification will almost certainly produce an underestimate, but if this is alternated with the exact value for other k, adjustments can be made.

This proposal is based on the work of Grove and Markvorsen (1992) on metric invariants for Riemannian manifolds.

X Case studies

Each case study in this chapter is treated separately, identified by name and a letter. The tables corresponding with each are numbered independently and include the indicating letter. The sequence of the studies is in about the same order as the concepts developed in the previous chapters, but not rigidly so because each study is meant to be more or less complete.

It will be apparent that the conclusions from most of the studies may not be free of controversy; when the subject matter is examined with external knowledge and experience, it is to be expected that disagreements will arise, but the function of clustering at the level of this book is to suggest hypotheses for further evaluation, not to proffer solutions.

A Butterflies and monocotyledon plants

The data used here (from Clifford 1975) are chosen to illustrate the covering reductions. They consist of the incidence of butterfly genera some of whose species are parasitic on species belonging to genera of monocotyledons. Clifford arranged the array in known taxonomic (familial) groupings but also speculated that simple rearrangements of the array (Table X.A1), i.e., permuting columns and rows, may suggest other insights. The application of the block–diagonal approximation algorithms (Chapter II) is clearly indicated but is left to the reader; an example of this sorting procedure is given in Case Study L. The objective here, perhaps of no more than passing interest, is to group the plants based on the butterflies parasitic on them, and to see how far one may proceed without anything more than the search for a minimum cover.

Table X.A1a indicates that some butterfly species are confined to a single genus of plants on which no other butterflies

Table X.A1 Known records of 34 butterfly (columns) and monocotyledon host plants (rows) in eastern Australia (after Clifford 1975)

(a) Incidence array:

Plant	Butterfly species 00000000011111111112222222222333333 12345678901234567890123456789012345
1	1.1.1....11..........1.111.....
2	1.....................1..........
3	1.....................1.1.....1...
4	1.....................................
5	....111.....
6	1.....
7	1.....................1.....
8	....1.....1.....1.....
9	.....11.1...11.1.....1.....
10	......1.....1.....1.....
11	......1.....1.....
12	.....1.....
13	.....1.....
14	.....1.....1.....
15	...1.....1.....1....
16	.....1.....
17	.....1.....111.....
18	.....1.....
19	.....1.....
20	.....11.....
21	.....1.....
22	.....1.....11...1.....
23	.....1.....
24	.....1.....
25	.....1.....
26	.1.....
27	.....1.....
28	.....1.....
29	.....1.....
30	.....1.....
31	.....1.....
32	111
33	.....1....1..
34	.....1..

Note: 1 = presence; . = absence.

(b) Isolated plant–butterfly associations: columns 2,8,12,15,21,22,29

(c) Mandatory columns in optimal solution:
columns 1,9,18,19,20,27,28,33

(d) Fully reduced array: (original columns 4,5,7,13,23,24,30)

Plant	Butterfly 0001223 4573340
5	.11....
8	.1..1..
10	..11.1.
11	..1.1..
15	1...1.1

are known (Table X.A1b). All that can be said of these associations is that the data give no evidence on the relationships among the plants (or, for that matter, the plants give no evidence on the relationships among the butterflies).

In the process of the reductions, a number of rows were seen to have a single unity (indicated in Table X.A1c); note that this set of columns is not necessarily unique. In the fully reduced array (Table X.A1d), there is just one minimal cover (the fourth and sixth columns), corresponding with original columns 13 and 24. The reader may be interested in determining a minimum cover of the butterflies, i.e., using the transpose of the data array.

B Plant frequencies

The objective here is to group sites based on percent presence data for 18 species of plants in 25 sites (Table X.B1). Estimates of the site probabilities, both from the matrix **B**, i.e., the proportions, and from **A** formed from **B** as a presence–absence array, are given in Table X.B2. Both arrays suggest the same eight species associations (Table X.B3), of which the first six are mandatory. The associations overlap considerably. The analyses based on presence–absence were repeated but using a threshold to eliminate infrequent species. All values in Table X.B1 exceeding 25% were scored as unity; all others were replaced by zero. From the result, only three associations are obtained (Table X.B4).

Because the role of clustering is to provide candidate groupings for further evaluation, this diversity of result shows the need for further ecological investigations to determine which if any association is more than random.

(Table X.B1 is on page 297.)
(Table X.B2 is on page 298.)

Table X.B3 Full data: species associations

Association	Species No.
1	7, 8, 12, 13, 15, 17
2	7, 8, 13, 17, 18
3	1, 2, 7, 8, 9, 10, 17
4	4, 7, 13, 17
5	1, 2, 6, 7, 8, 9, 13, 14, 15, 17
6	2, 7, 8, 9, 13, 16
7	1, 2, 3, 7, 8, 9, 13, 15, 17
8	3, 5, 7, 8, 11, 13, 15, 17

Table X.B1 Plant frequencies (after Dale 1971): percent presence of 18 species in 25 sites

Species		Site Nos.																								
Name	No.	1	2	17	22	24	3	18	4	20	9	13	23	11	12	14	10	7	8	15	16	5	21	6	25	19
Agrostis canina	1										13							7	36							
A. tenuis	2									11	8			23				9	20			38				28
Anthoxanthum odoratum	3									2									8							
Blechnum spicant	4															1										
Calluna vulgaris	5																			2		3				
Carex binervis	6																	64								
Deschampsia flexuosa	7	94	71	95	95	81	99	100	99	100	48	86	66	86	94	98	70	17	74	94	98	84	93	12	24	22
Festuca ovina	8	95	92	59	66	53	47	18	23	54	30	26	51					12	57		26	30	67	24		100
Galium saxatile	9								13	13	11	24			1		17	4	34					6		6
Holcus lanatus	10										2															
Juncus squarrosus	11								2					3							1	1	4			
Luzula campestris	12						1																			
Nardus stricta	13	5	1		8		31	33				27	36	13	3	16		59	80	81	59	53	46	4		2
Potentilla erecta	14																	2								
Pteridium aquilinum	15		2			17	17		7	1							2	2	13			1		14	17	
Rumex acetosa	16																									28
Vaccinium myrtillus	17	100	100	100	100	99	100	100	100	99	79	99	98	97	100	98	100	1	45	10	39	31		27	79	
V. vitis-idaea	18								9																	

Table X.B2 Plant frequencies: site probabilities based on percentages and incidences

Site No.	Percentage (B)	Incidence (A)
1	0.045305	0.013280
2	0.042030	0.024640
17	0.035498	0.006719
22	0.039485	0.013280
24	0.039247	0.018079
3	0.049168	0.049451
18	0.039671	0.038380
4	0.035869	0.049596
20	0.044891	0.064188
9	0.034586	0.073783
13	0.043759	0.024836
23	0.041188	0.013280
11	0.034090	0.041345
12	0.022838	0.020476
14	0.026803	0.034131
10	0.026498	0.025276
7	0.055979	0.113792
8	0.092059	0.089616
15	0.034728	0.032090
16	0.038020	0.033240
5	0.050708	0.080235
21	0.039598	0.030881
6	0.019060	0.036196
25	0.018900	0.013719
19	0.050020	0.059492

Table X.B4 Threshold data (25%): species associations

Association	Species No.
1	2, 8, 16
2	6, 13
3	1, 7, 8, 9, 13, 17

C Arctic grasses

This case study attempts to find a minimal diagnostic set of attributes for 42 species of grasses thought to occur in the Canadian high Arctic, chosen from 429 attributes (McLachlan et al. 1989). Of the 429, 117 attributes could not be used to distinguish any of the species, usually because their states are at present unknown for many taxa or are inapplicable. Thus the problem consists of finding an optimal set covering of 861 objects (i.e., pairs of species) based on a family of 312 subsets.

Six attributes were found to be sufficient to identify all but one taxon (see below). Because two of the six attributes are measurements difficult to use in the field, this class was eliminated. Further, if an attribute for a taxon is known to show more than one state, it usually cannot be used unequivocally for identification and is here ignored for that taxon. The final number of subsets was 301. The reductions reduced the number of constraints from 861 to 803, and the number of subsets from 301 to 137, 44 of which had a probability exceeding 0.01. The 803 by 137 array **A** contained 17 001 unities.

Using costs defined as $c_k = -\log_e p_k$, the covering solution using the modified Chvátal procedure (Chapter II) consisted of ten attributes (Table X.C1); a better solution was *not* found by the simulated annealing algorithm (Appendix 2). The cost of the obtained solution, $\mathbf{c}^T\mathbf{x}$ was 38.470, which compares with a lower bound of 10.771 obtained from the linear relaxation. Were the elements of **A** to be independent (clearly, they are not) and unity with constant probability, π, (there are no reasons for this assumption), Vercellis (1984) showed that for large numbers of subsets and constraints, as well as a number of other conditions, the ratio between the number of subsets in the optimal solution and $z = \log_e n/\log_e(1/(1-\pi))$ tends to unity. Thus for $n = 861$ and assuming $\pi = 0.5$, implies $z = 9.75$, which compares very well with the solution obtained. In the completely reduced array, assuming π to be 17 001/(803*137), the expected number of subsets in a minimal covering is 40, four times more than observed.

Table X.C2 lists the 42 species and the states shown by each selected attribute; note that *Poa alpigena* var. *colpodea* is apparently indistinguishable from the main form. Using the identification key-generating program of Dallwitz (1974) confined to the 10 attributes, the key necessarily used all 10 and had an average length of 4.6 steps (maximum 6) to achieve an identification. By contrast, a key based on 60 of the 429 attributes empirically thought to be most useful for these grasses (McLachlan et al. 1989) used 39 attributes and was only marginally shorter (average length 4.5 steps, maximum 6); it also failed to distinguish the same pair of subspecies.

Table X.C1 Arctic grasses: attributes, and their states, chosen from 429 to distinguish 42 species of grasses known to occur in the Canadian high Arctic

Attributes	States	
5: plants	1	rhizomatous
	2	lacking rhizomes
58: ligule	1	glabrous
	2	erose ciliate
61: ligule, shape at apex	1	acuminate
	2	acute
	3	obtuse
	4	truncate
133: primary branch surface	1	smooth
	2	scabrous
225: first glume, shape	1	linear
	2	oblong
	3	deltoid
	4	lanceolate
	5	ovate
	6	obovate
	7	oblanceolate
	8	transversely oblong
242: first glume apical shape	1	caudate
	2	acuminate
	3	acute
	4	obtuse
	5	truncate
	6	emarginate
272: second glume	1	with lateral keels
	2	with a central keel
	3	not keeled
341: lemma of fertile floret	1	keeled
	2	rounded on the back
374: palea of fertile floret	1	with glabrous keel nerves
	2	with scabrous nerves
	3	with hairy nerves
420: vernation	1	leaf blades rolled in bud
	2	leaf blades folded in bud

Table X.C2 Arctic grasses: species and attribute states

Species	Attributes * *									
	5	58	61	133	225	242	272	341	374	420
Alopecurus alpinus	1	1	4	2	*	3	2	1	*	1
Arctagrostis arundinacea	1	1	3	2	4	3	3	1	2	1
A. latifolia	1	1	3	2	4	3	2	1	2	1
Arctophila fulva	1	1	4	1	4	3	3	2	1	*
Calamagrostis canadensis var. *langsdorffii*	1	2	3	2	4	2	2	2	1	1
C. purpurascens	1	2	*	2	4	2	2	1	2	1
C. stricta ssp. *inexpansa*	1	2	2	2	4	3	2	2	*	1
Deschampsia caespitosa	2	1	2	1	4	3	2	1	2	1
D. caespitosa ssp. *brevifolia*	2	1	*	1	4	2	2	1	2	1
D. caespitosa ssp. *glauca*	2	1	*	1	4	3	2	2	2	1
Dupontia fisheri	1	1	4	1	4	*	2	1	1	2
Elymus alaskanus	2	2	4	*	4	2	3	2	2	1
Festuca baffinensis	2	1	*	2	4	2	3	2	2	2
F. brachyphylla	2	2	4	2	4	2	3	2	2	2
F. brevissima	2	1	4	2	4	2	3	1	2	2
F. hyperborea	2	2	4	1	4	2	3	2	2	2
F. lenensis	2	2	4	2	1	2	2	2	2	2
F. richardsonii	1	2	4	2	4	2	3	2	3	2
Hierochloë alpina	1	2	4	1	5	3	2	1	3	1

(continued)

Table X.C2 *(concluded)*

Species	Attributes * *									
	5	58	61	133	225	242	272	341	374	420
H. odorata	1	2	3	1	5	3	2	1	3	1
H. pauciflora	1	*	3	1	4	3	3	2	3	1
Leymus mollis	1	2	4	*	1	2	3	2	3	1
Phippsia algida	2	1	2	1	5	4	3	2	1	2
Pleuropogon sabinei	1	*	3	*	5	4	3	2	2	2
Poa abbreviata	2	1	2	1	4	2	3	1	3	2
P. alpigena	1	*	3	1	5	3	2	1	2	2
P. alpigena var. *colpodea?*	1	1	3	1	5	3	2	1	2	2
P. alpina	2	*	*	*	5	3	2	1	3	2
P. arctica	1	1	*	*	4	3	*	1	3	2
P. glauca	2	1	4	2	4	3	2	1	3	2
P. × *hartzii*	2	1	1	2	4	3	2	2	3	2
Puccinellia andersonii	2	1	3	1	4	3	3	2	2	1
P. angustata	2	1	2	2	7	3	3	2	2	1
P. bruggemannii	2	1	3	1	5	4	3	2	3	1
P. langeana	2	1	*	1	4	2	2	1	1	1
P. phryganoides	2	1	2	1	2	4	3	2	1	1
P. poacea	2	2	*	1	*	3	3	2	1	1
P. vaginata	2	1	2	*	4	3	3	2	2	1
P. vahliana	2	1	*	1	4	*	2	2	3	2
X Pucciphippsia vacillans	2	1	*	1	5	4	3	2	3	2
Trisetum sibericum	1	1	2	*	4	2	2	1	2	1
T. spicatum	2	2	2	*	4	2	2	1	2	1

** See Table X.C1.
* Indicates unknown, inapplicable, or variable.

D André's data

A constructed example of an incidence matrix of a site by species array (André 1984) is given here in transposed form in Table X.D1; this table also includes the site probabilities, as defined in Chapter II. The two objectives of this case study are, first, to illustrate the grouping of the sites by direct study of this array, and, second, to generate a family of subsets by use of vector dissimilarity (Chapter V).

Table X.D1 André's data: original incidence matrix transposed (A^T)

Sites	Species abcdefghij	Site probabilities
1	1.........	0.0119
2	1.........	0.0119
3	1.........	0.0119
4	11........	0.0198
5	11111.....	0.0382
6	11111....1	0.0502
7	11111.....	0.0382
8	11111....1	0.0502
9	11111.....	0.0382
10	11111....1	0.0582
11	11111..1..	0.0458
12	11111.....	0.0502
13	.1111.11..	0.0428
14	.11111111.	0.0635
15	.11111111.	0.0635
16	.11111111.	0.0635
17	...1111111	0.0592
18	....11111.	0.0402
19	111111	0.0521
20	111111	0.0521
21	111111	0.0521
22	......111.	0.0258
23	......111.	0.0258
24	.......11.	0.0258
25	.......11.	0.0170

For the first objective, the reduced matrix is given in Table X.D2; the optimal covering was obtained by the reductions without need of costs (Table X.D3). The unique solution is given by original sites {1–12}, and {11, 13–25}, with indicator species a and h, respectively. Only site 11 is in common.

Table X.D2 André's data: fully reduced incidence matrix transposed ($\mathbf{A}^T$)

Sites	Species acfj	Probabilities
4	1...	0.0198
6	11.1	0.0502
11	11..	0.0458
12	.1..	0.0428
14	.11.	0.0635
17	..11	0.0592
18	..1.	0.0402

Table X.D3a André's data: best covering solution

Species	Association 1 (site 6)	Association 2 (site 14)
a	1	.
b	1	1
c	1	1
d	1	1
e	1	1
f	.	1
g	.	1
h	.	1
i	.	1
j	1	.

Table X.D3b André's data: characteristic and common species

Species	Characteristic Association 1	2	Common
a	1	.	.
b	.	.	1
c	.	.	1
d	.	.	1
e	.	.	1
f	.	1	.
g	.	1	.
h	.	1	.
i	.	1	.
j	1	.	.

For the second objective, applying vector dissimilarity to a clustering problem, Table X.D4 gives the data for the 10 species after removing duplicates sites having the same species (which play no role in this procedure). Table X.D5 gives the 21 distinct subsets, generated as defined in Chapter V, confined to those initial pairs for which $\mathbf{g(i,j) \neq i + j}$. Table X.D6 gives both the subset probabilities, computed as described in Chapter II, and the measure of information. Table D7 gives the stepwise row reductions, the remaining subsets, and the optimal covering using both joint probability and information. The solution was identical for both objective functions and consisted of the two subsets {a,b,c,d,e,f,j} and {f,g,h,i,j} in which two objects are in common. No partition of the objects was found in these 21 subsets.

The covering solution for the vector dissimilarity can be compared with that obtained directly from the complete incidence matrix (Table X.D3), which consists of the two subsets {a,b,c,d,e,j} and {b,c,d,e,f,g,h,i}. Clearly, common ground but also differences exist in the two solutions; because the role of clustering is to

generate hypotheses, the differences are perhaps more informative than are the resemblances.

Table X.D4 André's data: initial incidence matrix

(10 species, 12 attributes)

Species	Sites
A	11111.....
B	.111111.....
C	..11111.....
D	..111111....
E	..11111111..
F	......1111..
G	.....111111.
H	11111111
I	111111
J	...1...11...

Table X.D5 André's data: 21 subsets generated

Species	Subsets
A	1.1..1........1.....
B	1111.11.1......1.....
C	1111111111.....11....
D	..11111111.....11....
E	.....11111..........
F	.....111111111...1111
G	........1111.1....11.
H	..........1.......1.
I	..........111.....11
J	.....11111....1111111

Table X.D6 André's data: subset probabilities and information

Subset	Probability	Information
1	0.036346	0.120474
2	0.023830	0.089047
3	0.048862	0.147502
4	0.036346	0.120474
5	0.026852	0.097136
6	0.086229	0.211326
7	0.073713	0.192212
8	0.064219	0.176311
9	0.087676	0.213412
10	0.078182	0.199264
11	0.026239	0.095522
12	0.053679	0.156996
13	0.024633	0.091233
14	0.038596	0.125614
15	0.027096	0.097772
16	0.050926	0.151625
17	0.041432	0.131907
18	0.026856	0.097145
19	0.040818	0.130563
20	0.068258	0.183237
21	0.039213	0.127001

Table X.D7 André's data: example of covering solution procedures

(a) Row reductions

H covered by I
H covered by G
E covered by F
B covered by C
E covered by D
E covered by J

(*continued*)

Table X.D7 (*concluded*)

(b) Remaining subsets and objects transposed and resequenced

Objects ABEI	Original subset numbers*
11..	{3,1}
.1..	{16,2,4}
111.	{6}
.11.	{9,7}
..1.	{10,8}
...1	{20,12,13,14,21}
11..	{15}

(c) Optimal covering

Species	Subsets 6	20
A	1	.
B	1	.
C	1	.
D	1	.
E	1	.
F	1	1
G	.	1
H	.	1
I	.	1
J	1	1

(i) Maximum joint probability: objective function = 0.005885

(ii) Maximum information: objective function = 0.394563

* The first in each set has maximum probability.

E ANOVA means

This series of numerical examples is intended to discuss the application of unidimensional clustering (Chapter VI) to the problem of grouping means in an analysis of variance (ANOVA) context.

Tables X.E1 and X.E2 refer to the same five data sets described by Calinski and Corsten (1985) (this publication is here referred to as CC) and also give the results of their two proposed procedures. The first two examples were also considered by McLachlan and Basford (1988).

The first of the CC procedures (method a) is based on the complete-linkage (furthest-neighbor) hierarchical-clustering method; the grouping process is terminated when the smallest range exceeds the upper α-point of the studentized range distribution for the contained means based on the degrees of freedom (DF). The second procedure (method b) is based upon Gabriel's (1964) method for testing homogeneity within a subset. It consists of obtaining the total sum of squares among the objects in the subsets around the latters' means, and terminating the grouping when this value, appropriately scaled by the error variance and the DF, exceeds the upper α-point of the F distribution. Further details are given by Gabriel (1964) and CC.

Some minor but relatively trivial differences occur in the groupings obtained for these four data sets. The fourth, however, is interesting in that the ratio of the smallest to the largest of the five variety mean weights exceeds 14, making it unlikely that the within-variety variance is homogeneous in the untransformed yields. Assuming that the logarithms of the reported means are reasonably homogeneous, it becomes clear that the grounds for separating beyond {A},{B–E} are weak.

According to CC, heterogeneity of variance is also exhibited by data set 5, but, in my opinion, this heterogeneity is not sufficient to understand the major disagreement in connection with

Table X.E1 ANOVA means, labels, generated subsets, probabilities, and the empirical x for the data used by Calinski and Corsten (1985) and compared with their solutions

Y_i	Generated subsets	Probability	x_k	Covering
Data set 1: (Keuls 1952; s^2=124.29, DF=24: g/head cabbages)				
97.7 A	H, I	1/6	0	Musters*: D–I, A–C, J–L, M
100.7 B				
111.3 C	E–I	1/3	1	McLachlan and Basford's solutions
120.7 D				(a) A–C, D–J, K–M
124.3 E	F, G	1/6	0	(b) A–C, D–I, J–L, M
128.7 F				
129.0 G	F–I	1/3	0	
131.0 H				
132.0 I	J–L	1.0	1	
141.7 J	D–E	1.0	1	Calinski and Corsten's solutions
150.7 K	A–C	1.0	1	(a) A–C, D–L, M
152.7 L				(b) A–C, D–I, J–L, M
176.0 M	M	1.0	1	
Data set 2: (Duncan 1955; s^2=79.64, DF=30: bushels/acre barley)				
	A	1.0	1	Musters: generated subsets are a
49.6 A	B–D	1.0	1	partition
58.1 B	E–G	1.0	1	
61.0 C				Calinski and Corsten's solution
61.5 D				(a) and (b) A–D, E–G
67.6 E				
71.2 F				McLachlan and Basford's solutions
71.3 G				(a) A–D, E–G
				(b) A, B–D, E–G
Data set 3: (Snedecor 1946; s^2=90.63, DF=30: bushels/acre potatoes)				
341.9 A	A–D	1.0	1	Musters: generated subsets are a
360.4 B	E–G	1.0	1	partition
360.6 C				
363.1 D				Calinski and Corsten's solution
379.9 E				(a) and (b) A–D, E–G
386.3 F				
387.1 G				
Data set 4: (Calinski and Corsten 1985; s^2=48.86, DF=12: kg/are tomatoes)				
12.5 A	A	1.0	1	Musters A, B, C, D, E
98.2 B	B–C	1.0	1	Calinski and Corsten's solutions
124.8 C	C–D	1.0	1	(a) A, B–D, E
140.4 D	E	1.0	1	(b) A, B, C–D, E
176.3 E				

* Union of overlapping subsets: see text.

Table X.E2 Data set 5: (Larmour 1941; s^2=342.62, DF=64:ml/loaf bread)

mL	label	sd*	Subset members	Probability	x_k	Solutions
(a) Based on original data						
654	A	24.5	O, P	0.054	0	Musters:
729	B	15.6	C, D	0.050	0	{A}, {B,C}, {D–I},
755	C	44.3	D–I	0.176	1	{J–M}, {N–P}, {Q}
801	D	47.2	E, F	0.042	0	
828	E	29.9	E–I	0.126	0	
829	F	82.4	G, H	0.042	0	
846	G	27.0	H, I	0.042	0	Calinski and
853	H	58.4	J, K	0.054	0	Corsten's solution
861	I	50.7	J–L	0.105	1	(a) and (b) {A}, {B,C}
903	J	63.9	L, M	0.102	1	{D–I}, {J–N}, {O–Q}
908	K	67.0	M, N	0.102	0	
922	L	98.6	N–P	0.105	1	
933	M	27.1	Q	1.0	1	
951	N	59.3	B, C	1.0	1	
977	O	81.3	A	1.0	1	
987	P	79.6				
1030	Q	40.3				
(b) Based on t-value distances						
Minimum spanning tree as linked list						
A	-		O, P	0.057	0	Musters: {A}, {B–I},
B	C		C, D, F	0.061	1	{J–M}, {N–P}, {Q}
C	F		D, F	0.061	0	
D	F		E, F	0.054	0	
E	F		E–I	0.162	1	
F	A		G, H	0.054	0	
G	H		H, I	0.054	0	
H	F		J, K	0.057	0	
I	H		J, K, L	0.111	1	
J	K		L, M	0.108	1	
K	L		M, N	0.108	0	
L	I		N, O, P	0.111	1	
M	L		Q	1.0	1	
N	M		B, C	1.0	1	
O	N		A	1.0	1	
P	O					
Q	P					

* sd = standard deviation.

objects M, N, O, P, and Q. It is a little surprising that CC's procedures place Q in a group with O and P, especially as the difference between P and Q is the third highest among the ordered values. However, it is also puzzling that N is grouped with O and P by the conditional clustering procedures described in Chapter VI and not with M, because the difference between M and N is less than half of that between N and O. Perhaps {J–P} ought to have been formed; it would have been if a weak muster (Appendix 6) based on neighborhood intersection (Lefkovitch 1982) had been used rather than the computationally simpler one based solely on nonempty set intersection. The solutions are roughly compared by the ANOVAs in Table X.E3; because one of the different groups contains just one object, it is not surprising that the residual mean squares are appreciably smaller for the conditional clustering solution. Nevertheless, the error mean square, calculated from CC's paper to be 343.6 with 64 DF, does not suggest that the within-group variance is appreciably large for any of the analyses. It does suggest heterogeneity in the 17 means (F = 9711.7/343.6 with 16 and 64 DF).

In the paper giving the original for data set 5 (Larmour 1941), the volumes of loaves for these 17 varieties of wheat were reported for five different levels of an additive (originally, there were 18 varieties, but there was a missing cell, presumably accounting for the subsequent use of 17). Thus the within–variety variance is really an estimate of that due to the additives and may have a genetic foundation; furthermore, these variances show a wide range (Table X.E2a). In addition, the "error" variance is really an estimate of the pooled interaction between the varieties and additives and random variation. As a result, these data are used to illustrate the t-value distance method described in Chapter VI. The dimensionality of the t-distance array (99% of the total distance) was found to be 4 (the first two principal coordinates accounted for 89%), and so the multidimensional procedures of Chapter VIII are needed; the relative neighborhood graph (Chapter VIII) was also

the minimum spanning tree and is given as a linked list (Table X.E2b). From the adjacent objects on this tree, 15 distinct subsets were obtained; Table X.E2b gives the musters formed by these from the optimal covering. Other than not separating {B,C} from {D–I}, these groupings are consistent with those of data set 5 in Table X.E2a, perhaps suggesting that the departures from homogeneity are not too serious.

Table X.E3 ANOVA means: analyses of variance for the groupings of data set 5

Source of variation	DF	Mean square
(a) All 17 means		
Varieties (V)	16	48 540
(a) Musters (M)	5	151 381
V within M	11	1 794
(b) CC's groups	4	186 916
V within C	12	2 415
Additives (A)	4	29 059
Residual	64	1 718
(a) A.M	20	1 426
Residual	44	1 850
(b) A.C	16	1 568
Residual	48	1 767
(b) Confined to the last 8 means, i.e., where the solutions differ		
Varieties (V)	7	9 664
(a) Musters (M)	2	30 706
V within M	5	1 248
(b) CC's groups	1	52 173
V within C	6	2 579
Additives (A)	4	25 996
Residual	28	1 588
(a) A.M	8	755
Residual	20	1 921
(b) A.C	4	655
Residual	24	1 744

F Caste skulls

This case study illustrates the nonmetric transformation of dissimilarities described in Chapter VII, and the conditional clustering methods of Chapter VIII. The data are the generalized squared morphological distances given by C.R. Rao (1971) for skulls of members of 12 Indian castes. Table X.F1a gives the empirical values below the diagonal, and the path distances derived from the relative neighborhood graph (RNG) above. The RNG has 12 edges; there is one cycle of six edges, and three vertices of degree 1. Table X.F1b, which gives the eigenvalues for both sets of distances expressed as a proportion of the largest, shows that the numerical rank of the transformed values is 7, one less than that of the original distances, and that the concentration of the distances in the first two dimensions is 57% in the original data and 73% after transformation. There is even further concentration if the edge distances are used to replace the empirical values. In contrast, there is diffusion brought about by the relative external graph (REG)-path transformation, as evidenced by the increased sphericity. Table X.F2a gives the first two principal coordinates, each normed to unity, for the original data and their three nonmetric transformations. Although the dominant principal coordinates of the RNG-path and edge distances can be said to coincide with the original (Table X.F2b), even though the angle between them is approximately 23°, the agreement is poorer for the subdominants, for which the angle is about 40°. The REG-path distances show considerably less resemblance to the original. Lefkovitch (1989) gave plots of the first two principal coordinates for the original and the RNG-transformed values, together with the edges in the RNG; the biggest changes brought about are in the apparent positions of D and Bh. Table X.F3 gives estimates of the effective dimensionalities for the original and transformed data.

Table X.F1 Caste skulls: matrix of generalized distances, eigenvalues, principal coordinates, and inner products

(a) Distances: below diagonal are original, above diagonal are RNG path-transformed

B_1	–	520	1403	1951	2297	2953	3902	4534	3620	3978	3674	2525
B_2	520	–	883	1431	1777	2433	3382	4014	3100	3458	3154	2005
A_1	1080	883	–	548	894	1548	2499	3131	2217	2575	2271	1122
A_2	1217	1015	548	–	346	1002	2583	1951	1669	2741	2819	1670
A_3	1459	1212	700	346	–	656	2237	1605	1323	2395	3165	2016
A_4	1817	1649	1233	762	656	–	1581	949	1979	3051	3821	2672
Ch	1746	1694	1838	1456	1649	1497	–	632	3560	4632	5402	4253
M	1691	1619	1565	1158	1204	949	632	–	2928	4000	4770	3621
Bh	2110	1954	1591	1493	1323	1497	2241	1778	–	1072	3674	2525
D	1691	1676	1706	1552	1520	1631	1960	1572	1072	–	2602	1453
C_1	1865	1900	1637	1726	1830	2049	2291	2112	2254	2126	–	1149
C_2	1493	1277	1122	1237	1292	1694	2163	1934	1862	1453	1149	–

(b) Eigenvalues as proportion of the largest

Eigenvalue	Original distances	RNG-path distances	RNG-edge distances	REG-path distances
1	1	1	1	1
2	0.7985	0.5552	0.4746	0.9729
3	0.6014	0.2487	0.1424	0.8553
4	0.4740	0.1275	0.1266	0.8222
5	0.1621	0.1026	–	0.8017
6	0.0607	0.0517	–	0.6637
7	0.0337	0.0344	–	0.6468
8	0.0227	–	–	0.5064
9	–	–	–	0.4547

Table X.F2a Caste skulls: first pairs of dominant principal coordinates

	Original distances		RNG-path distances		RNG-edge distances		REG-path distances	
Caste	1	2	1	2	1	2	1	2
B_1	1747	-3414	1457	-5736	3888	-4701	-593	-145
B_2	1430	-2891	1170	-4382	2890	-2838	5678	1847
A_1	1420	-878	681	-2004	1864	-1118	-1368	-972
A_2	-556	-878	-372	-924	435	-272	189	340
A_3	-841	486	-930	-84	-102	434	230	516
A_4	-2901	643	-2317	119	-2456	-192	1635	-3593
Ch	-4387	-3911	-5362	-1193	-5415	-1867	2709	-2980
M	-4237	-1480	-4628	2539	-3921	-959	1663	-3671
Bh	-1336	6436	-35	3374	-674	3262	-3620	-835
D	-603	4277	2671	4672	-356	5951	5774	-1559
C_1	5190	-392	4866	1291	3629	3454	1153	5472
C_2	4144	701	2946	488	2709	2430	1970	5342

Table X.F2b Caste skulls: inner products among principal coordinates

		RNG-path distances		RNG-edge distances		REG-path distances	
		1	2	1	2	1	2
Original distances	1	0.919863	-0.230199	0.920853	0.186857	-0.032001	0.920278
	2	0.286067	0.772989	-0.036788	0.810709	-0.231826	-0.001760

Table X.F3 Caste skulls: effective dimensionalities

α	Original distances	RNG-path distances	RNG-edge distances	REG-path distances
0	8	7	4	9
0.5	4.341	3.334	2.422	7.718
1	3.153	2.120	1.744	6.724
2	2.256	1.401	1.262	5.319

Using the original data, the procedures in Chapter VIII generated seven subsets, which formed five musters, namely, $\{C_1,C_2\}$, {Bh,D}, {Ch,M}, $\{A_1,A_2,A_3,A_4\}$, and $\{B_1,B_2\}$. The unique minimal covering divided the A_i into two, namely, $\{A_1,A_2,A_3\}$ and $\{A_2,A_3,A_4\}$. For the RNG-path transformed data, there were two musters in the 12 subsets which were generated, namely, {Bh,D} with the remainder forming the other. The optimal covering obtained from the 12 subsets was also a partition, namely, $\{C_1,C_2\}$, {Bh,D}, {Ch,M}, $\{A_1,A_2\}$, $\{A_3,A_4\}$, and $\{B_1,B_2\}$. Not surprisingly, the coverings obtained from the two sets of distances are consistent with each other, the only difference being the partition of the A_i into two groups.

The optimal covering obtained from a conditional clustering using the REG-path distance transformation, 15 subsets were generated; the optimal covering consisted of eight subsets, namely, {B,Ch}, $\{B_2,A_2,C_2\}$, $\{B_2,A_3,C_2\}$, $\{Bh,C_2\}$, $\{C_1\}$, $\{B_1,A_1,D\}$, $\{B_2,M\}$, and $\{B_2,A_4\}$, which combine to the three musters, $\{C_1\}$, $\{B_1,A_1,Ch,D\}$, and the remainder. This arrangement is consistent neither with a clustering based on the original data nor with an explanation based on historical information, i.e., the REG-path distance transformation here has resulted in unacceptable groupings, perhaps arising from the tendency to increased sphericity it brings.

It is perhaps remarkable that the principal coordinates based on 12 of the 66 distances agreed closely with those of the complete set. For larger numbers of objects (up to 187 have been investigated), the economy was even greater, because the number of edges in the RNG rarely exceeds $3n$ for random data (Lefkovitch 1984, Appendix) and tends to be less if there are distinct subsets.

G Letters

This case study is an application of the methods of Chapter VIII to psychometry. Published data on the similarities in shape of lower-case letters as perceived in Sweden (Kuennepas and Janson 1969) were used; after excluding w and three vowels represented by letters with diacritical marks, only 25 objects were studied.

Table X.G1 gives the generated distinct subsets in compact form, because the incidence matrix is sparse. This table also gives the estimated subset-covering and object-representation probabilities and the reduction patterns (e.g., "d covered by b"). Table X.G2 gives the best covering and the best approximation to a partition that were found, together with the values of the joint probabilities and entropy. General descriptions of the six musters formed from the optimal covering are given in Table X.G3; it is easy to see that the letters included in each muster tend to be alike rotationally (including reflections).

Table X.G1 Letters: generated subsets and representation and covering probabilities of letter-similarity data

Abbreviated incidence matrix					
Letter	Subsets	Representation probability	Subset	Covering probability	Reduction sequence
a	18	1.0	1	0.048	d covered by b
b	1, 3, 8	0.0	2	0.061	e c
c	2, 6, 7	0.0	3	0.198	f l
d	1, 3, 8	0.129	4	0.057	f t
e	2, 7	0.163	5	0.057	g q
f	16	0.0	6	0.047	i j
g	3, 9	0.149	7	0.108	k h
h	4, 11, 17	0.0	8	0.096	m n
i	16	0.0	9	0.055	r f
j	16	0.0	10	0.048	r t
k	17	1.0	11	0.113	s z
l	16	0.0	12	0.057	x v
m	4, 5, 11	0.152	13	0.057	x y
n	4, 5, 11, 12	0.0	14	1.0	
o	3, 6, 7	0.127	15	1.0	
p	3, 8, 10	0.129	16	1.0	
q	3, 8, 9, 10	0.0	17	1.0	
r	16	1.0	18	1.0	
s	15	1.0			
t	16	0.0			
u	11, 12, 13	0.152			
v	13, 14	0.0			
x	14	1.0			
y	14	0.0			
z	15	0.0			

Table X.G2 Letters: best covering, best approximation to a partition, joint probabilities, and entropy

Solution and content	Subsets	Musters	$-\log_e$(joint probability)	Entropy
Generated subsets*	18	5	34.691	3.530
Covering {3,7,11,14-18}	8	6	6.025	0.807
Near-partition {2,3,5,12,14-18}	9	8	10.167	0.816

* See Table X.G1.

Table X.G3 Letters: subsets, musters, and comments on optimal covering of Swedish letter data*

Musters**	Comments
bdgpqo ce	Circular, with or without a vertical stroke
unmh k	Parallel vertical linearity, with or without a vertical stroke
vxy	Angled letters, open above
sz	Zigzag letters
fijlrt	Vertical linearity
a	Roundness with a hook

* See Kuennepas and Janson (1969) for the letter shapes actually used.

** Subsets underlined.

H Angles and distances

This case study gives some numerical examples to illustrate the procedures described in Appendix 4 for measuring the distances and angles among subsets. The notation of Appendix 4 is assumed.

Example 1

(i) Let $X = \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}$ $a' = [\frac{1}{2} \quad 0]$ $b' = [-\frac{1}{2} \quad 0]$

hence $\delta = 2^{1/2}$ (Expression A4.13).

(ii) $L = \begin{bmatrix} \frac{1}{2} & -1 \\ -\frac{1}{2} & 1 \end{bmatrix}$ $M = \begin{bmatrix} \frac{1}{2} & -1 \\ \frac{1}{2} & 1 \end{bmatrix}$

$\|LL^T\| = \|MM^T\| = 4.25$
$\|LM^T\| = 3.125,$
$\cos \Theta_{LM} = 25/34$ (Expression A4.9)
$\Theta_{LM} LM = 0.7447$ rads.

(iii) At the centroid of the X, the cosine of the angle is -1, i.e., the two points are collinear with the centroid.

(iv) The cosine of the angles at both x_i between a and b is 3/5.

(v) $r_a = r_b = (5/8)^{1/2}$ (Expression A4.14);
$\Delta_{ab}|X = (5/8)^{1/2} \cos(25/34) \approx 0.5887$ (Expression A4.24).

(vi) The Euclidean distance between a and b, in units of δ, is $1/2^{1/2} \approx 0.7071$.

Example 2

(i) Let $X = \begin{bmatrix} 3 & 2 \\ 2 & 5 \\ -5 & -7 \end{bmatrix}$ $a' = [1 \quad 2]$ $b' = [-3 \quad 5]$

hence $\delta = 9.699$.

(ii) $L = \begin{bmatrix} 2 & 0 \\ 1 & 3 \\ -6 & -7 \end{bmatrix}$ $M = \begin{bmatrix} 6 & -3 \\ 5 & 0 \\ -2 & -12 \end{bmatrix}$

$\|LL^T\| = 17\,119$;
$\|MM^T\| = 16\,279$;
$\|LM^T\| = 26\,410$,
$\cos \Theta_{LM} LM = 0.9086$,
$\Theta_{LM} = 0.4308$.

(iii) At the centroid of the X, the cosine of the angle is $7/(5 \times 34)^{1/2} = 0.5368$, corresponding to 1.0041 rads.

(iv) The cosines of the angles at the three members of X are [0.8944 0.3162 0.9119].

(v) $r_a = 0.5492$; $r_b = 0.8205$; $\Delta_{ab}|X = 0.3980$.

(vi) The linear (Euclidean) distance between the ends of r_a and r_b is 0.3956.

(vii) Consider the relationship between two members of X with respect to X. Let a' and b' be the first two rows of X;
$\cos \Theta_{LM} = 0.9832$
$\Theta_{LM} = 0.1837$
$r_a = 0.5225$
$r_b = 0.5861$
$\Delta_{ab} = 0.1199$.

The Euclidean distance between a and b, in δ units, is 0.3260, and between the ends of r_a and r_b is 0.2242. Thus a and b become closer by virtue of being members of X.

Example 3

$$X = \begin{bmatrix} 3 & 2 \\ 2 & 5 \\ -5 & -7 \end{bmatrix} \quad A = \begin{bmatrix} 1 & 2 \\ -6 & -7 \end{bmatrix} \quad B = \begin{bmatrix} -3 & -6 \\ 2 & 5 \end{bmatrix}$$

$$L^T = \begin{bmatrix} 1 & 0 & -7 & 9 & 8 & 1 \\ -2 & 1 & -11 & 9 & 12 & 0 \end{bmatrix}$$

$\|LL^T\|^2 = 288\ 625$

$$M^T = \begin{bmatrix} 6 & 1 & -1 & 5 & 0 & -2 & -2 & -7 & -9 \\ 8 & -3 & 8 & 11 & 0 & 11 & -1 & -12 & -1 \end{bmatrix}$$

$\|MM^T\|^2 = 37\ 081,$
$\|LM^T\| = 277\ 679,$
$\cos \Theta_{AB} = 0.9213,$
$r_a = 0.7634,$
$r_b = 0.7982,$
$\Delta_{AB}|X = 0.3137.$

(There is no case study I.)

J GE interaction

This case study is chosen to illustrate the use of multicriteria clustering as described in Chapter IX. The data used by Lin (1982), originally published by Yates and Cochran (1938), are reconsidered; these data are given in Table X.J1 together with the means and among-environments variances. The pattern vectors (Chapter IX) are given in Table X.J2, the squared distances and angles in Table X.J3, Lin's estimates of the squared distances (which equate the means but which are not based on equal among-environments variances) in Table X.J4, and the squared Fréchet distances in Table X.J5. The differences are striking; the largest distance in Table X.J4 is between Trebi and Peatland, but the distance between this pair is ranked only fifth (out of 10) in Table X.J5. Considering just the cultivar means and variances, those for Trebi and Peatland suggest that these cultivars differ somewhat from each other and from the other three, without any need to consider the pattern of responses across environments; this grouping is essentially that obtained by Lin.

Table X.J1 GE interaction: cultivar means (based on three replications and two years data)

	Environments							
Cultivar	1	2	3	4	5	6	Mean	Variance
Manchuria	161.7	247.0	185.4	218.7	165.3	154.6	188.8	1349.82
Svansota	187.7	257.5	182.4	183.3	138.9	143.8	182.3	1810.20
Velvet	200.1	262.9	194.9	220.2	165.8	146.3	198.4	1685.55
Trebi	196.9	339.2	271.2	266.3	151.2	193.6	236.4	4664.80
Peatland	182.5	253.8	219.2	200.5	184.4	190.1	205.1	751.18
Mean	185.8	272.1	210.6	217.8	161.1	165.7	212.1	
Variance	230.8	1441.3	1335.8	962.1	293.4	588.2		1689.98*

* Variance of the 30 (cultivar × environments) values.
Source: Yates and Cochran (1938).

Table X.J2 GE interaction: environment pattern vectors for each cultivar

	Environments					
Cultivar	1	2	3	4	5	6
Manchuria	-0.3299	0.7084	-0.0414	0.3640	-0.2861	-0.4163
Svansota	0.0568	0.7904	0.0011	0.0105	-0.4562	-0.4046
Velvet	0.0185	0.7025	-0.0425	0.2374	-0.3550	-0.5674
Trebi	-0.2586	0.6731	0.2279	0.1958	-0.5579	-0.2802
Peatland	-0.3688	0.7946	0.2301	-0.0751	-0.3378	-0.2448

Table X.J3 GE interaction: matrix of squared distances among the cultivars based on the pattern vectors (below diagonal) and angles in radians among them (above diagonal)

Manchuria	-	0.5663	0.4048	0.4506	0.5618
Svansota	0.3122	-	0.3164	0.4785	0.5361
Velvet	0.1632	0.0993	-	0.5317	0.6719
Trebi	0.1996	0.2247	0.2761	-	0.3897
Peatland	0.3074	0.2806	0.4348	0.1500	-

Table X.J4 GE interaction: matrix of squared distances equating means but without equating among-environments variances

Manchuria	-				
Svansota	260.7	-			
Velvet	133.8	88.5	-		
Trebi	748.3	658.3	755.0	-	
Peatland	198.3	278.1	336.5	976.4	-

Source: Lin (1982).

Table X.J5 GE interaction: matrix of squared Fréchet distances based on cultivar means and among-environments variances

Manchuria	-				
Svansota	75.17	-			
Velvet	100.00	261.50	-		
Trebi	3268.26	3598.89	2196.82	-	
Peatland	353.66	748.37	229.96	2664.11	-

The Fréchet squared distances in Table X.J5 show that Trebi is unlike the other four, and that Peatland is also somewhat different from the others, although less so. Table X.J3 suggests that a grouping on pattern alone would give a different arrangement of the varieties. This difference is confirmed by the results of a separate conditional clustering of each set of distances (the square roots of the values in Tables X.J3–J5). Table X.J6, where the results are summarized, shows that the groupings obtained without using s_i^2 associates two cultivars having the highest and lowest among-environments variance (Table X.J2). The grouping obtained, corresponding with Tables X.J4 and J5, can be inferred almost by inspection of the among-environments variances in Table X.J2. A simultaneous conditional clustering using the Fréchet distances of Table X.J5 and the pattern distances of Table X.J3 yielded four distinct subsets (Table X.J6d) in which each of Trebi and Peatland are single-object subsets, and two overlapping subsets, namely, {Manchuria, Velvet} and {Svansota, Velvet}, form the three-object muster {Manchuria, Velvet, Svansota}. This solution, hinted at by combining the separate analyses of the data of Tables X.J3 and J5, seems to be the most suitable for these data.

Table X.J7, an analysis of variance for the three-group arrangement, shows that the variety groupings absorb more than 90% of the sums of squares estimated for differences among the cultivars. The sum of squares associated with the entry W.G.L. in Table X.J7 was further analyzed by a singular decomposition of the 5×6 array of residuals (Snee 1982); the rank of this array, which is confined to group 1 containing three cultivars, cannot exceed 2. The actual sum of squares for W.G.L. is 1606.85; the squared singular values are 1321.12 and 285.73; the singular vector associated with the larger of these is the contrast [0.754 –0.648 –0.106] corresponding to Manchuria, Svansota, and Velvet. These values suggest that Velvet is intermediate between the other two (note that conditional clustering placed Velvet in the intersection

Table X.J6 GE interaction: summary of results obtained from conditional clustering of the data in Tables J3–J5

Generated subsets of cultivars	Probability* of being in an optimal covering
(a) From Table J5 (Fréchet distances)	
{Manchuria,Svansota,Velvet,Trebi}	1.0
{Peatland}	1.0
(b) From Table J4 (Lin's distances)	
(a) {Manchuria,Peatland}	0.25
(b) {Manchuria,Svansota,Velvet,Peatland}	0.5
(c) {Svansota,Velvet}	0.25
(d) {Trebi}	1.0
Optimal solution** consists of subsets b and d	
(c) From Table J3 (Pattern distances)	
(a) {Manchuria,Velvet}	0.5
(b) {Manchuria,Trebi}	0.5
(c) {Trebi,Peatland}	1.0
(d) {Svansota,Velvet}	1.0
Optimal solutions** (i) {a,c,d} (ii) {b,c,d}	
Both optimal solutions generate the same musters, namely, {Manchuria, Svansota, Velvet} and {Trebi, Peatland}	
(d) From Tables J3 and J5 simultaneously	
{Manchuria,Velvet}	1.0
{Svansota,Velvet}	1.0
{Trebi}	1.0
{Peatland}	1.0
All four subsets required; three musters, namely, {Manchuria, Svansota, Velvet}, {Trebi}, and {Peatland}	

* Maximum entropy estimates (see Chapter II).
** Solutions that maximize the joint probability of the chosen subsets.

of the multi-object subsets); Manchuria also is perhaps somewhat different from the other pair. The Bartlett test for the equality of these two squared singular values (here regarded as estimates of variance) gives a value of 4.16, which as a chi-square with 5 DF is not significant. By contrast, a similar decomposition of the residuals without any grouping of the varieties (Table X.J7) gives a value of 56.57, which as a chi-square with 9 DF indicates heterogeneity in the interaction structure.

Table X.J7 GE interaction: analysis of variance for the data of Table X.J1 with and without cultivar groupings suggested by the simultaneous clustering based on the Fréchet and pattern distances, together with a decomposition of the residuals

Source	DF	SS%
Cultivars (C)	4	17.15
Groupings (G)	2	15.88
Within group 1	2	1.27
Environments (L)	5	68.53
Residual C.L	20	14.31
(a) Without grouping		
Singular vectors*		
1st	4	9.37
2nd	5	2.96
3rd	5	1.86
4th	5	0.12
(b) With grouping		
G.L.	10	11.72
W.G.L	10	2.59
Singular vectors		
1st	5	2.13
2nd	5	0.46
Total	29	100
	(actual value 61 927.8)	

* See text.

K Fescue grasses

This case study is a numerical clustering of the 44 taxa (species or subspecies) of *Festuca* known to occur in North America; the objective is to generate some hypotheses about possible species groups within the genus. The data were assembled by Dr. S.G. Aiken, Canadian Museum of Nature, prepared in DELTA format (Dallwitz and Paine 1986); the states of 91 attributes for the 44 taxa are available in Aiken and Darbyshire (1990) and Aiken and Dallwitz (1991). The present study is based on 46 of the attributes in the file; those excluded deal with geographical and nomenclatorial considerations.

Table X.K1a lists the full names of the species and the labels used subsequently; Table X.K1b is a possible infrageneric classification of the genus *Festuca* in North America based on Alexeev (1980, 1985). Included are two species not mentioned by Alexeev, namely, *F. dasyclada* and *F. ligulata*; these have been assigned by Dr. Aiken on the basis of her observations and on comments in the taxonomic literature.

Using the facilities provided by DELTA, the data were converted to dissimilarities. The effective rank (Chapter VII) of the array of dissimilarities generated by DELTA both before and after the nonmetric transformation described in Chapter VII are given in Table X.K2.

There were 15 more edges in the relative neighborhood graph (RNG) than in the minimum spanning tree (MST); from these two graphs, a 2 × 2 contingency table was formed. Lefkovitch (1985*c*) described a test based on this table for examining the hypothesis of stability of the clustering (Table X.K3); the value of X^2 obtained for these data is 5.482; as a chi-squared with 1 DF, this has a probability of 0.019. The conclusion is that any structure that may be suggested by a clustering procedure may not be stable.

Table X.K1a Fescue grasses: labels and names of taxa

Labels	Name of taxon
alta	*F. altaica*
ariz	*F. arizonica*
arun	*F. arundinacea*
baff	*F. baffinensis*
brac	*F. brachyphylla*
brbr	*F. brachyphylla* ssp. *breviculmis*
brco	*F. brachyphylla* ssp. *coloradensis*
brev	*F. brevissima*
cali	*F. californica*
call	*F. calligera*
camp	*F. campestris*
dasy	*F. dasyclada*
elme	*F. elmeri*
fili	*F. filiformis*
giga	*F. gigantea*
hall	*F. hallii*
hype	*F. hyperborea*
idah	*F. idahoensis*
idro	*F. idahoensis* var. *romeri*
king	*F. kingii*
lene	*F. lenensis*
ligu	*F. ligulata*
minu	*F. minutiflora*
occi	*F. occidentalis*
para	*F. paradoxa*
prat	*F. pratensis*
rich	*F. richardsonii*
rubr	*F. rubra*
rude	*F. rubra* ssp. *densiuscula*
rudi	*F. rubra* ssp. *diffusa*
saxi	*F. saximontana*
sapu	*F. saximontana* var. *purpusiana*
soro	*F. sororia*
sula	*F. subulata*
suli	*F. subuliflora*
subv	*F. subverticillata*
thur	*F. thurberi*
trac	*F. trachyphylla*
vale	*F. valesiaca*
vers	*F. versuta*
virid	*F. viridula*
vihi	*F. vivipara* ssp. *hirsuta*
vivi	*F. viviparoidea*
vikr	*F. viviparoidea* ssp. *krajinae*

Table X.K1b Fescue grasses: infrageneric classification in the genus *Festuca* based predominantly on Alexeev (1980, 1985) or Aiken assigned

---

[1. Subgenus 1 *Drymanthele* Krecz. & Bobr.: *F. versuta.*
(Not mentioned by Alexeev 1985)]

2. Subgenus *Subulatae* (Tzvelev) E. Alexeev:
Section *Subulatae: F. sororia* (Alexeev 1980), *F. subulata*
Section *Elmera: F. elmeri* (Alexeev 1980)

3. Subgenus *Subuliflorae* E. Alexeev: *F. subuliflora*

4. Subgenus *Schedonorus* (Beauv.) Peterm.:
Section *Schedonorus: F. arundinacea, F. pratensis*
Section *Plantynia* (Dumort.) Tzvelev: *F. gigantea*

5. Subgenus *Obtusae* E. Alexeev: *F. paradoxa, F. subverticillata*

6. Subgenus *Leucopoa* (Griseb.) Hack.: *F. kingii*
Section *Breviaristatae: F. altaica, F. campestris, F. hallii*
[*F. californica*, and *F. thurberi* (Alexeev 1980), *F. ligulata* Aiken assigned]

7. Subgenus *Festuca*:
Section *Festuca*: *F. arizonica, F. baffinensis*, F. brachyphylla*, F. brevissima*, F. calligera, F. dasyclada, F. filiformis, F. hyperborea*, F. idahoensis, F. lenensis, F. minutiflora*, F. occidentalis, F. richardsonii*, F. rubra, F. saximontana, F. trachyphylla, F. valesiaca, F. viridula, F. vivipara* ssp. *hirsuta*, F. viviparoidea*

---

* Arctic or high alpine species.

Table X.K2 Fescue grasses: estimates of rank of dissimilarity arrays

α	Original data	Transformed data
0	42	24
0.5	12.66	6.64
1.0	5.15	2.82
2.0	2.01	1.39

Table X.K3 Fescue grasses: test of clustering stability by comparing the relative neighborhood graph with the minimum spanning tree

		MST		
		in	not in	total
RNG	in	43	15	58
	not in	0	888	888
	total	43	903	946

Chi-squared = $(15^2 \times 43)/(43^2 - 3 \times 43 + 2) = 5.618$

Table X.K4 gives the first five principal coordinates computed from the empirical distances; they account for 51.1% of the total distance. The first two coordinates are plotted in Fig. X.K1. Table X.K5 gives the adjacent objects on the MST, the lengths of the edges, and the upper tolerance of its edges, defined as the length each edge would have to exceed for the spanning tree to have a different topology.

Table X.K4 Fescue grasses: first five principal coordinates—empirical data

Labels	1 2.343377*	2 1.472337	3 1.019022	4 0.737053	5 0.593209
alta	0.0800	-0.3103	0.0212	-0.2126	0.0076
ariz	0.0627	-0.1864	-0.0234	-0.1527	0.1670
arun	0.1904	0.0804	-0.3062	0.0642	-0.2194
baff	-0.1202	-0.0264	0.3404	0.1053	-0.0225
brac	-0.2030	0.3319	0.1101	-0.0709	-0.0597
brbr	-0.3443	0.1190	0.0956	-0.1434	-0.1809
brco	-0.3104	0.2054	0.1450	-0.1376	-0.1166
brev	-0.2603	0.2377	0.1574	-0.1209	-0.1052
cali	0.0485	-0.3449	-0.0673	-0.0173	-0.1257
call	-0.0198	-0.0231	0.1523	0.1445	0.0542
camp	-0.0455	-0.3917	0.0163	-0.1366	-0.0231
dasy	0.1874	-0.0032	0.1608	0.1152	0.2236
elme	0.3665	0.0121	0.0286	0.0022	0.0239
fili	-0.3417	-0.0495	-0.0412	0.1432	0.0613
giga	0.1802	0.0880	-0.3170	0.1207	-0.2927
hall	-0.0548	-0.3334	0.0227	-0.1603	-0.0784
hype	-0.1642	0.2996	0.1460	-0.1511	-0.0620
idah	-0.2120	-0.1564	-0.2377	0.0607	0.0737
idro	-0.1578	-0.0800	-0.2413	0.0748	-0.0073
king	0.0712	-0.2881	0.0194	-0.2014	-0.0389
lene	-0.3406	-0.0811	-0.0678	0.0540	0.0840
ligu	0.1957	-0.0903	0.1759	0.1203	0.2288
minu	-0.1536	0.0284	0.3211	-0.0672	-0.1482
occi	0.0624	-0.0066	0.1740	0.1804	0.0117
para	0.3240	0.1048	0.1199	0.0559	-0.0511
prat	0.1972	0.1701	-0.2873	0.0883	-0.1287
rich	0.0199	0.2405	-0.1259	-0.1595	0.2117
rubr	0.0735	0.2629	-0.1365	-0.1654	0.1846
rude	0.0952	0.2426	-0.1441	-0.2428	0.1459
rudi	0.1212	0.2628	-0.1489	-0.1459	0.1769
saxi	-0.3231	0.0061	-0.0492	0.1376	-0.0119
sapu	-0.3317	0.0039	-0.0469	0.1314	-0.0088
soro	0.3778	0.0518	0.0427	0.0586	-0.0227
sula	0.3266	-0.0421	0.0603	0.0679	-0.0048
suli	0.3558	-0.0125	0.0147	-0.0109	-0.0493
subv	0.2833	0.0696	0.1597	0.1025	-0.1161
thur	0.1003	-0.2930	-0.0196	-0.2292	-0.0121
trac	-0.2833	-0.0879	-0.1765	0.1126	0.0906
vale	-0.3325	-0.0513	-0.0903	0.1378	0.0711
vers	0.3409	0.0166	0.0724	0.1452	-0.0113
vird	0.1452	-0.0382	0.0325	0.1276	0.0744
vihi	-0.2974	-0.0806	-0.1022	0.0472	0.0178
vivi	-0.2723	-0.1084	0.1068	0.1387	0.0815
vikr	-0.0589	0.2774	-0.0263	0.1065	0.0673

* Eigenvalue.

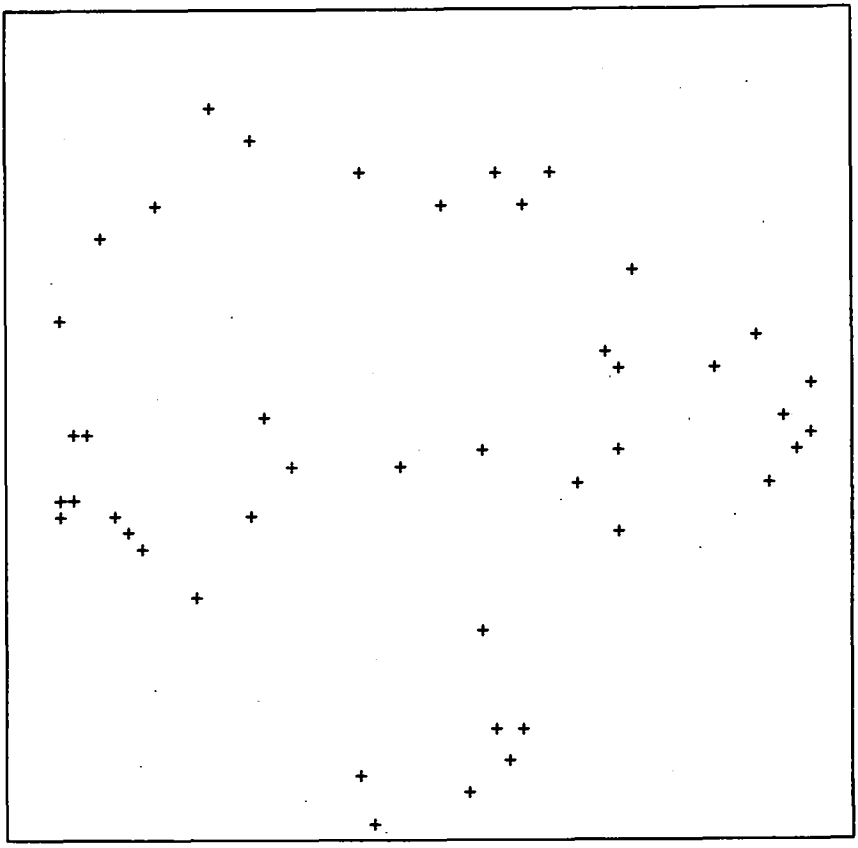


Fig. X.K1 Fescue grasses: empirical distances; coordinate 1 is horizontal, and 2 is vertical; horizontal and vertical limits are both {–0.39168, 0.39168}.

The 58 pairs of species in the RNG were used to generate subsets. The relationships in the reduction process are given in Table X.K6. Table X.K7 gives the remaining generated subsets and indicates those objects that form the rows of the reduced array. Table X.K8 lists the membership of the musters formed from Table X.K7, while Table X.K9 gives the subset and object probabilities (note that a probability of unity indicates a mandatory object or

Table X.K5 Fescue grasses: minimum spanning tree as linked list, and printed to show relationships

Object	Object	Distance	Upper tolerance
alta	camp	0.54610	0.58405
camp	cali	0.59254	0.68079
camp	hall	0.56824	0.58405
camp	vivi	0.64333	0.66321
vivi	baff	0.53464	0.65450
baff	call	0.58469	0.65450
call	occi	0.46475	0.65450
call	vird	0.51276	0.66241
vird	vers	0.59598	0.66241
vers	dasy	0.61073	0.64757
dasy	ligu	0.52932	0.75561
vers	para	0.57091	0.64757
para	soro	0.54144	0.62707
soro	elme	0.54989	0.58937
elme	suli	0.48389	0.58937
soro	sula	0.56379	0.64757
para	subv	0.54167	0.62707
vivi	lene	0.46937	0.65450
lene	vale	0.41553	0.65450
vale	fili	0.38463	0.74641
vale	saxi	0.35907	0.65450
saxi	sapu	0.25839	0.65450
sapu	brbr	0.53813	0.65450
brbr	brco	0.42259	0.65450
brco	brev	0.44625	0.66772
brev	brac	0.43719	0.66772
brac	hype	0.46824	0.71263
brac	vikr	0.54249	0.69223
vikr	rubr	0.59623	0.69223
rubr	prat	0.66153	0.71997
prat	arun	0.54392	0.85506
prat	giga	0.60973	0.66046
rubr	rude	0.51631	0.54311
rubr	rudi	0.50903	0.54311
rudi	rich	0.50146	0.69223
brco	minu	0.52157	0.66772
vale	trac	0.41669	0.60154
trac	idah	0.46202	0.60154
idah	idro	0.40395	0.60154
vale	vihi	0.48566	0.55718
alta	king	0.55281	0.73870
alta	thur	0.53907	0.70910
thur	ariz	0.60408	0.70910

Notes: There are 18 vertices of degree unity.

Length of minimum spanning tree = 22.10061
Sum of tolerances = 28.23793

Let x = (upper tolerance/distance)
Mean (logit(x)) = 1.52436
Variance (logit(x)) = 0.76417

In complete graph: smallest dissimilarity = 0.25839
average dissimilarity = 0.74073
largest dissimilarity = 0.91696.

subset; a probability of zero indicates an object eliminated by the reductions of Table X.K6). Table X.K10 gives the subsets in the maximum joint probability covering, and Table X.K11 gives the musters formed from the covering; the latter musters are nested within those of Table X.K8. Tables X.K12 and K13 give the covering and musters obtained from the maximum information solution for the same subsets and probabilities as in Tables X.K7 and K9.

The process was then repeated after the RNG (nonmetric) transformation (Chapter VII) of the dissimilarities, but because most of the generated subsets contained just two objects, the results are omitted from this summary. The rejection of the transformed data for these purposes emphasizes again that clustering is a hypothesis-generating procedure; note below, however, the result obtained using a dendrogram procedure with the transformed data.

Table X.K6 Fescue grasses: conditional clustering (55 initiating pairs were used)

Reduction relationships			
{alta king}*	{ariz ariz}	{arun giga}	{baff minu}
{brac hype}	{brbr brbr}	{brco brbr}	{brev brev}
{cali cali}	{call occi}	{camp cali}	{dasy dasy}
{elme elme}	{fili lene}	{giga giga}	{hall cali}
{hype hype}	{idah vihi}	{idro idah}	{king king}
{lene vihi}	{ligu dasy}	{minu minu}	{occi occi}
{para subv}	{prat arun}	{rich rich}	{rubr rude}
{rude rude}	{rudi rich}	{saxi vihi}	{sapu saxi}
{soro sula}	{sula sula}	{suli elme}	{subv subv}
{thur ariz}	{trac vihi}	{vale fili}	{vers vers}
{vird vird}	{vihi vihi}	{vivi vihi}	{vikr vikr}

* The second name in each pair is the retained object in the reduced array.

Table X.K7 Fescue grasses: empirical data (39 subsets generated and 26 remain after reductions)

	Label	Subset number 000000000111111111122222222 123456789012345678901234567
	alta	.....................1.....
>	ariz	1
	arun	....................1...
	baff	..1.................1.....
	brac	...11................1....
>	brbr	.....................1.
	brco	......1..............1.
>	brev	...1..1..................
>	cali	........................1..
	call	1.1....11................
	camp	........................1..
>	dasy	.........11..............
>	elme	..........111............
	fili	..................1.....
>	giga	.....................1...
	hall	......................1..
>	hype	....................1....
	idah	..................1.....
	idro	..................1.....
>	king	...................1....
	lene	..................1.....
	ligu	.........11..............
>	minu	..................1.....
>	occi	1.1....1................
	para	............1.1..........
	prat	....................1...
>	rich	...............11.......
	rubr	.1..............1.......
>	rude	.1..............1.......
	rudi	...............11.......
	saxi	..................1.....
	sapu	..................1.....
	soro	............111..........
>	sula	...........111...........
	suli	...........111...........
>	subv	............1.1..........
	thur	....................1....1
	trac	..................1.....
	vale	..................1.....
>	vers	.....1...1..1............
>	vird	1....1..1................
>	vihi	..................1.....
	vivi	..1..............1.....
>	vikr	1...1....................

> Indicates presence in the reduced array.

Table X.K8 Fescue grasses: musters (the union of intersecting subsets)

Muster	Size	Content
1	30	baff brac brbr brco brev call dasy elme fili hype idah idro lene ligu minu occi para saxi sapu soro sula suli subv trac vale vers virdvihi vivi vikr
2	4	rich rubr rude rudi
3	4	alta ariz king thur
4	3	arun giga prat
5	3	cali camp hall

Table X.K9 Fescue grasses

(a) Estimated subset covering probabilities

entropy = 4.006292 bits (i.e., using $\log_2$)

Subset	Probability									
1	0.111451	2	0.040046	3	0.037068	4	0.041717	5	0.038681
6	0.068441	7	0.041717	8	0.037068	9	0.035702	10	0.040351
11	0.073090	12	0.034473	13	0.068947	14	0.139160	15	0.034473
16	0.037475	17	0.040046	18	0.080093	19	1.000000	20	1.000000
21	1.000000	22	1.000000	23	1.000000	24	1.000000	25	1.000000
26	1.000000								

(b) Estimated object representation probabilities

entropy = 3.455464 bits

Species	Probability									
alta	0.000000	ariz	1.000000	arun	0.000000	baff	0.000000	brac	0.000000
brbr	1.000000	brco	0.000000	brev	0.101065	cali	1.000000	call	0.000000
camp	0.000000	dasy	0.097757	elme	0.083517	fili	0.000000	giga	1.000000
hall	0.000000	hype	1.000000	idah	0.000000	idro	0.000000	king	1.000000
lene	0.000000	ligu	0.000000	minu	1.000000	occi	0.089802	para	0.000000
prat	0.000000	rich	0.097018	rubr	0.000000	rude	0.097018	rudi	0.000000
saxi	0.000000	sapu	0.000000	soro	0.000000	sula	0.083517	suli	0.000000
subv	0.090788	thur	0.000000	trac	0.000000	vale	0.000000	vers	0.079315
virdvihi	0.086493	vivi	1.000000	vivi	0.000000	vikr	0.093711		

Table X.K10 Fescue grasses: maximum joint probability solution

Label	Solution subset numbers* 0000000001111 1234567890123
alta	.....1.....
ariz	1
arun	.........1...
baff	......1......
brac	.1.....1....
brbr	...........1.
brco	...........1.
brev	.1...........
cali	.........1..
call	1............
camp	.........1..
dasy	..1..........
elme	...1.........
fili	.....1.......
giga	.........1...
hall	.........1..
hype	........1....
idah	.....1.......
idro	.....1.......
king	.......1.....
lene	....1........
ligu	..1..........
minu	......1......
occi	1............
para	...1.........
prat	.........1...
rich	....1........
rubr	....1........
rude	....1........
rudi	....1........
saxi	.....1.......
sapu	.....1.......
soro	...1.........
sula	...1.........
suli	...1.........
subv	...1.........
thur	.......1....1
trac	.....1.......
vale	.....1.......
vers	..11.........
vird	1............
vihi	.....1.......
vivi	.....1.......
vikr	1............

Notes: –log joint probability (all subsets) = 53.856590
–log joint probability (irredundant covering) = 12.483777
* This irredundant cover contains 13 subsets; it is not a partition.

Table X.K11 Fescue grasses

(a) Probabilities of subsets in the cover (Table X.K10)

Entropy = 1.507503 bits

1	0.111451	2	0.041717	3	0.073090	4	0.139160	5	0.080093
6	1.000000	7	1.000000	8	1.000000	9	1.000000	10	1.000000
11	1.000000	12	1.000000	13	1.000000				

(b) Musters formed from the irredundant covering

Muster	Size	Content
1	4	call occi vird vikr
2	3	brac brev hype
3	9	dasy elme ligu para soro sula suli subv vers
4	4	rich rubr rude rudi
5	10	fili idah idro lene saxi sapu trac vale vihi vivi
6	2	baff minu
7	4	alta ariz king thur
8	3	arun giga prat
9	3	cali camp hall
10	2	brbr brco

Table X.K12 Fescue grasses: maximum information (= minimum entropy) solution

	Solution subset numbers*
	000000000111111
Label	123456789012345
alta	.........1.....
ariz	1
arun	..........1...
baff	1.......1.....
brac	.11.......1....
brbr	..............1.
brco	.............1.
brev	.1.............
cali	............1..
call	1..1...........
camp	............1..
dasy	....1..........
elme	.....1.........
fili	.......1.......
giga	...........1...
hall	............1..
hype	..........1....
idah	.......1.......
idro	.......1.......
king	.........1.....
lene	........1......
ligu	....1..........
minu	........1......
occi	1..............
para	.....1.........
prat	..........1...
rich	......1........
rubr	......1........
rude	......1........
rudi	......1........
saxi	.......1.......
sapu	.......1.......
soro	.....1.........
sula	.....1.........
suli	.....1.........
subv	.....1.........
thur	.........1....1
trac	.......1.......
vale	......1........
vers	.....1.........
vird	...1...........
vihi	.......1.......
vivi	1......1.......
vikr	..1............

* Note: This irredundant cover contains 15 subsets; it is not a partition.

Table X.K13 Fescue grasses

(a) Probabilities of subsets in the cover (Table X.K12)

Entropy = 1.595083 bits

1	0.037068	2	0.041717	3	0.038681
4	0.035702	5	0.040351	6	0.139160
7	0.080093	8	1.000000	9	1.000000
10	1.000000	11	1.000000	12	1.000000
13	1.000000	14	1.000000	15	1.000000

(b) Musters formed from the irredundant covering

Muster	Size	Content
1	15	baff call fili idah idro lene minu occi saxi sapu trac vale vird vihi vivi
2	4	brac brev hype vikr
3	2	dasy ligu
4	7	elme para soro sula suli subv vers
5	4	rich rubr rude rudi
6	4	alta ariz king thur
7	3	arun giga prat
8	3	cali camp hall
9	2	brbr brco

Table X.K14 is a summary of the musters formed from the subsets using the untransformed data, and after obtaining the optimal covering for both the maximum joint probability and information solutions, together with the subgeneric and sectional groupings of Table X.K1b. Using the labels of the groups in Table X.K14, groups 1a, 1b, 1d, 1e, 1f, 2, 4, and 5 can be regarded as consistent with the classification in Table X.K1b, even though there is more resolution; for example, perhaps group 1f "ought" to be together with group 1b to bring together the two subspecies and the main form. The two groups inconsistent with the Table X.K1b classification are 1c and 3. In considering group 1c, Aiken had

Table X.K14 Fescue grasses: musters formed from the generated subsets (M), from the optimal covering based on joint probability (P) and from the optimal covering based on maximum information, compared with the subgeneric (Sg) and sectional classification within each subgenus (Se) proposed by Aiken (pers. comm. 1992).

Group	Species	M	P	I	Sg	Se
1a						
	F. calligera	1	1	1	7	1
	F. occidentalis	1	1	1	7	1
	F. viridula	1	1	1	7	1
	F. viviparoidea ssp. *krajinae*	1	1	2	7	1
1b						
	F. brachyphylla	1	2	2	7	1
	F. brevissima	1	2	2	7	1
	F. hyperborea	1	2	2	7	1
1c						
	F. versuta	1	3	4	1	1
	F. sororia	1	3	4	2	1
	F. subulata	1	3	4	2	1
	F. elmeri	1	3	4	2	2
	F. subuliflora	1	3	4	3	1
	F. paradoxa	1	3	4	5	1
	F. subverticillata	1	3	4	5	1
	F. ligulata	1	3	3	6	2
	F. dasyclada	1	3	3	7	1
1d						
	F. filiformis	1	5	1	7	1
	F. idahoensis	1	5	1	7	1
	F. idahoensis var. *romeri*	1	5	1	7	1
	F. lenensis	1	5	1	7	1
	F. saximontana	1	5	1	7	1
	F. saximontana var. *purpusiana*	1	5	1	7	1
	F. trachyphylla	1	5	1	7	1
	F. valesiaca	1	5	1	7	1
	F. vivipara ssp. *hirsuta*	1	5	1	7	1
	F. viviparoidea	1	5	1	7	1
1e						
	F. baffinensis	1	6	1	7	1
	F. minutiflora	1	6	1	7	1

(continued)

Table X.K14 (*concluded*)

Group	Species	M	P	I	Sg	Se
1f						
	F. brachyphylla ssp. *breviculmis*	1	10	9	7	1
	F. brachyphylla ssp. *coloradensis*	1	10	9	7	1
2						
	F. richardsonii	2	4	5	7	1
	F. rubra	2	4	5	7	1
	F. rubra ssp. *densiuscula*	2	4	5	7	1
	F. rubra ssp. *diffusa*	2	4	5	7	1
3						
	F. kingii	3	7	6	6	1
	F. thurberi	3	7	6	6	2
	F. altaica	3	7	6	6	2
	F. arizonica	3	7	6	7	1
4						
	F. arundinacea	4	8	7	4	1
	F. pratensis	4	8	7	4	1
	F. gigantea	4	8	7	4	2
5						
	F. californica	5	9	8	6	2
	F. campestris	5	9	8	6	2
	F. hallii	5	9	8	6	2

assigned *F. dasyclada* to subgenus *Festuca* because previous analyses suggested that affiliation, and because it also has the subgenus *Festuca* seed protein band (this set of attributes was not included in the present study). However, the species has been put into a monophyletic genus, *Argillochloa*, by Weber (1984). All group 1c, other than *F. dasyclada*, have flat leaves, most (including this species) have hairy ovaries, rhizomes, and long anthers, as well as other character states in common; if this group is genuinely

more heterogeneous than the others, the explanation may be that it is the result of the particular subset of attributes used in calculating the dissimilarities. The presence of *F. arizonica* in group 3 (the other three of this group are consistent with Table X.K1b) raises another interesting point, because this species may be a hybrid between *F. calligera* (group 1a) and *F. idahoensis* (group 1d), so that they form a natural group together with *F. occidentalis*. Again, the set of attributes used may be the reason for this separation.

The conditional clustering of the data has generated some interesting questions about the relationships in the genus and, if the Table X.K1b classification is the truth, has demonstrated the importance of the choice of attributes for a taxonomic study. Conversely, if the groupings in Table X.K14 are a better approximation to the truth, this conclusion raises the need to reconsider the relationships within this genus, at least for the North American flora.

In this book, I have not used any other clustering methods to the data in any of the case studies. For interest, two familiar methods are now used for the *Festuca* data. Neither the single-linkage nor unweighted average linkage clustering methods suggested much structure in the untransformed data, although there was more apparent resolution into hierarchical levels for the average-linkage procedure applied to the transformed data than any other of the "classical," agglomerative, sequential clustering (i.e., dendrogram-forming) methods. Fig. X.K2 gives the single linkage dendrogram (which is by necessity identical in both the empirical and transformed data). Fig. X.K3 gives the average linkage dendrogram for the transformed data. Comparison of them with the conditional subsets solution requires that the dendrogram be cut at an appropriate level; this choice is left to the reader.

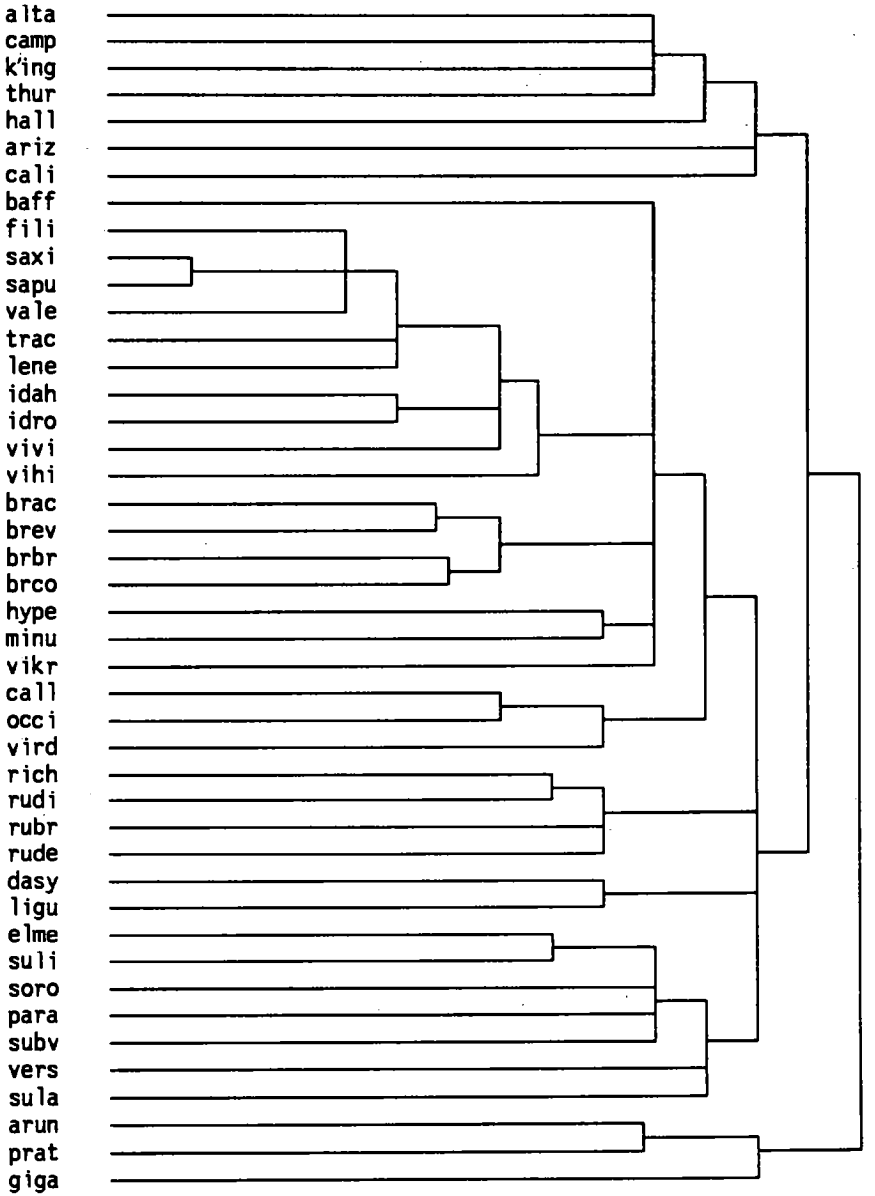


Fig. X.K2 Fescue grasses: single linkage dendrogram.

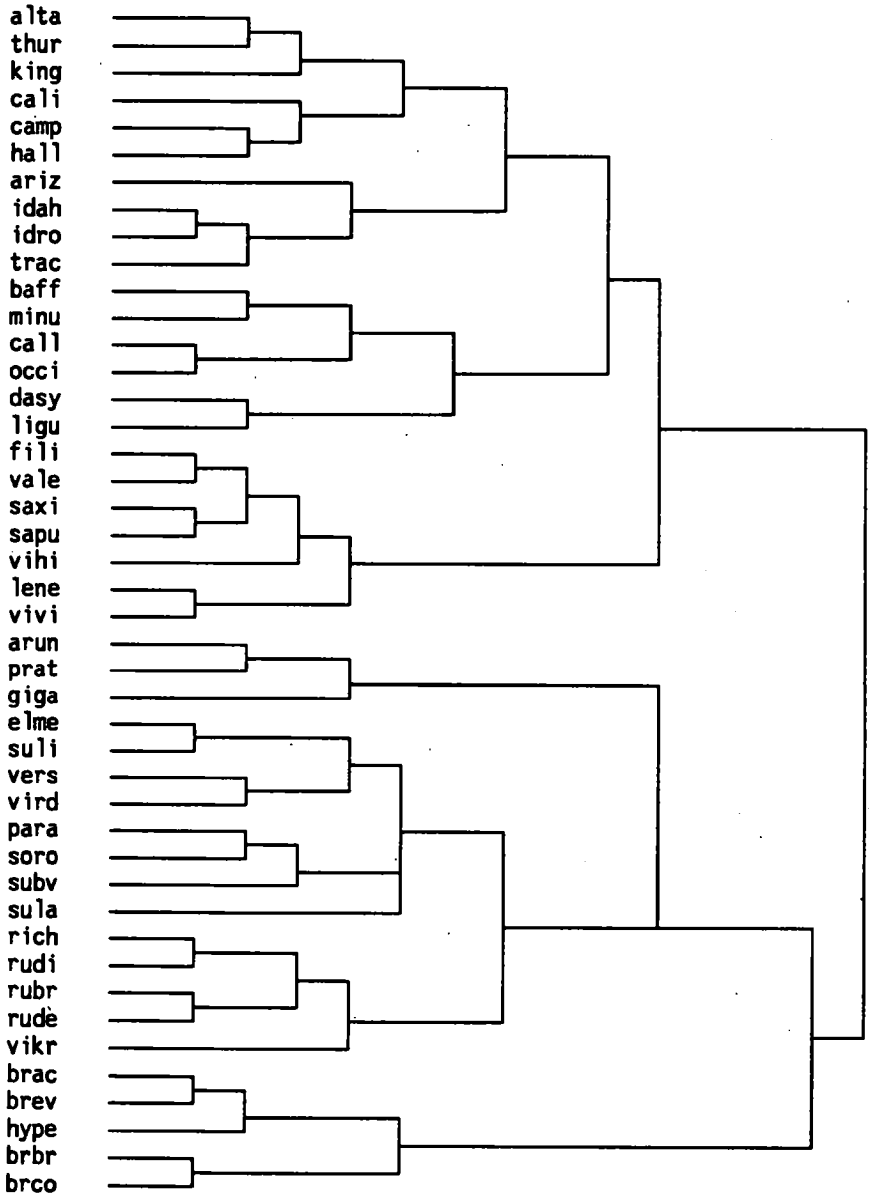


Fig. X.K3 Fescue grasses: transformed data, average linkage dendrogram.

L Cabbages

These data consist of observations made available to me by Dr. S.I. Warwick, Agriculture Canada, and are included in Warwick and Black (1991). The data consist of 223 restriction-site polymorphisms of the chloroplast DNA measured on 15 taxa belonging to *Brassica*, *Sinapis*, and *Raphanus* (Table X.L1). The objective of this case study is to illustrate the effects on the groupings based on dissimilarities determined from the Jaccard similarity (see Chapter VI), initially using all 223 attributes, and then just the 37 distinct patterns. A secondary objective is to illustrate the use of some consensus procedures.

The frequency and incidence of the 37 distinct patterns are given in Table X.L2a and b. Table X.L2b, which has been sorted first by rows and then on the columns, presents a clear picture of potential relationships. After the reductions on the pattern array, nine species remain (Table X.L2c and d).

The two sets of Jaccard similarities were computed and converted to dissimilarities by the negative of their natural logarithms, with similarities of zero assigned a (large) value (of 9.9999); the two sets of values are given in Tables X.L2e and L2f. Conditional clustering (Chapter VIII) was performed on both sets, both before and after the nonmetric transformation described in Chapter VII. The results, which are summarized in Table X.L3, show that one covering solution is also a partition, and also that partitions exist among the generated subsets for the remaining three.

The consensus method based on adjoining the incidence arrays of the optimal coverings (Chapter II) was then used to obtain final groupings. Two virtually identical solutions were obtained after reductions performed on this array (Table X.L4), which not only are consistent with each other but also suggest that two of the three subsets in the solutions may be more alike than is either to the third. A comparison among the covering and

consensus solutions, the musters and the sectional taxonomy in Table X.L1 is not without interest and is left to the reader. Other analyses of these data are given by Warwick and Black (1991).

Table X.L1 Cabbages: abbreviations, names, sectional taxonomy and chromosome number (*n*, not 2*n*) of taxa studied

Code	Genus, section, and species	*n*
	Sinapis section *Eriosinapis* (perennials)	
PUB1	*S. pubescens* ssp. *pubescens*	9
PUB2	*S. pubescens* ssp. *virgata*	9
ARIS	*S. aristidis*	9
INBV	*S. indurata* and *S. boivinii*	18
	Sinapis section *Sinapis* (annuals)	
ALBA	*S. alba* ssp. *alpa* and ssp. *mairei*	6
FLEX	*S. flexuousa*	6
	Sinapis section *Ceratosinapis* (annuals)	
ARVN	*S. arvensis* ssp. *arvensis* and ssp. *nilotica*	9
	Sinapis section *Chondrosinapis*	
AUCH	*S. aucheri*	7
	Brassica	
NIGR	*B. nigra*	8
RAPA	*B. campestris*	10
ORIN	*B. campestris*, oriental type	10
OLER	*B. oleracea*	9
ALBO	*B. alboglabra*	9
	Raphanus	
RRAP	*R. raphanistrum*	9
RSAT	*R. sativus*	9

Table X.L2 Cabbages: restriction site polymorphisms of chloroplast DNA for some *Brassica*, *Sinapis*, and *Raphanus* species

(a) Frequencies of the 37 distinct patterns in (b)

```
9,1,3,11,1,13,1,1,1,1,1,11,1,1,1,7,38,1,
1,4,5,1,10,2,1,21,7,1,9,1,1,19,3,1,31,1,1
```

(b) Species abbreviations and partly sorted incidence matrix

```
                    Distinct patterns
          ------------------------------------
          000000000111111111122222222233333333
          123456789012345678901234567890123456 7

ARVN      111111.............................
ALBA      1.....111111.......................
FLEX      1.....111111.......................
NIGR      1111..1.....1111...................
AUCH      ....1.......11..1111111............
RAPA      .1..1..1....1...1111...1111........
ORIN      .1..1..11...1...1111...1111........
ALBO      .1..1.......1...1111...111.111.....
OLER      .1..1.......1...1111...111.11......
RRAP      .1..1.......1...1...1..1......111...
RSAT      .1..1....1..1...11..1..1......11.1..
PUB1      ..1.1.1...1...1...1..1..1..1......1.
INBV      ..1.1.1...1...1...1..1..1..1......11.
PUB2      ..1.1.1...1...1...1..1..1..1......111
ARIS      ..1.1.1...1...1...1..1..1..1...11.
```

(c) Reductions

1. Delete ALBA: identical with FLEX.
2. Delete ORIN: covered by RAPA.
3. Delete ALBO: covered by OLER.
4. Delete PUB2, INVB, ARIS: covered by PUB1.
5. Delete 13, 34, 37: empty.

(d) Remaining species: {ARVN, FLEX, NIGR, AUCH, RAPA, OLER, RRAP, RSAT, PUB1}.
Remaining patterns: {1-12, 14-33, 35, 36}.

Table X.L2(e) Matrix of distances based on all the data (maximum value 9.9999 ≡ similarity = 0)

1													
2	2.4849												
3	2.4849	0.0000											
4	1.0116	1.9459	1.9459										
5	2.7081	9.9999	9.9999	2.1401									
6	2.0794	2.8904	2.8904	2.2513	0.9808								
7	2.1401	2.1972	2.1972	2.3026	1.0415	0.0800							
8	2.1401	9.9999	9.9999	2.3026	1.0415	0.4055	0.4700						
9	2.0794	9.9999	9.9999	2.2513	0.9808	0.3365	0.4055	0.0800					
10	1.8718	9.9999	9.9999	2.0794	1.3218	1.1632	1.2238	1.2238	1.1632				
11	2.0149	2.8332	2.8332	2.1972	1.1632	1.0415	1.0986	1.0986	1.0415	0.4055			
12	1.9459	2.0149	2.0149	1.6740	1.7346	1.8458	1.8971	1.5581	1.5041	2.8904	2.9957		
13	2.0149	2.0794	2.0794	1.7346	1.7918	1.8971	1.9459	1.6094	1.5581	2.9444	3.0445	0.0953	
14	2.0794	2.1401	2.1401	1.7918	1.8458	1.9459	1.9924	1.6582	1.6094	2.9957	3.0910	0.1823	0.0870
15	2.0794	2.1401	2.1401	1.7918	1.8458	1.9459	1.9924	1.6582	1.6094	2.2513	2.3514	0.1823	0.0870	0.1671

Table X.L2(f) Matrix of distances based on the distinct patterns

1													
2	1.7918												
3	1.7918	0.0000											
4	0.7138	1.6094	1.6094										
5	4.6052	9.9999	9.9999	3.8712									
6	4.0518	4.6347	4.6347	4.0254	0.7357								
7	4.0604	3.9416	3.9416	4.0342	0.7461	0.0126							
8	4.0775	9.9999	9.9999	4.0518	0.7665	0.2371	0.2482						
9	4.0690	9.9999	9.9999	4.0431	0.7563	0.2260	0.2371	0.0123					
10	3.9797	9.9999	9.9999	3.9512	0.6820	0.9116	0.9209	0.9393	0.9301				
11	3.9797	4.5539	4.5539	3.9512	0.6487	0.8792	0.8886	0.9072	0.8979	0.0846			
12	2.9444	3.4812	3.4812	2.6672	3.5264	3.6721	3.6805	3.4012	3.3928	4.7185	4.7185		
13	2.9575	3.4965	3.4965	2.6810	3.5361	3.6805	3.6889	3.4095	3.4012	4.7274	4.7274	0.0235	
14	2.9704	3.5115	3.5115	2.6946	3.5458	3.6889	3.6972	3.4177	3.4095	4.7362	4.7362	0.0465	0.0230
15	2.9704	3.5115	3.5115	2.6946	3.5458	3.6889	3.6972	3.4177	3.4095	4.0342	4.0342	0.0465	0.0230	0.0455

Table X.L3 Cabbages: various covering solutions for the chloroplast DNA data

Transformation	All data		Distinct patterns	
	Without	With	Without	With
Rank	5	4	6	2
Edges in RNG	19	19	16	16
Subsets generated and remaining	9	10	11	11
Musters	1. {RAPA,ORIN, ALBO,ARIS} 2. {PUB2,ARIS} 3. {PUB1,INVB} 4. {AUCH}ALBO, 5. {RRAP,RSAT} 6. {ALBA,FLEX} 7. {ARVN,NIGR}	1. {ARVN,NIGR PUB1,INVB PUB2,ARIS} 2. {RAPA,ORIN OLER} 3. {RRAP,RSAT} 4. {AUCH} 5. {ALBA,FLEX}	1. {ARVN,ALBA, FLEX,NIGR, PUB1,INVB, PUB2,ARIS} 2. {AUCH,RAPA, ORIN,ALBO, OLER,RRAP, RSAT}	1. {ARVN,NIGR, PUB1,INVB, PUB2,ARIS} 2. {AUCH,RAPA, ORIN,ALBO, OLER,RRAP, RSAT} 3. {ALBA,FLEX}
Covering solution	as the musters	2-5 above 1. {ARVN,NIGR} 6. {NIGR,PUB1, INBV,PUB2, ARIS}	1 above 3. {AUCH,RRAP, RSAT} 4. {AUCH,RAPA, ORIN,ALBO, OLER}	1 and 3 above 4. {AUCH,RRAP, RSAT} 5. {AUCH,RAPA, ORIN,ALBO, OLER}
Partition	as above	1-5 above 7. {PUB1,INVB} 8. {PUB2,ARIS}	1,4 above 5. {RRAP,RSAT}	1,3,4 above 6. {RAPA,ORIN, ALBO,OLER}
$-\log_e$(joint probability)				
All subsets	3.466	11.090	27.901	24.108
Covering	0.693	2.773	5.456	5.394
Partition	0.693	5.545	6.087	5.821

Table X.L4 Cabbages: subsets forming the optimal coverings

	All data		Transformed data		Consensus solutions	
Species	Without	With	Without	With	1	2
ARVN	1	1	1..	1...	1..	1..
ALBA	....1.	...1.	1..	...1	1..	1..
FLEX	....1.	...1.	1..	...1	1..	1..
NIGR	1	1...1	1..	1...	1..	1..
AUCH	...1..	..1..	.11	.11.	..1	..1
RAPA	1.....	.1....	..1	..1.	..1	.1.
ORIN	1.....	.1....	..1	..1.	..1	.1.
ALBO	1.....	.1....	..1	..1.	..1	.1.
OLER	1.....	.1....	..1	..1.	..1	.1.
RRAP	...1...	..1...	.1.	.1..	.1.	..1
RSAT	...1...	..1...	.1.	.1..	.1.	..1
PUB1	..1....	1.....	1..	1...	1..	1..
INBV	..1....	1.....	1..	1...	1..	1..
PUB2	.1.....	1.....	1..	1...	1..	1..
ARIS	.1.....	1.....	1..	1...	1..	1..

M Beetles

This case study shows that it is possible to obtain a unique minimal covering without recourse to the probabilities. The data consist of a set of 17 attributes coded by Sharkey (1989) for 24 species of the beetle genus *Hoplicnema* from published data (Table X.M1). At the end of the reduction process, which required about 5 minutes to do by hand and which is given in Table X.M2, the array is reduced to two rows and three columns (Table X.M3a). Clearly, the most parsimonious choice from this array is given by column 15. The unique minimal covering (Table X.M3b), which consists of the subsets indicated by the elements of **x** equal to unity, is not a partition. It should be understood that these attributes are not necessarily of special importance, but only that the species covered by them show some logical patterns in their attributes. The last step in the reduction process (Table X.M2) states that attributes 4 and 16 are identical; this remark refers to the array after the previous reductions, because this statement is not true of the original data. If a least cost (rather than a minimal) covering is to be chosen, the costs of including these attributes need to be obtained, so that the attribute of the two with the lower cost is selected. However, since attribute 15 appears to make either of attributes 4 and 16 redundant, perhaps the issue is moot.

Table X.M1 Beetles: data matrix for *Hoplicnema* (24 species, 17 attributes), after Sharkey (1989)

Code	Species	Attribute numbers 00000000011111111 12345678901234567
A	*amplissima*	111.11.....
B	*debrae*	111.11.....
C	*fundita*	11.....1.....
D	*lata*	11.....11.....
E	*brasilensis*	11.....1.....
F	*panamensis*	11.....1.....
G	*insularis*	11.....1...1.1..
H	*puertoricensis*	11.....1...1.1..
I	*maya*	11.....1.....1..
J	*matthewsi*	11.....1.1...1..
K	*woldai*	11.....1.....11.
L	*darlingtoni*	11.....11....11.
M	*sallaei*	11.....1111.1111
N	*spiniventer*	11.....1111.1111
O	*aquilonaria*	11....1.1111..11.
P	*cubensis*	11....1.11.1..11.
Q	*thomasi*	11....1.11.1..11.
R	*jamaicensis*	11.....1..1..
S	*affluens*	11.....1..11.
T	*impunctata*	11....1....1..11.
U	*media*	11.1.....11.
V	*schwarzi*	11.1.....11.
W	*hesperia*	11.1...1.....11.
X	*minima*	11.1...1.....11.

Table X.M2 Beetles: reduction sequence

Condition	Action
1 and 2 uniform	Delete 1 and 2
3,5, and 6 identical	Delete 5 and 6
A covers B	Delete B
C covers D	Delete D
E covers {F-Q}	Delete {F-Q}
R covers S and T	Delete S and T
U covers V, W, and X	Delete V, W, and X
7,8,10,13,14,17 emptied	All deleted
A, C, and E in one subset each	Set $x_3 = x_{11} = x_9 = 1$; delete 3,9,11,A,C,E
4 and 16 identical	Delete 16

Table X.M3 Beetles:

(a) Reduced array

Code	Species	Attribute numbers 011 425
R	*jamaicensis*	.11
U	*media*	1.1

(b) Minimal covering

x_3 = {A, B} : x_9 = {E–Q}
x_{11} = {C, D, J, M, N, O}: x_{15} = {G–X}

N Blood and language

This case study looks at the average genetic distances among speakers of different language groups. The distances were calculated from frequency data for blood antigens, enzymes, and proteins of 26 genetic systems, from among 3369 localities across Europe which had been assigned to one of 12 language-family affiliations by Harding and Sokal (1988). They considered the estimates of the first nine (Table X.N1a) of the twelve language-groups to be more reliable than those of the remaining three, which they described as furnishing "unreliable estimates of distances" because they are based on 2, 3, and 7 systems, respectively. These three (Baltic, Albanian, and Semitic) are omitted from Table X.N1; in the original data for Baltic and Albanian, the corresponding rows were complete, but there were missing data for the Baltic–Semitic, the Albanian–Semitic, and within-Semitic distances. The first series of analyses is confined to the nine; the second gives an example of the unfolding procedure (Chapter VII) for replacing missing values.

Table X.N1 Blood

(a) matrix of dissimilarities (based on blood groups) among speakers of different language families transformed from those given by Harding and Sokal (1988)

Language family									
Germanic	0.000								
Romance	0.142	0.000							
Slavic	0.099	0.179	0.000						
Finnic	0.389	0.375	0.419	0.000					
Ugric	0.178	0.272	0.027	0.410	0.000				
Greek	0.414	0.246	0.329	0.487	0.371	0.000			
Celtic	0.037	0.094	0.089	0.445	0.275	0.454	0.000		
Basque	0.428	0.457	0.510	0.906	0.785	0.782	0.340	0.000	
Turkic	0.260	0.075	0.157	0.348	0.121	0.037	0.330	0.778	0.000

(b) squared lengths of the principal coordinates

Coordinate	Original data	After RNG transformation
1	0.5972	0.4482
2	0.1573	0.2219
3	0.1364	0.0300
4	0.0321	0.0151
5	0.0286	
6	0.0209	

The original data include a within-family distance, which can be regarded as a measure of internal heterogeneity, and were transformed as follows. Let g_{ij} be the tabulated values reported by Harding and Sokal; it is postulated that variances and covariances are proportional to $\exp(-g_{ij})$. Since the values of $\exp(-g_{ii})$ are not unities, to convert the array to similarities, i.e., s_{ij}, the whole array was transformed in the same way as a covariance matrix is

converted to correlations. Dissimilarities were obtained from these values as $-\log(s_{ij})$; the complete transformation can be summarized as

$$d_{ij} = g_{ij} - \tfrac{1}{2}(g_{ii} + g_{jj}).$$

For the first nine language groups, these values are given in Table X.N1a. The squared lengths of the principal coordinates (Table X.N1b) show that the transformation has essentially made the data two-dimensional.

Subsets were generated (Chapter VIII) using the values given in Table X.N1a; after the reductions, there were nine subsets, of which two (Finnic and Basque) consisted of the speakers of one language family each; the remaining seven (Table X.N2) form a single muster. The subset probabilities are given in Table X.N3; the representation probabilities (i.e., **q** of Chapter II) are given in Table X.N2. The interpretation of the latter values is that they estimate the importance of that object as an indicator of which subsets participate in an optimal covering. The optimal covering, using joint probability, consisted of

1. {Germanic, Romance, Slavic, Celtic}
2. {Slavic, Ugric}
3. {Greek, Turkic}
4. {Basque}
5. {Finnic}

in which it can be seen that the first two form a muster, having Slavic speakers in common.

Table X.N2 Blood: generated subsets using the data from Table X.N1a

Language family	Original data: Object probabilities	Original data: Subset numbers	Transformed data: Object probabilities	Transformed data: Subset numbers
		123456789		123456789
Germanic	0.280	1.1.1....	0.175	1.1.11...
Romance	0.337	.11.....	0.250	.11..1...
Slavic	0.000	.1111....	0.250	...111...
Finnic	1.000	1	1.000	1
Ugric	0.384	..1.1....	0.000	...111...
Greek	1.000	......1..	0.325	.....11..
Celtic	0.000	1.1.1....	0.000	1.1.11...
Basque	1.000	......1..	1.000	.......1.
Turkic	0.000	.1...1.1.	0.000	.1...11..

Table X.N3 Blood: subset probabilities using the data from Table X.N2

Original data: Subset	Original data: Probability	Transformed data: Subset	Transformed data: Probability
1	0.123	1	0.062
2	0.148	2	0.088
3	0.270	3	0.149
4	0.168	4	0.088
5	0.123	5	0.149
6	0.168	6	0.351
7	1.000	7	0.114
8	1.000	8	1.000
9	1.000	9	1.000

The relative neighborhood graph (Table X.N4) included one more edge than in the minimum spanning tree. After the nonmetric transformed dissimilarities based on this graph, the number of subsets remaining after reductions was also nine; they are given in Table X.N2, together with the representation probabilities of the

Table X.N4 Blood: relative neighbor graph for the data in Table X.N1

Ugric-Slavic	Turkic-Finnic	Celtic-Slavic
Celtic-Romance	Turkic-Ugric*	Turkic-Romance
Basque-Celtic	Celtic-Germanic	Turkic-Greek

* Not on minimum spanning tree.

language groups. Table X.N3 gives the subset probabilities. The optimal covering consisted of

1. {Germanic, Romance, Slavic, Ugric, Greek, Celtic, Turkic}
2. {Basque}
3. {Finnic}.

This grouping corresponds with the musters formed from the subsets generated from the untransformed data, consistent with an hypothesis that the speakers of Basque and Finnic differ sufficiently from others in Europe that they should be excluded from any further attempts at resolving the relationships based on these data. A conditional clustering of the remaining seven produced the same musters (as did the untransformed data), namely, a group formed of the Greek and Turkic speakers, and a second of the speakers of all the remaining languages. Remembering that the children and grandchildren of migrants often have no knowledge of their ancestral language, these results speak to the relative isolation of the Basque and Finnic speakers, the links between the Greek and Turkic speakers, and the general mixing among the remainder of speakers of European languages.

To include the data for speakers of Baltic, Albanian, and Semitic, the 12 × 9 array given by Harding and Sokal (1988) was unfolded as described in Chapter VII for missing data. A singular

decomposition of the $\{-\exp(g_{ij})\}$ yielded nine singular values; the 12×9 matrix of singular vectors was formed into a symmetric array of order 12 as described in Chapter VII. This array was then converted into a correlation matrix, from which an array of distances was computed as already described; the distances (Table X.N5) were scaled to be no greater than 1.0. Using these distances, 12 subsets were generated; the musters formed from the generated subsets consisted of Baltic, Basque, and Finnic each alone, with the remaining nine forming a single group. The following subsets were obtained as the optimal covering:

1. {Germanic, Romance, Slavic, Ugric, Greek, Celtic, Turkic, Semitic}
2. {Greek, Albanian}
3. {Baltic}
4. {Basque}
5. {Finnic}.

Table X.N5 Blood: matrix of unfolded transformed genetic distances

Language family											
Germanic											
Romance	0.088										
Slavic	0.057	0.121									
Finnic	0.239	0.233	0.274								
Ugric	0.109	0.155	0.022	0.278							
Greek	0.190	0.094	0.146	0.241	0.163						
Celtic	0.022	0.068	0.067	0.270	0.145	0.201					
Basque	0.262	0.267	0.315	0.558	0.500	0.357	0.190				
Turkic	0.151	0.075	0.093	0.227	0.102	0.054	0.153	0.449			
Baltic	0.327	0.309	0.246	0.334	0.184	0.157	0.406	0.543	0.149		
Albanian	0.465	0.326	0.415	0.491	0.401	0.101	0.488	0.999	0.192	0.223	
Semitic	0.175	0.112	0.170	0.392	0.228	0.064	0.204	0.274	0.116	0.332	0.292

Of these subsets, 1 and 2 form a muster. After the relative neighborhood graph nonmetric transformation (which produced the same musters as above), the first of these subsets was divided into three in the optimal covering:

1. {Germanic, Slavic, Ugric, Celtic}
2. {Romance, Turkic}
3. {Greek, Semitic}
4. {Greek, Albanian}
5. {Baltic}
6. {Basque}
7. {Finnic}.

Here, the musters are subsets 3 and 4 combined and the others as given above. Further discussion of these data using other procedures is given by Harding and Sokal (1988).

(There is no case study O)

P Pollution in a river ecosystem

This example is meant to show how similarities, described in Chapter VII, may be used to generate hypotheses without formal clustering methods. Karayiannis and Venetsanopoulos (1990) reported the incidence of 25 chemicals on 15 consecutive days in a river (Table X.P1); the possible sources are given in Table X.P2, and the chemicals identified in Table X.P3. The absence of a unit in the arrays implies that the chemical did not exceed a (detection) threshold. The objective is to determine the probable sources, confined to the six listed in Table X.P2, of the pollution seen on each of the 15 days.

Table X.P1 Pollution: excess chemicals present on 15 days in a river ecosystem

Day	Chemicals* 000000000111111111122222212345678901234567890123456789012345
1	1.111..11111111111.1...1.
2	..11.1.111.11.1111.1.1...
3	1111.1111.111.11111111111
4	111..1111.11..11111111111
5	111.1111.111111111111.111
6	1.1111.1..111111.1.1.1.1.
7	11...111111...111.1.11111
8	..11.1.1...11.1..1.1.1.1.
9	11..1111111.11111.1.11111
10	111111111.111111111111111
11	..1....111.1..1111.1...1.
12	11111111.11111111111111.1
13	1..1..11...11111..1..1..1
14	.1..1.11....111...1...111
15	11.11......1...1.....111.

* See Table X.P3 for identification.

Table X.P2 Possible pollution sources

Source	Chemicals* 000000000111111111122222212345678901234567890123456789012345
Sewage plant	1.111..1..1111.1.1.1...1.
Plating plant	.......111....111........
Milk products	..11.1.....11.1..1.1.1...
Papermill	11...1111.1...111.1.11111
Slaughterhouse	..1....1...1.......1...1.
Textile mill	11..1.11.1111111111.1.1.1

* See Table X.P3 for identification.

Table X.P3 Pollution: chemicals indicated in Tables X.P1 and X.P2

Number	Pollutant
1	alum
2	ammonia
3	B.O.D.
4	carbohydrate
5	carbonaceous compounds
6	casein
7	cellulose
8	chlorine
9	copper compounds
10	cyanides
11	detergents
12	fats, grease, oil
13	ferric chloride
14	ferric sulfate
15	lime
16	mineral acids
17	mineral alkalies
18	nitrogenous compounds
19	phosphates
20	proteins
21	starch
22	sugars
23	sulfites
24	suspended organic acids
25	suspended inorganic acids

This *inverse inference* problem may be investigated in many ways; here, the emphasis is on the use of a similarity coefficient, *without* clustering. Although the data in Tables X.P1 and P2 form 0–1 arrays, and each may be considered in that context (e.g., which days are alike, which sources are alike, which chemicals tend to occur together), together they do not form an obvious example of the circumstances discussed in Chapter II.

For simplicity, I assume first, that no pollution is carried over from day to day within the river; and second, the various sources are assumed to be releasing effluent, at least potentially, on

each day without any single source necessarily exceeding the threshold of detection. A simple inspection of the data shows that the pollution on any single day does not match exactly that of any single source. In consequence, the possibility that there may be an exact coincidence with each day for the union of two or more sources, must be investigated. The computational procedure adopted here is to determine which pairs, trios, etc. of sources may be responsible for each day based on the Jaccard similarity (Chapter VII) between each day and each combination of sources, and to propose that the highest similarity involving the fewest sources be considered as a solution to the problem.

Table X.P4 gives the similarities between each day and each source, and between each day and the 15 pairs of sources; the table has been abbreviated from the complete set of 62 (the null and improper subsets of sources were not considered) because the highest similarities involving the fewest sources were confined to the pairs.

Table X.P5 gives the sources associated with the maximum similarity for each day for the single and the pairs of sources extracted from Table X.P4. Note that source 5 was never implicated alone, and that 4 and 6, in descending order, are the most frequently involved on either a single or paired source basis.

Because similarity can be interpreted as a probability (Chapter VII), the geometric means of the similarities for the single sources over the 15 days may be regarded as giving a measure of the importance of the source for pollution; in descending order, the first two are sources 6 and 4, the same pair as by the other (ad hoc) procedure. The interpretation of the similarity coefficients as probabilities can be taken somewhat further. The similarity between the union of sources i and j either for each individual day or for the geometric mean may be represented as

$$\Pr(i \cup j) = \Pr(i) + \Pr(j) - \Pr(i \cap j)$$

so that $\Pr(i \cap j)$ may be considered as measuring the probability of sources i and j polluting simultaneously. From the matrix of these values, with the diagonal set to $\Pr(i)$, the Perron–Frobenius eigenvector gives an ordering of the sources (Table X.P6a) in which sources 6 and 4 (followed closely by 1) appear to be most implicated.

The probability that source i is polluting given that source j is polluting is given by

$$\Pr(i \mid j) = \Pr(i \cap j)/\Pr(j)$$

(Table X.P6b). This matrix, which by definition has unities on the diagonal, is not symmetric, but, because it is a matrix of positive elements, the singular vectors associated with the largest singular value can be recognized as being those associated with the Perron–Frobenius eigenvalue of the matrix. From the way this array is presented in the table, it is apparent that the left singular vector is likely to be informative about the sources of pollution; it places source 6 at the head, followed by sources 1 and 4. It is interesting (but not mathematically surprising because of the way this array is formed) that the right singular (Perron–Frobenius) vector ranks the sources in approximately the reverse order to that of the left.

The intersection and conditional probabilities can be obtained for the individual days in the same way as illustrated for the geometric means. It seems reasonable to conclude that if the conditional probability of the second source of the pair (in comparison with the single source having the highest probability) is appreciably larger than 0.5, the second source should also be considered as a polluter.

More formal statistical methods, based upon generalized linear models and assuming that the entries in the data tables are Bernoulli random variables, are beyond the present scope.

Table X.P4 Pollution: similarities between days and single sources, and between days and pairs of sources

Single sources

Days	1	2	3	4	5	6
1	0.8745	0.5941	0.5659	0.4851	0.5423	0.6860
2	0.5930	0.6547	0.8018	0.4677	0.4781	0.5040
3	0.6504	0.4352	0.6396	0.8528	0.4767	0.7538
4	0.5582	0.4564	0.5217	0.8944	0.5000	0.7379
5	0.7096	0.4352	0.4975	0.7462	0.4767	0.9045
6	0.9014	0.3062	0.7500	0.5000	0.5590	0.5893
7	0.3363	0.5941	0.2425	0.9701	0.2169	0.7432
8	0.6690	0.2462	0.9045	0.3769	0.6742	0.3553
9	0.4961	0.5477	0.2981	0.8944	0.2000	0.8433
10	0.7360	0.4167	0.6124	0.8165	0.4564	0.8179
11	0.5854	0.7385	0.5025	0.4523	0.6742	0.4975
12	0.6940	0.4256	0.6255	0.7298	0.3730	0.8847
13	0.5604	0.3536	0.4811	0.5774	0.2582	0.6804
14	0.4181	0.2462	0.2010	0.6030	0.2697	0.7107
15	0.5547	0.1361	0.3333	0.5000	0.2981	0.4714
Geometric mean						
	0.6042	0.4047	0.4905	0.6305	0.4018	0.6579

Pairs of sources

Days	1,2	1,3	1,4	1,5	1,6	2,3	2,4
1	1.0000	0.8489	0.7921	0.8745	0.8273	0.7778	0.5294
2	0.7778	0.7350	0.7092	0.5930	0.6268	1.0000	0.5186
3	0.7239	0.7462	0.9574	0.6504	0.8636	0.7407	0.8273
4	0.6508	0.6708	0.9129	0.5582	0.8104	0.6574	0.8677
5	0.7756	0.7462	0.9139	0.7096	0.9545	0.6268	0.7756
6	0.8489	1.0000	0.8165	0.9014	0.7462	0.7350	0.4851
7	0.5294	0.4851	0.7921	0.3363	0.7239	0.5186	1.0000
8	0.6581	0.8292	0.6770	0.6690	0.5785	0.8058	0.3656
9	0.6508	0.6149	0.8672	0.4961	0.8104	0.5379	0.9220
10	0.7921	0.8165	1.0000	0.7360	0.9139	0.7092	0.7921
11	0.8044	0.6030	0.6155	0.5854	0.6428	0.8058	0.5119
12	0.7586	0.7819	0.9364	0.6940	0.9336	0.7245	0.7586
13	0.5601	0.6495	0.7071	0.5604	0.6770	0.5401	0.5601
14	0.4388	0.4523	0.6770	0.4181	0.7071	0.2417	0.5850
15	0.4851	0.5833	0.6124	0.5547	0.5685	0.3563	0.4851
Geometric mean							
	0.6811	0.6897	0.7894	0.6042	0.7490	0.6193	0.6394

(continued)

Table X.P4 *(concluded)*

Days	Pairs of sources (*continued*)							
	2,5	2,6	3,4	3,5	3,6	4,5	4,6	5,6
1	0.7670	0.7233	0.7239	0.6581	0.7586	0.6121	0.7239	0.7939
2	0.7606	0.5518	0.7407	0.8058	0.7245	0.6131	0.6268	0.5832
3	0.6068	0.7826	1.0000	0.7071	0.8891	0.9293	0.8636	0.8374
4	0.6364	0.7695	0.9535	0.6068	0.8393	0.9747	0.8581	0.8295
5	0.6068	0.8804	0.8636	0.5785	0.9336	0.8315	0.9091	0.9770
6	0.5534	0.5735	0.7462	0.8292	0.7819	0.6309	0.6929	0.7092
7	0.5369	0.7790	0.8273	0.3656	0.7586	0.8903	0.8790	0.7410
8	0.5721	0.3459	0.7071	1.0000	0.6287	0.5534	0.5143	0.5264
9	0.4950	0.8721	0.8104	0.4045	0.8393	0.8208	0.9535	0.8295
10	0.5809	0.8429	0.9574	0.6770	0.9364	0.8898	0.9139	0.8909
11	0.9535	0.5534	0.6428	0.6364	0.5658	0.6225	0.5785	0.6580
12	0.5275	0.8611	0.8891	0.6287	1.0000	0.8132	0.8891	0.9100
13	0.3651	0.6623	0.6770	0.5222	0.7223	0.5960	0.6770	0.6299
14	0.2860	0.6917	0.5785	0.3636	0.6287	0.5534	0.7071	0.7237
15	0.3162	0.4588	0.5685	0.4020	0.5560	0.5353	0.5685	0.5092
Geometric mean	0.5449	0.6690	0.7679	0.5864	0.7592	0.7089	0.7433	0.7301

Table X.P5 Pollution: possible single and paired sources of pollution

(a) Day, sources and maximum similarity

Day	Single		Pairs	
	Source	Similarity	Sources	Similarity
1	1	0.875	1,2	1.0
2	3	0.802	2,3	1.0
3	4	0.853	3,4	1.0
4	4	0.894	4,5	0.975
5	6	0.905	5,6	0.977
6	1	0.901	1,3	1.0
7	4	0.970	2,4	1.0
8	3	0.905	3,5	1.0
9	4	0.894	4,6	0.954
10	6	0.818	1,4	1.0
11	2	0.739	2,3	0.806
12	6	0.885	3,6	1.0
13	6	0.680	3,6	0.722
14	6	0.711	5,6	0.724
15	1	0.555	1,4	0.612

(b) Frequency of sources implicated

Source	Single	Pairs	Total
1	3	4	7
2	1	4	5
3	2	7	9
4	5	6	11
5	0	4	4
6	5	5	10

Table X.P6 Pollution: source probabilities derived from geometric means (Table X.P4)

(a) $Pr(i \cap j)$

1	0.6042					
2	0.3278	0.4047				
3	0.4050	0.2759	0.4905			
4	0.4453	0.3958	0.3531	0.6305		
5	0.4018	0.2616	0.3059	0.3234	0.4018	
6	0.5149	0.3936	0.3892	0.5451	0.3296	0.6579

Perron-Frobenius eigenvector

0.4517	0.3411	0.3666	0.4536	0.3325	0.4784

(b) $Pr(i|j)$

i \ j	1	2	3	4	5	6
1	1.0000	0.8100	0.8257	0.7062	1.0000	0.7826
2	0.5425	1.0000	0.5625	0.6278	0.6511	0.5983
3	0.6703	0.6817	1.0000	0.5600	0.7713	0.5926
4	0.7370	0.9780	0.7199	1.0000	0.8049	0.8285
5	0.6650	0.6464	0.6236	0.5129	1.0000	0.5010
6	0.8522	0.9725	0.7949	0.8646	0.8203	1.0000

Left dominant singular vector

0.4496	0.3520	0.3740	0.4437	0.3493	0.4639

Right dominant singular vector

0.3962	0.4468	0.3969	0.3803	0.4401	0.3843

Appendixes

Appendix 1 Postulates for the operation *join*

The subject consists of the elements and subsets of a basic set N; elements are denoted by a, b, c, ..., and subsets by A, B, C, ...; the empty set is represented by $\varnothing$; $\{a_1, ..., a_n\}$ denotes a finite set; $\bigcup$ denotes union, $\bigcap$ denotes intersection, $A \setminus B$ is the set of elements in A but not in B. If $A \bigcap B \neq \varnothing$, then A "meets" or "intersects" B; otherwise they are disjoint. $A \subseteq B$ means A is a subset of B. Suppose $A = \{a\}$ and $A \subseteq B$; then $\{a\} \subseteq B$ or, equivalently, $a \in B$, so that it is convenient to write $\{a\}$ as a.

Let A and B be any subsets of N:

DEFINITION A.1.1. *The **join** of set A to set B, denoted either by A.B or by AB, is defined as*

$$A.B = \bigcup_{a \in A, b \in B} (ab).$$

This operation is clearly distinct from union. The postulates for join are

N1 (existence): $ab \neq \varnothing$.
N2 (commutation): $ab = ba$.
N3 (association): $(ab)c = a(bc)$.
N4 (idempotency): $aa = a$ (or better, $aa = \{a\}$).
N5 (identity): $\forall a \exists! u : au = ua = a$.
N6 (no inverse): $\forall a \nexists a' : aa' = a'a = u$.

The lack of an inverse shows that **join** is a commutative monoid, i.e., a commutative semi-group with an identity.

A number of theorems are relevant (the proofs are omitted).

THEOREM A.1.1 (monotonicity). *$A \subseteq B$ implies $AC \subseteq BC$ and $CA \subseteq CB$ for any set C.*

COROLLARY A.1.1. *$A' \subseteq A$, $B' \subseteq B$ imply $A'B' \subseteq AB$.*

THEOREM A.1.2 (existence for sets). *$A \neq \emptyset$ and $B \neq \emptyset$ imply $AB \neq \emptyset$.*

THEOREM A.1.3 (commutation for sets). *$AB = BA$.*

THEOREM A.1.4 (association for sets). *$(AB)C = A(BC)$.*

THEOREM A.1.5 (idempotency for sets). *$AA \supseteq A$.*

Remark A.1.1. Suppose $A \supseteq B$ and $A \supseteq C$ (abbreviated as $A \supseteq B,C$) and let A satisfy the condition $A \supseteq x,y$ implies $A \supseteq xy$, then A is **convex** or **closed** under the operation **join.**

THEOREM A.1.6. *Each of (1) $A \supseteq AA$, and (2) $A = AA$ is equivalent to the convexity of A.*

THEOREM A.1.7. *Let A be convex; then $A \supseteq X,Y$ implies $A \supseteq XY$.*

THEOREM A.1.8. *If two sets are convex, so is their join.*

THEOREM A.1.9. *The intersection of two convex sets is also convex.*

DEFINITION A.1.2. *Let S be a set, and p an element of S. Suppose $x \subseteq S$ implies $px \subseteq S$. Then S is **star-shaped** relative to p, and p is called a **focus** of S. The set of all foci of S is called its **kernel**.*

Remark A.1.2. Each point of $A \bigcap B$ is a focus of $A \bigcup B$.

DEFINITION A.1.3. *Let A be a convex set. An element p is called **interior** if $p \subseteq A$ satisfies for each $x \subseteq A$, there exists $y \subseteq A$ such that $p \subseteq xy$. The interior of A, $\ddot{I}(A)$, is the set of all interior elements of A. The **frontier** of A is $\ddot{Y}(A) = A \setminus \ddot{I}(A)$.*

Much of the above is implicit in the procedures used in this monograph for generating subsets. The reason for joining two or more subsets depends on various criteria, as discussed in several of the chapters.

Appendix 2 A simulated annealing algorithm for solving nonlinear and multi-objective set-covering problems

The following proposal is a stochastic procedure, based on a version of the simulated annealing algorithm suggested as suitable for some combinatorial problems by Lundy (1985) and Aarts and van Laarhoven (1987). This algorithm has two main phases: a cooling (local) phase, which searches for a better solution in the neighborhood of the current solution, and a heating (global) phase, which chooses a new start position for a search. It is assumed that a minimum represents the optimum for each function.

In more detail, let $f_\alpha(1)$ be the value of the α^{th} objective function for each value of the unknowns set to unity:

(1) Initial phase. Define $g_\alpha = f_\alpha(1)$; select ß as a positive parameter for controlling the rate of convergence.

(2) Cooling phase. Given a **current** feasible solution, find a random feasible neighbor, the **candidate**. If all function values of the **candidate** are lower than the **best** so far, replace **best** by **candidate**; if all function values of the **candidate** are lower than the **current** set, replace **current** by **candidate**.

(3) Heating phase. Otherwise, replace **current** by **candidate** with probability

$$\exp[\ -\max_\alpha\ (f_\alpha(\textbf{current}) - f_\alpha(\textbf{candidate}))/g_\alpha\ ];$$

in addition

$$\text{replace } g_\alpha \text{ by } g_\alpha/(1 + ßg_\alpha).$$

(4) Repeat phases 2 and 3 until $\max_\alpha g_\alpha$ is sufficiently small.

If W different minima have been found after K random (heating) replacements, the Bayesian estimate of the number of local minima is the integer I nearest $W(K-1)/(K-W-2)$, so that the search may also be

terminated when $I = W$ (Rinnooy Kan et al. 1985). The essential component of this algorithm, which for a single objective function converges to the global optimum with probability unity (Lundy 1985; Rinnooy Kan et al. 1985; Bohachevsky et al. 1986), is the possibility of escape from a local optimum by the random acceptance of a solution poorer than that currently being considered. I conjecture that efficient (Pareto) solutions are obtained for multi-objective programs with probability unity, including the utopia point, if one exists, although I have not been able to prove it. Zeleny (1982) provided a detailed discussion of multicriteria decision making.

This algorithm is appropriate for multiple linear and nonlinear functions, but it is *not* recommended for a single linear objective function, because it is slower than Chvátal's (1979) heuristic (Chapter II), although faster than the exact procedure described by Garfinkel and Nemhauser (1972). For single nonlinear functions, e.g., fractional set-covering, it may be competitive with other methods.

Appendix 3 Topological spaces for clustering

This appendix summarizes what I consider to be the mandatory requirements of a distance space in a clustering context. This appendix can be omitted if the interest is in clustering rather than in its underlying theory.

The following two conditions define a Fréchet neighborhood space:

(1) There is an abstract set, S.

(2) For each element of S, x, there is a nonempty class $\{V_x\}$ of subsets of S. Each member of $\{V_x\}$ is a neighborhood of x.

Loosely speaking, an element x of S is "near" a subset E of S if every neighborhood V_x of x contains an element of E. If in addition

(3a) each neighborhood V_x of an element x of S contains x,

(3b) if U_x and V_x are neighborhoods of x, there exists W_x such that $W_x \subset U_x \cap V_x$,

(3c) if $x,y \in S$, $x \neq y$, then $\exists V_x$ such that $y \notin V_x$, and

(3d) if $x \in V_y$, then $\exists V_x \in V_y$,

a topological space is defined. If (3c) is replaced by

(3c') if $x,y \in S$, $x \neq y$, then $\exists V_x$ and $\exists V_y$ such that $V_x \cap V_y = \varnothing$,

a Hausdorff topological space is defined. Additional distance conditions specializing the topological space are at the discretion of the user, but each has to be justified.

Appendix 4 The geometry of multifocal ellipsoids

This appendix discusses the geometry of the multifocal regions of Chapter VIII so as to clarify some of their properties.

In clustering, a common starting point is provided by some data that allow objects to be treated as if located in Euclidean space. Usually, either no account is taken of possible changes in this space brought about by the recognition (or decision) that certain objects belong together, or if it is, the changed space is homeomorphic with the original one (i.e., if the original space is Euclidean, then the transformed space is also Euclidean) but perhaps with a changed metric. An example of this changed metric is use of generalized distances: if the simple Euclidean squared distance between the m-element vectors μ_i and μ_j is

$$d^2_{ij} = (\mu_i - \mu_j)^T(\mu_i - \mu_j) = \|\mu_i - \mu_j\|^2, \qquad \text{(A4.1)}$$

the generalized squared distance with respect to the $m \times m$ p.d. matrix, Σ, defining the space (the covariances among the coordinates), is

$$\mathbf{D}^2_{ij} = (\mu_i - \mu_j)\Sigma^{-1}(\mu_i - \mu_j). \qquad \text{(A4.2)}$$

If Σ is p.s.d, a generalized inverse may be used in (A4.2) in place of Σ^{-1}; this complication can be ignored here. Since generalized distances of this sort are also Euclidean, the locus of a point φ such that

$$\varphi^T\Sigma^{-1}\varphi = k,\ k > 0 \qquad \text{(A4.3)}$$

defines a hypersphere around the origin in the transformed space equivalent to a hyperellipse in the original space. $\mathbf{D}^2_{ij}$ is independent of the choice of origin, since its location vanishes in $\mu_i - \mu_j$, and is not a component of Σ, which is assumed to apply at all points in the space.

In geometry, Expression A4.2 is an example of a Minkowski distance function, MDF, which in analysis defines a Banach space. Let k be the boundary of any convex body having a single centre of symmetry, i.e., it bisects all chords passing through it. The MDF between any point and this centre is defined as follows:

DEFINITION A4.1. *On the line through the centre and the point, the MDF is the ratio of the length of the segment between the centre and the boundary of k.*

DEFINITION A4.2. *The MDF between any two points is defined by translating the segment between them (without any rotations) so that one or other of the two ends coincides with the centre of symmetry, and then computing the ratio as above. (The importance of assuming symmetry is apparent for this definition in order for the value to be unique).*

It follows that the original boundary becomes the surface of a unit ball in the transformed space. The boundary of k is known as the indicatrix or absolute of the space (Fig. A4.1). As is easily shown, distances in the transformed space are Euclidean. It is interesting to note that an ellipse, of eccentricity about 0.9 and with major axis along the line of sight (Hagino and Yoshioka 1976), appears to be the indicatrix of human perception of distance.

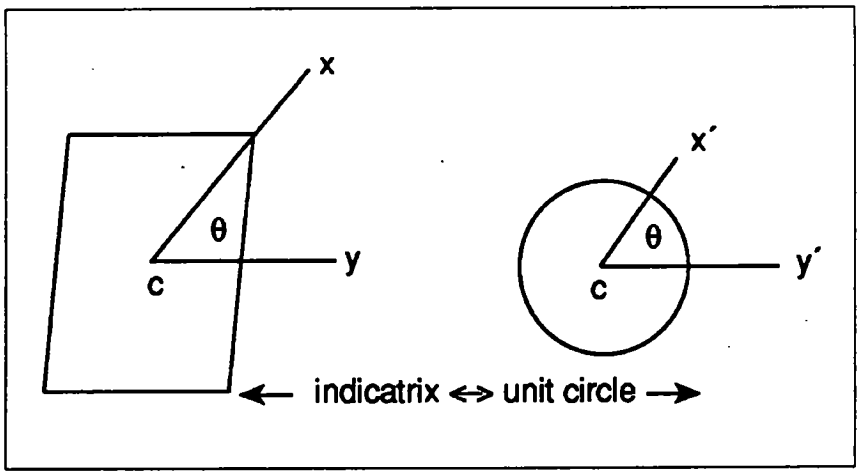


Fig. A4.1 Example of the use of the Minkowski distance function.

Although Euclidean spaces have "nice" computational properties, which make them attractive for many purposes, these properties do not seem to be those of the subjectively perceived world (Battro et al. 1976; Watson 1977). In a gestalt-type experiment (Lefkovitch 1978), the

membership of a candidate object to a set of points on a plane could be explained, among other things, by the subjectively inferred shape properties of the set, suggesting that a mathematical model for assigning membership should recognize shape. Subsequent analysis of the same data has shown that the shape component is consistent with the use of a MDF. It is also apparent that sets remote from mutually proximal nonmembers should not greatly influence the relationship among the latter, so that transformations that change the metric of the space uniformly (including the making of no changes) are inappropriate for clustering, and expressions such as Expression A4.2 are likely to be misleading except in special circumstances. The implicit recognition of this risk has been made so often that it is part of the folklore of numerical clustering.

The key component of the model developed in Chapter VIII is that nonmembers of a set are considered from the viewpoint of the members. From this, two requirements follow:

(1) The origin should be the set and not be discretionary as in Euclidean space, i.e., there are some essential singularities.

(2) The measurement of distance with respect to this set should depend on its internal structure.

It is not a requirement that equal distances in the original space be equal in the transformed space. These assumptions and requirements together lead to a non-Euclidean geometry.

In common with Watson's (1977) model to explain visual illusions, the notion of a "force field," in which the lines of force are parallel prior to any clustering, is a useful concept. Three simple assumptions are made:

> ASSUMPTION A4.1. *The ungrouped objects are located in Euclidean space (this assumption is without loss of generality).*
>
> ASSUMPTION A4.2. *The dimensionality of the space is unchanged by recognizing or defining subsets of the objects (this assumption may not be needed but seems innocuous).*

ASSUMPTION A4.3. *Recognizing or defining subsets of the objects containing more than one member changes the force field within the space both by displacement and by curvature.*

Although in what follows it is the geometry of the space that is considered variable, there is a completely equivalent model in which the geometry remains constant but the lines of force are displaced, as in Assumption A4.3.

It is convenient to consider the relationship among two nonmembers with respect to the set, although there is no distinction between members and nonmembers. This discussion considers the problem as having three components:

- the angle between two objects subtended at the set
- the radial distance between each object and the set
- the distance between any two objects with respect to the set.

The angle subtended by n_a and n_b points at n_x points

Let n_u denote the number of objects in subset u. Consider two points, whose coordinates with respect to any origin in Euclidean m-space are the m-element column vectors $\mathbf{a}$ and $\mathbf{b}$, and a set of $n_x \geq 1$ other points, whose coordinates with respect to the same origin are given by the $n_x \times m$ matrix $\mathbf{X}$. If the n_x points of $\mathbf{X}$ coincide at $\mathbf{x}$, then the angle, $\Theta_{ab}|\mathbf{x}$, between rays joining this position to $\mathbf{a}$ and $\mathbf{b}$ is given by

$$\cos \Theta_{ab}|\mathbf{x} = (\mathbf{a} - \mathbf{x})^{\mathrm{T}}(\mathbf{b} - \mathbf{x})/(\|\mathbf{a} - \mathbf{x}\| \ \|\mathbf{b} - \mathbf{x}\|). \qquad \text{(A4.4)}$$

If the n_x points do not coincide but are very closely grouped in comparison with the distances of $\mathbf{a}$ and $\mathbf{b}$ from them, it can be imagined that some position replaces the n_x points, and there the angle is defined. However, this simplification can be misleading if they are not close; in Fig. A4.2, where $n = 2$, the most reasonable choice for a common position is the centroid of $\mathbf{x}_1$ and $\mathbf{x}_2$, indicated by $\mathbf{c}$, where the angle subtended by $\mathbf{a}$ and $\mathbf{b}$ is 180° for any value of ϕ. This choice is unsatisfactory, because it seems that the various ϕ should play a role.

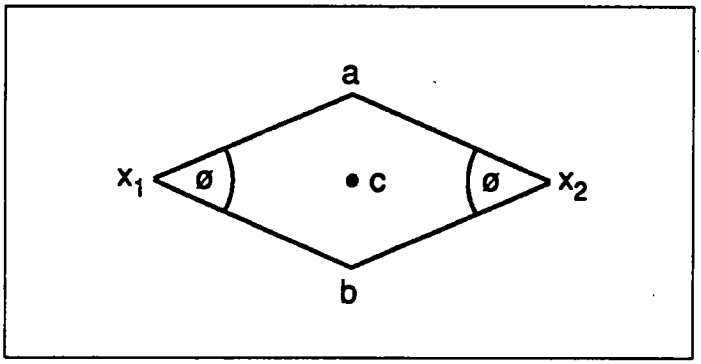


Fig. A4.2 To illustrate the inappropriateness of the centroid as replacement for x_1 and x_2.

The approach adopted here is to define the position of **a** with respect to **X** by treating the n_x points in **X** as a set of origins, and then to define the angle between **a** and **b** as the coefficient of inclination between vector subspaces (Afriat 1957). The location of **a** with respect to **X** is given by

$$\mathbf{L} = \mathbf{X} - \mathbf{1a}^{\mathrm{T}}, \tag{A4.5}$$

where **1** is a n_x-element column vector of unities, and $\mathbf{1a}^{\mathrm{T}}$ is a $n_x \times m$ array. In like manner,

$$\mathbf{M} = \mathbf{X} - \mathbf{1b}^{\mathrm{T}}. \tag{A4.6}$$

Increasing generality, suppose that the single point of **a** is replaced by $n_a \geq 1$ points, whose coordinates are the $n_a \times m$ matrix **A**; then Expression A4.5 can be written as

$$\mathbf{L} = \mathbf{1}_{n_a} \otimes \mathbf{X} - \mathbf{1}_{n_a} \otimes \mathbf{A}, \tag{A4.7}$$

where $\mathbf{1}_k$ represents a column vector of k unities, and $\otimes$ denotes the Kronecker (direct) product, so that **L** is a matrix of $n_a n_b$ rows and m columns. Similarly, if **B** is $n_b \times m$, then Expression A4.6 can be rewritten as

$$\mathbf{M} = \mathbf{1}_{n_b} \otimes \mathbf{X} - \mathbf{1}_{n_b} \otimes \mathbf{B} \tag{A4.8}$$

Geometrical inversion (Coxeter 1969) allows **L** and **M**, which represent **X** with origins **A** and **B**, respectively, to be considered as having a common origin; the coefficient of inclination between the vector subspaces represented by them is

$$\cos \Theta_{LM} = [(\|\mathbf{LM}^T\| \ \|\mathbf{ML}^T\|)/(\|\mathbf{LL}^T\| \ \|\mathbf{MM}^T\|)]^{1/2} \quad \text{(A4.9)}$$

Since the magnitude of angles is unchanged by geometric inversion,

$$\Theta_{AB}|\mathbf{X} = \Theta_{LM}. \quad \text{(A4.10)}$$

This measure of the angle between **A** and **B** sometimes coincides with measures based on $n_x^{-1}\Sigma(\cos \Theta_{ab}|\mathbf{x}_i)$ for $n_a = n_b = 1$, $n_x > 1$ if the $\mathbf{x}_i$ are symmetrically arranged around their centroid; it will be true, for example, in Fig. A4.2, where $\Theta_{AB}|\mathbf{X}$ is the common angle at the $\mathbf{x}_i$.

The distance between A and B with respect to X

Radial distance

In considering the angle between **A** and **B** with respect to **X**, a single origin is not needed and all n_x members of **X** are treated equally. This approach is also followed by considering the distance between a point and a set of points as being based on the average of the distances it has from them, ignoring direction. This criterion is that of average linkage sometimes used in sequential clustering algorithms but differs in an important respect.

The average distance between **a** and **X** can be written as

$$n_x^{-1}\Sigma\|\mathbf{x}_i - \mathbf{a}\| = n_x^{-1}\mathrm{tr}(\{\mathrm{diag}(\mathbf{LL}^T)\}). \quad \text{(A4.11)}$$

This definition is preferred to $n_x^{-1}\|\mathbf{L}\|$, since there is evidence that it represents what people actually do, in contrast to such measures as the nearest neighbor distance and that just cited (Lefkovitch 1978). The family of curves (Expression A4.1), apparently first noted by Maxwell (1846), appears not to have found an application until now.

However, Expression A4.11 does not completely satisfy the requirements noted at the beginning of this appendix, because it does not

recognize that **X** forms a subset; no property of the subset is used. An expression is now obtained that can be used as the indicatrix of a MDF. Consider the locus of a point, **y**, such that

$$n_x^{-1}\Sigma \| \mathbf{x}_i - \mathbf{y} \| = \text{constant}; \qquad \text{(A4.12)}$$

this expression defines a convex body with respect to the **X**, which, except in special circumstances, need not be symmetric in the original space (for $n_x = m = 2$, Expression A4.12 defines an ellipse) and may have singularities, just as is possible in the boundary of the indicatrix in a MDF (Fig. A4.1). For an appropriately chosen constant, Expression A4.12 defines a convex body that can also be used as an indicatrix. This constant can be some measure of the scatter among the members of **X**, such as the average distance, δ, among them, which is

$$\delta = 2[\Sigma_{i=1\ldots j-1}\Sigma_{j=1\ldots n} \| \mathbf{x}_i - \mathbf{x}_j \|]/[n_x(n_x - 1)]. \qquad \text{(A4.13)}$$

The radial distance of **a** from **X** is now defined as

$$r_a = (n_x\delta)^{-1}\Sigma_{i=1\ldots n} \| \mathbf{x}_i - \mathbf{a} \| . \qquad \text{(A4.14)}$$

In the original space, the boundary is given by those values of **y** for which

$$r_y = 1 \qquad \text{(A4.15)}$$

and so defines a unit ball generated by the subset, i.e., the basis indicatrix, in the transformed space. Other measures of scatter may be used; $\| \mathbf{X} \|$ is inappropriate since its value depends on the choice of origin. The only difficulty with Expression A4.14 occurs if $n_x = 1$, which implies $\delta = 0$; but if $n_x = 1$, there is no grouping of individuals in **X** and so the space is unchanged; hence radial distance may be defined as

$$r = \begin{cases} \tfrac{1}{2}(n_x - 1)\, \Sigma \| \mathbf{x}_i - \mathbf{a} \| \; / \; \Sigma_{j>i}\Sigma_i \| \mathbf{x}_i - \mathbf{x}_j \| \;, & n_x > 1 \\ \| \mathbf{x} - \mathbf{a} \|, & n_x = 1. \end{cases} \qquad \text{(A.16)}$$

This transformation is to a space homeomorphic with the original one and so is Euclidean if the original is also. Since the distances r_a and r_b and the angle $\Theta_{ab}|\mathbf{X}$ are known, the Euclidean distance between a and b is easily obtained.

If the single point, a, is replaced by $n_a > 1$ points, Expression A4.11 is replaced by

$$(n_a n_x)^{-1}\Sigma_p\Sigma_i \|\mathbf{x}_i - \mathbf{a}_p\|, \qquad \text{(A4.11a)}$$

the first option in Expression A4.16 becomes

$$r_a = \tfrac{1}{2}n_a^{-1}(n_x - 1)\Sigma_p\Sigma_i \|\mathbf{x}_i - \mathbf{a}_p\| \;/\; \Sigma_{j>i}\Sigma_i \|\mathbf{x}_i - \mathbf{x}_j\| \qquad \text{(A4.16a)}$$

while the condition for the second option becomes $n_x = n_a = 1$.

Nonradial distance

A Euclidean distance, however, does not reflect the requirement that the presence of a set introduces curvature, and so nonradial distances are not linear. Because symmetric convex bodies can be transformed into hyperspheres by the MDF and an analogue for this transformation has just been proposed for discrete sets of points, the problem is reduced to the consideration of concentric hyperspheres. The model now postulated is that a spherical object induces spherical contours in the space, that two points on the same contour are separated by $r\Theta$ (where r is the radius to that contour and Θ is the smallest angle they subtend at the centre), and that the centre is an essential singularity in the space. The problem is to define the distance between two points on different contours. The method now used is conformal and is related to geometric inversion and to the Schwarz–Christoffel transformation; in two dimensions, the interior of the indicatrix becomes one side of a plane, the exterior the other side, and the boundary is the line of infinite length between them.

Let the two points, a and b, be at distances r_a and r_b from the centre, and let the angle they there subtend be Θ (Fig. A4.3); then the conformal transformation

$$f(r,\Theta) = \log_e(r\exp(i\Theta)) = \log_e r + i\Theta \qquad \text{(A4.17)}$$

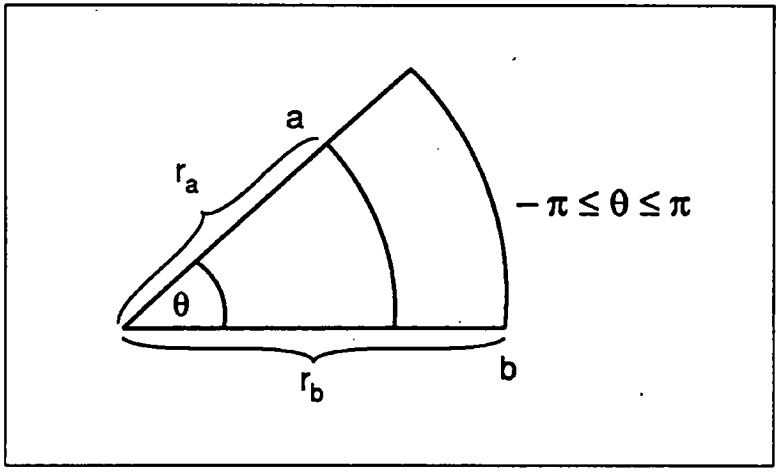


Fig. A4.3 The location of points a and b in relation to the centre of concentric circles having radius of r_a and r_b.

converts Fig. A4.3 to Fig. A4.4. The length of the diagonal from **a** to **b**, now defined as the distance between **a** and **b** with respect to the centre, can be found as follows. Any line parallel to $\log_e r_b$, e.g., the broken horizontal line in Fig. A4.4, can be represented by

$$\phi = p \log_e r + q, \tag{A4.18}$$

where p is the slope of the line connecting **a** to **b** and q is the point on the line corresponding with p and $\log_e r$. Thus

$$d\phi/dr = \phi' = p/r. \tag{A4.19}$$

Let ℓ denote the length of the diagonal from **a** to **b** in Fig. A4.4. To calculate ℓ, three cases need to be considered.

(1) If $r_a = r_b = r$, then

$$\ell = \int_0^{\infty} \{ (dr/d\phi)^2 + r^2 \}^{1/2} \, d\Theta, \tag{A4.20}$$

but since $dr/d\phi = 0$, it follows that

$$\ell = r\Theta, \tag{A4.21}$$

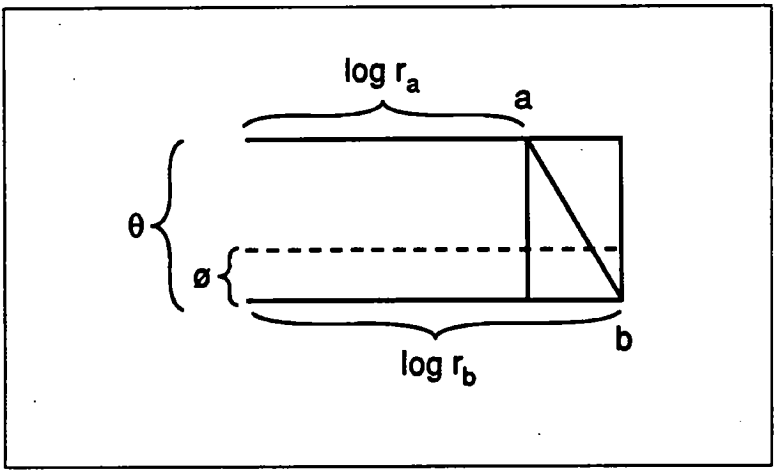


Fig. A4.4 Conformal transformation of Fig. A4.3.

which corresponds to the objects on the same contour.

(2) If $r_a \neq r_b \neq 0$, then

$$\ell = \int_{r_b}^{r_a} (1 + \phi'^2 r^2)^{1/2}\, dr$$

$$= \int (1 + p^2)\, dr = (r_b - r_a)\,(1 + p^2)^{1/2}$$

$$= (r_b - r_a)\,(1 + [\Theta/\log_e(r_a/r_b)]^2)^{1/2}. \qquad \text{(A4.22)}$$

As r_b becomes large relative to r_a, ϕ approaches zero, so that the distances among points one of which is much further from the centre than the other are virtually the same as their Euclidean distance; if a and b are proximal but remote from the centre, $\Theta \to 0$, so their distance is again approximately Euclidean.

(3) If $r_a = 0$ and $r_b \neq 0$,

then, since a coincides with the origin, Θ is indeterminate; but since $\log_e(r_a/r_b)$ will be equal to $-\infty$, $\phi = 0$ and hence

$$\ell = r_b. \qquad \text{(A4.23)}$$

By combining Expressions A4.16 (or A4.16a) with A4.21–A4.23, the distance between **a** and **b** with respect to **X**, denoted by $\Delta_{ab}|\mathbf{X}$, is

$$\Delta^2_{ab}|\mathbf{X} = \begin{cases} (r_a - r_b)^2(1 + \{\Theta_{LM}/\log_e(r_a/r_b)\}^2), & r_a \neq r_b \neq 0 \\ r^2\Theta^2_{LM}, & r_a = r_b \neq 0 \\ r^2_b, & r_a = 0, r_b \neq 0 \\ 0, & r_a = r_b = 0. \end{cases} \quad \text{(A4.24)}$$

Distance as defined in Expression A4.24 is a semi-metric, since

$$\Delta_{ab}|\mathbf{X} \geq 0$$

with equality if $\mathbf{a} = \mathbf{b}$, and

$$\Delta_{ab}|\mathbf{X} = \Delta_{ba}|\mathbf{X}. \quad \text{(A4.25)}$$

It is not a metric since the triangle inequality need not be satisfied in the vicinity of the centre; conceptually, this departure is not a problem in clustering, since it is usually radial distances (Expression A4.14) that are used, nor is it a problem computationally. Note that

$$\Delta_{ab}|\mathbf{X} \neq \Delta_{ab}|\mathbf{Y} \quad \text{(A4.26)}$$

unless $\mathbf{X} = \mathbf{Y}$, since these refer to different sets of essential singularities.

According to Weiman and Chaikin (1979), human intuitive assessment of distance often appears to measure it along the path of a spiral. They cited evidence for this based on the exponential arrangement of receptor cells in the eye, in fields of ganglion cells, and on experimental observations. Expression A4.24 is consistent with a spiral distance. Assume that the curve from *a* to *b* forms part of an equiangular spiral, with magnification

$$\mu = r_b/r_a, \; r_b \geq r_a \quad \text{(A4.27)}$$

for an angle Θ; the spiral angle, ϕ, which is defined by

$\cot \phi = (\log_e \mu)/\Theta$, gives the squared length of the curve from a to b as

$$(r_b - r_a)^2 \sec^2 \phi = (r_b - r_a)^2 (1 + \tan^2 \phi), \qquad \text{(A4.28)}$$

which is easily shown to be identical with Expression A4.24. That A4.24 is the length of a segment of a spiral also follows from the properties of the conformal transformation in Expression A4.17.

Characterizing the whole space

Suppose that N objects have been assigned to s subsets, with $n_j \geq 1$ $(j = 1 \ldots s)$ objects in each, $\Sigma n_j \geq N$. The force field of the original space prior to any assignments is changed by the presence of these subsets in ways that follow from the previous descriptions. This section sketches something of these changes.

Assume $s = m = 2$; denote the points corresponding to one set by $\mathbf{X}_1$, with coordinates for each point denoted by $\mathbf{x}_{1i}$, and similarly for the second subset. Then the radial distance of any point $\mathbf{y}$ from $\mathbf{X}$ is

$$r_y|\mathbf{X}_1 = (n_1 \delta_1)^{-1} \Sigma_{i=1 \ldots n_1} \| \mathbf{x}_{1i} - \mathbf{y} \| \qquad \text{(A4.29)}$$

and from $\mathbf{X}_2$ is

$$r_y|\mathbf{X}_2 = (n_2 \delta_2)^{-1} \Sigma_{i=1 \ldots n_2} \| \mathbf{x}_{2i} - \mathbf{y} \| . \qquad \text{(A4.30)}$$

If $r_y|\mathbf{X}_1 = r_y|\mathbf{X}_2$, $\mathbf{y}$ is equidistant from the two in their own semimetrics; thus the locus of all points, $\mathbf{y}$, for which

$$r_y|\mathbf{X}_1 = r_y|\mathbf{X}_2, \qquad \text{(A4.31)}$$

defines the *neutral line* between the two subsets. Although in the transformed space, the neutral line can be considered as straight, it need not be in the original space. Clearly, other lines may be of interest, such as those determined by the locus of all points for which

$$r_y|\mathbf{X}_1 \,/\, r_y|\mathbf{X}_2 = k, \text{ a constant.} \qquad \text{(A4.32)}$$

Expression A4.32 in a Euclidean space is the circle of Apollonius (which, when $k = 1$, is the straight line of Expression A4.31) and so corresponds with a closed contour in the original space.

The pairwise neutral lines between subsets may be used to investigate the accepted ensemble of subsets, because if the line for any pair intersects the indicatrices, they must be good candidates for fusion.

If $s = 3$ and $m = 2$, the three pairwise neutral lines may either intersect at what can be called the neutral point, or may not intersect at all. The neutral line between two subsets divides the plane into two open halves; if $s > 2$, open convex polygons, some bounded and others perhaps extending to infinity, are formed by their intersection. Thus the neutral lines, in the transformed space, divide the plane into polygons, forming a Dirichlet tessellation. Each polygon represents a region in which points are closer to one subset than they are to any other. This partition of the space permits the assignment of newly found objects to existing subsets. Green and Sibson (1978) described an efficient algorithm for finding the planar tessellation given the coordinates of points; for higher dimensional spaces, in which the space is divided into open polyhedra, an efficient algorithm has not yet been described.

In terms of the original space, the sides of the polygons of the tessellation need no longer be straight, and so the partition of the plane is into regions which are not necessarily convex, and the Delaunay triangulation dual to the tessellation appears as triangles whose sides are not straight.

The space of the set of subsets can be described in the same way as that for a single subset; the distance of y from all the subsets is defined as in Expression A4.30 as

$$R_y = s^{-1} \Sigma_{j=1\ldots s} r_y | \mathbf{X}_j$$

$$= s^{-1} \Sigma_{j=1\ldots s} (\delta_j n_j)^{-1} \Sigma_{i=1\ldots n(j)} \| \mathbf{x}_{ji} - \mathbf{y} \| . \qquad \text{(A4.33)}$$

The angle between two points with respect to the subsets is obtained as in Expression A4.16, taking care to distinguish between the angle subtended at a single set of all objects from that subtended at a set of subsets, because the relationship among the latter must also play a role. Were each subset to have been replaced by a single point, the angle

between any two points subtended there could be obtained as in Expression A4.16 without further ado, but in the absence of a clear choice for its location, it is necessary to define the location of each member of a set with reference to the subsets of which it is *not* a member, and then express **a** and **b** with reference to this new set of origins using Expression A4.9. Combining the angle obtained, Θ, with Expressions A4.33 and A4.24, the squared distance between **a** and **b** with reference to all the subsets is

$$\Delta^2_{ab}|\mathbf{X}_j = (R_a - R_b)^2 \left(1 + \{\Theta/\log_e(R_a/R_b)\}^2\right), (j = 1 \ldots s). \tag{A4.34}$$

A measure of change

The effect of forming subsets, in the present model, is to transform the pairwise distances, Expression A4.1, into those of Expression A4.34; since a semi-metric or metric space consists of an underlying set, here the N original points, and a measure of distance (Expression A4.34), an interesting question is to find whether the underlying set and the new distances can be represented in Euclidean space. Gower (1966) showed that, if each distance is transformed as

$$-\Delta^2_{ab}/2, \tag{A4.35}$$

followed by a double centroid transformation of the matrix of these values, the resulting symmetric $N \times N$ matrix, $\mathbf{T}$, has non-negative eigenvalues iff a coordinate system exists that reproduces the original distances when treated in Euclidean space. The number of negative eigenvalues, and their sizes in relationship to the positive ones, indicate the departure from a Euclidean representation (Chapter VII). Let an eigenvalue decomposition of $\mathbf{T}$ be

$$\mathbf{T} = \mathbf{U}\Lambda\mathbf{U}^T; \tag{A4.36}$$

some of the diagonal elements of Λ may be positive, some may be negative, and others may be zero. Because of the double centroid transformation, the sum of the columns of $\mathbf{U}$ will be zero. A set of

$N - 1$ principal coordinates for the N objects can be written as

$$\mathbf{Z} = \mathbf{U}\Lambda^{1/2}, \tag{A4.37}$$

where some columns contain only real elements (the real subspace), some others imaginary (the complex subspace), and others zero (the null subspace). If $\mathbf{W}$ is a set of principal coordinates corresponding to the original distances (Expression A4.1), the effect of forming the subsets can be measured by generalizing the distance measure of Gower (1971*b*), which becomes

$$\omega^2 = \text{tr}(\mathbf{W} - \mathbf{ZK})(\mathbf{W} - \mathbf{ZK})^*, \tag{A4.38}$$

where $(.)^*$ denotes transposition taking complex conjugates, and $\mathbf{K}$ is unitary, chosen to minimize ω^2. If $\mathbf{V} = \mathbf{W}^*\mathbf{Z}$, then $\mathbf{K}$ is obtained from a singular decomposition of this complex matrix as

$$\mathbf{V} = \mathbf{PSQ}^* \tag{A4.39}$$

from which

$$\mathbf{K} = \mathbf{QP}^*. \tag{A4.40}$$

Expanding Expression A4.38 gives

$$\omega^2 = \text{tr } \mathbf{WW}^* + \text{tr } \mathbf{ZZ}^* - \text{tr } \mathbf{ZKW}^* - \text{tr } \mathbf{WK}^*\mathbf{Z}^*, \tag{A4.41}$$

which is real and non-negative. Interpreting ω^2, it measures the change in position brought about by the assignment of the objects to subsets; if $\|\mathbf{W}\| = \|\mathbf{Z}\| = 1$ by a prior normalization (using the Frobenius quadratic norm) then tr $\mathbf{WK}^*\mathbf{Z}^*$ is the coefficient of inclination between the vector subspaces. The consequences of ignoring the negative eigenvalues of $\mathbf{T}$, e.g., by setting them to zero and thereby increasing the dimensionality of the null space, can be evaluated similarly.

Although identifying a single point to represent a subset in the original space is not required in the above model, it is of interest to consider if one is implied. With respect to the definition of cos $\Theta_{ab}|\mathbf{X}$,

there appears to be no such point except in special circumstances. In Fig. A4.2, $\Theta_{ab}|\mathbf{X}$ is the common angle ϕ, and hence the two positions for the centre coincide with each of the two objects. If the space is three-dimensional, there is an infinity of positions for the centre of a circle.

However, there is such a point if distance is defined as the average of the distances to the members of the set. Consider Expression A4.12, now repeated

$$n^{-1}\Sigma \| \mathbf{x}_i - \mathbf{y} \| = \text{constant}, \qquad \text{(A4.12)}$$

and allow the constant to approach zero. The real position(s) of y for which this constant is a minimum can be regarded as the centre; this position, known variously as the mediancentre or the generalized Steiner point, GSP, is unique (unless $n = 2$, where all points on the line joining $\mathbf{x}_1$ to $\mathbf{x}_2$ minimize Expression A4.12; conventionally, the midpoint is chosen). However the GSP does not replace the essential singularities of the set and is not located at the position at which the angles are measured.

The adaptative distance just described clearly arises from the conditional clustering algorithm(s) described elsewhere in this book. The use of the proposed radial distance is also a natural candidate for use in sequential agglomerative hierarchical procedures, which is the family of directed tree-forming algorithms:

step 1: find the closest pair of subsets, and join them.

step 2: find the distance between the newly formed subset and the others.

step 3: repeat steps 1 and 2 until some terminating condition is attained (e.g., a single tree is formed from the objects).

In step 2, the use of the radial distance (Expression A4.16a) seems indicated, but as soon as a subset contains more than one object, because of Expression A4.26 an asymmetric matrix of distances is obtained; because of this asymmetry, some single-valued function, e.g.,

$$f(AB) = f(r_A|B, r_A|A),\ f \in \{\text{min,max,sum,etc.}\}$$

needs to be chosen to represent the relationship. More appropriate is a single-valued measure of the distance between A and B conditional on the current assignment of the subsets, such as has been described above.

The MDF has been defined here somewhat more restrictively than it need be. Let K be the boundary of a star body, i.e., a closed body for which there is at least one point, ϕ, such that the straight line to any other point in K is contained entirely within K. Then the MDF of any point x with respect to ϕ is the ratio of the distance

$$d(x,\phi)/d(x_0,\phi),$$

where x_0 is the point that the line (x,ϕ) crosses K. If K is also convex, and has a centre of symmetry, ω, i.e., any chord through ω is bisected by ω, ω is called the centre.

These remarks on distances have wider application than in clustering. When the relationship between two entities is to be measured, the requirements almost always state or imply the metric axioms, which define distance as a function, f, having two arguments:

AXIOM A4.1: $f(i,j) \geq 0$;

AXIOM A4.2: $f(i,j) = 0 \Leftrightarrow i = j$,

where $i = j$ means that either i and j are the same object, or that they are identical with respect to the function.

AXIOM A4.3: $f(i,j) = f(j,i)$;

AXIOM A4.4: $f(i,k) \leq f(i,j) + f(j,k)$.

These axioms accord with our elementary notions of Euclidean geometry. Thus if experimental observations are made for which Axiom A4.3 is false, e.g., two measurements are made, d_{ij}^* and d_{ji}^*, where the order of the subscripts is important (e.g., the antibody–antigen relationship, the offspring of males of one population and the females of another, in

contrast with the reciprocal cross), adjustments such as defining

$$d'_{ij} = (d^*_{ij} + d^*_{ji})/2,$$

or choosing the smaller (or larger) to be "the" distance, are made to satisfy Axiom A4.3, with a sense of achievement. But what has been achieved is the discarding of information; in fact, their consequences may be more serious than is usually realized. In contrast, this appendix considers the distance between two objects as being a single-valued function of two ordered arguments, which can be expressed as

the distance between i and j with reference to i,

which is not necessarily equal to

the distance between i and j with reference to j.

Although Axiom A4.4 is supposedly independent of Axiom A4.3, the deletion of Axiom A4.3 from the set of axioms makes Axiom A4.4 somewhat artificial. However, Axiom A4.3 in this appendix has been replaced by the requirement that if the distance between i and j is considered with reference to k, it should be the same as that between j and i from the same point of view. The axioms can now be rewritten with an indication of the reference object:

AXIOM A4.1′: $f_k(i,j) \geq 0, \forall\, k;$

AXIOM A4.2′: $f_k(i,j) = 0 \leftrightarrow i = j, \forall\, k;$

AXIOM A4.3′: $f_k(i,j) = f_k(j,i), k \neq i,j;$

and Axiom A4.4 in any form is not a requirement. This set of axioms defines a **conditional semi-metric**; they are the assumptions of the measure described in this appendix.

For further discussion of metrics for convex bodies, see Shephard and Webster (1965). A set of toy examples of the calculations is given in Chapter X, "Angles and distances."

Appendix 5 Further comments on subset generation

Consider a set N of $n = |N|$ objects containing an unknown number of as yet undescribed "true" populations of interest. Let $P(N)$ denote the power set, $m = |P(N)|$, $\mathbf{a}_k$ a 0,1 n-element column vector denoting the k^{th} subset, $\mathbf{a}_k \in P(N)$, where $a_{ik} = 1$ if object i is a member of the k^{th} subset and is zero otherwise.

The process of recognizing which of the $P(N)$ are the "true" populations is sequential; at each stage, some subsets are selected, and others discarded. At stage t of the selection process, the choice is random and according to a probability

$$\mathbf{p}(t), \text{ where } \Sigma_{k=1\ldots m} p_k(t) = 1. \tag{A5.1}$$

At the final stage, for there is one, although each of the remaining subsets need not correspond with one of the unknown populations, they may do so. Note that, based on the chosen set of descriptors, some members of one true population can be closer to the centroid of another than they are to that of their own population. Certainly, it is hoped that distinct populations are represented by disjoint sets of subsets.

At stage t, let $\mathbf{q}(t)$ be the probability of success in identifying a "true" population, and $\mathbf{1} - \mathbf{q}(t)$ the probability of failure. The distribution of $\mathbf{q}(t)$ is not known to the decision maker, but the objective is to maximize the probability of success in identifying the true populations. Since there is no knowledge about $\mathbf{q}$, consider a learning algorithm of the form

$$\mathbf{p}(t + 1) = \mathbf{T}(\mathbf{p}(t)), \tag{A5.2}$$

where, omitting the empty subset, $p_k(0) = 1/(m - 1)$, $\forall \mathbf{a}_k \in P(N)$, and $\mathbf{T}$ is an operator to be discussed. If $q(t + 1)/q(t) > 1$, the decision is called a "reward," while if this ratio is less than unity, it is called a "penalty." A ratio of unity is neutral. Expression A5.2 is a nonlinear reward-penalty algorithm, in which much (all?) depends on the choice of $\mathbf{T}$; here, it represents the subset-generating procedure described in Chapter VIII. For subset k at stage t, denoted by $b(k,t)$, algorithms based on C-neighbors transform it into another by a process, which can be

represented as

$$\mathbf{b}(k, t + 1) = \phi(\mathbf{b}(k,t)), \tag{A5.3}$$

where $\mathbf{b}(k,0) = \mathbf{a}_k$, and $\mathbf{b}(k, t + 1)$ is another member of $P(N)$. Let $p(k, t_{(t+1)})$ denote the probability of subset $\mathbf{b}(k,t)$ at stage $t + 1$; under a number of mild regularity conditions, which will be exposed in the process of describing ϕ, the following conclusions arise:

(1) If $\mathbf{b}(k, t + 1) \neq \mathbf{b}(k, t)$, then $p(k, t_{(t+1)}) = 0$. (A5.4)

It follows that if the subset $\phi(\mathbf{b}(k,t))$ at stage t has zero probability, then so should $\mathbf{b}(k,t)$.

(2) There is a stage, t', such that $\mathbf{b}(k, t' + 1) = \mathbf{b}(k,t')$. (A5.5)

In other words, for any initial vector, $\mathbf{b}(k,0)$, the process represented by ϕ has an **absorbing state**, α_t, which is also a member of $P(N)$. Since ϕ is not (usually, or even ever) a linear operator and is not even continuous, the process corresponds with identifying the **fixed points** (the absorbing states) in a discrete topological space. Different initial vectors may give rise to the same absorbing state; just one absorbing state for all initial vectors indicates that there is just one "true" population in the N (the hope, of course, is that there is more than one, but not too many).

Although this is not a computational protocol, at stage t, three operations are performed:

- each subset for which $p_k(t) > 0$ is operated on by ϕ
- those subsets that are transformed by this operation are assigned a probability of zero
- the vector of probabilities $\mathbf{p}(t + 1)$ is estimated.

The estimation procedure, described below, is based on the minimum cross-entropy principle. There is a stage, tc, for which

$$\mathbf{p}(tc + 1) = \mathbf{p}(tc), \tag{A5.6}$$

which is advantageous if most $p_k(tc) = 0$. It is also necessary that the chosen subsets form a covering of the objects, i.e.,

$$N = \{\bigcup \mathbf{a}_k | p_k(tc) \neq 0\}. \tag{A5.7}$$

Using the definition of ϕ given below, which implies that the clustering space is a generalized metric, with associated space given by the dissimilarities, it is easy to show that $tc \leq n$.

The distance between any subset $\mathbf{a}_k$ and any absorbing state α_i can be measured as the minimum number of ϕ steps needed to convert the first to the second, i.e.,

$$\sigma(k,t) = \min (\infty, s | \phi^s(\mathbf{a}_k) = \mathbf{a}_t, s \text{ a minimum}). \tag{A5.8}$$

The operation of ϕ on a particular subset is independent of its operation on any other, although the choice of subsets on which to operate is not. As noted above, the choice is based on the vector $\mathbf{p}$, which, other than stating an initial condition and also that certain of the elements become zero, has not been defined. Consider the situation at state $t > 0$, and assemble the matrix $\mathbf{A}(t)$, which consists of those m' subsets for which $p_k(t) > 0$. Dropping the suffix t to simplify the notation,

- $\mathbf{A}$ is a $n \times m'$ incidence matrix of a subset system as in Chapter II
- a covering is indicated by any binary vector $\mathbf{x}$ satisfying $\mathbf{Ax} \geq \mathbf{1}$
- those $\mathbf{x}$ for which $\mathbf{Ax} = \mathbf{1}$ indicate partitions.

If a particular object belongs to precisely one subset, this subset must be part of every covering, and the corresponding element in $\mathbf{x}$ can be set to unity. Referring to the reductions described in Chapter II and considering all members of $P(N)$, these remarks can be expressed as

$$p_k = \Pr(x_k = 1) = \begin{cases} 0 \text{ if subset } k \text{ has been eliminated} \\ \quad \text{by } \phi \text{ or is emptied by the reductions} \\ \\ 1 \text{ if subset } k \text{ is part of every covering.} \end{cases} \quad \text{(A5.9)}$$

Although those subsets for which $p_k = 1$ may form a covering, this is unlikely. As described in Chapter II and in Lefkovitch (1982), each remaining subset has a probability p_k of participating in a covering, namely,

$$0 \leq p'_k < 1. \quad \text{(A5.10)}$$

The p_k are to be considered more as logical probabilities than as frequencies; these probabilities, including those whose prior value is unity, can be renormalized to standardize the total to unity. An optimal covering can now be regarded as the conjunction of individual hypotheses that a subset participates in an optimal solution, and so the desired solution is one for which the joint probability is a maximum. This representation allows $q(t)$, the probability of success in identifying a "true" population, to be $p_k(t) = 1$, defined as $[1 - \Pi_k(1 - p_k(t))]$. Determining the remaining elements of $p_k(t)$ has been described in Chapter II.

Although estimating the nonzero $p_k(t)$ for each t is possible, the choice of subsets is essentially based on $p_k(t) > 0$, so that it need be performed only once, namely, at stage tc; as a result, only $q(tc)$ is obtained (other than $q(0)$). The comparison between these two is of lesser interest than that between the $q(tc)$ for different ϕ. The information gain and other statistics are subject to similar comment.

If $\mathbf{T}$, namely, the process by which $\mathbf{p}(t)$ becomes $\mathbf{p}(t + 1)$, is to represent a practical procedure, it must eliminate many of the subsets from consideration very rapidly. In fact, almost all can be eliminated *a priori* even though which these are is not known until stage tc. This elimination depends on the regularity conditions for conditional clustering and is expressed in the following theorems, the first of which concerns fixed points in generalized metric spaces; the second, which depends on the first, is fundamental to the whole method.

A fixed-point theorem given by Collatz (1966, p. 208–210), recast for the present purposes, can be expressed (with comments) as follows:

THEOREM A5.1. *If in an operator equation* $\phi(u) = u$,

(1) the domain, B, of the operator ϕ *lies in a complete generalized metric space R, (here formed by the subset space P(N) with associated partially ordered space D, the dissimilarity space); and associated with* ϕ *is another operator G on D, which is positive and continuous but not necessarily linear, with an element* $z \in R$ *such that for any two elements* $v, w, \in B$

$$d(T(v), T(w)) < G(d_{vw} + d_{vz}) - G(d_{vz})$$

(here G is the operation of defining set membership; z is the null element or a fixed point, as long as there is at least one element for which this condition is true for v); and

(2) the distances d, d', s, s' in D where

$$\Theta_D \leq d \leq d', \text{ and } \Theta_D \leq s \leq s'$$

satisfy

$$\Theta_D \leq G(d + s) - G(d) \leq G(d' + s') - G(d')$$

(i.e., for $d = d' = \Theta_D$ *it follows that* $\Theta_D \leq G(s) \leq G(s')$ *for* $\Theta_D \leq s \leq s'$*); and*

(3) there is another iteration S

$$s_{t+1} = Ss_t, \ (t = 0,1\ldots)$$

given by the operator S and distances $s_t \in D$ *and these distances satisfy*

$$s_0 \geq d(u_0, z)$$

$$s_1 \geq s_0 + d(u_0, u_1);$$

(i.e., S compares the ratios of distances); and

(4) there exists a fixed element $\chi \in D$ *such that*

$$S_\tau = G_\tau + \chi \text{ for } \tau \in D; \text{ and}$$

(5) the sequence s_t *converges to a limit element* s *(here* $s = 1$*); and*

(6) the sphere K of all elements that satisfy

$$d(v, u) \leq s - s_1$$

belongs to the domain B, where s is a limit element;

then at least one solution exists to the equation

$$\phi(u) = u,$$

and the sequence $u_{t+1} = \phi(u_t)$ *converges to the solution. All* u_t *and u lie in the sphere K, and the error estimate*

$$d(u, u_t) \leq s - s_t$$

holds $(t = 0, 1\ldots)$.

Proof. See Collatz (1966); note that Collatz uses "pseudometric space" for what is usually called a generalized metric space.

The next theorem considers more than one fixed point and shows that the operator ϕ finds them all with probability approaching 1.

THEOREM A5.2. Assuming ϕ satisfies the regularity conditions for conditional clustering, all absorbing states are obtained by the repeated action of ϕ on each of the $\binom{n}{2}$ two-object subsets, on each single-object subset and on N.

Proof. There are several parts: parts (a), (b), and (c) consider the simple cases; part (d), the important case, itself consists of two parts; and part (e) completes the proof.

(a) The action of ϕ on the null set yields the null set.

(b) The action of ϕ on a single-object subset either leaves it unchanged, or changes it to another. If the latter, then this subset is either N, considered in (c), or a proper subset of N, considered in (d).

(c) The action of ϕ on N leaves it unchanged.

(d) Consider any subset of two (distinct) objects; ϕ either leaves it unchanged, in which case it is an absorbing state, or changes it. If it is changed, then it is used at the next stage. Thus some subsets used at the first and subsequent stages of the modified process are present at stage 0 for the complete process. Thus the members of P(N) can be divided into two subsets, namely, those explicitly used at some stage, and those that are not. Of the latter, some are considered implicitly, and perhaps some are not. Each is considered in turn.

(1) Implicitly considered subsets. If a subset is changed by ϕ, objects are either included or excluded, or both.

(i) Consider the inclusion of objects. If just one object is included, the changed subset is used explicitly at the next stage; if more than one object is included, first consider two such objects; the challenge is to determine if the new subset excluding one or other of the newly included objects when transformed by ϕ yields the same subsets, and, in particular, the same absorbing state as the complete new subset.

This situation is guaranteed by the first regularity condition, which implies that each subset at stage $t > 0$ contains all objects within a convex region of the dissimilarity component of the domain of ϕ. By induction, it is also true for three newly included objects, and so on.

(ii) Consider the exclusion of objects. If the cases that yield the null subset and subsets of cardinality 1 or 2 are eliminated (note that a cycle through pairs is excluded by the convexity consequence of the regularity conditions), then arguments analogous to those for inclusions, but relying on both regularity conditions, are easily made. No additional comments are required if there are both exclusions and inclusions.

(2) Subsets neither explicitly nor implicitly considered are of interest iff they are associated with an absorbing state different from the others. The assumption that such absorbing states exist results in a contradiction. Suppose $\mathbf{a}_j$ is such a subset and $\alpha_{t'} \neq N$ is the corresponding absorbing state (note that $|\mathbf{a}_j| > 2$, which in this case is assumed not to be generated by the explicitly or implicitly used subsets and is therefore not found by ϕ. Thus the members of $\mathbf{a}_j$ belong to more than one absorbing state (other than $\alpha_{t'}$ or N); if they had belonged to the same $\alpha_{t'}$ they would have generated it. Consider two members of $\mathbf{a}_j$ belonging to different absorbing states (other than $\alpha_{t'}$ or N); this pair was explicitly used, and therefore generated an absorbing state. Since this absorbing state must be different from $\alpha_{t'}$, which is obtained only by subsets neither explicitly nor implicitly used, it could be only $\varnothing$ (the empty set) or N. In either case, it follows that $\alpha_{t'}$ has been obtained, i.e., a contradiction to the assumption. Thus there can be no absorbing states different from those generated

by the pairs, by the null set, by the single objects, or by N.

(e) Since all distinct pairs are used, the whole dissimilarity space is straddled in the vicinity of the objects, and thus the probability of identifying all absorbing states approaches unity. Q.E.D.

Excluding N and the single-object absorbing states, those remaining maynot form a covering of the objects. The absent objects must therefore be unlike any of the others to such an extent that their participation in a multi-object subset implies that the corresponding absorbing state is N. Thus, in the context of the objective of the clustering, each of these represents a different "true" population, and, other than noting this fact, they need no longer be considered. It is assumed, therefore, that each absorbing state contains at least two objects.

The operator ϕ, which is fundamental to the whole process, implies that the absorbing states are contained within convex regions of the dissimilarity space; even though these regions may differ in size, shape, orientation, and location, there is no reason to assume that a "true" population coincides with one and only one of the absorbing states. (In fact, for subsets containing *many* objects, the asphericity of their convex hull is almost unity in the dissimilarity space.) If a true population is sufficiently isolated from all others, then a one-to-one relationship may hold; but any lack of isolation needs further study or additional empirical data. Of more immediate concern is the possibility that a single true population cannot be contained within one absorbing state without containing others.

Conditional on the initial membership and subsequent history, the absorbing states ϕ obtains are isolated from others insofar as possible. Excluding N from consideration, if two or more absorbing states are nondisjoint, then their union, called a **muster** (Appendix 6), is a first approximation to one true population (Lefkovitch 1982). Furthermore, if none of the participating absorbing states in a muster is equal to the muster, the containing regions in the dissimilarity space is not part of the ϕ family but may be defined as the union of the separate containing regions. Thus the convexity assumption, which is so important for ϕ, is removed for subsequent processing. A weaker muster, based upon the

objects in the union of nondisjoint containing regions, is also possible.

Because a single muster thus formed can include two or more true populations each of which is a proper subset of it, there is advantage in postponing muster formation until after a further selection ofabsorbing states is made using an optimal covering procedure (Chapter II). If the chosen subsets form a partition, each can be regarded as a second approximation to a true population; if they do not, the musters formed from them serve this role. Whatever the case, the clustering at this stage is completed, and the conjectured populations may be studied, compared, joined, further subdivided, and so on, preferably with independently obtained data.

While there are many ways to perform these last processes, the same general algorithm can be used to form musters. Let $\hat{A}$ denote the subsets selected by an optimal covering from which musters are to be selected. Each member of $\hat{A}$ can be considered as an object, and so the musters are members of $P(\hat{A})$. Given a measure of dissimilarity between subsets (Chapter VII), the musters are precisely the absorbing states corresponding to the operation of ϕ on each of the pairs of objects belonging to $\hat{A}$. If the sole absorbing state is their union, then it can be assumed that the musters are identical with the subsets forming $\hat{A}$, and the process stops; if not, the subsets in the optimal second phase covering can be used for a third phase, and repeated until the sole absorbing state is their union.

Defining ϕ for conditional clustering, extreme value model

Let

D denote the $n \times n$ matrix of pairwise distances (dissimilarities);

$\mathbf{x}^+$ denote a vector whose elements are $1/x_i$, or zero if $x_i = 0$;

* denote the matrix operation analogous to multiplication, in which {min, max} replaces {×, +}, but the rules for combining matrices and vectors are otherwise unchanged;

a denote a 0,1 vector describing a subset.

The definition of some familiar quantities in these terms is useful.

(1) The total distance between each of the n objects and the members of subset **a** is given by the corresponding element of **v** where

$$\mathbf{v} = \mathbf{Da}.$$

(2) The average distance between each of the n objects and the members of **a** is

$$\mathbf{w} = |\mathbf{a}|^{-1}\mathbf{v} = \mathbf{Da}/\mathbf{1}^{\mathrm{T}}\mathbf{a} = \mathbf{Da}/\mathbf{a}^{\mathrm{T}}\mathbf{a}.$$

(3) The total distance among members of subset **a** is

$$\Delta = \mathbf{a}^{\mathrm{T}}\mathbf{Da} = \mathbf{a}^{\mathrm{T}}\mathbf{v}.$$

(4) The average distance among members of subset **a** is

$$\delta = |\mathbf{a}|^{-2}\Delta = \mathbf{a}^{\mathrm{T}}\mathbf{Da}/\mathbf{a}^{\mathrm{T}}\mathbf{11}^{\mathrm{T}}\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{Da}/\mathbf{a}^{\mathrm{T}}\mathbf{aa}^{\mathrm{T}}\mathbf{a} = \mathbf{a}^{\mathrm{T}}\mathbf{w}/\mathbf{a}^{\mathrm{T}}\mathbf{a}.$$

(5) The maximum distance among members of subset **a** is

$$\mu = \mathbf{a}^{\mathrm{T}}\mathbf{*D*a}.$$

(6) The minimum distance between each of the n objects and the members of **a** is given by the corresponding element in

$$\mathbf{b} = \mathbf{D*a}.$$

The decision criterion of conditional clustering (Chapter VIII) is to compare **w** with μ; if $w_i \leq \mu$, i.e., if the dissimilarity of object i to the members of the subset does not exceed the maximum among the members, it is to be included in the subset at the next stage. Suppose **z** is the vector defined as

$$\mathbf{z} = \mu\mathbf{w}^{+} = \mathbf{a}^{\mathrm{T}}\mathbf{a}(\mathbf{a}^{\mathrm{T}}\mathbf{*D*a})(\mathbf{Da})^{+},$$

then if $z_i \geq 1$, a new vector is to be formed with an element in the i^{th} position equal to unity and is otherwise zero. This formulation can be expressed as

$$\mathbf{1 - I*[(I*z)^+ - 1]},$$

and hence the operator ϕ is given by

$$\phi(\mathbf{a}) = (\mathbf{1 - I*((I*(a^Ta(a^T*D*a)(Da)^+))^+ - 1)}).$$

Three operators, namely, ordinary matrix multiplication (including addition and subtraction), *, and $(.)^+$ are involved in these definitions. Because the distributive law need not apply in operations involving * and either or both of the others, there seems to be no obvious way to simplify the definition of $\phi(\mathbf{a})$.

Appendix 6 Subset homogeneity and musters

Denote a subset of the N objects by S, and the dissimilarities among the N by $\mathbf{D}$. The average dissimilarity among the members of S is denoted by $\delta(S)$, and the neighborhood of S is denoted by $V(S)$. It is assumed that

$$i \in S \leftrightarrow i \in V(S).$$

A subset generated by ϕ (Appendix 5) is denoted by

$$T_k(i,j) = \phi(i,j;V,\mathbf{D})$$

and is called a ϕ-subset.

The concept of subset homogeneity can now be considered in terms of ϕ-subsets, or, more concisely, in terms of their inverse images. Consider the subset

$$B_k = T_k \setminus \{i \mid \exists j,(i,j) \in T_k,\ i \neq j;\ \phi(i,j;V,\mathbf{D}) \neq T_k\} \tag{A6.1}$$

defined by the point-to-set function that omits those members of T_k which do not generate it. Then

DEFINITION A6.1. *A subset T_k is called ϕ-restricted if $B_k = T_k$, and ϕ-diffuse if $B_k \subset T_k$.*

There are many differences between these two kinds of subsets, of which one is now described in terms of the dissimilarities among their members. Let $r(T_k)$ be the ratio of the smallest to the largest dissimilarity between distinct pairs; this order statistic tends to unity in homogeneous subsets, and to zero in the heterogeneous.

THEOREM A6.1. *If L and M are respectively ϕ-restricted and ϕ-diffuse subsets, where $|L| = |M|$, and $\delta(L) = \delta(M)$, then*

$$r(L) > r(M).$$

Proof. Since M is diffuse and $\delta(L) = \delta(M)$; because there is at least one pair of members of M that do not generate the subset, i.e., $\min(d_{ij} \mid i,j \in M) < \min d(_{ij} \mid i,j \in L)$, coupled with the assumption of equal δ, the proof is completed. Q.E.D.

COROLLARY A6.1. $max(d_{ij} \mid i,j \in M) > max(d_{ij} \mid i,j \in L)$.

Proof. This assertion follows at once from equality of mean dissimilarities and the inequality of Theorem A6.1. Q.E.D.

Thus with respect to $r(T_k)$ and related measures, ϕ-restricted subsets can be said to be more homogeneous than ϕ-diffuse. It is of interest that ϕ-restricted subsets are reminiscent of maximal cliques, as the following easily proved consequences of Expression A6.1 illustrate.

THEOREM A6.2. *(a) If the union of two or more ϕ-restricted subsets is a ϕ-subset, it is ϕ-diffuse.*
(b) If the union of two or more ϕ-diffuse subsets is a ϕ-subset, it may be ϕ-diffuse or ϕ-restricted.
(c) If a proper subset of a ϕ-restricted subset contains more than one object, it cannot be a ϕ-subset.
(d) A ϕ-subset that is a subset of a ϕ-diffuse subset may be ϕ-restricted or ϕ-diffuse.

The proofs of these are not difficult and are omitted.

The following lemmas are useful with respect to optimal coverings and prepare the ground for the definition of isolated ϕ-restricted subsets.

LEMMA A6.1. *The intersection two ϕ-restricted subsets contains no more than one object.*

Proof. Suppose L_1 and L_2 are both ϕ-restricted, $L_1 \neq L_2$, $|L_1 \bigcap L_2| \geq 2$; choose $S_1 = \{i,j \in L_1 \bigcap L_2, i \neq j\}$. By definition, both L_1 and L_2 are obtained from S_1, which implies $L_1 = L_2$, which contradicts the supposition that they are distinct.
Q.E.D.

LEMMA A6.2. *Those ϕ-subsets that are proper subsets of ϕ-restricted subsets are not generated by $\phi(.)$, as defined above.*

Proof. This assertion is a simple consequence of $\phi(.)$ and Expression A6.1. Q.E.D.

DEFINITION A6.2. *A ϕ-restricted subset disjoint from any other ϕ-subset is called strongly ϕ-restricted.*

THEOREM A6.3. *If each of the final ensemble of subsets is strongly ϕ-restricted, together they form an optimal partition with respect to ϕ.*

Proof. Since they are disjoint, they form a partition, and since they are ϕ-restricted, by Theorem A6.1 they are more homogeneous than if they were to have been ϕ-diffuse, and since no other subsets can be formed by the specified ϕ, the proof is complete. Q.E.D.

THEOREM A6.4. *If the final ensemble includes ϕ-restricted subsets whose union forms a covering, then only these need to be considered for an optimal covering, and the ϕ-diffuse subsets may be deleted.*

Proof. Deleting the ϕ-diffuse subsets, which by Theorem A6.1 are more heterogeneous than the ϕ-restricted, does not create infeasibility; combining this statement with Lemma A6.2 completes the proof. Q.E.D.

The main consequences of these theorems are of practical interest and can be expressed as

COROLLARY A6.2. *If the final ensemble includes strongly ϕ-restricted subsets, they participate in the optimal covering.*

Determining if a subset is ϕ-restricted is best done after completing any logical reductions (Chapter II) on the ensemble. Lemma A6.1 shows that

only those subsets whose intersection with any other is no greater than unity need be investigated. The determination may be made either by repeating some computation, or by recording the pairs of objects generating each subset.

Although $r(T_k)$ is a useful description of the homogeneity of a subset, it rarely takes a value of unity even for ϕ-restricted subsets. There may be advantage, on occasions, in describing heterogeneity by the ratio

$$|T_k \setminus B_k| / |T_k|,$$

which will always be zero for ϕ-restricted subsets.

In Chapters II and VIII, I have emphasized that each subset in an optimal solution need not bear a one-to-one relationship with an unknown true population, since the "shapes" that the latter have in the dissimilarity space are not necessarily part of the V-family and may not even be convex. Consequently, the convexity assumption needs to be weakened and the V-family enlarged to define a form of subset homogeneity that exhibits continuity rather than the compactness of ϕ-restricted subsets.

DEFINITION A6.3. *A subset R is called a ϕ-muster when it satisfies the three conditions:*

(1) $R \neq \varnothing$ *[i.e., R is not empty],*

(2) $\forall T_k \in (T_k \subseteq R \text{ or } T_k \subseteq (N \setminus R)$ *[i.e., a ϕ-subset belongs either to R or to the complement of R],*

(3) $W \subseteq R \Rightarrow \exists\, T_k : (T \not\subset W) \wedge (T_k \not\subset N \setminus W)$ *[i.e., R has no proper subset which satisfies (1) and (2)].*

This definition is equivalent with one proposed by Tutte (1979).

THEOREM A6.5. *The family of ϕ-musters forms a partition.*

Proof. Definition A6.3 implies that ϕ-musters are pairwise disjoint. Q.E.D.

COROLLARY A6.3. *If* $min|R| = n$, *the only set of* ϕ*-musters covering* N *consists of the partition formed by the improper subset.*

Proof. The assumption implies that there is only one ϕ-muster. Q.E.D.

DEFINITION A6.4. *The neighborhood of a* ϕ*-muster is the union of the neighborhoods of its component* ϕ*-subsets.*

The neighborhood of a ϕ-muster need not be convex, nor its boundary in the dissimilarity space be smooth, nor for it to be without holes.

The weakest form of homogeneity can now be defined: if R_u is the u^{th} ϕ-muster, and $Z_u = Z(R_u)$ is its neighborhood, then

DEFINITION A6.5. *The subset,* $H = \bigcup_{Z_u \cap Z_v \neq \varnothing} R_u \bigcup R_v$ *is called a weak* ϕ*-muster.*

Thus if the neighborhoods of two musters intersect, form the subset that is the union of the objects they contain.

Linking with Theorem A6.3, the following definitions characterize special clusters and partitions:

DEFINITION A6.6. *A* ϕ*-muster that is also strongly* ϕ*-restricted is called a* ϕ***-restricted cluster.***

DEFINITION A6.7. *If all* ϕ*-musters are* ϕ*-restricted clusters, the partition is called* ϕ***-regular.***

DEFINITION A6.8. *If the subsets forming a* ϕ*-regular partition are also weak* ϕ*-musters, the partition is called* ϕ***-isolated.***

The finding of a ϕ-isolated partition implies that large changes in the definition of V would be needed to achieve a different partition; it also implies the mutual isolation important in the context of numerical taxonomy (Cormack 1971). In consequence, there is a high probability

that the partition is optimal.

If there is only one muster for a given set of objects, there may be some inadequacy in the definition of V or $\mathbf{D}$; equally, the objects may truly belong to one group, suggesting that an ordination may be preferable to a clustering.

References

Aarts, E.H.L.; van Laarhoven, P.J.M. 1987. Simulated annealing: a pedestrian review of the theory and some applications. Pages 179–192 *in* Devijver, P.A.; Kittler, J., eds. Pattern recognition theory and applications. Springer, Berlin, Germany.

Abadie, J. 1978. Advances in nonlinear programming. Pages 900–930 *in* Haley, K.B. ed. OR'78. North-Holland, Amsterdam.

Afriat, S.N. 1957. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. Proc. Camb. Phil. Soc. 53:800–816.

Agresti, A. 1984. Analysis of ordinal categorical data. Wiley, New York, N.Y. 287 pp.

Aiken, S.G.; Dallwitz, M.J. 1991. *Festuca* (Poaceae) of North America: interactive identification and information retrieval. TAXACOM, FLORA ONLINE 26 (electronic publication; Zander, R.H., ed.): Systematic biology. Clinton Herbarium, Buffalo Museum of Science, Buffalo, N.Y. (Available via telephone (716) 896-7581 in N. America.)

Aiken, S.G.; Darbyshire, S.J. 1990. Fescue grasses of Canada. Agriculture Canada, Biosystematics Research Centre, Publication 1844. 113 pp.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 *in* Petrov, B.N.; Czáki, F., eds. Second international symposium on information theory. Akademiai Kiadó, Budapest.

Alexandroff (Aleksandrov), P.S. 1956. Combinatorial topology. Grayback Press, Rochester, N.Y. 225 pp.

Alexeev, E.B. 1980. *Festuca* L. Subgenera et sections novae ex America borelai et Mexica. Novst. Sist. Vyssh. Rast. 17:42–53. [In Russian]

Alexeev, E.B. 1985. *Festuca* L. (Poaceae) in Alaska and Canada. Novst. Sist. Vyssh. Rast. 22:5–35. [In Russian]

Almeida, M.T.; Bisby, F.A. 1984. A simple method for establishing taxonomic characters from measurement data. Taxon 33:405–409.

Anderberg, M.R. 1973. Cluster analysis for applications. Academic, New York. 359 pp.

Andersen, E.B. 1980. Discrete statistical models with social science applications. North-Holland, Amsterdam, The Netherlands. 383 pp.

Anderson, J.W. 1958. An introduction to multivariate statistical analysis. Wiley, New York, N.Y. 374 pp.

André, H.M. 1984. Overlapping recurrent groups: an extension of Fager's concept and algorithm. Biometrie–Praximetrie 24:49–65.

Arabie, P.; Boorman, S.A.; Levitt, P.R. 1978. Constructing blockmodels: how and why. J. Math. Psychol. 17:21–63.

Arabie, P.; Carroll, J.D. 1980. Mapclus: a mathematical programming approach to fitting the Adclus model. Psychometrika 45:211–235.

Archie, J.W. 1985. Methods for coding variable morphological features for numerical taxonomic analysis. Syst. Zool. 34:326–345.

Ashton, E.H.; Healy, M.J.R.; Lipton, S. 1957. The descriptive use of discriminant functions in physical anthropology. Proc. R. Soc. Lond. Biol. Sci. 146:552–592.

Atkinson, A.C. 1985. Plots, transformations, and regression. Clarendon, Oxford, U.K. 282 pp.

Auguston, J.G.; Minker, J. 1970. An analysis of some graph theoretical clustering techniques. J. Assoc. Comput. Mach. 17:571–588.

Ayles, P.S.; Beale, E.M.L.; Blues, R.C.; Wild, S.J. 1978. Mathematical models for the location of government. Math. Program. Study 9:59–74.

Babaev, D.A. 1978. The method of hyperspheres for solving problems of Boolean programming. USSR Comput. Math. Math. Phys. 16:32–44.

Bacelar-Nicolau, H. 1987. On the distribution equivalence in cluster analysis. Pages 73–79 *in* Devijver, P.A.; Kittler, J., eds. Pattern recognition theory and applications. Springer, Berlin, Germany.

Baker, F.B. 1992. Item response theory. Dekker, New York. N.Y. 439 pp.

Balas, E.; Ho, A. 1980. Set covering algorithms using cutting planes, heuristics, and subgradient optimization: a computational study. Math. Program. 12:37–60.

Balas, E.; Padberg, M.W. 1975. Set partitioning. Pages 207–258 *in* Roy, B., ed. Combinatorial programming—methods and applications. Reidel, Boston, Mass.

Barankin, E.W.; Takahasi, K. 1978*a*. Betweenness for real vectors and lines. I. Basic generalities. Ann. Inst. Stat. Math. 30(A):125–162.

Barankin, E.W.; Takahasi, K. 1978*b*. Betweenness for real vectors and lines. II. Relatedness of betweenness. Ann. Inst. Stat. Math. 30(A):443–464.

Barnett, V.; Lewis, T. 1984. Outliers in statistical data (2nd edition). Wiley, New York, N.Y. 463 pp.

Basford, K.E.; McLachlan, G.J. 1985. Cluster analysis in a randomized complete block design. Comm. Stat.-Theor. Methods. 14:451–463.

Battro, A.M.; di Pierro Netto, S.; Rozestraten, J.A. 1976. Riemannian geometries of variable curvature in visual space: visual alleys, horopters, and triangles in big open fields. Perception 5:9–23.

Baulieu, F.B. 1989. A classification of presence/absence based dissimilarity coefficients. J. Classif. 6:233–246.

Bazaraa, M.S.; Goode, J.J. 1975. A cutting-plane algorithm for the quadratic set-covering problem. Oper. Res. 23:150–158.

Beyer, W.A.; Stein, M.L.; Smith, T.F.; Ulam, S.M. 1974. A molecular sequence metric and evolutionary trees. Math. Biosci. 19:9–25.

Binder, D.A. 1978. Bayesian cluster analysis. Biometrika 65:31–38.

Binder, D.A. 1981. Approximations to Bayesian clustering rules. Biometrika 68:275–285.

Bitran, G.R. 1977. Linear multiple objective programs with zero–one variables. Math. Program. 13:121–139.

Bitran, G.R. 1979. Theory and algorithms for linear multiple objective programs with zero–one variables. Math. Program. 17:362–290.

Blumenthal, L.P. 1952. Boolean geometry I. Rend. Circ. Mat. Palermo Ser. II 1:343–360.

Bock, H.H. 1989. Probabilistic aspects of cluster analysis. Pages 12–44 *in* Opitz, O., ed. Conceptual and numerical analysis of data. Springer, Berlin, Germany.

Bohachevsky, I.O.; Johnson, M.B.; Stein, M.L. 1986. Generalized simulated annealing for function optimization. Technometrics 28:209–217.

Bölter, M.; Meyer, M. 1986. Structuring of ecological data sets by methods of correlation and cluster analysis. Ecol. Modell. 32:1–13.

Bollobas, B. 1985. Random graphs. Academic, London, U.K. 447 pp.

Bondy, J.A. 1989. Trigraphs. Discrete Math. 75:69–79.

Boros, E.; Hammer, P.L. 1989. On clustering problems with connected optima in Euclidean spaces. Discrete Math. 75:81–88.

Bradfield, G.E.; Kenkel, N.C. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology 68:750–753.

Bremer, K.; Wanntorp, H.E. 1979. Hierarchy and reticulation in systematics. Syst. Zool. 31:25–34.

Brown, P.J. 1977. Functions for selecting tests in diagnostic key construction. Biometrika 64:589–596.

Buckley, F.; Harary, F. 1990. Distance in graphs. Addison-Wesley, Redwood City, Calif. 335 pp.

Buser, M.W. 1983. Testing significant homogeneities with binary information strings. Biom. J. 25:167–180.

Buser, M.W.; Baroni-Urbani, C. 1982. A direct nondimensional clustering method for binary data. Biometrics 38:351–360.

Calinski, T.; Corsten, L.C.A. 1985. Clustering means in ANOVA by simultaneous testing. Biometrics 41:39–48.

Cattell, R.B. 1966. The screen test for the numbers of factors. J. Mult. Behav. Res. 1:245–276.

Cavalli-Sforza, L.L.; Edwards, A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570.

Chittineni, C.B. 1980. Efficient feature-subset selection with probabilistic distance criteria. Inform. Sci. 22:19–35.

Christofides, N.; Brooker, P. 1976. The optimal partitioning of graphs. SIAM J. Appl. Math. 30:55–69.

Chvátal, V. 1979. A greedy heuristic for the set covering problem. Math. Operations Res. 4:233–235.

Chen, C.K.; Andrews, H.C. 1974. Nonlinear intrinsic dimensionality computations. IEEE (Inst. Electr. Electron. Eng.) Trans. Comput. C-23:178–184.

Chesser, R.K.; Van Den Bussche, R.A. 1988. Contiguous clustering: a method for identification of nonrandom aggregates within population samples. Occas. Pap. Mus. Texas Tech. Univ. 122:1–13.

Clarke, R.J. 1990. Covering a set by subsets. Discrete Math. 81:147–152.

Claus, A. 1973. Hyperspheric integer programming. Stud. Appl. Math. 52:153–162.

Clifford, H.T. 1975. Host-parasite relationships. Lecture notes in mathematics 452:79–82. Springer, Berlin, Germany.

Clifford, P.; Green, N.J.B. 1985. Distances in Gaussian point sets. Math. Proc. Camb. Phil. Soc. 97:515–524.

Clymo, R.S. 1980. Preliminary survey of the peat-bog Hummell Knowe Moss using various numerical methods. Vegetatio 42:129–148.

Collatz, L. 1966. Functional analysis and numerical mathematics. Academic, New York, N.Y. 473 pp.

Cormack, R.M. 1971. A review of classification. J. R. Stat. Soc. A 134:321–367.

Coxeter, H.S.M. 1969. Introduction to geometry (2nd edition). Wiley, New York, N.Y. 469 pp.

Craddock, J.M.; Floud, C.R. 1969. Eigen vectors for representing the 500 mb. geopotential surface over the Northern Hemisphere. Q. J. R. Meteorol. Soc. 95:576–593.

Crisci, J.V. 1982. Parsimony in evolutionary theory: law or methodological prescription? J. Theor. Biol. 97:35–41.

Csiszár, I. 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. 19:2032–2066.

Czekanowski, J. 1932. "Coefficient of racial likeness" und "durchschnittliche Differenz." Anthropol. Anz. 9:227–249.

Dagnelie, P. 1966. A propos des differentes méthodes de classification automatique. Rev. Stat. Appl. 14:55–75.

Dagnelie, P.; Merckz, A. 1991. Using generalized distances in classification of groups. Biom. J. 33:683–695.

Dale, M.B. 1971. Information analysis of quantitative data. Stat. Ecol. 3:133–148.

Dallwitz, M.J. 1974. A flexible computer program for generating identification keys. Syst. Zool. 27:50–57.

Dallwitz, M.J.; Paine, T.A. 1986. User's guide to the Delta system: a general system for processing taxonomic descriptions. Division of Entomology Report 13, CSIRO, Australia.

David, H.A. 1981. Order statistics (2nd edition). Wiley, New York, N.Y. 360 pp.

Davison, M.L. 1983. Multidimensional scaling. Wiley, New York, N.Y. 242 pp.

Dewdney, A.K. 1979. A fast, approximate gap minimization algorithm. Pages 349–365 *in* Proceedings, 10th south-eastern conference on combinatorics, graph theory and computing 1.

Dixon, W.J. 1950. Analysis of extreme values. Ann. Math. Stat. 21:488–506.

Dowson, D.C.; Landau, B.V. 1982. The Fréchet distance between multivariate normal distributions. J. Multivar. Anal. 12:450–455.

Dice, L.R. 1945. Measures of the amount of ecologic association between species. J. Ecol. 26:297–302.

Dumas, F.N. 1955. Manifest structure analysis. Montana State University Press, Missoula, Mont. 200 pp.

Duncan, D.B. 1955. Multiple range and multiple F tests. Biometrics 11:1–42.

Eckes, T.; Orlick, P. 1991. An agglomerative method for two-mode hierarchical clustering. Pages 3–8 *in* Bock, H.-H.; Ihm, P., eds. Studies in classification, data analysis and knowledge organization. Springer, Berlin, Germany.

Edmonds, J. 1971. Matroids and the greedy algorithm. Math. Programming 1:127–136.

Efron, B. 1979. Computers and the theory of statistics: thinking the unthinkable. SIAM Rev. 21:460–480.

Eilbert, R.F.; Christensen, R.A. 1982. Contrivedness: the boundary between pattern recognition and numerology. Pattern Recogn. 15:253–261.

Ellis, D. 1951. Autometrized Boolean algebras. Can. J. Math. 3:145–147.

Emptoz, H.; Fages, R. 1980. A set function to cluster analysis. Pages 511–516 *in* Kunt, M.; de Coulon, F., eds. Signal processing: theories and application. North-Holland, Amsterdam, The Netherlands.

Epanechnikov, V.A. 1969. Non-parametric estimation of a multivariate probability probability density. Theor. Prob. Appl. 14:153–158.

Ernvall, J.; Nevalainen, O. 1982. An algorithm for unbiased random sampling. Comput. J. 25:45–47.

Erlander, S. 1981. Entropy in linear programs. Math. Programming 21:137–151.

Esterbrook, G.; Johnson, C.; McMorris, F. 1975. An idealized concept of the true cladistic character. Math. Biosci. 23:263–272.

Etcheberry, J. 1977. The set-covering problem: a new implicit enumeration algorithm. Oper. Res. 25:760–772.

Faith, D.P.; Minchin, P.R.; Belbin, L. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57–68.

Fiala, F. 1973. Solution of linear programming problems in 0–1 variables. Comm. Assoc. Comput. Mach. 16:445–447.

Fienberg, S.E. 1984. Towards a comprehensive approach to the analysis of categorical data. Pages 393–422 *in* David, H.A.; David, H.T., eds. Statistics: an appraisal. Springer, New York, N.Y.

Fillenbaum, S.; Rapoport, A. 1971. Structures in the subjective lexicon. Academic, New York, N.Y. 266 pp.

Filus, L. 1977. A combinatorial lemma related to the search for fixed points. Bull. Acad. Pol. Sci. Ser. Sci. Math. Astr. Phys. 25:615–616.

Filus, L. 1978. A combinatorial lemma for fixed point algorithms. Pages 407–427 *in* Mangasanrian, O.L.; Meyer, R.R.; Robinson, S.M., eds. Nonlinear programming 3. Academic Press, New York, N.Y.

Filus, L. 1978. Combinatorial fixed point algorithms. *in* Proceedings, seminar on gram theory and related topics. North-Holland, Amsterdam, The Netherlands.

Fine, T.L. 1973. Theories of probability. Academic, New York, N.Y. 263 pp.

Fishburn, P.C.; Gehrlein, W.V. 1988. Pick-and-choose heuristics for partial set covering. Discrete Appl. Math. 22:119–132.

Fisher, L.; Van Ness, J.W. 1971. Admissible clustering procedures. Biometrika 58:91–106.

Fisher, W.D. 1958. On grouping for maximum homogeneity. J. Am. Stat. Assoc. 53:789–798.

Fisher, R.A. 1958. The genetical theory of natural selection. (2nd edition) Dover, New York, N.Y. 291 pp.

Fitch, W.M. 1981. A non-sequential method for constructing trees and hierarchical classifications. J. Mol. Evol. 18:30–37.

Fortier, J.J.; Solomon, H. 1966. Clustering procedures. Pages 493–506 *in* Proceedings, 1st international symposium multivariate analysis.

Fréhel, J. 1975. Le problème de partition sous contrainte. Pages 269–274 *in* Roy, B., ed. Combinatorial programming—methods and applications. Reidel, Boston, Mass.

Frieze, A.M. 1987. On the exact solution of random symmetric travelling salesman problems with medium-sized integer costs. SIAM J. Comput. 16:1052–1072.

Fulkerson, D.R.; Gross, O.A. 1965. Incidence matrices and interval graphs. Pac. J. Math. 15:835–855.

Gabbani, D.; Magazine, M. 1986. An interactive heuristic approach for multi-objective integer-programming problems. J. Operational Res. Soc. 37:285–291.

Gabriel, K.R. 1964. A procedure for testing the homogeneity of all sets of means in analysis of variance. Biometrics 20:459–477.

Gabriel, K.R. 1978. Least squares approximation of matrices by additive and multiplicative models. J. R. Stat. Soc. B 40:186–196.

Gabriel, K.R.; Sokal, R.R. 1969. A new statistical approach to geographic variation analysis. Syst. Zool. 18:259–278.

Gantmacher, F.R. 1959. The theory of matrices, Vol. 2. Chelsea, New York, N.Y. 276 pp.

Garfinkel, R.; Nemhauser, G.L. 1972. Integer programming. Wiley, New York, N.Y. 427 pp.

Gentleman, J.F. 1975. Algorithm AS 88. Generation of all ${}_NC_R$ combinations by simulating nested Fortran *DO* loops. Appl. Stat. 24:374–376.

Gnanadesikan, R.; Kettenring, J.R., eds. 1989. Discriminant analysis and clustering: panel on discriminant analysis, classification, and clustering. Stat. Sci. 4:34–69.

Goldberg, D.E. 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, Reading, Mass. 412 pp.

Goldman, N. 1988. Methods for discrete coding of morphological characters for numerical analysis. Cladistics 4:59–71.

Golomb, S.W. 1961. A mathematical theory of discrete classification. Pages 404–425 *in* Cherry, C., ed. Information theory. Butterworth, London, U.K.

Goodman, J.E.; Pollack, R. 1983. Multidimensional sorting. SIAM J. Comput. 12:484–507.

Goodman, L.A. 1971. The analysis of multidimensional contingency tables: stepwise procedure and direct estimation methods for building models for multiple classification. Technometrics 13:33–61.

Gordon, A.D. 1973. Classification in the presence of constraints. Biometrics 29:821–827.

Gordon, A.D.; Henderson, J.T. 1977. An algorithm for Euclidean sums of squares classification. Biometrics 33:355–362.

Gourlay, A.R. 1979. A heuristic algorithm for the solution of seriation problems. Appl. Math. Modell. 3:303–306.

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 23:623–637.

Gower, J.C. 1971*a*. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

Gower, J.C. 1971*b*. Statistical methods of comparing different multivariate analyses of the same data. Pages 138–149 *in* Hodson, F.R.; Kendall, D.G.; Tautu, P., eds. Mathematics in the archaeological and historical sciences. University Press, Edinburgh, U.K.

Gower, J.C. 1974. Maximal predictive classification. Biometrics 30:643–654.

Gower, J.C. 1983. Comparing classifications. Pages 137–155 *in* Felsenstein, J., ed. Numerical taxonomy. Springer, Berlin, Germany.

Gower, J.C. 1984. Distance matrices and their Euclidean approximation. Pages 3–21 *in* Diday, E., ed. Data analysis and informatics. Elsevier, Amsterdam, The Netherlands.

Gower, J.C.; Legendre, P. 1986. Metric and Euclidean properties of dissimilarity coefficients. J. Classif. 3:5–48.

Gower, J.C.; Ross, G.J.S. 1969. Minimum spanning trees and single-linkage cluster analysis. Appl. Stat. 18:54–64.

Granot, D.; Granot, F. 1977. On integer and mixed integer and mixed integer fractional programming problems. Ann. Discrete Math. 1:221–231.

Green, P.J.; Sibson, R. 1978. Computing Dirichlet tessellations in the plane. Comput. J. 21:168–173.

Greenacre, M.J.; Browne, M.W. 1986. An efficient alternating least-squares algorithm to perform multidimensional unfolding. Psychometrika 51:241–250.

Grove, K.; Markvorsen, S. 1992. Curvature, triameter and beyond. Bull. (New Ser.) Am. Math. Soc. 27:261-265.

Gumbel, E.J. 1958. Statistics of extremes. Columbia, New York, N.Y. 375 pp.

Gusfield, D. 1991. Efficient algorithms for inferring evolutionary trees. Networks 21:19-28.

Hagino, G.; Yoshioka, I. 1976. A new method for determining the personal constants in the Luneburg theory of binocular visual space. Percept. Psychophys. 19:499-509.

Hansen, P. 1975. Fonctions d'évaluation et pénalities pour les programmes quadratiques en variables 0-1. Pages 361-370 *in* Roy, B., ed. Combinatorial programming—Methods and applications. Reidel, Boston, Mass.

Hansen, P.; Delattre, M. 1978. Bicriterion cluster analysis as an exploration tool. Pages 247-273 *in* Zionts, S., ed. Multiple criteria problem solving. Lect. Notes Econ. Math. Syst. 155.

Harding, R.M.; Sokal, R.R. 1988. Classification of the European language families by genetic distance. Proc. Natl. Acad. Sci. USA 85:9370-9372.

Hartigan, J.A. 1972. Direct clustering of a data matrix. J. Am. Stat. Assoc. 67:123-129.

Hartigan, J.A. 1975. Clustering algorithms. Wiley, New York, N.Y. 351 pp.

Hawkins, D.M.; Merriam, D.F. 1973. Optimal zonation of digitized sequential data. Math. Geol. 5:389-395.

Hawkins, D.M.; Muller, M.W.; ten Krooden, J.A. 1982. Cluster analysis. Pages 303–356 *in* Hawkins, D.M., ed. Topics in applied multivariate analysis. University Press, Cambridge, U.K.

Hey, A.M. 1987. Algorithms for the set covering problem. Ph.D. thesis, Imperial College, London, U.K.

Hinkley, D.V. 1988. Bootstrap methods. J. R. Stat. Soc. B 50:321–337, 355–370.

Höhle, U. 1988. Quotients with respect to similarity relations. Fuzzy Sets and Systems 27:31–44.

Hoffmann, A. 1989. Arguments on evolution: a palaeontologist's perspective. University Press, Oxford, U.K. 274 pp.

Hubert, L. 1974. Some applications of graph theory to clustering. Psychometrika 39:283–309.

Humphries, C.J. 1983. Primary data in hybrid analysis. Pages 89–111 *in* Platnick, N.I.; Funk, V.A., eds. Advances in cladistics 2. University Press, New York, N.Y.

Ichino, M.; Sklansky, J. 1985. The relative neighbourhood graph for mixed feature variables. Pattern Recogn. 18:161–167.

Izenman, A.J. 1991. Recent developments in nonparametric density estimation. J. Am. Stat. Assoc. 86:205–224.

Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaudoise Sci. Nat. 44:223–270.

Jaccard, P. 1912. The distribution of the flora in the alpine zone. New Phytol. 11:37.

Jardine, N.; R. Sibson, R. 1971. Mathematical taxonomy. Wiley, London, U.K.

Jaynes, E.T. 1983. Papers on probability, statistics and statistical physics Rosenkrantz, R.D. ed. Reidel, Boston, Mass. 434 pp.

Johnson, D.S. 1974. Approximation algorithms for combinatorial problems. J. Comput. System. Sci. 9:256–278.

Johnson, N.L.; Kotz, S. 1970. Continuous univariate distributions, Vol. 1. Houghton Mifflin, Boston, Mass. 300 pp.

Johnson, R.W. 1979. Determining probability distributions by maximum entropy and minimum cross-entropy. Pages 24–29 *in* APL79 conference proceedings.

Jones, L.K.; Byrne, C.L. 1990. General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis. IEEE (Inst. Electr. Electron. Eng.) Trans. Inf. Theory IT-36:23–30.

Juhász, F. 1981. On the spectrum of a random graph. Pages 313–316 *in* Algebraic methods in graph theory, Szeged, 1978. Colloq. Math. Soc. Janos Bolyai 25. North-Holland, Amsterdam, The Netherlands.

Karayiannis, N.B.; Venetsanopoulos, A.N. 1990. Applications of neural networks to environmental protection. Pages 334–337 *in* International neural network conference 1. Kluwer, Dordrecht, The Netherlands.

Karp, R.M. 1972. Reducibility among combinatorial problems. Pages 85–103 *in* Miller, R.E.; Thatcher, J.W., eds. Complexity of computer computations. New York, Plenum, N.Y.

Kautz, W.H. 1968. Fault testing and diagnosis in combinatorial digital circuits. IEEE (Inst. Electr. Electron. Eng.) Trans. Comput. C17:332–366.

Kendall, M.G.; Moran, P.A.P. 1963. Geometrical probability. Griffin, London, U.K. 125 pp.

Kenney, C.; Laub, A.J. 1991. Rational iterative methods for the matrix sign function. SIAM J. Matrix Anal. Appl. 12:273-291.

Kernighan, B.W.; Lin, S. 1970. An efficient heuristic procedure for partitioning graphs. Bell System Tech. J. 49:291-307.

Keuls, M. 1952. The use of the studentized range in connection with an analysis of variance. Euphytica 1:12-122.

Kirkpatrick, D.G.; Radke, J.D. 1985. A framework for computational morphology. Pages 217-248 *in* Toussaint, G.T., ed. Computational geometry. Elsevier, North Holland.

Kiziltan, G.; Yucaoglu, E. 1983. An algorithm for multiobjective zero-one linear programming. Manage. Sci. 29:1444-1453.

Klein, R.W.; Dubes, R.C. 1989. Experiments in projection and clustering by simulated annealing. Pattern Recogn. 22:213-220.

Klix, F. 1979. On interrelationships between natural and artificial intelligence research. Pages 1-9 *in* Fundamentals of computer science 8. North-Holland, Amsterdam, The Netherlands.

Kohonen, T. 1989. Self-organization and associative memory. Springer, New York, N.Y. 312 pp.

Kruskal, J.B. 1964. Nonmetric multi-dimensional scaling: a numerical method. Psychometrika 29:115-129.

Kruskal, J.B. 1972. Linear transformation of multivariate data to reveal clustering. Pages 179-191 *in* Shepard, R.N.; Romney, A.K.; Nerlove, S.B., eds. Multidimensional scaling 1. Seminar Press, New York, N.Y.

Kuennapas, T.; Janson, A.-J. 1969. Multidimensional similarity of letters. Percept. Motor Skills 28:3–12.

Kuenzi, H.; Tzschach, H.G.; Zehnder, C.A. 1971. Numerical methods of mathematical optimization (2nd edition). Academic, New York, N.Y. 219 pp.

Kulczynski, S. 1927. Die Pflanzenassoziatonen der Pinien. Bull. Int. Acad. Pol. Sci. Lett. Ci. Sci. Math. Nat. B:57–203.

Larmour, R.K. 1941. A comparison of hard red spring and hard red winter wheats. Cereal Chem. 18:778–789.

Lau, H.T. 1986. Combinatorial heuristic algorithms with Fortran. Pages 52–80 *in* Lect. Notes Econ. Math. Syst. 280. Springer, New York, N.Y.

Ledley, R.S. 1973. Logic and Boolean algebra in medical science. *In* Proceedings, conference on application of undergraduate mathematics, Atlanta, Ga.

Lefkovitch, L.P. 1975. Choosing levels for non-hierarchical clustering. Pages 132–142 *in* Proceedings, 8th conference on numerical taxonomy. Freeman, San Francisco, Calif.

Lefkovitch, L.P. 1976. Hierarchical clustering from principal coordinates: an efficient method for small to very large numbers of objects. Math. Biosci. 31:157–174.

Lefkovitch, L.P. 1978. Cluster generation and grouping using mathematical programming. Math. Biosci. 41:91–110.

Lefkovitch, L.P. 1979. Consensus coordinates from qualitative and quantitative attributes. Biom. J. 20:679–691.

Lefkovitch, L.P. 1980. Conditional clustering. Biometrics 36:43–58.

Lefkovitch, L.P. 1982. Conditional clusters, musters and probability. Math. Biosci. 60:207–234.

Lefkovitch, L.P. 1984. A nonparametric method for comparing dissimilarity matrices, a general measure of biogeographical distance, and their application. Am. Nat. 123:484–499.

Lefkovitch, L.P. 1985*a*. Entropy and set covering. Inf. Sci. 36:283–294.

Lefkovitch, L.P. 1985*b*. Multi-criteria clustering in genotype–environment interaction problems. Theor. Appl. Genet. 70:585–589.

Lefkovitch, L.P. 1985*c*. Further nonparametric tests for comparing dissimilarity matrices based on the relative neighborhood graph. Math. Biosci. 73:71–88.

Lefkovitch, L.P. 1987*a*. Species associations and conditional clustering: clustering with or without pairwise resemblances. Pages 309–331 *in* Legendre, P. and L., eds. Developments in numerical ecology. Springer, Berlin, Germany.

Lefkovitch, L.P. 1987*b*. Clustering from ordination. Math. Biosci. 84:17–30.

Lefkovitch, L.P. 1987*c*. Optimal attribute sets for identification and diagnosis. Math. Biosci. 84:69–83.

Lefkovitch, L.P. 1989. A non-metric procedure for transforming dissimilarities to Euclidean distances useful in numerical taxonomy and ecology. Biom. J. 31:525–543.

Lefkovitch, L.P. 1991*a*. Vector dissimilarity and clustering. Math. Biosci. 104:39–48.

Lefkovitch, L.P. 1991*b*. Individual samples and the bootstrap. Biom. J. 33:299–303.

Legendre, P. 1987. Constrained clustering. Pages 291–307 *in* Legendre, P. and L., eds. Developments in numerical ecology. Springer, Berlin, Germany.

Leuschner, D.; Sviridov, A.V. 1986. The mathematical theory of taxonomic keys. Biom. J. 28:109–113.

Liepins, G.E.; Hilliard, M.R.; Palmer, M.; Morrow, M. 1987. Greedy genetics. Pages 90–99 *in* Grefenstette, J.J. ed. Genetic algorithms and their applications: proceedings, 2nd international conference on genetic algorithms. Lawrence Erlbaum, Hillsdale, N.J.

Likeš, J. 1966. Distribution of Dixon's statistics in the case of an exponential distribution. Metrika 11:46–54.

Lim, T.M.; Khoo, H.W. 1985. Sampling properties of Gower's general coefficient of similarity. Ecology 6:1682–1685.

Lin, C.S. 1982. Grouping genotypes by a cluster method directly related to genotype-environment interaction mean square. Theor. Appl. Genet. 62:277–280.

Ling, R.F.; Killough, G.G. 1976. Probability tables for cluster analysis based on a theory of random graphs. J. Am. Stat. Assoc. 71:293–300.

Little, T.M. 1985. Analysis of percentage and rating scale data. HortScience 20:642–644.

Liu, R.Y. 1990. On a notion of data depth based on random simplices. Ann. Stat. 18:405–414.

Lukes, J.A. 1974. Efficient algorithm for the partitioning of trees. IBM J. Res. Devel. 18:217–224.

Lukes, J.A. 1975. Combinatorial solutions to the partitioning of general graphs. IBM J. Res. Dev. 19:170–190.

Lundy, M. 1985. Applications of the annealing algorithm to combinatorial problems in statistics. Biometrika 72:191–198.

Luria, A.R. 1976. Cognitive development: its cultural and social foundations. Harvard University Press, Cambridge, Mass.

Magurran, A.E. 1988. Ecological diversity and its measurement. Croom Helm, London, U.K. 179 pp.

Malvestuto, F.M. 1989. Computing the maximum-entropy extension of given discrete probability distributions. Computat. Stat. Data Anal. 8:299–311.

Marcotorchino, F. 1987. Block seriation problems: a unified approach. Appl. Stoch. Models Data Anal. 3:73–91.

Matula, D.W.; Sokal, R.R. 1980. Properties of Gabriel graphs relevant to geographic variation research and the clustering of points in the plane. Geogr. Anal. 12:205–222.

Maxwell, J.C. 1846. On the description of oval curves, and those having a plurality of foci; with remarks by Professor Forbes. Proc. Roy. Soc. Edinburgh II:1–3.

McArdle, B. 1991. Testing for cluster structure before a cluster analysis. SAS Users Group International, 16th annual conference 17–20 Feb. 1991, New Orleans, La. 4 pp.

McCullagh, P. 1980. Regression models for ordinal data. J. R. Stat. Soc. B 42:109–142.

McCullagh, P.; Nelder, J.A. 1989. Generalized linear models (2nd edition). Chapman and Hall, London, U.K. 511 pp.

McLachlan, G.J.; Basford, K.E. 1988. Mixture models: inference and applications to clustering. Dekker, New York, N.Y. 253 pp.

McLachlan, K.I.; Aiken, S.G.; Lefkovitch, L.P.; Edlund, S.A. 1989. Grasses of the Queen Elizabeth Islands. Can. J. Bot. 67:2088-2105.

McNeill, J. 1982. Phylogenetic reconstruction and phenetic taxonomy. Zool. J. Linn. Soc. 74:337-344.

Menger, K. 1942. Statistical metrics. Proc. Nat. Acad. Sci. USA 28:535-537.

Menger, K. 1979. Geometry and positivism, a probabilistic microgeometry. Pages 225-234 *in* Selected papers in logic and foundations, didactics, economics. Reidel, Boston, Mass.

Milligan, G.W.; Cooper, M.C. 1988. A study of standardization of variables in cluster analysis. J. Classif. 5:181-204.

Minchin, P.R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69:89-107.

Moore, G.W. 1976. Proof of the maximum parsimony ("red king") algorithm. Pages 117-137 *in* Goodman, M.; Tashian, R.E., eds. Molecular anthropology. Plenum, New York, N.Y.

Moreau, J.V.; Jain, A.K. 1987. The bootstrap approach to clustering. Pages 63-71 *in* Devijver, P.A.; Kitler, J., eds. Pattern recognition theory and applications. Springer, Berlin, Germany.

Moret, B.M.E. 1982. Decision trees and diagrams. ACM Comput. Surv. 14:593-623.

Moret, B.M.E.; Shapiro, H.D. 1985. On minimizing a set of tests. SIAM J. Sci. Stat. Comput. 6:983-1003.

Mossbrugger, V. 1989. Phylogenetic systematics in palaeobotany and botany. Abh. Naturwiss. Ver. Hamburg (NF) 28:227-245.

Mulvey, J.M.; Crowder, H.P. 1979. Cluster analysis: an application of Lagrangian relaxation. Manage. Sci. 25:329-340.

Needham, R.M. 1961. The theory of clumps II. Camb. Lang. Res. Unit Pap. M.L.:189.

Needham, R.M. 1965. Applications of the theory of clumps. Mechanic. Transl. 8:113-127.

Nosofsky, R.M. 1989. Further tests of an examplar-similarity approach to relating identification and categorization. Percept. Psychophys. 45:279-290.

Ochai, A. 1957. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. Bull. Jpn. Soc. Sci. Fish. 22:526-530.

Oldford, R.W. 1987. On the n-dimensional geometry of regression diagnostics. Commun. Stat.-Theory Meth. 6:2517-2540.

O'Callaghan, J.F. 1976. A model for recovering perceptual organization from dot patterns. IEEE (Inst. Electr. Electron Eng.) Pages 294-298 *in* Proceedings, 3rd international joint conference on pattern recognition.

O'Neill, R.; Wetherill, G.B. 1971. The present state of multiple comparison methods (with discussion). J. R. Stat. Soc. B 33:218-250.

Opitz, O.; Wiedemann, R. 1989. An agglomerative algorithm of overlapping clustering. Pages 201-211 *in* Opitz, O., ed. Conceptual and numerical analysis of data. Springer, Berlin, Germany.

Neyman, J.; Scott, E.L. 1948. Consistent estimates based on partially consistent observations. Econometrika 16:1–32.

North, P.M. 1977. A novel clustering method for estimating numbers of bird territories. Appl. Stat. 26:149–155.

Padberg, M.W. 1975. Characterizations of totally unimodular, balanced and perfect matrices. Pages 275–284 *in* Combinatorial programming: methods and applications. Reidel, Boston, Mass.

Padberg, M.W. 1979. Covering, packing and knapsack problems. Ann. Discrete Math. 4:265–287.

Palmer, E.M. 1985. Graphical evolution. Wiley, New York, N.Y. 177 pp.

Panayirci, E.; Dubes, R. 1987. Spatial point processes and clustering tendency in exploratory data analysis. Pages 81–97 *in* Devijver, P.A.; Kittler, J., eds. Pattern recognition theory and applications. Springer, Berlin, Germany.

Pankhurst, R.J. 1970. A computer program for generating diagnostic keys. Comput. J. 13:145–151.

Pankhurst, R.J. 1978. Biological identification, the principles and practice of identification methods in biology. Edward Arnold, London, U.K. 99 pp.

Pankhurst, R.J. 1983. An improved algorithm for finding diagnostic taxonomic descriptions. Math. Biosci. 65:209–218.

Paris, J.B.; Vencovská, A. 1990. A note on the inevitability of maximum entropy. Int. J. Inexact Reason. 4:183–223.

Payne, R.W. 1981. Selection criteria for the construction of efficient diagnostic keys. J. Stat. Plan. Infer. 5:27–36.

Payne, R.W.; Preece, D.A. 1980. Identification keys and diagnostic tables (with discussion). J. R. Stat. Soc. A 143:253-292.

Pettis, K.W.; Bailey, T.A.; Jain, A.K.; Dubes, R.C. 1979. An intrinsic dimensionality estimator from near-neighbor information. IEEE (Inst. Electr. Electron. Eng.) Trans. Pattern Anal. Mach. Intell. PAMI-1:25-37.

Pollard, D. 1981. Strong consistency of k-means clustering. Ann. Stat. 9:135-140.

Rao, C.R. 1971. Taxonomy in anthropology. Pages 19-29 *in* Hodson, F.R.; Kendall, D.G.; Tautu, P., eds. Mathematics in the archaeological and historical sciences. University Press, Edinburgh, U.K.

Rao, M.M. 1971. Cluster analysis and mathematical programming. J. Am. Stat. Assoc. 66:622-626.

Rasch, G. 1960. Probabilistic models for some intelligence and attainment tests. Danmarks pædagogiske Institut, Copenhagen, Denmark. 184 pp.

Rinnooy Kan, A.H.G.; Boender, C.G.E.; Timmer, G.T. 1985. A stochastic approach to global optimization. Pages 281-288 *in* Schittkowski, K., ed. NATO ASI Series Vol. F15, Computational Mathematical Programming. Springer, Berlin, Germany.

Riordan, J. 1958. An introduction to combinatorial analysis. Wiley, New York, N.Y. 244 pp.

Ripley, B.D. 1981. Spatial statistics. Wiley, New York, N.Y. 252 pp.

Roberts, S.J. 1984. A branch and bound algorithm for determining the optimal feature subset of given size. Appl. Stat. 33:236-241.

Rogers, D.J.; Tanimoto, T.T. 1960. A computer program for classifying plants. Science 132:1115–1118.

Rohrer, J.R. 1988. Incongruence between gametophytic and sporophytic classifications in mosses. Taxon 37:838–845.

Rosenthal, A.; Pino, J.A. 1989. A generalized algorithm for centrality problems on trees. J. Assoc. Comput. Mach. 36:349–361.

Rothkopf, E. 1957. A measure of stimulus similarity and errors in some paired-associate learning tasks. J. exp. Psychol. 53:94–101.

Royaltey, H.H.; Astrachan, E.; Sokal, R.R. 1975. Tests for patterns in geographic variation. Geogr. Anal. 7:369–395.

Russell, P.F.; Rao, T.R. 1940. On habitat and association of species of anopheline larvae in south-eastern Madras. J. Malaria Inst. India 3:153–178.

Rypka, E.W.; Clapper, W.E.; Bowen, I.G.; Babb, R. 1967. A model for the identification of bacteria. J. Microbiol. 46:407–424.

Saaty, T.L. 1981. Modern nonlinear equations. Dover, New York, N.Y. 473 pp.

Sager, T.W. 1978. Estimation of a multivariate mode. Ann. Stat. 6:802–812.

Schiffman, S.S.; Reynolds, M.L.; Young, F.W. 1981. Introduction to multidimensional scaling. Academic, New York, N.Y. 413 pp.

Scott, E.L. 1965. Subclustering. Pages 33–44 *in* Patil, G.P., ed. Classical and contagious discrete distributions. Statistical Publishing Society, Bombay, India.

Sharkey, M.J. 1989 A hypothesis-independent method for character weighting for cladistic analysis. Cladistics 5:63–86. (Appendix by Lefkovitch, L.P., pp. 84–86)

Shepard, R.N.; Arabie, P. 1979. Additive clustering: representation of similarities as combinations of discrete overlapping properties. Psychol. Rev. 86:87–123.

Shephard, G.C.; Webster, R.J. 1965. Metrics for sets of convex bodies. Mathematika 12:73–88.

Shore, J.E.; Johnson, R.W. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross entropy. IEEE (Inst. Electr. Electron. Eng.) Trans. Inf. Theory IT-26:26–37.

Sibley, C.G.; Alquist, J.E. 1984. The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. J. Mol. Evol. 20:2–15.

Sjöberg, L. 1975. Models of similarity and intensity. Psychol. Bull. 82:191–206.

Sneath, P.H.A. 1983. Philosophy and method in biological classification. Pages 22–37 *in* Felsenstein, J., ed. Numerical taxonomy. Springer, Berlin, Germany.

Sneath, P.H.A.; Sokal, R.R. 1973. Numerical taxonomy. Freeman, San Francisco, Calif. 573 pp.

Snedecor, G.W. 1946. Statistical methods. Iowa State College Press, Iowa. 485 pp.

Snee, R.D. 1982. Nonadditivity in a two-way classification: is it interaction or nonhomogeneous variance? J. Am. Stat. Assoc. 77:515–519.

Sokal, R.R. 1985. The continuing search for order. Am. Nat. 126:729–749.

Sokal, R.R.; Michener, C.D. 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38:1409–1438.

Sokal, R.R.; Sneath, P.H.A. 1963. Principles of numerical taxonomy. Freeman, San Francisco, Calif. 450 pp.

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. Biol. Skr. 5:1–34.

Stanfel, L.E. 1982. An algorithm using Lagrangean relaxation and column generation for one-dimensional clustering problems. TIMS/Stud. Manage. Sci. 19:165–185.

Stepanov, V.E. 1970*a*. On the probability of connectedness of a random graph I. Theor. Prob. Appl. XV:55–67.

Stepanov, V.E. 1970*b*. On the probability of connectedness of a random graph II. Theor. Prob. Applic. XV:187–203.

Streng, R. 1991. Classification and seriation by iterative reordering of a data matrix. Pages 121–130 *in* Bock, H.-H.; Ihm, P., eds. Studies in classification, data analysis and knowledge organization. Springer, Berlin, Germany.

Stuart, A.; Ord, J.K. 1987. Kendall's advanced theory of statistics, Vol. 1, 5th edition. University Press, Oxford, U.K. 604 pp.

Supowit, K.J. 1983. The relative neighbourhood graph, with an application to minimum spanning trees. J. Assoc. Comput. Mach. 30:428–441.

Sutcliffe, J.P. 1986. Differential ordering of objects and attributes. Psychometrika 51:209-240.

Thisted, R.A. 1982. Decision-theoretic regression diagnostics. Pages 363-382 *in* Statistical decision theory and related topics III, 2. Academic, New York, N.Y.

Thu, H.M. 1978. A short algorithm to transform dissimilarities into distances. Pages 392-394 *in* 11th annual symposium on the interface. Institute of Statistics, North Carolina State University, Raleigh, N.C.

Tjur, T. 1982. A connection between Rasch's item analysis model and a multiplicative Poisson model. Scand. J. Stat. 9:23-30.

Törn, A.; Žilinskas, A. 1989. Global optimization. Lect. Notes Comput. Sci. 350:95-116.

Toussaint, G.T. 1980. The relative neighbourhood graph of a finite planar set. Pattern Recogn. 12:261-268.

Toussaint, G.T. 1988. A graph-theoretical primal sketch. Pages 229-260 *in* Toussaint, G.T., ed. Computational morphology. Elsevier, Amsterdam, The Netherlands.

Tutte, W.T. 1979. All the king's horses (a guide to reconstruction). Pages 15-33 *in* Bondy, J.S.; Murthy, U.S.R., eds. Graph theory and related topics. Academic, New York, N.Y.

Urquhart, R.B. 1982. Graph theoretical clustering based on limited neighbourhood sets. Pattern Recogn. 15:173-187.

Urquhart, R.B. 1983. Some properties of the planar Euclidean relative neighborhood graph. Pattern Recogn. Lett. 317-322.

Vasko, F.J.; Wilson, G.R. 1986. Hybrid heuristics for minimum cardinality set covering problems. Nav. Res. Logist. Q. 33:241-249.

Vercellis, C. 1984. A probabilistic analysis of the set covering problem. Ann. Oper. Res. 1:255-271.

Vinod, H.D. 1969. Integer programming and the theory of grouping. J. Am. Stat. Assoc. 64:506-519.

Wang, J.C.; Mickle, M.H.; Hoelzeman, R.G. 1977. Simplification of Boolean-valued data by minimum covering. Inf. Sci. 12:163-178.

Wagner, H.-J. 1981. The minimum number of mutations in an evolutionary network. J. Theor. Biol. 91:621-636.

Warwick, S.I.; Black, L.D. 1991. Molecular systematics of *Brassica* and allied genera (subtribe Brassicinae, Brassiceae)—chloroplast genome and cytodeme congruence. Theor. Appl. Genet. 82:81-92.

Watanabe, S. 1969. Knowing and guessing. Wiley, New York, N.Y. 592 pp.

Watanabe, S. 1981. Pattern recognition as a quest for minimum entropy. Pattern Recogn. 13:381-387.

Watson, A.I. 1977. A theory of the visual illusions. Br. J. Math. Stat. Psychol. 30:43-59.

Watson, L.; Aiken, S.G.; Dallwitz, M.J.; Lefkovitch, L.P.; Dubé, M. 1986. Canadian grass genera: keys and descriptions in English and French from an automated data bank. Can. J. Bot. 64:53-70 (with microfiches).

Weber, W.A. 1984. New names and combinations, principally in the Rocky Mountain flora—IV. Phytologia 55:1.

Weiman, C.F.R.; Chaikin, G.M. 1979. Logarithmic spiral grids for image processing. Pages 25–31 *in* Proceedings, IEEE conference on pattern recognition and image processing. Chicago, Ill.

West, M. 1991. Kernel density estimation and marginalization consistency. Biometrika 78:421–425.

White, L.J.; Gillenson, M.L. 1975. An efficient algorithm for minimum k-covers in weighted graphs. Math. Program. 8:20–42.

Williams, W.T.; Dale, M.B.; Lance, G.N. 1971. Two outstanding ordination problems. Aust. J. Bot. 19:251–258.

Williams, W.T.; Lambert, J.M. 1959. Multivariate methods in plant ecology. I. Association analysis in plant communities. J. Ecol. 47:83–101.

Williams, W.T.; Lambert, J.M. 1960. Multivariate methods in plant ecology. II. The use of an electronic digital computer for association analysis. J. Ecol. 48:689–710.

Williamson, M.H. 1978. The ordination of incidence data. Ecology 66:911–920.

Wilson, E.O. 1985. The search for faunal dominance. Pages 489–493 *in* Ball, G.E., ed. Taxonomy, phylogeny and zoogeography of beetles and ants. Junk, Dordrecht, The Netherlands.

Winsberg, S.; Ramsey, J.O. 1983. Monotone spline transformations for dimension reduction. Psychometrika 48:575–595.

Wolda, H. 1981. Similarity indices, sample size and diversity. Oecologia (Berl.) 50:296–302.

Woodward, P.M. 1953. Probability and information theory, with applications to radar. Pergamon, London, U.K. 136 pp.

Yao, A.C. 1982. On constructing minimum spanning trees in k-dimensional spaces and related problems. SIAM J. Comput. 11:721–736.

Yates, F.; Cochran W.G. 1938. The analysis of groups of experiments. J. Agric. Sci. 28:556–580.

Zadeh, L.A. 1975. Calculus of fuzzy restrictions. Pages 1–39 *in* Proceedings, US–Japan seminar on fuzzy sets and their applications, University of California, Berkeley, 1974; Zadeh, L.A.; Fu, D.S.; Tanaka, K.; Shimura, M., eds. Fuzzy sets and their applications to cognitive and decision processes. Academic Press, New York, N.Y.

Zeleny, M. 1982. Multiple criterion decision making. McGraw-Hill, New York, N.Y. 563 pp.

Zemel, E. 1981. Measuring the quality of approximate solutions to zero–one programming problems. Math. Oper. Res. 6:319–339.

Zubin, T. 1938. A technique for measuring like-mindedness. J. Abnorm. Soc. Psychol. 33:508–516.

Index

Canada