3

Measurement of proximity

3.1 Introduction

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

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

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

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

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

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

Data from the United States Statistical Abstract (1970).

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

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

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

3.2 Similarity measures for categorical data

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

3.2.1 Similarity measures for binary data

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

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

Table 3.2 Counts of binary outcomes for two individuals.

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

Table 3.3 Similarity measures for binary data.

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

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

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

3.2.2 Similarity measures for categorical data with more than two levels

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

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

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

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

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

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

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

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

3.3 Dissimilarity and distance measures for continuous data

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

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

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

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

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

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

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

Table 3.4 Dissimilarity measures for continuous data.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

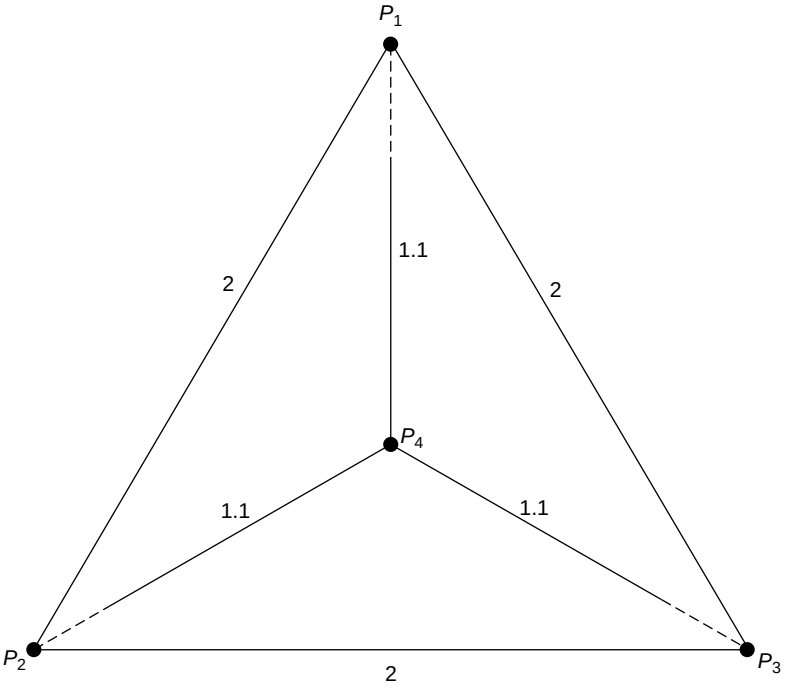

Figure 3.1 *An example of a set of distances that satisfy the metric inequality but which have no Euclidean representation.* (Reproduced with permission from Gower and Legendre, 1986.)

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

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

or

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

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

3.4 Similarity measures for data containing both continuous and categorical variables

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

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

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

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

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

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

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

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

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

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

$$= 0.8. \tag{3.11}$$

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

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

Part of the dissimilarity matrix is shown in Table 3.6.

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

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

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

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

3.5 Proximity measures for structured data

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

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

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

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

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

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

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

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

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

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

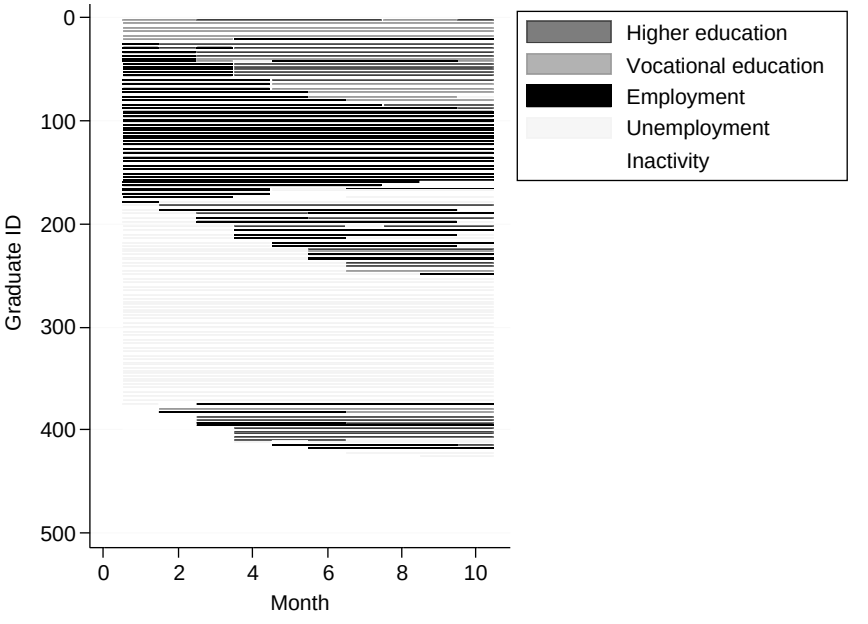

Figure 3.2 *Sequence index plot for employment.* (Reproduced with permission from Brinsky-Fay *et al.*, 2006.)

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

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

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

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

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

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

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

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

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

3.6 Inter-group proximity measures

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

3.6.1 Inter-group proximity derived from the proximity matrix

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

dissimilarity index of the form

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

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

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

$$D_p^2 = (\mathbf{p}_A - \mathbf{p}_B)' \mathbf{W}_p^{-1} (\mathbf{p}_A - \mathbf{p}_B), \tag{3.20}$$

where $\mathbf{p}_A = (p_{A11}, p_{A12}, \ldots, p_{A1c_1}, p_{A21}, p_{A22}, \ldots, p_{A2c_2}, \ldots, p_{Ak1}, p_{Ak2}, \ldots, p_{Akc_k})'$ contains sample proportions in group A and $\mathbf{p}_B$ is defined in a similar manner, and $\mathbf{W}_p$ is the $m \times m$ common sample covariance matrix, where $m = \sum_{k=1}^{p} c_k$. Various alternative forms of this dissimilarity measure may be derived, depending on how the elements of $\mathbf{W}_p$ are calculated. Kurczynski (1970), for example, shows that if each variable has a multinomial distribution, and the variables are independent of one another, then the dissimilarity measure in (3.20) is equal to the dissimilarity measure defined in (3.19). Kurczynski (1969) also demonstrated some important applications of inter-group distance measures for categorical data where the variables are gene frequencies.

3.7 Weighting variables

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

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

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

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

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

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

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

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

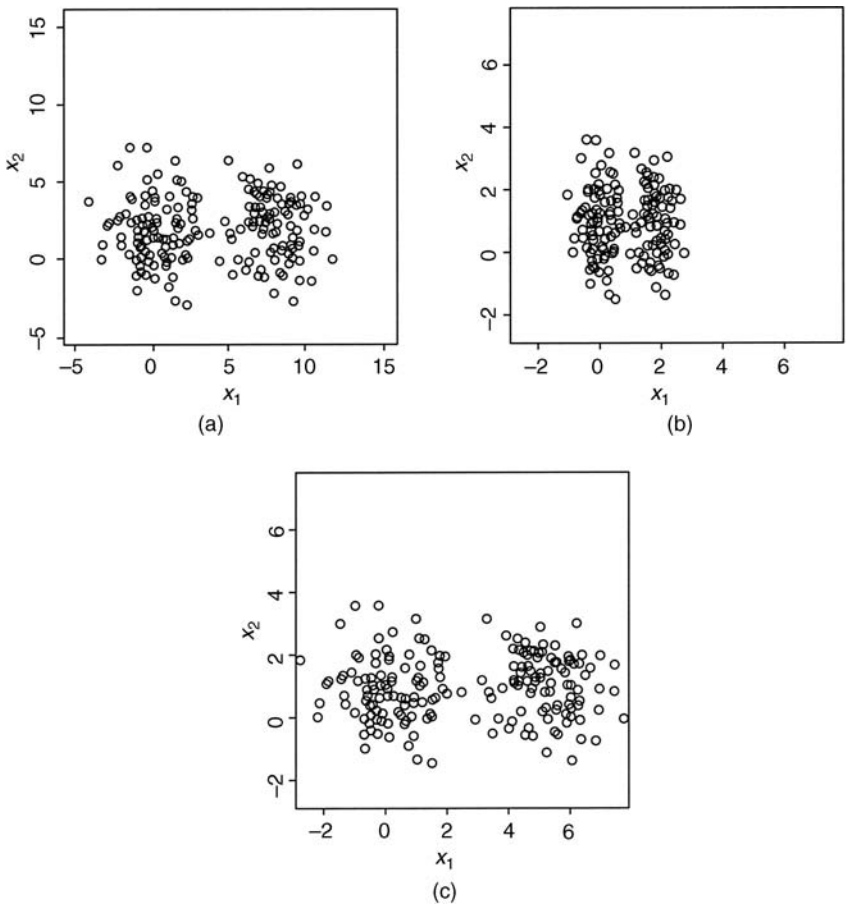

Figure 3.3 *Illustration of standardization problem. (a) Data on original scale. (b) Undesirable standardization: weights based on total standard deviations. (c) Desirable standardization: weights based on within-group standard deviations.*

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

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

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

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

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

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

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

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

3.8 Standardization

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

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

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

3.9 Choice of proximity measure

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

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

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

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

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

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

3.10 Summary

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