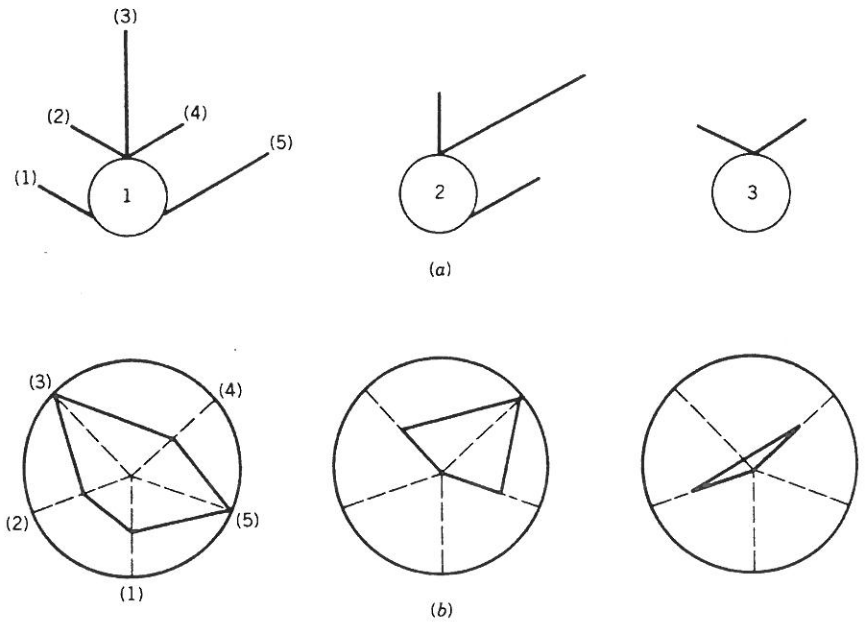

Fig. 4.1 Graphical representation of three four-dimensional points. (*a*) Glyphs. (*b*) STARS.

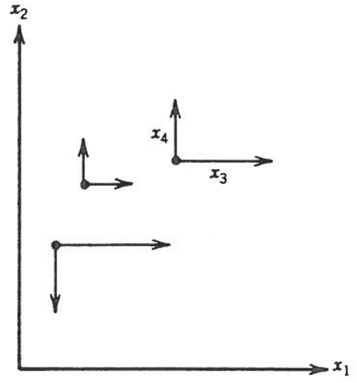

Fig. 4.2 Graphical representation of four-dimensional data.

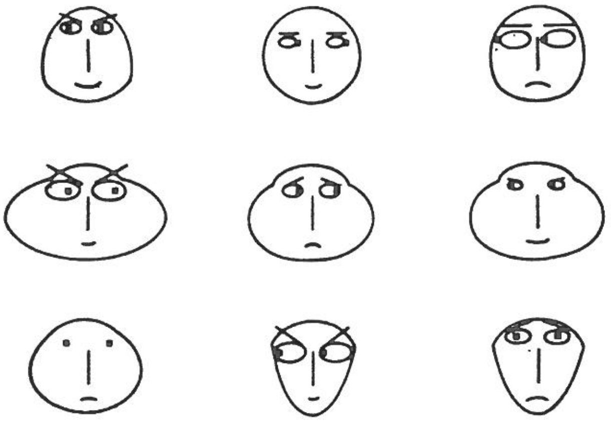

Fig. 4.3 Chernoff's faces for measurements on permanent first lower premolar of various groups of humans and apes. From Fienberg [1979].

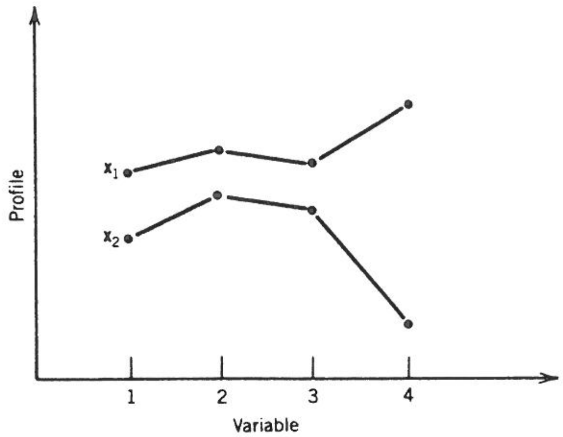

Fig. 4.4 Profiles for two four-dimensional observations.

TABLE 4.1 Percentage of Republican Votes in United States Presidential Elections in 6 Southern States, 1932–1940, 1960–1968[a]

State	1932	1936	1940	1960	1964	1968
Missouri	35	38	48	50	36	45
Maryland	36	37	41	46	35	42
Kentucky	40	40	42	54	36	44
Louisiana	7	11	14	29	57	23
Mississippi	4	3	4	25	87	14
South Carolina	2	1	4	49	59	39

[a] From Kleiner and Hartigan [1981: Table 1].

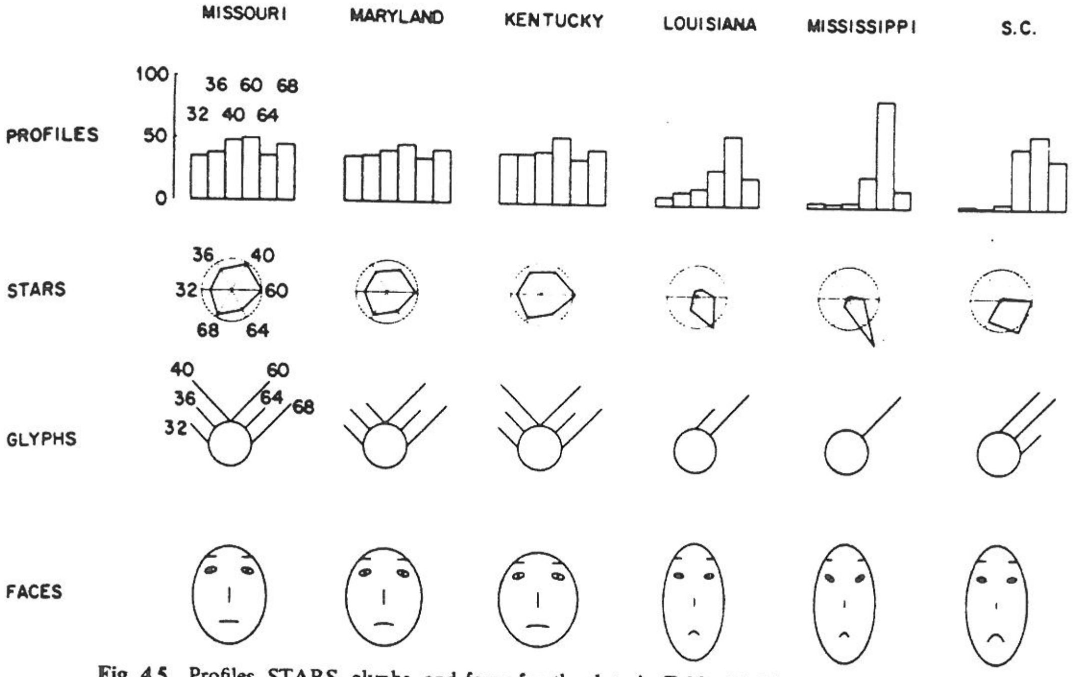

Fig. 4.5 Profiles, STARS, glyphs, and faces for the data in Table 4.1. The circles in the STARS are drawn at 50%. The assignment of the variables to facial features is: 1932—shape of face; 1936—length of nose; 1940—curvature of mouth; 1960—width of mouth; 1964—slant of eyes; 1968—length of eyebrows. From Kleiner and Hartigan [1981].

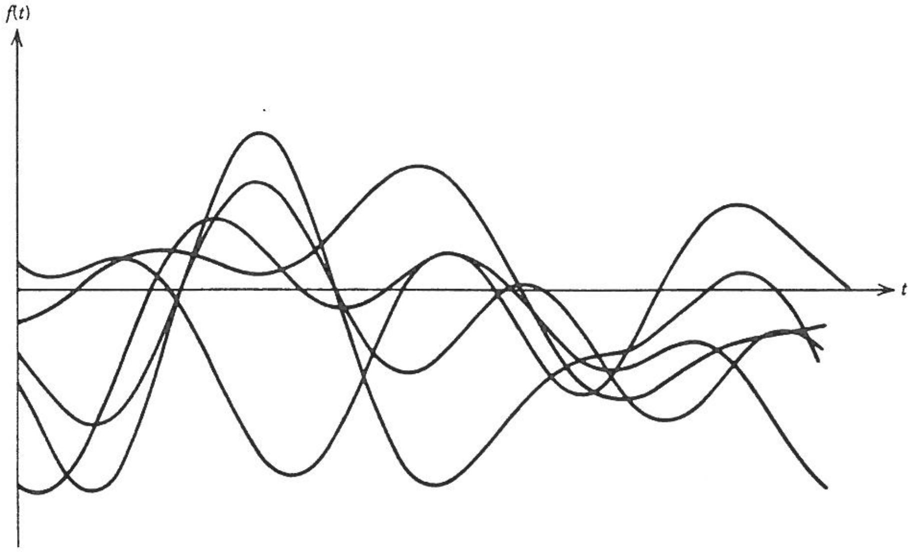

Fig. 4.6 Andrews' Fourier plot for five seven-dimensional observations.

The Fourier plot method consists of representing a d-dimensional vector $\mathbf{x}' = (x_1, x_2, \ldots, x_d)$ by the finite Fourier series

$$f_{\mathbf{x}}(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots$$

and plotting $f_{\mathbf{x}}(t)$ for a grid of t-values in the range $-\pi < t < \pi$ (or else replacing t by $2\pi t$ and using the range $0 < t < 1$). An example showing the corresponding curves for five seven-dimensional observations is given in Fig. 4.6. Andrews [1972] gives the following properties for such plots.

1. The function preserves means. We have

$$f_{\bar{\mathbf{x}}}(t) = \frac{1}{n} \sum_{i=1}^{n} f_{\mathbf{x}_i}(t)$$

so that the curve representing the mean looks like an "average" curve.

2. The function preserves distance. A natural measure of distance between two functions is the L_2 norm. From Exercise 4.1 we have

$$\begin{aligned} \int_{-\pi}^{\pi} [f_{\mathbf{x}}(t) - f_{\mathbf{y}}(t)]^2 \, dt &= \pi \sum_{i=1}^{d} (x_i - y_i)^2 \\ &= \pi \|\mathbf{x} - \mathbf{y}\|^2, \end{aligned} \tag{4.1}$$

so that points $\mathbf{x}$ and $\mathbf{y}$ that are close together lead to curves which are close together.

3. The function preserves linear relationships. If $\mathbf{y}$ lies on the line joining $\mathbf{x}$ and $\mathbf{z}$, then $f_{\mathbf{y}}(t)$ lies between $f_{\mathbf{x}}(t)$ and $f_{\mathbf{z}}(t)$ for all t.

4. The representation yields one-dimensional projections. For a particular value of $t = t_0$, the function value $f_{\mathbf{x}}(t_0)$ is proportional to the length of the projection of the vector $\mathbf{x}$ on the vector

$$\mathbf{a}_0 = \left(1/\sqrt{2}, \sin t_0, \cos t_0, \sin 2t_0, \cos 2t_0, \ldots\right),$$

since $f_{\mathbf{x}}(t_0) = \mathbf{x}'\mathbf{a}_0$. This projection onto a one-dimensional space may show up clusterings or any data peculiarities that occur in this subspace and which may be otherwise obscured by other dimensions. The plot, therefore, provides a continuum of such one-dimensional projections all on the one graph. We note that $\mathbf{a}_0/\|\mathbf{a}_0\|$ represents a point on the d-dimensional sphere of unit radius and one would hope that, in the course of the plot, as many of these points as possible were covered as t_0 ranged from $-\pi$ to π. Andrews [1972] demonstrated that a better coverage can be achieved using more complex functions of t_0 in $\mathbf{a}_0$ above, but at the expense of having a curve with more "wiggles."

5. The representation preserves variances. If the components of $\mathbf{x}$ are uncorrelated with common variance σ^2, then

$$\mathrm{var}[f_{\mathbf{x}}(t)] = \sigma^2(\tfrac{1}{2} + \sin^2 t + \cos^2 t + \sin^2 2t + \cos^2 2t + \cdots).$$

If d is odd, this reduces to a constant, $\frac{1}{2}d\sigma^2$; if d is even, the variance lies between $\frac{1}{2}(d-1)\sigma^2$ and $\frac{1}{2}(d+1)\sigma^2$. For all d the variance is therefore either independent of t, or else the relative dependence on t is slight and decreases as d increases. This implies that the variance of $f_{\mathbf{x}}(t)$ is almost constant along the graph. Unfortunately, the components of $\mathbf{x}$ are invariably correlated with unequal variances. However, the above conditions can be approximately satisfied by transforming $\mathbf{x}$ to its vector of standardized principal components (Section 5.2.1).

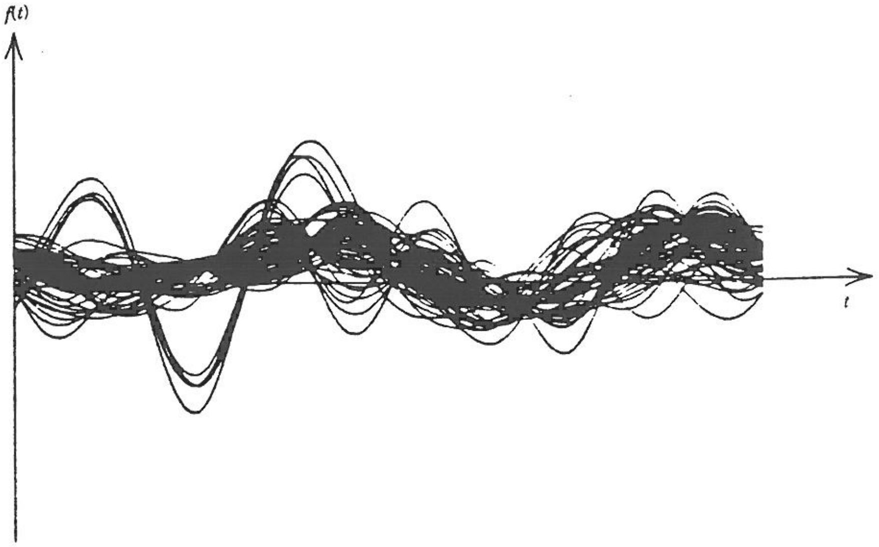

Fig. 4.7 Andrews' curves for a cluster of 86 people. From Morgan [1981].

4.2 TRANSFORMING TO NORMALITY

As the multivariate normal (MVN) distribution has played such a central role in multivariate analysis, it is appropriate that we should consider transformations that help to normalize the data. However, there are some pitfalls associated with the transformations discussed below (see Hernandez and Johnson [1980], Bickel and Doksum [1981], Box and Cox [1982]). Also, the parameters associated with the transformed data may not be as meaningful as those associated with the original data; for example, $\mu_1 - \mu_2$ may be more appropriate than $\log(\mu_1/\mu_2)$. To set the scene we consider the univariate case first.

4.2.1 Univariate Transformations

A useful family of transformations is the following:

$$x^{(\lambda)} = \begin{cases} x^\lambda, & \lambda \neq 0, \\ \log x, & \lambda = 0 \quad \text{and} \quad x > 0. \end{cases} \tag{4.2}$$

This particular family, studied in detail by Tukey [1957] for $|\lambda| \leq 1$, contains the well-known log, square root, and inverse transformations. To avoid a discontinuity at $\lambda = 0$, Box and Cox [1964] considered the modification

$$x^{(\lambda)} = \begin{cases} \dfrac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log x, & \lambda = 0 \quad \text{and} \quad x > 0. \end{cases} \tag{4.3}$$

Using this modification, if we assume that the transformed observations $x_i^{(\lambda)}$ are i.i.d. $N_1(\mu, \sigma^2)$, the likelihood function for the untransformed data is

$$(2\pi\sigma^2)^{-n/2}\left[\exp\left\{-\sum_{i=1}^{n}\frac{\left(x_i^{(\lambda)}-\mu\right)^2}{2\sigma^2}\right\}\right]\left[\prod_{i=1}^{n}x_i^{\lambda-1}\right]. \tag{4.4}$$

Since the last term in square brackets, the Jacobian of the transformation, does not contain μ or σ^2, the maximum likelihood estimates of μ and σ^2 for given λ are

$$\bar{x}(\lambda)=\sum_{i=1}^{n}\frac{x_i^{(\lambda)}}{n} \quad \text{and} \quad \hat{\sigma}_x^2=\sum_{i=1}^{n}\frac{\left(x_i^{(\lambda)}-\bar{x}^{(\lambda)}\right)^2}{n}. \tag{4.5}$$

If $\dot{x}$ $[=(x_1x_2\cdots x_n)^{1/n}]$ is the geometric mean of the x_i, then the maximum value of the log likelihood is (apart from a constant)

$$L_{\max}(\lambda) = -\tfrac{1}{2}n\log\hat{\sigma}_x^2 + n\log\dot{x}^{(\lambda-1)} \tag{4.6}$$

$$= -\tfrac{1}{2}n\log\hat{\sigma}_z^2, \tag{4.7}$$

where $z_i^{(\lambda)} = x_i^{(\lambda)}/\dot{x}^{\lambda-1}$. Box and Cox [1964] then suggested choosing $\lambda = \hat{\lambda}$, where $\hat{\lambda}$ maximizes $L_{\max}(\lambda)$. The maximization can be carried out directly using a standard numerical procedure such as solving $dL_{\max}(\lambda)/d\lambda = 0$ iteratively, or by simply plotting $L_{\max}(\lambda)$ against λ. A plot is always useful, as the local behavior of $L_{\max}(\lambda)$ in the neighborhood of $\hat{\lambda}$ can be considered. For example, if $\hat{\lambda} = 0.2$, it may be quite reasonable, for a fairly flat likelihood function, to set $\lambda = 0$, that is, a log transformation. More formally, an approximate $100(1-\alpha)$% confidence region for the true value of λ is the set of all λ satisfying

$$L_{\max}(\hat{\lambda}) - L_{\max}(\lambda) \le \tfrac{1}{2}\chi^2_{1,\alpha},$$

where $\text{pr}[\chi_1^2 \ge \chi^2_{1,\alpha}] = \alpha$. To test $H_0: \lambda = \lambda_0$ we simply treat $2[L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_0)]$ as being approximately distributed as χ_1^2. Andrews [1971] proposed an "exact" more robust test for H_0, though the empirical study of Atkinson [1973] suggests that the test is less powerful than the likelihood ratio test.

If some of the x_i are negative, we can add a positive constant ξ to all the observations to make them positive. Alternatively, we can include ξ in the above likelihood function, now written $L_{\max}(\xi, \lambda)$ to indicate that x_i is replaced by $x_i + \xi$, and find the maximum likelihood estimates of ξ and λ (Box and Cox [1964]).

John and Draper [1980] provided an alternative family of transformations

$$x^{(\lambda)} = \begin{cases} \text{sign}\left[\dfrac{(|x|+1)^{\lambda}-1}{\lambda}\right], & \lambda \ne 0, \\ \text{sign}\left[\log(|x|+1)\right], & \lambda = 0, \end{cases} \tag{4.8}$$

where the sign of $x^{(\lambda)}$ is that associated with the observation x, called the modulus transformation. The power transformation (4.3) is effective in making skewed distributions more symmetrical and, hopefully, more normal. For example, the effect of a logarithmic or square root transformation is to pull in one tail of the distribution. John and Draper [1980] noted that the "modulus transformation, on the other hand, is effective on a distribution that already possesses approximate symmetry about some central point and alters each half of the distribution through the same power transformation in an attempt to make the shape more normal." If all the data are positive, the modulus and power transformations are equivalent. However, an alternative is still provided by the modulus family, since data of the form $x - a$, for some constant a such as a robust estimator of location, can be transformed. The maximum likelihood method described above also applies to the modulus family: Equation (4.7) still holds with $\dot{x}$ in $z^{(\lambda)}$ now being the geometric mean of the $|x_i| + 1$. John and Draper [1980] gave an example where the best power transformation is inadequate (the normal plot is S shaped), while the best modulus transformation gives a linear residual plot. This might have been expected, as the residual plot for the untransformed data was S shaped but symmetric, indicating that a modulus transformation, which treats the tails symmetrically, would be better than a skew-correcting power transformation.

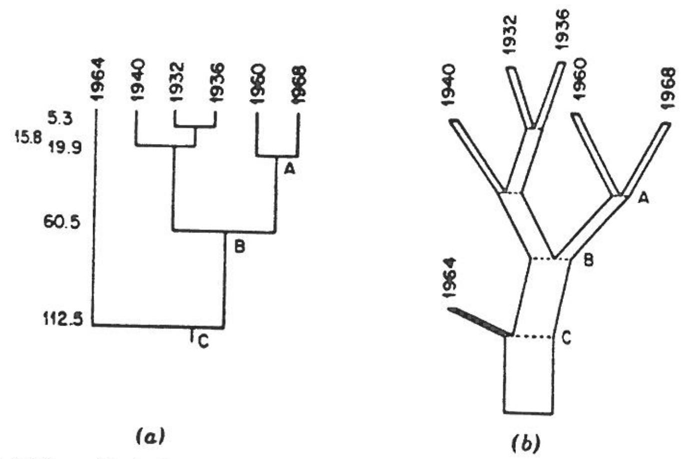

Fig. 4.8 (*a*) Hierarchical clustering by complete linkage for the data in Table 4.1. (*b*) Tree representation of data in Table 4.1 for Missouri. From Kleiner and Hartigan [1981].

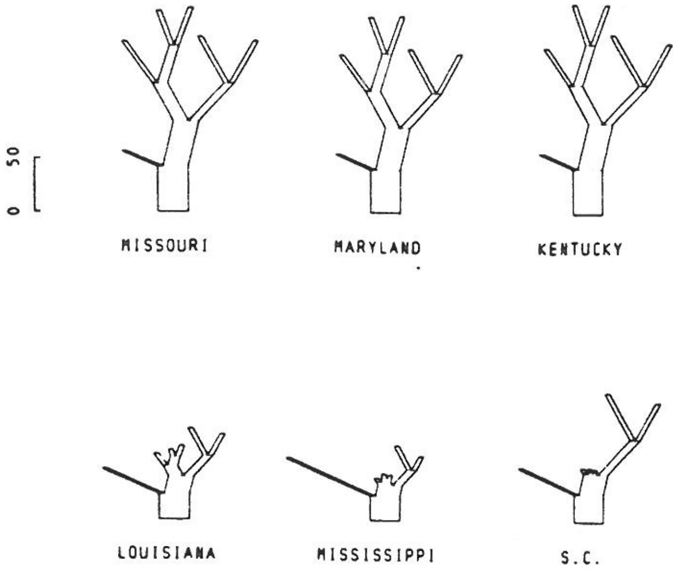

Fig. 4.9 Tree representation of the data in Table 4.1. From Kleiner and Hartigan [1981].

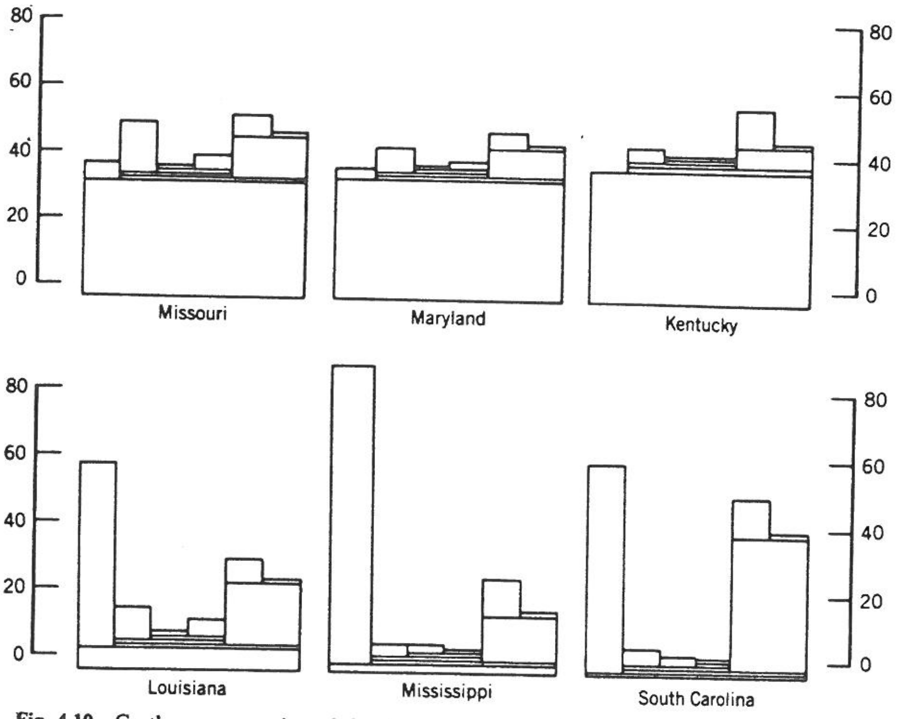

Fig. 4.10 Castle representation of the data in Table 4.1. From Kleiner and Hartigan [1981].

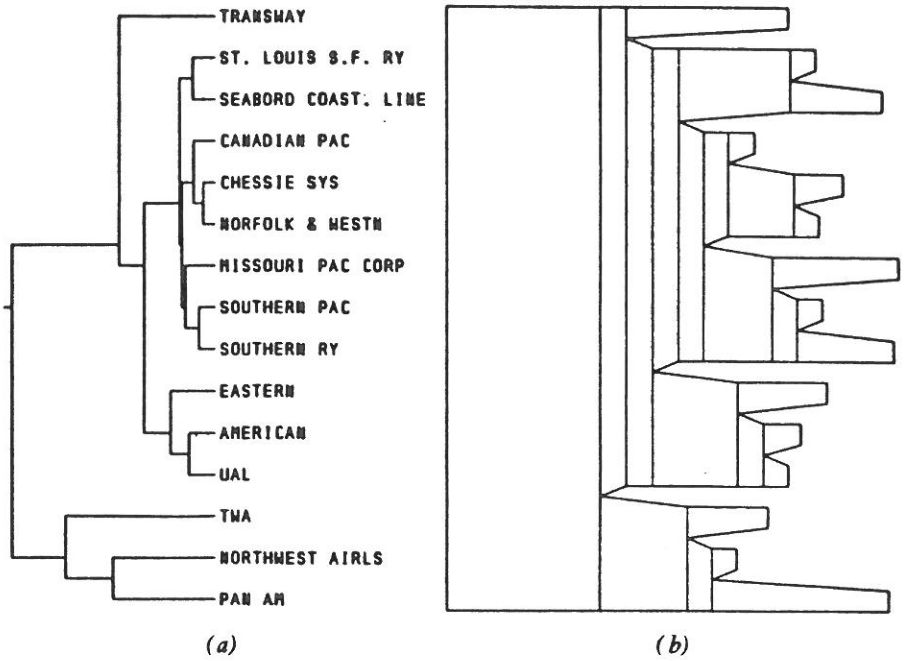

Fig. 4.11 (*a*) Hierarchical clustering for yields of 15 transport companies over 25 years by complete linkage. (*b*) Tapered castle representing yields in 1953. From Kleiner and Hartigan [1981].

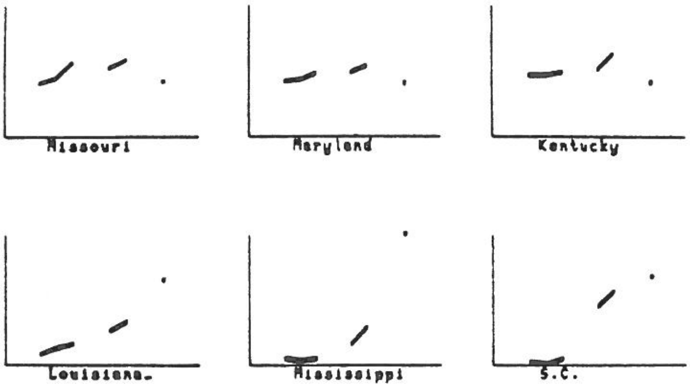

Fig. 4.12 Profiles with clustered coordinates.

TABLE 4.2 Adults Who "Really Like to Watch"; Correlations to 4 Decimal Places (Programs Ordered Alphabetically with Channel)[a]

		PrB	ThW	Tod	WoS	GrS	LnU	MoD	Pan	RgS	24H
ITV	PrB	1.0000	0.1064	0.0653	0;5054	0.4741	0.0915	0.4732	0.1681	0.3091	0.1242
"	ThW	0.1064	1.0000	0.2701	0.1424	0.1321	0.1885	0.0815	0.3520	0.0637	0.3946
"	Tod	0.0653	0.2701	1.0000	0.0926	0.0704	0.1546	0.0392	0.2004	0.0512	0.2437
"	WoS	0.5054	0.1474	0.0926	1.0000	0.6217	0.0785	0.5806	0.1867	0.2963	0.1403
BBC	GrS	0.4741	0.1321	0.0704	0.6217	1.0000	0.0849	0.5932	0.1813	0.3412	0.1420
"	LnU	0.0915	0.1885	0.1546	0.0785	0.0849	1.0000	0.0487	0.1973	0.0969	0.2661
"	MoD	0.4732	0.0815	0.0392	0.5806	0.5932	0.0487	1.0000	0.1314	0.3267	0.1221
"	Pan	0.1681	0.3520	0.2004	0.1867	0.1813	0.1973	0.1314	1.0000	0.1469	0.5237
"	RgS	0.3091	0.0637	0.0512	0.2963	0.3412	0.0969	0.3261	0.1469	1.0000	0.1212
"	24H	0.1242	0.3946	0.2432	0.1403	0.1420	0.2661	0.1211	0.5237	0.1212	1.0000

[a] From Ehrenberg [1977: Table 4].

TABLE 4.3 The Correlations in Table 4.2 Rounded and Reordered[a]

Programs		WoS	MoD	GrS	PrB	RgS	24H	Pan	ThW	Tod	LnU
World of Sport	ITV		.6	.6	.5	.3	.1	.2	.1	.1	.1
Match of the Day	BBC	.6		.6	.5	.3	.1	.1	.1	0	0
Grandstand	BBC	.6	.6		.5	.3	.1	.2	.1	.1	.1
Prof. Boxing	ITV	.5	.5	.5		.3	.1	.2	.1	.1	.1
Rugby Special	BBC	.3	.3	.3	.3		.1	.1	.1	.1	.1
24 Hours	BBC	.1	.1	.1	.1	.1		.5	.4	.2	.2
Panorama	BBC	.2	.1	.2	.2	.1	.5		.4	.2	.2
This Week	ITV	.1	.1	.1	.1	.1	.4	.4		.3	.2
Today	ITV	.1	0	.1	.1	.1	.2	.2	.3		.2
Line-Up	BBC	.1	0	.1	.1	.1	.2	.2	.2	.2	

[a] From Ehrenberg [1977: Table 5].

4.2.2 *Multivariate Transformations*

For multivariate data, the above univariate procedures can be applied to each dimension with a separate λ or separate pair (ξ, λ) for each of the d variables. However, Andrews et al. [1971] have given the following multivariate generalization of Box and Cox's technique. We now have a vector of parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_d)$, one λ_i for each dimension, and the transformed vectors $\mathbf{x}_i^{(\boldsymbol{\lambda})} = (x_{i1}^{(\lambda_1)}, \ldots, x_{id}^{(\lambda_d)})'$ are assumed to be i.i.d. $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Corresponding to (4.6), the likelihood function for $\boldsymbol{\lambda}$ is [see (3.6)]

$$L_{\max}(\boldsymbol{\lambda}) = -\tfrac{1}{2}n \log|\hat{\boldsymbol{\Sigma}}| + \sum_{j=1}^{d} (\lambda_j - 1) \sum_{i=1}^{n} \log x_{ij}, \tag{4.9}$$

where x_{ij} is the jth element of $\mathbf{x}_i$ and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{x}_i^{(\boldsymbol{\lambda})} - \bar{\mathbf{x}}^{(\boldsymbol{\lambda})}\right)\left(\mathbf{x}_i^{(\boldsymbol{\lambda})} - \bar{\mathbf{x}}^{(\boldsymbol{\lambda})}\right)'. \tag{4.10}$$

We now choose $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$, where $\hat{\boldsymbol{\lambda}}$ maximizes $L_{\max}(\boldsymbol{\lambda})$. To test $H_0 : \boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ we calculate $2[L_{\max}(\hat{\boldsymbol{\lambda}}) - L_{\max}(\boldsymbol{\lambda}_0)]$, which is approximately χ_d^2 when H_0 is true. This statistic can also be used for constructing a confidence region for $\boldsymbol{\lambda}$.

Andrews et al. [1971] give a transformation procedure for improving normality in certain directions. Some interesting plots demonstrating the transformations described above and in the previous section are given in Gnanadesikan [1977: pp. 144–150].

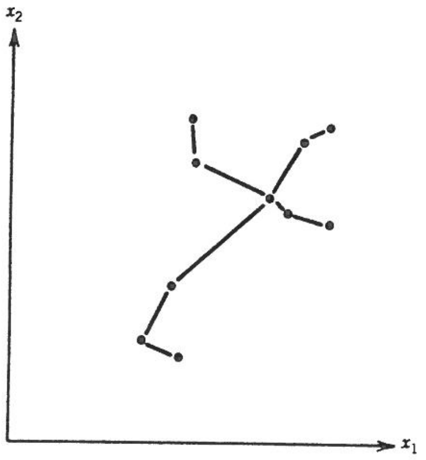

Fig. 4.14 Minimum spanning tree for 10 two-dimensional observations.

A number of graphical methods based on the scaled differences $\mathbf{y}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$ are available. Under the null hypothesis H_0 of multivariate normality, the $\mathbf{z}_i = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu})$ are i.i.d. $N_d(\mathbf{0}, \mathbf{I}_d)$ and $\mathbf{z}_i'\mathbf{z}_i = (\mathbf{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) \sim \chi_d^2$ [Theorem 2.1(vi)]. Clearly the $\mathbf{y}_i$ will have similar properties and the Mahalanobis distances squared (Section 1.5),

$$D_i^2 = \mathbf{y}_i'\mathbf{y}_i = (\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \tag{4.25}$$

will be approximately i.i.d. χ_d^2 under H_0. We note that D_i^2 is unchanged if we work with the correlation matrix $\mathbf{R}$ instead of $\mathbf{S}$, as D_i^2 is invariant under linear transformations of the $\mathbf{x}_i$.

7.2.2 *Similarities*

Similarity coefficients have a long history and in the early literature were usually known as association coefficients. For example, suppose the d variables or characteristics are dichotomous, indicating the presence (+) or absence (−) of each characteristic. Then, for each pair of the n objects, we can form the usual two-way contingency table given by Table 7.1, where, for example, β is the number of characteristics present in object r but absent in object s. Numerous measures of association c_{rs}, satisfying $0 \le c_{rs} \le 1$, have been proposed and Clifford and Stephenson [1975: pp. 54–55] list 11 (see also Anderberg [1973: Table 4.53]). The most popular measures are Jaccard's [1908] coefficient $\alpha/(\alpha + \beta + \gamma)$, Czekanowski's [1913] coefficient $2\alpha/(2\alpha + \beta + \gamma)$, and the simple matching proportion $(\alpha + \delta)/d$. The choice of a coefficient depends very much on the relative weights given to positive matches (α) and negative matches (δ). Clearly there are situations, in numerical taxonomy, for example, where the joint absence of a characteristic would carry little or no weight in comparison with the joint presence, and the matching coefficient, with equal emphasis on both categories, would not be appropriate. However, the matching coefficient would be appropriate if the variables were all nominal with two states, the states simply being alternatives with equal weight. The above measures of association can be extended to nominal variables with more than two states (Anderberg [1973: Section 5.4]). For example, we can calculate the matching coefficient as the proportion of the nominal variables that match for two objects.

With quantitative variables, one measure of similarity between $\mathbf{x}_r$ and $\mathbf{x}_s$, the observations on objects r and s, is the correlation of the pairs (x_{rj}, x_{sj}), $j = 1, 2, \ldots, d$, namely,

$$c_{rs} = \frac{\sum_j (x_{rj} - \bar{x}_{r\cdot})(x_{sj} - \bar{x}_{s\cdot})}{\left\{\sum_j (x_{rj} - \bar{x}_{r\cdot})^2 \sum_j (x_{sj} - \bar{x}_{s\cdot})^2\right\}^{1/2}}, \tag{7.7}$$

TABLE 7.1 Number of Characteristics Occurring in, or Absent from, Two Objects: α Common to Both Objects; β and γ Occurring in Only One Object; δ Absent from Both

		Object s		
		Present (+)	Absent (−)	Sum
Object r	Present (+)	α	β	$\alpha + \beta$
	Absent (−)	γ	δ	$\gamma + \delta$
	Sum	$\alpha + \gamma$	$\beta + \delta$	$\alpha + \beta + \gamma + \delta = d$

and $-1 \le c_{rs} \le 1$. Apart from not satisfying axiom (1) below, this measure, however, has certain disadvantages. For example, if $c_{rs} = 1$, it does not follow that $\mathbf{x}_r = \mathbf{x}_s$, only that the elements of $\mathbf{x}_r$ are linearly related to those of $\mathbf{x}_s$ (see Exercise 7.9). Also, what meaning can we give $\bar{x}_{r\cdot}$, the mean over the different variables for object r? For these and other reasons, the correlation coefficient has been criticized by a number of authors (e.g., Fleiss and Zubin [1969]). Although there is some difference of opinion, the evidence would suggest that dissimilarities based on metrics are better proximity measures than correlations. Cormack [1971] states that the "use of the correlation coefficient must be restricted to situations in which variables are uncoded, comparable measurements or counts; it is not invariant under scaling of variables, or even under alterations in the direction of coding of some variables (Minkoff [1965])."

A large number of similarity measures have been proposed in the literature and they can be categorized mathematically in several ways (e.g., using trees in Hartigan [1967]; see also Duran and Odell [1974: Chapter 4]). If $\mathscr{P}$ is the population of objects, then we can define a similarity as a function that maps $\mathscr{P} \times \mathscr{P}$ into R^1 and satisfies the following axioms:

(1) $0 \le C(r, s) \le 1$ for all r, s in $\mathscr{P}$.
(2a) $C(r, r) = 1$.
(2b) $C(r, s) = 1$ only if $r = s$.
(3) $C(r, s) = C(s, r)$.

We shall write $c_{rs} = C(r, s)$ and use the notation $C(\mathbf{x}_r, \mathbf{x}_s)$ for vector data. The Jaccard and Czekanowski coefficients satisfy the above axioms.

Gower [1971a]) has proposed an all-purpose measure of similarity

$$c_{rs} = \left(\sum_{j=1}^{d} c_{rsj} \right) \Big/ \sum_{j=1}^{d} w_{rsj}, \tag{7.8}$$

where c_{rsj} is a measure of similarity between objects r and s for variable j. Here w_{rsj} is unity except when a comparison is not possible, as with missing observations or negative matches of dichotomous variables, in which case we set $c_{rsj} = w_{rsj} = 0$. In Table 7.2 we have the appropriate coefficients for a dichotomous variable or a two-state qualitative variable. With a multistate variable (ordinal or nominal) of more than two states, we set $c_{rsj} = 1$ if objects r and s agree in variable j, and $c_{rsj} = 0$ otherwise: In both cases $w_{rsj} = 1$. For a quantitative variable, $w_{rsj} = 1$ and

$$\begin{aligned} c_{rsj} &= 1 - |x_{rj} - x_{sj}|/R_j \\ &= 1 - |x'_{rj} - x'_{sj}|, \end{aligned}$$

where R_j is the range of variable j and $x'_{rj} = x_{rj}/R_j$. Thus if we have d_1 quantitative variables, d_2 dichotomous variables, and d_3 multistate variables, then

$$c_{rs} = \frac{\sum_{j=1}^{d_1} \left(1 - |x'_{rj} - x'_{sj}|\right) + \alpha_2 + m_3}{\left[d_1 + (d_2 - \delta_2) + d_3\right]},$$

where α_2 and δ_2 are the number of positive and negative matches, respectively, for the dichotomous variables, and m_3 are the number of matches for the multistate variables. If all the variables are dichotomous, then c_{rs} reduces to Jaccard's coefficient, whereas if all the variables are two-state, then c_{rs} reduces to the matching coefficient. Williams and Lance [1977] do not recommend its use if the continuous data are highly skewed, as the range is very sensitive to skewness.

Gower [1971a] showed that $[(c_{rs})]$ is positive semidefinite for his coefficient. From (5.73), $d_{rs} = (2 - 2c_{rs})^{1/2}$ satisfies the triangle inequality, being the Euclidean distance measure for some configuration of points (the factor 2 may be omitted).

A dissimilarity coefficient can always be obtained from a similarity by setting $d_{rs} = 1 - c_{rs}$, though d_{rs} will not be a metric unless c_{rs} satisfies Axiom (2b) above and d_{rs} satisfies the triangle inequality. For example, if we apply the scaled Euclidean metric $\|\mathbf{x}_r - \mathbf{x}_s\|/d$ or the Canberra metric (7.6) to binary 0–1 data, we get $(\beta + \gamma)/d$, the "one-complement" of the matching coefficient c_{rs}, so that $1 - c_{rs}$ is a metric. Similarly, we find that the one-complements of the Jaccard and Czekanowski coefficients satisfy the triangle inequality (Ihm [1965]) so that they are also metrics.

TABLE 7.2 Gower's Similarity Coefficients for (a) Dichotomous and (b) Two State Qualitative Variables [see Equation (7.8)]

(a) Presence/absence of dichotomous variable j

Object r	+	+	−	−
Object s	+	−	+	−
c_{rsj}	1	0	0	0
w_{rsj}	1	1	1	0

(b) Two state qualitative variable j

Object r	1	1	2	2
Object s	1	2	1	2
c_{rsj}	1	0	0	1
w_{rsj}	1	1	1	1

If we are interested in clustering *variables* rather than objects, the correlation coefficient for variables j and k is

$$r_{jk} = \frac{\sum_i (x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k})}{\left\{\sum_i (x_{ij} - \bar{x}_{\cdot j})^2 \sum_i (x_{ik} - \bar{x}_{\cdot k})^2\right\}^{1/2}} \tag{7.9}$$

$$= \sum_{i=1}^{n} \tilde{x}_{ij}\tilde{x}_{ik}, \tag{7.10}$$

where $\tilde{x}_{ij}$ is x_{ij} suitably "standardized." If we define the Euclidean distance d_{jk} between standardized variables j and k by

$$\begin{aligned} d_{jk}^2 &= \sum_i (\tilde{x}_{ij} - \tilde{x}_{ik})^2 \\ &= \sum_i \tilde{x}_{ij}^2 + \sum_i \tilde{x}_{ik}^2 - 2\sum_i \tilde{x}_{ij}\tilde{x}_{ik} \\ &= 2(1 - r_{jk}), \end{aligned}$$

then we can use $d_{jk} = [2(1 - r_{jk})]^{1/2}$ to transform the similarity measure r_{jk} into a distance; the factor 2 can be dropped. However, there are problems with using a correlation coefficient if one or both of the variables j and k are nominal (disordered multistate) variables. One solution has been proposed by Lance and Williams [1968] (see also Anderberg [1973: pp. 96–97]). Dichotomous variables can be handled using the values 0 and 1 (see Exercise 7.12). Some measures of association between nominal and ordinal variables are described by Agresti [1981]. Methods for estimating missing values are discussed by Wishart [1978b] and Gordon [1981: Section 2.4.3].

7.3 HIERARCHICAL CLUSTERING: AGGLOMERATIVE TECHNIQUES

The agglomerative methods all begin with n clusters each containing just one object, a proximity matrix for the n objects (we assume, for the moment, that this is an $n \times n$ matrix $\mathbf{D} = [(d_{rs})]$ of dissimilarities), and a measure of distance between two clusters, where each cluster contains one or more objects. The first step is to fuse the two nearest objects into a single cluster so that we now have $n - 2$ clusters containing one object each and a single cluster of two objects. The second step is to fuse the two nearest of the $n - 1$ clusters to form

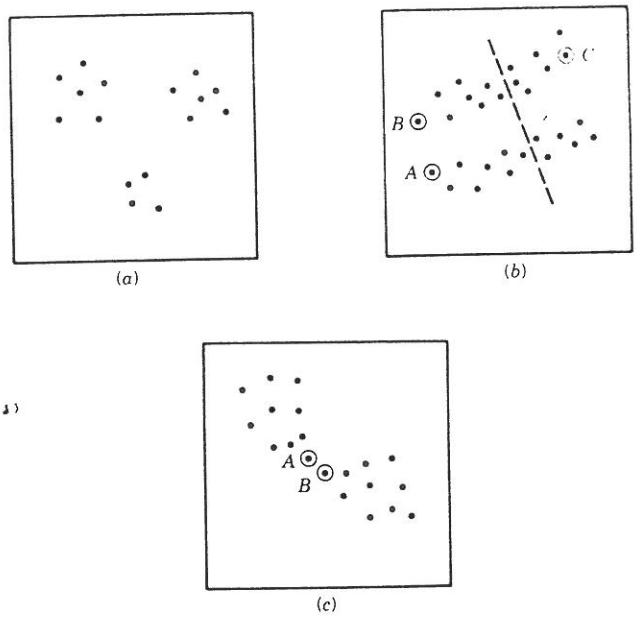

Fig. 7.1 Three types of clustering. (*a*) Spherical clusters. (*b*) Two or four clusters. (*c*) Chaining.

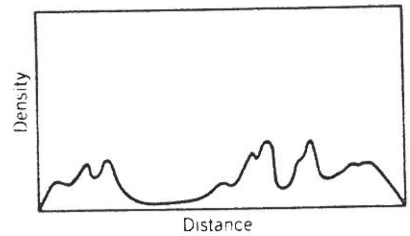

Fig. 7.2 Density of points along a line.

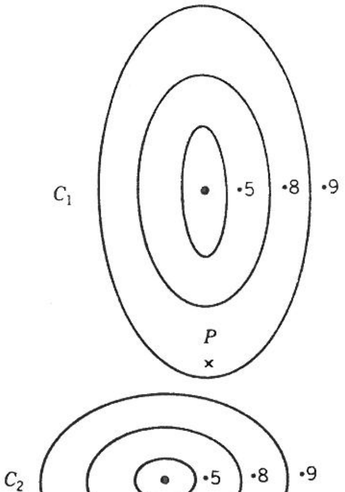

Fig. 7.3 Probability contour ellipses for two bivariate normal distributions.

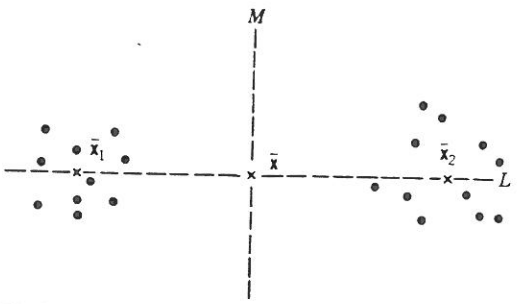

Fig. 7.4 Two similar well-separated spherical clusters.

a SINGLE LINKAGE (NEAREST NEIGHBOR) METHOD

If C_1 and C_2 are two clusters, then the distance between them is defined to be the smallest dissimilarity between a member of C_1 and a member of C_2 (Sneath [1957], Sokal and Sneath [1963], Johnson [1967]), namely,

$$d_{(C_1)(C_2)} = \min\{d_{rs} : r \in C_1, s \in C_2\}, \tag{7.11}$$

where r denotes "object r." We demonstrate the fusion process with the following simple example. Let

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \!\!\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \left(\begin{array}{ccccc} 0 & 7.0 & 1.0 & 9.0 & 8.0 \\ 7.0 & 0 & 6.0 & 3.0 & 5.0 \\ 1.0 & 6.0 & 0 & 8.0 & 7.0 \\ 9.0 & 3.0 & 8.0 & 0 & 4.0 \\ 8.0 & 5.0 & 7.0 & 4.0 & 0 \end{array}\right) \end{array}. \tag{7.12}$$

The minimum d_{rs} is $a_1 = d_{13} = 1.0$ so that objects 1 and 3 are joined and our clusters are (1, 3), (2), (4), and (5). Now

$$d_{(2)(1,3)} = \min\{d_{21}, d_{23}\} = d_{23} = 6.0,$$

$$d_{(4)(1,3)} = \min\{d_{41}, d_{43}\} = d_{43} = 8.0,$$

$$d_{(5)(1,3)} = \min\{d_{51}, d_{53}\} = d_{53} = 7.0,$$

and the distance matrix for the clusters is

$$\mathbf{D}_1 = \begin{array}{c|cccc} & (1,3) & 2 & 4 & 5 \\ \hline (1,3) & 0 & 6.0 & 8.0 & 7.0 \\ 2 & 6.0 & 0 & 3.0 & 5.0 \\ 4 & 8.0 & 3.0 & 0 & 4.0 \\ 5 & 7.0 & 5.0 & 4.0 & 0 \end{array}.$$

The smallest entry is $a_2 = d_{24} = 3.0$, so that objects 2 and 4 are joined and our clusters become (1, 3), (2, 4), and (5), with

$$d_{(1,3)(2,4)} = \min\{d_{(2)(1,3)}, d_{(4)(1,3)}\} = 6.0,$$

$$d_{(5)(2,4)} = \min\{d_{52}, d_{54}\} = d_{54} = 4.0$$

and

$$\mathbf{D}_2 = \begin{array}{c|ccc} & (1,3) & (2,4) & 5 \\ \hline (1,3) & 0 & 6.0 & 7.0 \\ (2,4) & 6.0 & 0 & 4.0 \\ 5 & 7.0 & 4.0 & 0 \end{array}.$$

The smallest entry is $a_3 = d_{(5)(2,4)} = 4.0$, so that object 5 is joined to cluster (2, 4) and the clusters are now (1, 3) and (2, 4, 5). Finally, these two clusters are fused to give the single cluster (1, 2, 3, 4, 5). We note that

$$\begin{aligned} d_{(1,3)(2,4,5)} &= \min\{d_{(1,3)(2,4)}, d_{(1,3)(5)}\} \\ &= d_{32} = 6.0 \qquad (= a_4, \text{ say}). \end{aligned}$$

The above process can be described diagrammatically in the form of a dendrogram as shown in Fig. 7.5. The vertical scale gives a measure of the size of a cluster; tight clusters tend to have lower values. In constructing dendrograms, some relabeling is generally needed so that each cluster is a contiguous sequence of objects, for example, the interchange of 2 and 3 as in Fig. 7.5 (for algorithms, see Exercises 7.13 and 7.19). Although the above technique of spelling out $\mathbf{D}, \mathbf{D}_1, \mathbf{D}_2$, and so on, is a general one and can be applied to other agglomerative methods, it can be simplified for single linkage using the ordered d_{rs} as in Table 7.3.

TABLE 7.3 Single-Linkage Clustering for Dissimilarity Matrix (7.12)

Ordered Distances	Clusters
$d_{13} = 1.0$	(1, 3), (2), (4), (5)
$d_{24} = 3.0$	(1, 3), (2, 4), (5)
$d_{45} = 4.0$	(1, 3), (2, 4, 5)
$d_{25} = 5.0$	(1, 3), (2, 4, 5)
$d_{23} = 6.0$	(1, 2, 3, 4, 5)
$d_{35} = 7.0$	(1, 2, 3, 4, 5)
$d_{15} = 8.0$	(1, 2, 3, 4, 5)
$d_{14} = 9.0$	(1, 2, 3, 4, 5)

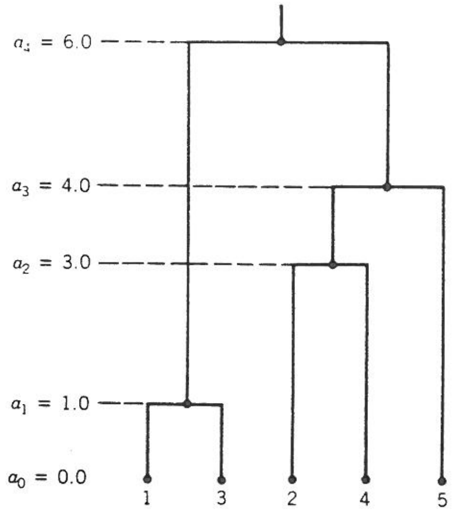

Fig. 7.5 Single linkage dendrogram for dissimilarity matrix (7.12).

g LANCE AND WILLIAMS FLEXIBLE METHOD

Lance and Williams [1967a] showed that the preceding methods labeled a–c, e, and f (with $P(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2) = \|\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\|^2$ in the centroid method) are special cases of the following formula for the distance between clusters C_3 and $C_1 \cup C_2$:

$$d_{(C_3)(C_1 \cup C_2)} = \alpha_1 d_{(C_3)(C_1)} + \alpha_2 d_{(C_3)(C_2)} + \beta d_{(C_1)(C_2)} + \gamma |d_{(C_3)(C_1)} - d_{(C_3)(C_2)}|. \tag{7.19}$$

Wishart [1969a] then showed that the incremental sum of squares method also satisfies the above formula (see Exercise 7.17), and the values of the parameters are given in Table 7.4; n_i is the number of objects in cluster C_i ($i = 1, 2, 3$). Lance and Williams [1967a] suggested using a flexible scheme satisfying the constraints $\alpha_1 + \alpha_2 + \beta = 1$, $\alpha_1 = \alpha_2$, $\beta < 1$, and $\gamma = 0$, and recommended a small negative value of β such as $\beta = -0.25$. Sibson [1971] noted that (7.19) is symmetric with regard to C_1 and C_2 for all the methods in Table 7.4 so that these methods are independent of labeling.

TABLE 7.4 Parameters for Lance and Williams [1967] Recurrence Formula (7.19)

	Parameter		
	α_i	β	γ
1. Nearest neighbor	$\frac{1}{2}$	0	$-\frac{1}{2}$
2. Farthest neighbor	$\frac{1}{2}$	0	$\frac{1}{2}$
3. Centroid	$\dfrac{n_i}{n_1 + n_2}$	$\dfrac{-n_1 n_2}{(n_1 + n_2)^2}$	0
4. Incremental	$\dfrac{n_i + n_3}{n_1 + n_2 + n_3}$	$\dfrac{-n_3}{n_1 + n_2 + n_3}$	0
5. Median	$\frac{1}{2}$	$-\frac{1}{4}$	0
6. Group average	$\dfrac{n_i}{n_1 + n_2}$	0	0
7. Flexible	$\frac{1}{2}(1 - \beta)$	$\beta (< 1)$	0