variable time on a CRT display system. A noise field was immediately displayed on the CRT so that the digit was not clearly visible. The subjects had to respond what digit was present in the noisy image. Each student looked at only 50 stimuli, so Table 2.2 shows the aggregate response of all eight students. The table shows the confusion for each possible stimulus–response combination. The entry in the second row of the matrix in Table 2.2 indicates that of the 400 stimuli presented for digit 1, 269 correct responses were made by the subjects. Levine (1977a) defined the frequency of confusion between stimuli to be the measure of similarity. Thus digit pair 9 and 3 are considered more similar than digit pair 9 and 1. Notice that this similarity matrix is nonsymmetric. Multidimensional scaling and hierarchical clustering algorithms were applied to this matrix by Levine to study the evidence of hierarchical structure in the organization of visual stimuli.

2.1.3 Data Types and Scales

Now that the two primary formats for representing data—the pattern matrix and the proximity matrix—have been established, we turn to the characteristics of the data themselves. Anderberg (1973) outlines a categorization of data types and data scales appropriate for cluster analysis that is summarized below. Recognizing the type and scale of data will help in selecting a clustering algorithm.

Data *type* refers to the degree of quantization in the data. A single feature can be typed as binary, discrete, or continuous. *Binary* features have exactly two values and occur, for example, in "yes–no" responses on a questionnaire. A *discrete* feature has a finite, usually small, number of possible values. For example, samples of a speech signal can be quantized to 16, or 2^4, levels, so a feature representing the sample can be coded into 4 bits. All measurements and all numbers stored in computers have a finite number of significant digits, so, strictly speaking, all features are discrete. However, it is often convenient to think of a feature value as a point on the real line that can take on any real value in a fixed range of values. Such a feature is called *continuous*.

Proximity indices can also be binary, discrete, or continuous. For example, suppose that a set of objects is partitioned into mutually exclusive, all-inclusive subsets. One binary index of similarity assigns zero to a pair of objects that fall in different subsets and one to a pair in the same subset. A rank order proximity index is an integer from 1 to $n(n-1)/2$, where n is the number of objects. The integers represent the relative order of the proximities. Such an index is discrete. The Euclidean distance proximity index, defined for patterns in a pattern space, is typed continuous.

The second trait of a feature and of a proximity index is the data *scale*, which indicates the relative significance of numbers. Data scales can be dichotomized into *qualitative* (nominal and ordinal) scales and *quantitative* (interval and ratio) scales. A *nominal* scale is not really a scale at all because numbers are simply used as names. For example, a (yes, no) response could be coded as (0, 1) or (1, 0) or (50, 100); the numbers themselves are meaningless in any quantitative sense. The other qualitative scale, and the weakest numerical scale, is the *ordinal* scale; the numbers have meaning only in relation to one another. For example,

the scales (1, 2, 3), (10, 20, 30), and (1, 20, 300) are all equivalent from an ordinal viewpoint. Binary and discrete features and proximity indices can be coded on these qualitative scales.

The separation between numbers has meaning on an *interval* scale. A unit of measurement exists, and the interpretation of the numbers depends on this unit. For example, a person can be asked to judge satisfaction with politicians on a scale from 0 to 100. The pair of scores (45, 55) and the pair (10, 90) on two politicians would indicate very different perceptions. Before the number 10 could be interpreted, one would need to know that the scale was 0 to 100 or 1 to 10 or 10 to 100. Temperature provides another example of an interval scale. A reading of 90° Fahrenheit has a very different implication for comfort than does a temperature of 90° Celsius.

The strongest scale is the *ratio* scale, on which numbers have an absolute meaning. This implies that an absolute zero exists along with a unit of measurement, so the ratio between two numbers has meaning. For example, the distance between two cities can be measured in meters, miles, or inches, but doubling the distance always has the same significance when driving from one to the other. Similarly, doubling one's income should double purchasing power, no matter what unit of currency is used. Degrees Kelvin establishes a ratio temperature scale because it has a natural zero. All three data types can be coded on the two quantitative scales.

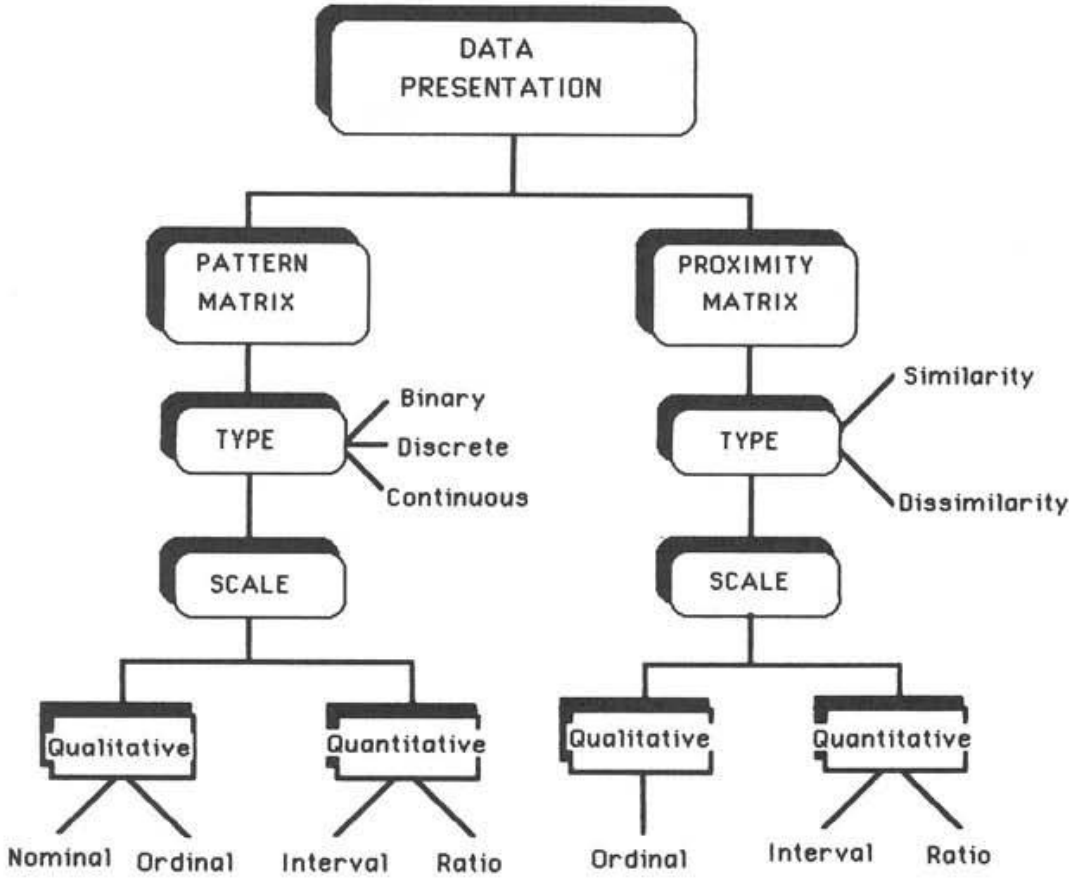

Figure 2.2 Formats, types, and scales for data.

Data type and scale are not always of one's choosing. Recognizing type and scale is important in both forming proximity indices and interpreting the results of a cluster analysis. For example, one should realize that human subjects are good at generating binary, qualitative data but that instruments are required to produce continuous, quantitative data. A human subject required to generate discrete, interval data will be under greater stress than one asked to provide binary, ordinal data, so the reliability of data can depend on type and scale. Anderberg (1973) explains conversions from one scale to another. Clustering methods (Chapter 3) use quantitative indices of proximity to assign a cluster label, or name, to each object, so a nominal scale can be generated from a quantitative scale. Multidimensional scaling (Section 2.7) changes ordinal scales into ratio scales. The various formats, types, and scales for data are summarized in Figure 2.2.

2.2 PROXIMITY INDICES

This section explains some of the more common proximity indices. Anderberg (1973) provides a thorough review of measures of association and their interrelationships. A proximity index between the ith and kth patterns is denoted $d(i, k)$ and must satisfy the following three properties:

1. (a) For a dissimilarity: $d(i, i) = 0$, all i

   (b) For a similarity: $d(i, i) \geq \max_k d(i, k)$, all i

2. $d(\mathrm{i}, k) = d(k, i)$, all (i, k)
3. $d(i, k) \geq 0$, all (i, k)

Ratio and nominal proximity indices are discussed in separate sections.

2.2.1 Ratio Types

A proximity index can be determined in several ways. Suppose that we begin with a pattern matrix $[x_{ij}]$, where x_{ij} is the jth feature for the ith pattern. All features are continuous and measured on a ratio scale. The most common proximity index for such patterns is the Minkowski metric, which measures dissimilarity. The ith pattern, which is the ith row of the pattern matrix, is denoted by the column vector $\mathbf{x}_i$.

$$\mathbf{x}_\mathrm{i} = (x_{i1}\, x_{i2}\, .\ .\ .\ x_{id})^\mathrm{T},\ i = 1, 2, .\ .\ .\ , \mathrm{n}$$

Here d is the number of features, n the number of patterns, and T denotes vector transpose. The Minkowski metric is defined by

$$d(i, k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^r \right)^{1/r} \qquad \text{where } r \geq 1$$

All Minkowski metrics satisfy the additional metric properties stated below. Property 5 is called the *triangle inequality*.

4. $d(i, k) = 0$ only if $\mathbf{x}_i = \mathbf{x}_k$
5. $d(i, k) \leqq d(i, m) + d(m, k)$, all (i, k, m)

Gower and Legendre (1986) show that for a metric dissimilarity matrix $[d(i, j)]$, only properties 1 and 4 are required; other properties can be derived from these two.

The three most common Minkowski metrics are defined below and are illustrated in Figure 2.3.

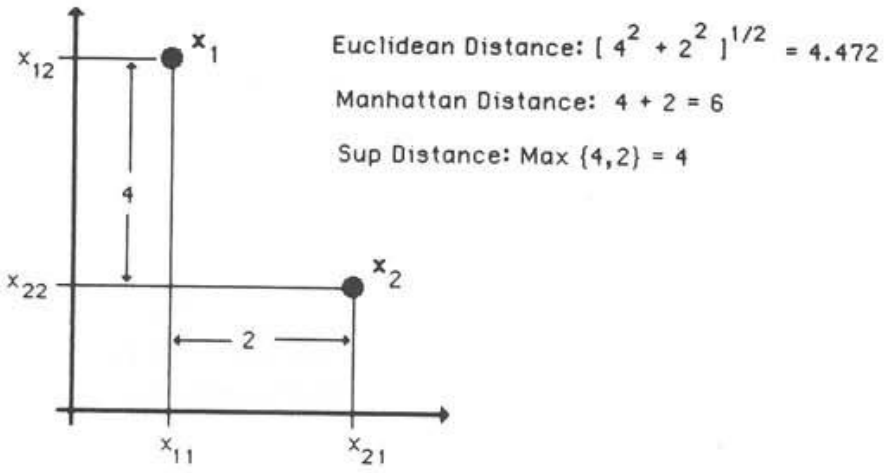

Figure 2.3 Minkowski metrics.

1. $r = 2$ (Euclidean distance)

$$d(i, k) = \left[\sum_{j=1}^{d} (x_{ij} - x_{kj})^2\right]^{1/2} = [(\mathbf{x}_i - \mathbf{x}_k)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_k)]^{1/2}$$

2. $r = 1$ (Manhattan, or taxicab, or city block distance)

$$d(i, k) = \sum_{j=1}^{d} |x_{ij} - x_{kj}|$$

3. $r \to \infty$ ("sup" distance)

$$d(i, k) = \max_{1 \leq j \leq d} |x_{ij} - x_{kj}|$$

Euclidean distance is the most common of the Minkowski metrics. The familiar geometric notions of invariance to translations and rotations of the pattern space are valid only for Euclidean distance. Accepted practice in the application area strongly affects the choice of proximity index. Euclidean distance seems to be preferred in engineering work. When all features are binary, the Manhattan metric is called the *Hamming distance*, or the number of features in which two patterns differ. Not all proximities encountered in applications are metrics. Tversky (1977) gives several examples to illustrate why a similarity is not always symmetric or transitive.

The squared Mahalanobis distance has also been used as a distance measure in cluster analysis (Everitt, 1974). The expression for the squared Mahalanobis distance between patterns $\mathbf{x}_i$ and $\mathbf{x}_k$ is

$$d(i, k) = (\mathbf{x}_i - \mathbf{x}_k)^{\mathrm{T}} \mathcal{S}^{-1} (\mathbf{x}_i - \mathbf{x}_k)$$

where the matrix $\mathcal{S}$ is the pooled sample covariance matrix, defined in Appendix D. The Mahalanobis distance incorporates the correlation between features and standardizes each feature to zero mean and unit variance. If $\mathcal{S}$ is the identity matrix, the squared Mahalanobis distance is the same as the squared Euclidean distance.

The sample correlation coefficient defined below is an index of similarity for continuous, ratio data that can be used with patterns but is more frequently used to measure the degree of linear dependency between two features.

$$d(j, r) = \left| \frac{(1/n) \sum_{i=1}^{n} (x_{ij} - m_j)(x_{ir} - m_r)}{s_j s_r} \right|$$

where m_j and s_j^2 are the sample mean and sample variance, respectively, for feature j and are defined in Section 2.3. The absolute value is required because a negative and a positive correlation that differ in sign but not in absolute value have the same significance when measuring similarity. If $d(j, r) = 0$, then features j and r are linearly independent. One of the features is usually discarded if $d(j, r)$ is close to 1. When data are on an ordinal scale, measures of rank correlation (Conover, 1971; Anderberg, 1973; Goodman and Kruskal, 1954) can be applied.

2.2.2 Nominal Types

If continuous, ratio-scaled data are considered to be the "strongest" type of data, then binary, nominal-scaled data are the "weakest" type. Many actual measurements, especially data collected from human subjects, are binary and nominal. Matching coefficients are proximity indices for such data. For convenience, all feature values are taken to be either 0 or 1. These symbols should be assigned consistently; if "1" means "large" for the first feature and "0" means "small," "1" must also denote "large" for all other features measuring size. Proximity indices between the ith and kth patterns are derived from the following contingency table. For example, a_{11} is the number of features that are 1 for both patterns, and a_{10} is the number of features that are 1 for pattern $\mathbf{x}_i$ and zero for pattern $\mathbf{x}_k$. The four entries sum to d, the number of features.

		$\mathbf{x}_k$: 1	$\mathbf{x}_k$: 0
$\mathbf{x}_i$	1	a_{11}	a_{10}
	0	a_{01}	a_{00}

Several measures of proximity can be defined from the four numbers $\{a_{00}, a_{01}, a_{10}, a_{11}\}$ in the contingency table for two binary vectors. Anderberg (1973) reviews most of them and puts them into context. Gower (1971) discusses the properties of general coefficients based on weighted combinations of these four numbers and shows the conditions under which proximity matrices formed from them are positive-definite matrices. Gower's index can also be used with a mixture of binary, qualitative, and quantitative features. Measures of proximity for discrete data have been proposed by Hall (1967), who described a heterogeneity function, and Bartels et al. (1970), who introduced the Calhoun distance as the percentages of patterns "between" two given patterns. Many other proximity measures have been defined for particular problems. Hubalek (1982) summarizes and evaluates proximity measures for binary vectors.

Two common matching coefficients between $\mathbf{x}_i$ and $\mathbf{x}_k$ are defined below:

1. Simple matching coefficient

$$d(i, k) = \frac{a_{00} + a_{11}}{a_{00} + a_{11} + a_{01} + a_{10}} = \frac{a_{00} + a_{11}}{d}$$

2. Jaccard coefficient

$$d(i, k) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} = \frac{a_{11}}{d - a_{00}}$$

The simple matching coefficient weights matches of 0's the same as matches of 1's, whereas the Jaccard coefficient ignores matches of 0's. The value 1 means "presence of effect" in some applications, so 1–1 matches are much more important than 0 – 0 matches. One example is that of questionnaire data. These two matching coefficients take different values for the same data and their meanings and interpretations are not obvious. Accepted practice in the area of application seems to be the best guide to a choice of proximity index.

Example 2.3

Suppose that two individuals are given psychological tests consisting of lists of 20 questions to which "yes" (1) and "no" (0) responses are required. Assuming that the questions are phrased so that "yes" and "no" have consistent interpretations, meaningful matching coefficients can be computed from the two patterns.

	Feature Number																			
	1	2	3	4	5				10				15							20
Pattern 1 ($\mathbf{x}_1$)	0	1	1	0	0	1	0	0	1	0	0	1	1	1	0	0	1	0	1	0
Pattern 2 ($\mathbf{x}_2$)	0	1	1	1	0	0	0	0	1	1	1	1	1	1	0	1	1	0	1	0

The matching coefficients are derived from the following table:

		x_2	
		1	0
x_1	1	8	1
	0	4	7

Simple matching coefficients: 15/20 = 0.75

Jaccard coefficient: 8/13 = 0.615

A value of 1 for either coefficient would mean identical patterns. However, other values are not as easily interpreted.

Example 2.4

Suppose that two partitions of nine numerals are given and a measure of their proximity is desired.

$$\mathcal{C}_1 = \{(1, 3, 4, 5), (2, 6), (7), (8, 9)\}$$

$$\mathcal{C}_2 = \{(1, 2, 3, 4), (5, 6), (7, 8, 9)\}$$

The characteristic function for a partition assigns the number 1 or 0 to a pair of numerals as follows.

$$T(i, j) = \begin{cases} 1 & \text{if numerals } i \text{ and } j \text{ are in the same subset in the partition} \\ 0 & \text{if not} \end{cases}$$

The characteristic functions T_1 and T_2, for partitions $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, are listed below in matrix form; T_1 is shown above the diagonal and T_2 is shown below the diagonal.

	1	2	3	4	5	6	7	8	9
1	—	0	1	1	1	0	0	0	0
2	1	—	0	0	0	1	0	0	0
3	1	1	—	1	1	0	0	0	0
4	1	1	1	—	1	0	0	0	0
5	0	0	0	0	—	0	0	0	0
6	0	0	0	0	1	—	0	0	0
7	0	0	0	0	0	0	—	0	0
8	0	0	0	0	0	0	1	—	1
9	0	0	0	0	0	0	1	1	—

The two characteristic functions are matched term by term to obtain the following table and coefficients. The relative significance of these values is discussed in the next section.

		T_1	
		1	0
T_2	1	4	6
	0	4	22

Simple matching coefficient: 26/36 = 0.722
Jaccard coefficient: 4/14 = 0.286

2.2.3 Missing Data

The problem of missing observations occurs often in practical applications. Suppose that some of the pattern vectors have missing feature values, as in

$$\mathbf{x}_i = (x_{i1}\ x_{i2}\ ?\ x_{i4}\ ?\ x_{i6})^{\mathrm{T}}$$

where the third and fifth features have not been recorded for the ith pattern. Missing values occur because of recording error, equipment failure, the reluctance of subjects to provide information, carelessness, and unavailability of information. Should incomplete pattern vectors be discarded? Should missing values be replaced by averages or nominal values? Answers to these questions depend on the size of the data set and the type of analysis. Sneath and Sokal (1973), Kittler (1978), Dixon (1979), and Zagoruiko and Yolkina (1982) all treat the problem of missing data.

Dixon (1979) describes several simple, inexpensive, easy to implement, and general techniques for handling missing values. These techniques either eliminate part of the data, estimate the missing values, or compute an estimated distance between two vectors with missing values. We summarize some of these techniques here.

1. Simply delete the pattern vectors or features that contain missing values. This technique does not lead to the most efficient utilization of the data and should be used only in situations where the number of missing values is very small.
2. Suppose that the jth feature value in the ith pattern vector is missing. Find the K nearest neighbors of $\mathbf{x}_i$ and replace the missing value x_{ij} by the average of the jth feature of the K nearest neighbors. The value of K should be a function of the size of the pattern matrix.
3. The distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_k$ containing missing values is computed as follows. First define the distance d_j between the two patterns along the jth feature.

$$d_j = \begin{cases} 0 & \text{if } x_{ij} \text{ or } x_{kj} \text{ is missing} \\ x_{ij} - x_{kj} & \text{otherwise} \end{cases}$$

Then the distance between $\mathbf{x}_i$ and $\mathbf{x}_k$ is written as

$$d(i, k) = \frac{d}{d - d_0} \sum d_j^2$$

where d_0 is the number of features missing in $\mathbf{x}_i$ or $\mathbf{x}_k$ or both. Note that if there are no missing values, then $d(i, k)$ defined above is the squared Euclidean distance.

4. Let $\bar{d}_j$ denote the average distance between all pairs of patterns along the jth feature defined as follows:

$$\bar{d}_j = \frac{2}{n(n-1)} \sum_{i=2}^{n} \sum_{k=1}^{i-1} |x_{ij} - x_{kj}|$$

where n is the number of patterns. Now define the distance between two patterns along the jth feature as

$$d_j = \begin{cases} \bar{d}_j & \text{if } x_{ij} \text{ or } x_{kj} \text{ is missing} \\ x_{ij} - x_{kj} & \text{otherwise} \end{cases}$$

Finally, the distance between patterns $\mathbf{x}_i$ and $\mathbf{x}_k$ is written as

$$d(i, k) = \sum d_j^2$$

Based on experimental results, Dixon (1979) recommends method 3 as the best overall method.

2.2.4 Probabilistic Indices

Goodall (1966) proposed an index of similarity that has a uniform distribution when the data are "random." The idea of using a probability scale to assess the significance of a proximity measure appears in Hamdan and Tsokos (1971), who define an information measure for a contingency table, and Brockett et al. (1981), who used the asymptotic distribution of an information-theoretic measure on questionnaire data. Li (1984) provided the most recent example of this type of measure. Before explaining the proximity measure, we reexamine the simple matching and Jaccard coefficients in light of their distributions under "random" data.

Matching coefficients measure the degree of similarity between objects. We know that their value is between 0 and 1 but do not know how large a value is required before two objects can be called "close." We now examine baseline distributions for the simple matching coefficient and the Jaccard coefficient. A baseline distribution describes a state of "randomness," or the absence of structure, for gauging the magnitude of a matching coefficient. Baseline distributions are used extensively in Chapter 4. Two vectors will be called "close" if a similarity as large as the one observed is unlikely under a baseline distribution.

The simple matching coefficient between two d-position binary vectors **a** and **b** can be expressed as

$$\text{SMC}(\mathbf{a}, \mathbf{b}) = (1/d)(\text{number of positions in which } \mathbf{a} \text{ and } \mathbf{b} \text{ match})$$

A value z of SMC can be considered to be "unusually" large if the probability of achieving a value of z or more is sufficiently low under some baseline distribution, such as the distribution of SMC for two randomly selected d-vectors. The choice of a baseline distribution is a matter of taste and depends on the application. The population Ω_{01} is defined as the set of all 4^d possible pairs of d-vectors. The probability function P_{01} assigns probability 4^{-d} to each pair in Ω_{01} and provides an obvious baseline distribution. This is equivalent to filling in the d-vectors by choosing the 1's and 0's independently with probability 1/2, as in d flips of a true coin. It is easy to show that the distribution of SMC follows the binomial distribution (Appendix B). Thus the probability that SMC is k/d or more can be written as

$$P_{01}[\text{SMC}(\mathbf{a}, \mathbf{b}) \geq k/d] = \sum_{j=k}^{d} \binom{d}{j} (1/2)^d$$

The notation

$$\binom{d}{m}$$

denotes the binomial coefficient or

$$\binom{d}{m} = \frac{d!}{m!\,(d-m)!}$$

For example, the probability that two six-position, randomly chosen binary vectors match in four or more positions (i.e., SMC $\geq 2/3$) is 0.3438, while the chance that SMC is 5/6 or more is 0.1094. Thus a value as large as 2/3 for SMC when $d = 6$ is not too unlikely even when there is no inherent correspondence between the vectors. A value of 5/6 might be required before calling the vectors unusually close under this baseline distribution. Even a perfect match has probability 0.0156, so one can never be absolutely sure that a large similarity is not a purely random event.

The Jaccard coefficient for two d-position vectors is given below, where $\mathbf{0}$ is the vector containing all zeros.

$$J(\mathbf{a}, \mathbf{b}) = \frac{\text{number of 1–1 matches}}{d - \text{Number of 0–0 matches}} \qquad \text{if } (\mathbf{a}, \mathbf{b}) \neq (\mathbf{0}, \mathbf{0})$$

$$J(\mathbf{0}, \mathbf{0}) = 1$$

The baseline distribution for J cannot be stated as compactly as that for SMC.

$$P_{01}(J \geq z) = \sum_{x \geq z} \sum\nolimits^{*} P_{01}(k, m)$$

where

$$P_{01}(k, m) = \binom{d}{m}\binom{d-m}{k}(1/2)^{d+m+k} \qquad \text{if } 0 \le m \le d \quad \text{and} \quad 0 \le k \le d-m$$

The starred sum is over the set

$$\{(k, m) : \frac{k}{d-m} = x \text{ and } 0 \le m \le d \text{ and } 0 \le k \le d-m\}$$

When $d = 6$, the chance that J is 5/6 or more is 0.0186, while the chance that SMC is 5/6 or more is 0.1094. Thus a Jaccard value of 5/6 or more is more unusual than an SMC value of 5/6 or more. Figure 2.4 shows the probabilities that the two coefficients are x or more when $d = 6$. The Jaccard coefficient can take on more values than the simple matching coefficient.

One can argue that choosing two binary vectors purely at random does not provide a sharp test of SMC or J because the population of all pairs of vectors is too large. For example, if the number of 1's is fixed in each vector, the vectors [10111011] and [01001111] have three 1–1 matches no matter how the entries of the two vectors are rearranged. The permutation statistic proposed by Li (1984) overcomes this limitation. It measures the correspondence between two binary vectors, just as SMC and J. However, it can be interpreted directly because it has a uniform distribution over the interval [0, 1] under a baseline distribution, shown as S in Figure 2.4. Note that the distributions of SMC and J in Figure 2.4 are under P_{01}, while the distribution of S is under a baseline distribution based on random permutations and described below.

Suppose that a measure of correspondence between binary d-vectors **a** and **b** is to be defined and the vectors are treated as dichotomous, so 0–0 matches

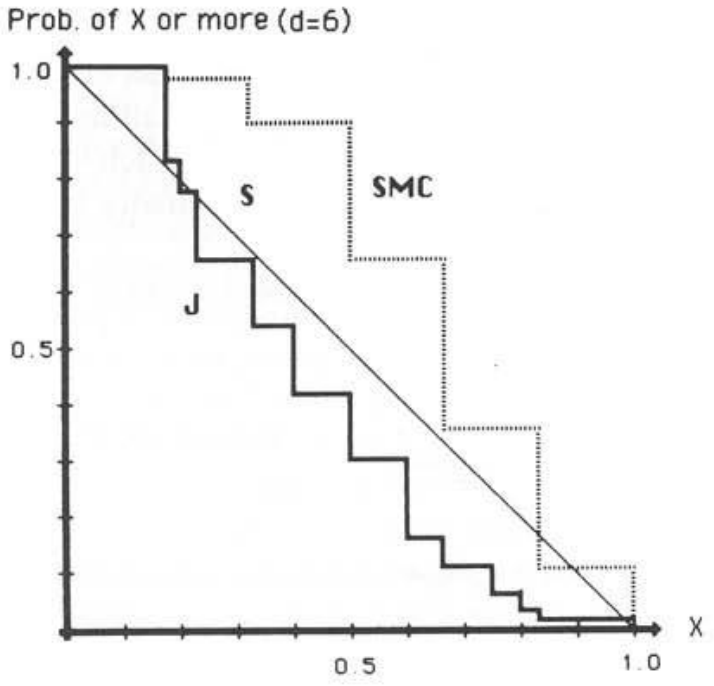

Figure 2.4 Baseline distributions of matching coefficients.

are not as important as 1–1 matches. Consider the population Ω_{02} of all $d!$ pairs of vectors that can be obtained by permuting the entries of one of the vectors. Not all pairs of vectors are distinct. Probability function P_{02} assigns each pair of vectors probability mass $1/d!$, thus establishing a new baseline distribution.

Let A_{11} be the number of 1–1 matches in a randomly selected pair of vectors from population Ω_{02}. Let N_a be the number of 1's in **a** and let N_b be the number of 1's in **b.** All pairs of vectors in Ω_{02} have N_a and N_b 1's. For example, there are six 1's in [10111011]. The probability that $A_{11} = k$ can be obtained from the hypergeometric distribution (Appendix B) under P_{02}. In the notation of Appendix B, we have a population of size d with N_b defectives and we take a sample of size N_a. Of course, the roles of N_a and N_b can be reversed. The probability of exactly k matches between pairs of 1's is

$$P_{02}(A_{11} = k) = \frac{\binom{N_b}{k}\binom{d - N_b}{N_a - k}}{\binom{d}{N_a}} = H(k, N_a, N_b, d)$$

This probability expression requires that

$$\max\{0, N_a + N_b - d)\} \leq k \leq \min\{N_a, N_b\}$$

The S-statistic defined below is essentially the inverse of the hypergeometric cumulative density function. Such statistics have been used elsewhere (Kempthorne, 1952). The additive factor ensures that S has a (continuous) uniform distribution over the unit interval since U is a continuous uniform random variable over the unit interval. If t is the number of 1–1 matches observed between d-vectors **a** and **b,** the S-measure of proximity is

$$S(\mathbf{a}, \mathbf{b}) = \sum_{k<t} H(d, N_a, N_b, k) + H(d, N_a, N_b, t)U$$

Since the distribution of S is uniform under P_{02}, the value of S is implicitly meaningful. For example, the probability that S is z or more is $1 - z$ for z between 0 and 1, as shown in Figure 2.4. This proximity has been used in the analysis of questionnaire data (Li and Dubes, 1984) and in a template-matching problem (Li and Dubes, 1985). The additive factor does not contribute much to the value of S except when d is small.

2.3 NORMALIZATION

Suppose that the raw data consist of an $n \times d$ pattern matrix in which all features are continuous and on a ratio scale. Raw data, or the actual measurements, are seldom used just as they are recorded unless a probabilistic model for pattern generation is available. Some normalization is usually employed based on the requirements of the analysis. Preparing the data for a cluster analysis requires

some sort of normalization that takes into account the measure of proximity. For example, Euclidean distance is a popular and familiar index of dissimilarity, but it implicitly assigns more weighting to features with large ranges than to those with small ranges. Scaling one feature in miles and a second feature in inches makes the second feature numerically overpower the first. We present a normalization scheme that remedies some of these problems.

As explained earlier in this section, the basic unit of data is called a pattern, denoted by a d-vector, whose components are scalars called features. The ith pattern is denoted by the (column) vector $\mathbf{x}_i^*$ in this section and the jth feature value for the ith pattern is denoted by x_{ij}^*. The asterisk denotes "raw" or unnormalized data. If n is the number of patterns in the analysis, the pattern matrix is the $n \times d$ matrix $\mathcal{A}^*$:

$$\mathcal{A}^* = [\mathbf{x}_1^* \quad \mathbf{x}_2^* \quad \cdots \quad \mathbf{x}_n^*]^{\mathrm{T}} = \begin{bmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1d}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2d}^* \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1}^* & x_{n2}^* & \cdots & x_{nd}^* \end{bmatrix}$$

Each row of $\mathcal{A}^*$ is a pattern. Each point in the pattern space is a potential pattern. We treat the case when $n > d$, so the patterns are visualized as a number of points scattered around the pattern space.

The jth feature average, m_j, and jth feature variance, s_j^2, are defined as the sample mean and the sample variance for the jth feature.

$$m_j = (1/n) \sum_{i=1}^{n} x_{ij}^*$$

$$s_j^2 = (1/n) \sum_{i=1}^{n} (x_{ij}^* - m_j)^2$$

The simplest type of normalization subtracts the feature means:

$$x_{ij} = x_{ij}^* - m_j \tag{2.1}$$

This normalization makes feature values invariant to rigid displacements of the coordinates. The second type of normalization translates and scales the axes so that all the features have zero mean and unit variance:

$$x_{ij} = \frac{x_{ij}^* - m_j}{s_j} \tag{2.2}$$

Removing the asterisk indicates that the pattern has been normalized, but the type of normalization must be clear from the context. Other types of normalization include scaling by the range (Carmichael et al., 1968) and a heterogeneity measure (Hall, 1969). Lumelsky (1982) incorporates the normalization into the clustering procedure. Normalization or scaling is not always desirable. For example, if the spread among the patterns is due to the presence of clusters, the normalization in Eq. (2.2) can change the interpoint distances and can alter the separation between natural clusters as demonstrated in Figure 2.5.

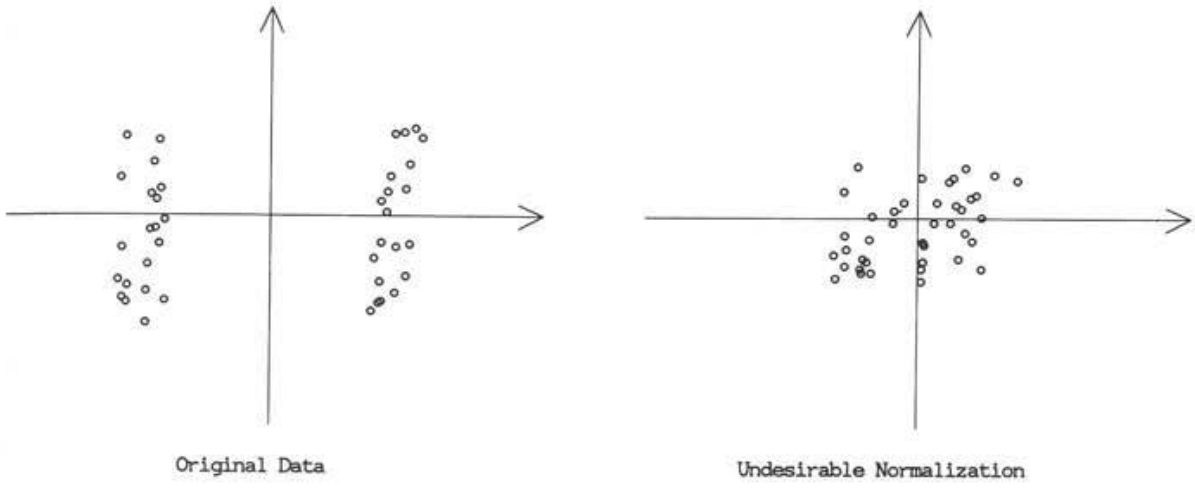

Figure 2.5 Effect of normalization.

The $d \times d$ matrix $\mathcal{R} = [r_{ij}]$ is defined below in terms of normalized data.

$$\mathcal{R} = (1/n)\, \mathcal{A}^{\mathrm{T}} \mathcal{A}$$

where

$$r_{ij} = (1/n) \sum_{k=1}^{n} x_{ki} x_{kj} \tag{2.3}$$

Under Eq. (2.1), $\mathcal{R}$ is a (sample) *covariance matrix*. Under Eq. (2.2), r_{ij} is the sample correlation coefficient between features i and j and $r_{jj} = 1$ for all j; $\mathcal{R}$ is then called a *correlation matrix*. The entries of $\mathcal{R}$ can be interpreted as relative spreads in the following sense. Each pattern is pictured as a unit of mass placed in the pattern space. The matrix $\mathcal{A}$ is pictured as a swarm of points in the pattern space (Figure 2.5), each point having the same mass and representing a row of $\mathcal{A}$. The normalizations of Eqs. (2.1) and (2.2) make the diagonal elements of $\mathcal{R}$ the moments of inertia of the swarm about the coordinate axes and force the origin of the coordinate system to coincide with the sample mean. Equation (2.2) makes all the moments of inertia unity. Another type of normalization that rotates the coordinate axes is discussed in Appendix C and is used in linear projection of the data (Section 2.4).

2.4 LINEAR PROJECTIONS

Projection algorithms map a set of n d-dimensional patterns onto an m-dimensional space, where $m < d$. The main motivation for studying projection algorithms in the context of cluster analysis is to permit visual examination of multivariate data, so $m = 2$ in our discussion. When a reasonably accurate two-dimensional representation of a set of patterns can be obtained, one can cluster by eye and qualitatively validate conclusions drawn from clustering algorithms. This search

for a two-dimensional projection is closely related to problems in multivariate analysis of variance and factor analysis (see Appendices E and F).

A linear projection expresses the m new features as linear combinations of the original d features.

$$\mathbf{y}_i = \mathcal{H}\mathbf{x}_i \qquad \text{for } i = 1, \ldots, n$$

Here, $\mathbf{y}_i$ is an m-place column vector, $\mathbf{x}_i$ is a d-place column vector, and $\mathcal{H}$ is an $m \times d$ matrix. Linear projection algorithms are relatively simple to use, tend to preserve the character of the data, and have well-understood mathematical properties. The type of linear projection used in practice is influenced by the availability of category information about the patterns in the form of labels on the patterns. If no category information is available, the eigenvector projection (also called the Karhunen–Loeve method, or principal component method) is commonly used. Discriminant analysis is a popular linear mapping technique when category labels are available. We now describe these two popular linear projection algorithms. Readers will find some useful background information on linear algebra and scatter matrices in Appendices C and D.

2.4.1 Eigenvector Projection

The eigenvectors of the covariance matrix $\mathcal{R}$ in Eq. (2.3) define a linear projection that replaces the features in the raw data with uncorrelated features. These eigenvectors also provide a link between cluster analysis and factor analysis (Appendix E). Since $\mathcal{R}$ is a $d \times d$ positive definite matrix, its eigenvalues are real and can be labeled so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$$

A set of corresponding eigenvectors, $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_d$, is labeled accordingly. The $m \times d$ matrix of transformation $\mathcal{H}_m$ is defined from the eigenvectors of the covariance matrix (or correlation matrix) $\mathcal{R}$ as follows. The eigenvectors are also called principal components.

$$\mathcal{H}_m = \begin{bmatrix} \mathbf{c}_1^{\mathrm{T}} \\ \mathbf{c}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{c}_m^{\mathrm{T}} \end{bmatrix}$$

The rows of $\mathcal{H}_m$ are eigenvectors, as are the rows of $\mathcal{C}_R$ defined in Appendix C. This matrix projects the pattern space into an m-dimensional subspace (hence the subscript m on $\mathcal{H}_m$) whose axes are in the directions of the largest eigenvalues of $\mathcal{R}$ as follows. The derivation is given in Section 2.4.2.

$$\mathbf{y}_i = \mathcal{H}_m\mathbf{x}_i \qquad \text{for } i = 1, \ldots, n \tag{2.4}$$

The projected patterns can be written as follows:

$$\mathcal{B}_m = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \mathcal{H}_m^T = \mathcal{A}\mathcal{H}_m^T$$

Note that $\mathbf{x}_i$ is the original pattern and $\mathbf{y}_i$ is the corresponding projected pattern. Equation (2.4) will be called the *eigenvector transformation*.

The covariance matrix in the new space can be defined with Eq. (C.1) as follows:

$$(1/n)\mathcal{B}_m^T\mathcal{B}_m = (1/n)\sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i^T = \mathcal{H}_m\mathcal{R}\mathcal{H}_m^T = \mathcal{H}_m\mathcal{C}_R^T\Lambda_R(\mathcal{H}_m\mathcal{C}_R^T)^T$$

The matrix $\mathcal{H}_m\mathcal{C}_R^T$ can be partitioned as follows, where $\mathcal{I}$ is an $m \times m$ identity matrix and $\mathcal{O}$ is an $m \times (d - m)$ zero matrix.

$$\mathcal{H}_m\mathcal{C}_R^T = [\mathcal{I}|\mathcal{O}]$$

Thus the covariance matrix in the new space, Λ_m, becomes a diagonal matrix as shown below.

$$(1/n)\mathcal{B}_m^T\mathcal{B}_m = \Lambda_m = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_m) \tag{2.5}$$

This implies that the m new features obtained by applying the linear transformation defined by $\mathcal{H}_m$ are uncorrelated.

Techniques for choosing an appropriate value for m in Eq. (2.4) are based on the eigenvalues of $\mathcal{R}$. Comparing Eqs. (C.2) and (2.5) shows that the sum of the first m eigenvalues is the "variance" retained in the new space. That is, the eigenvalues of $\mathcal{R}$ are the sample variances in the new space, while the sum of the d eigenvalues is the total variance in the original pattern space. Since the eigenvalues are ordered largest first, one could choose m so that

$$r_m = \sum_{i=1}^{m} \lambda_i \Big/ \sum_{i=1}^{d} \lambda_i \geq 0.95$$

which would assure that 95% of the variance is retained in the new space. Thus a "good" eigenvector projection is that which retains a large proportion of the variance present in the original feature space with only a few features in the transformed space. Krzanowski (1979) provides a table for the distribution of this ratio for $m = 1$, $d = 3$ and 4, and several values of n under the assumption that all components of all patterns are samples from independent standard normal distributions. These tables should help determine whether it is reasonable to say that the size of the largest eigenvalue could have been achieved by chance. Another technique for choosing m is to plot r_m as a function of m and look for a "knee" in the plot.