Similarity Measures

Most efforts to produce a rather simple group structure from a complex data set require a measure of "closeness," or "similarity." There is often a great deal of subjectivity involved in the choice of a similarity measure. Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (nominal, ordinal, interval, ratio), and subject matter knowledge.

When *items* (units or cases) are clustered, proximity is usually indicated by some sort of distance. By contrast, *variables* are usually grouped on the basis of correlation coefficients or like measures of association

Distances and Similarity Coefficients for Pairs of Items

We have already discussed the notion of distance. Recall that the Euclidean (straight-line) distance between two p-dimensional observations (items) $\mathbf{x}' = [x_1, x_2, \ldots, x_p]$ and $\mathbf{y}' = [y_1, y_2, \ldots, y_p]$ is

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \\ &= \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})} \end{aligned}$$

The statistical distance between the same two observations is of the form

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'\mathbf{A}(\mathbf{x} - \mathbf{y})}$$

Ordinarily, $\mathbf{A} = \mathbf{S}^{-1}$, where $\mathbf{S}$ contains the sample variances and covariances. However, without prior knowledge of the distinct groups, these sample quantities cannot be computed. For this reason, Euclidean distance is often preferred for clustering.

Another distance measure is the Minkowski metric

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m \right]^{1/m}$$

For $m = 1$, $d(\mathbf{x}, \mathbf{y})$ measures the "city-block" distance between two points in p dimensions. For $m = 2$, $d(\mathbf{x}, \mathbf{y})$ becomes the Euclidean distance. In general, varying m changes the weight given to larger and smaller differences.

Two additional popular measures of "distance" or dissimilarity are given by the Canberra metric and the Czekanowski coefficient. Both of these measures are defined for nonnegative variables only. We have

$$\text{Canberra metric:} \qquad d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

$$\text{Czekanowski coefficient:} \qquad d(\mathbf{x}, \mathbf{y}) = 1 - \frac{2 \sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}$$

Whenever possible, it is advisable to use "true" distances—that is, distances satisfying the "usual" (or "customary") distance properties—for clustering objects. On the other hand, most clustering algorithms will accept subjectively assigned distance numbers that may not satisfy, for example, the triangle inequality.

When items cannot be represented by meaningful p-dimensional measurements, pairs of items are often compared on the basis of the presence or absence of certain characteristics. Similar items have more characteristics in common than do dissimilar items. The presence or absence of a characteristic can be described mathematically by introducing a *binary variable*, which assumes the value 1 if the characteristic is present and the value 0 if the characteristic is absent. For $p = 5$ binary variables, for instance, the "scores" for two items i and k might be arranged as follows:

	Variables				
	1	2	3	4	5
Item i	1	0	0	1	1
Item k	1	1	0	1	0

In this case, there are two 1–1 matches, one 0–0 match, and two mismatches.

Let x_{ij} be the score (1 or 0) of the jth binary variable on the ith item and x_{kj} be the score (again, 1 or 0) of the jth variable on the kth item, $j = 1, 2, \ldots, p$. Consequently,

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{if } x_{ij} = x_{kj} = 1 \quad \text{or} \quad x_{ij} = x_{kj} = 0 \\ 1 & \text{if } x_{ij} \neq x_{kj} \end{cases}$$

and the squared Euclidean distance, $\sum_{j=1}^{p} (x_{ij} - x_{kj})^2$ provides a count of the number of mismatches. A large distance corresponds to many mismatches—that is, dissimilar items. From the preceding display, the square of the distance between items i and k would be

$$\begin{aligned} \sum_{j=1}^{5} (x_{ij} - x_{kj})^2 &= (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 \\ &= 2 \end{aligned}$$

Although a distance based on such scoring might be used to measure similarity, it suffers from weighting the 1–1 and 0–0 matches equally. In some cases, a 1–1 match is a stronger indication of similarity than a 0–0 match. For instance, in grouping people, the evidence that two persons both read ancient Greek is stronger evidence of similarity than the absence of this ability. Thus, it might be reasonable to discount the 0–0

matches or even disregard them completely. To allow for differential treatment of the 1–1 matches and the 0–0 matches, several schemes for defining similarity coefficients have been suggested.

To introduce these schemes, let us arrange the frequencies of matches and mismatches for items i and k in the form of a contingency table:

		Item k		
		1	0	Totals
Item i	1	a	b	$a+b$
	0	c	d	$c+d$
Totals		$a+c$	$b+d$	$p=a+b+c+d$

In this table, a represents the frequency of 1–1 matches, b is the frequency of 1–0 matches, and so forth. Given the foregoing five pairs of binary outcomes, $a = 2$ and $b = c = d = 1$.

The following table lists common similarity coefficients defined in terms of the frequencies from the table above.

Similarity Coefficients for Clustering Items*		
	Coefficient	Rationale
1.	$\frac{a+d}{p}$	Equal weights for 1–1 matches and 0–0 matches.
2.	$\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for 1–1 matches and 0–0 matches.
3.	$\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4.	$\frac{a}{p}$	No 0–0 matches in numerator.
5.	$\frac{a}{a+b+c}$	No 0-0 matches in numerator or denominator. (The 0–0 matches are treated as irrelevant)
6.	$\frac{2a}{2a+b+c}$	No 0–0 matches in numerator or denominator. Double weight for 1–1 matches.
7.	$\frac{a}{a+2(b+c)}$	No 0–0 matches in numerator or denominator. Double weight for unmatched pairs.
8.	$\frac{a}{b+c}$	Ratio of matches to mismatches with 0–0 matches excluded

* [p binary variables; see the preceding table.]

Coefficients 1, 2, and 3 in the table are monotonically related. Suppose coefficient 1 is calculated for two contingency tables, Table I and Table II. Then if

$$\frac{a_I + d_I}{p} \geq \frac{a_{II} + d_{II}}{p}$$

we also have

$$\frac{2(a_I + d_I)}{2(a_I + d_I) + b_I + c_I} \geq \frac{2(a_{II} + d_{II})}{2(a_{II} + d_{II}) + b_{II} + c_{II}}$$

and coefficient 3 will be at least as large for Table I as it is for Table II. Coefficients 5, 6, and 7 also retain their relative orders.

Monotonicity is important, because some clustering procedures are not affected if the definition of similarity is changed in a manner that leaves the relative orderings of similarities unchanged. The single linkage and complete linkage hierarchical procedures (these will be discussed later) are not affected. For these methods, any choice of the coefficients 1, 2, and 3 in the similarity coefficients table will produce the same groupings. Similarly, any choice of the coefficients 5, 6, and 7 will yield identical groupings.

Example: Calculating the values of a similarity coefficient Suppose five individuals possess the following characteristics:

	Height	Weight	Eye color	Hair color	Handedness	Gender
Individual 1	68 in	140 lb	green	blond	right	female
Individual 2	73 in	185 lb	brown	brown	right	male
Individual 3	67 in	165 lb	blue	blond	right	male
Individual 4	64 in	120 lb	brown	brown	right	female
Individual 5	76 in	210 lb	brown	brown	left	male

Define six binary variables, $X_1, X_2, X_3, X_4, X_5, X_6$ as

$$X_1 = \begin{cases} 1 & \text{height } \geq 72 \text{ in.} \\ 0 & \text{height } < 72 \text{ in.} \end{cases} \qquad X_4 = \begin{cases} 1 & \text{blond hair} \\ 0 & \text{not blond hair} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{weight } \geq 150 \text{ lb} \\ 0 & \text{weight } < 150 \text{ lb} \end{cases} \qquad X_5 = \begin{cases} 1 & \text{right handed} \\ 0 & \text{left handed} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{brown eyes} \\ 0 & \text{otherwise} \end{cases} \qquad X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

The scores for individuals 1 and 2 on the $p = 6$ binary variables are

		X_1	X_2	X_3	X_4	X_5	X_6
Individual	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

and the number of matches and mismatches are indicated in the two-way array

		Individual 2		
		1	0	Total
Individual 1	1	1	2	3
	0	3	0	3
	Totals	4	2	6

Employing similarity coefficient 1, which gives equal weight to matches, we compute

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}$$

Continuing with similarity coefficient 1, we calculate thee remaining similarity numbers for pairs of individuals. These are displayed in the 5×5 symmetric matrix

$$\begin{array}{c} \text{Individuals} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 1 & & & & \\ \frac{1}{6} & 1 & & & \\ \frac{4}{6} & \frac{3}{6} & 1 & & \\ \frac{4}{6} & \frac{3}{6} & \frac{2}{6} & 1 & \\ 0 & \boxed{\frac{5}{6}} & \frac{2}{6} & \frac{2}{6} & 1 \end{bmatrix} \end{array}$$

Based on the magnitudes of the similarity coefficient, we should conclude that individuals 2 and 5 are most similar and individuals 1 and 5 are least similar. Other pairs fall between these extremes. If we were to divide the individuals into two relatively homogeneous subgroups on the basis of the similarity numbers, we might form the subgroups (1 3 4) and (2 5).

Note that $X_3 = 0$ implies an absence of brown eyes, so that two people, one with blue eyes and one with green eyes, will yield a 0–0 match. Consequently, it may be inappropriate to use similarity coefficient 1, 2, or 3 because these coefficients give the same weights to 1–1 and 0–0 matches. □

We have described the construction of distances and similarities. It is always possible to construct similarities from distances. For example, we might set

$$\widetilde{s}_{ik} = \frac{1}{1 + d_{ik}}$$

where $0 < \widetilde{s}_{ik} \leq 1$ is the similarity between items i and k, and d_{ik} is the corresponding distance.

However, distances that must satisfy the distance properties cannot always be constructed from similarities. As Grower has shown, this can be done only if the matrix of similarites is nonnegative definite. With the nonnegative definite condition, and with the maximum similarity added so that $\widetilde{s}_{ij} = 1$,

$$d_{ik} = \sqrt{2(1 - \widetilde{s}_{ik})}$$

has the properties of a distance.

Similarities and Association Measures for Pairs of Variables

Thus far, we have discussed similarity measures for items. In some applications, it is the variables, rather than the items, that must be grouped. Similarity measures for variables often take the form of sample correlation coefficients. Moreover, in some clustering applications, negative correlations are replaced by their absolute values.

When the variables are binary, the data can again be arranged in the form of a contingency table. This time, however, the variables, rather than the items, delineate the categories. For each pair of variables, there are n items categorized in the table. With the usual 0 and 1 coding, the table becomes as follows:

		Variable k		
		1	0	Totals
Variable i	1	a	b	$a+b$
	0	c	d	$c+d$
Totals		$a+c$	$b+d$	$p=a+b+c+d$

For instance, variable i equals 1 and variable k equals 0 for b of the n items.

The usual product moment correlation formula applied to the binary variables in the contingency table above, gives:

$$r = \frac{ad - bc}{[(a+b)(c+d)(a+c)(b+d)]^{1/2}}$$

This number can be taken as a measure of the similarity between the two variables.

The correlation coefficient r is related to the chi-square statistic ($r^2 = \chi^2/n$) for testing the independence of two categorical variables. For n fixed, a large similarity (or correlation) is consistent with the presence of dependence.

Given the previous table, measures of association (or similarity) exactly analogous to the ones listed in the similarity coefficients table can be developed. The only change required is the substitution of n (the number of items) for p (the number of variables).

Concluding Comments on Similarity

To summarize, we note that there are many ways to measure the similarity between pairs of objects. It appears that most practitioners use distances or coefficients, to cluster *items* and correlations to cluster *variables*. However, at times, inputs to clustering algorithms may be simple frequencies.

Example: Measuring the similarities of 11 languages The meanings of words change with the course of history. However, the meaning of the numbers 1, 2, 3,... represents one conspicuous exception. Thus, a first comparison of languages might be based on the numerals alone. The following table gives the first 10 numbers in English, Polish, Hungarian, and eight other modern European languages. (Only languages that use the Roman alphabet are considered, and accent marks, cedillas, diereses, etc., are omitted.) A cursory examination of the spelling of the numerals in the table suggests that the first five languages (English, Norwegian, Danish, Dutch, and German) are very much alike. French, Spanish, and Italian are in even closer agreement. Hungarian and Finnish seem to stand by themselves, and Polish has some of the characteristics of the languages in each of the larger subgroups.

Numerals in 11 Languages

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuattro	quattro	cztetry	negy	nelja
five	fem	fem	vijf	fuenf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

Concordant First Letters for Numerals in 11 Languages											
	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du	3	5	4	10							
G	4	6	5	5	10						
Fr	4	4	4	1	3	10					
Sp	4	4	5	1	3	8	10				
I	4	4	5	1	3	9	9	10			
P	3	3	4	0	2	5	7	6	10		
H	1	2	2	2	1	0	0	0	0	10	
Fi	1	1	1	1	1	1	1	1	1	2	10

The words for 1 in French, Spanish, and Italian all begin with *u*. For illustrative purposes, we might compare languages by looking at the *first letters* of the numbers. We call the words for the same number in two different languages *concordant* if they have the same first letter and *discordant* if they do not. From the table with the whole words for naming numbers, the table of concordances (frequencies of matching first initials) for the numbers 1–10 is given in the table following it. For example, we see that English and Norwegian have the same first letter for 8 of the 10 word pairs.

The results confirm our initial visual impression, that English, Norwegian, Danish, Dutch, and German seem to form a group. French, Spanish, Italian, and Polish might be grouped together, whereas Hungarian and Finnish appear to stand alone. □

In our examples so far, we have used our visual impression of similarity or distance measures to form groups. We now discuss less subjective schemes for creating clusters.

Hierarchical Clustering Methods

We can rarely examine all grouping possibilities, even with the largest and fastest computers. Because of this problem, a wide variety of clustering algorithms have emerged that find "reasonable" clusters without having to look at all configurations.

Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. *Agglomerative hierarchical methods* start with the individual objects. Thus, there are initially as many clusters as objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster.

Divisive hierarchical methods work in the opposite direction. An initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" the objects in the other. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many subgroups as objects—that is, until each object forms a group.

The results of both agglomerative and divisive methods may be displayed in the form of a two-dimensional diagram known as *dendrogram*. As we shall see, the dendrogram illustrates the mergers or divisions that have been made at successive levels.

We shall concentrate on agglomerative hierarchical procedures and, in particular, *linkage methods*.

Linkage methods are suitable for clustering items, as well as variables. This is not true for all hierarchical agglomerative procedures. We shall discuss, in turn, *single linkage* (minimum distance or nearest neighbor), *complete linkage* (maximum distance or farthest neighbor) and *average linkage* (average distance). The merging of clusters under the three linkage criteria is illustrated schematically in the figure that follows.

From the figure, we shall see that single linkage results when groups are fused according to the distance between their nearest members. Complete linkage occurs when groups are fused according to the distance between their farthest members. For average linkage, groups are fused according to the average distance between pairs of members in the respective sets.

The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects (items or variables):

1. Start with N clusters, each containing a single entity and $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = \{d_{ik}\}$.

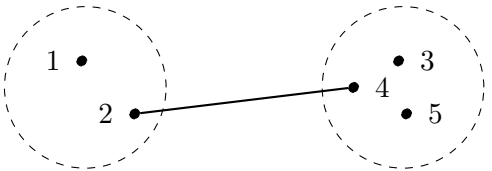
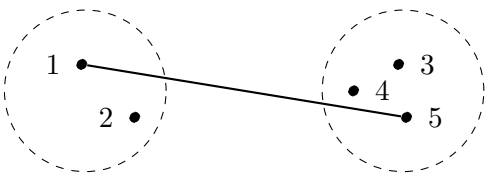
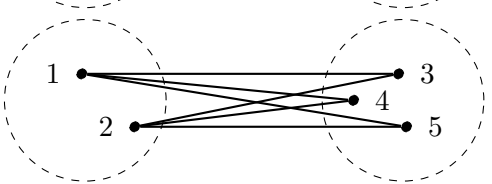
2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{UV}.

3. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by

   (a) deleting the rows and columns corresponding to clusters U and V and

   (b) adding a row and column giving the distances between cluster (UV) and the remaining clusters

4. Repeat steps 2 and 3 a total of $N - 1$ times. (All objects will be in a *single* cluster after the algorithm terminates.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

		Cluster distance
1.		d_{24}
2.		d_{15}
3.		$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$

Intercluster distance (dissimilarity) for

1. single linkage
2. complete linkage
3. average linkage

The ideas behind any clustering procedure are probably best conveyed through examples, which we shall present after brief discussions of the input and algorithmic components of the linage methods.

Single Linkage

The inputs to a single linkage algorithm can be distances or similarities between pairs of objects. Groups are formed from the individual entities by merging nearest neighbors, where the term *nearest neighbor* connotes the smallest distance or largest similarity.

Initially, we must find the smallest distance in $\mathbf{D} = \{d_{ik}\}$ and merge the corresponding objects, say, U and V, to get the cluster (UV). For Step 3 of the general algorithm, the distances between (UV) and any other cluster W are computed by

$$d_{(UV)W} = \min\{d_{UW}, d_{VW}\}$$

Here the quantities d_{UW} and d_{VW} are the distances between the nearest neighbors of clusters U and W and clusters V and W, respectively.

The results of single linkage clustering can be graphically displayed in the form of a *dendrogram*, or tree diagram. The branches in the tree represent clusters. The branches come together (merge) at nodes whose positions along a distance or similarity) axis indicate the level at which the fusions occur. Dendrograms for some specific cases are considered in the following examples.

Example: Clustering using single linkage To illustrate the single linkage algorithm, we consider a hypothetical distances between pairs of five objects as follows:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

Treating each object as a cluster, we commence clustering by merging the two closest items. Since

$$\min_{i,k}(d_{ik}) = d_{5\,3} = 2$$

objects 5 and 3 are merged to from the cluster (3 5). To implement the next level of clustering, we need the distances between the cluster (3 5) and the remaining objects, 1, 2 and 4. The nearest neighbor distances are

$$\begin{aligned} d_{(35)1} &= \min\{d_{31}, d_{51}\} = \min\{3, 11\} &= 3 \\ d_{(35)2} &= \min\{d_{32}, d_{52}\} = \min\{7, 10\} &= 7 \\ d_{(35)4} &= \min\{d_{34}, d_{54}\} = \min\{9,\ 8\} &= 8 \end{aligned}$$

Deleting the rows and columns of **D** corresponding to objects 3 and 5, and adding a row and column for the cluster (3 5), we obtain the new distance matrix

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ \textcircled{3} & 0 & & \\ 7 & 9 & 0 & \\ 8 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

The smallest distance between pairs of clusters is now $d_{(35)1} = 3$, and we merge cluster (1) with cluster (3 5) to get the next cluster, (1 3 5). Calculating

$$\begin{aligned} d_{(135)2} &= \min\{d_{(35)2}, d_{12}\} = \min\{7, 9\} &= 7 \\ d_{(135)4} &= \min\{d_{(35)4}, d_{14}\} = \min\{8, 6\} &= 6 \end{aligned}$$

we find that the distance matrix for the next level of clustering is

$$\begin{array}{c} \\ (135) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (135) & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

The minimum nearest neighbor distance between pairs of clusters is $d_{42} = 5$, and we merge objects 4 and 2 to get the cluster (42).

At this point we have two distinct clusters, (1 3 5) and (2 4). Their nearest neighbor distance is

$$d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7, 6\} = 6$$

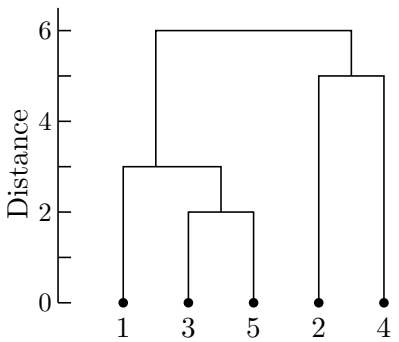
The final distance matrix becomes

$$\begin{array}{c} \\ (135) \\ (24) \end{array} \begin{array}{c} \begin{array}{cc} (135) & (24) \end{array} \\ \begin{bmatrix} 0 & \\ \textcircled{6} & 0 \end{bmatrix} \end{array}$$

Consequently, clusters (1 3 5) and (2 4) are merged to form a single cluster of all five objects, (1 2 3 4 5), when the nearest neighbor distance is 6.

The dendrogram picturing the hierarchical clustering just concluded is shown in the figure.

The groupings and the distance levels at which they occur are clearly illustrated by the dendrogram. □



Single linkage dendrogram for distances between five objects

In typical applications of hierarchical clustering, the intermediate results—where the objects are sorted into a moderate number of clusters—are of chief interest.

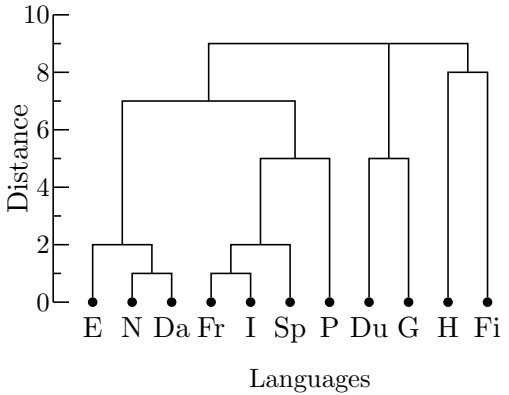
Example: Single linkage clustering of 11 languages Let us return to our table of concordant first letters for numbers, from the example before last, and consider the array of concordances, representing the closeness between the numbers 1–10 in 11 languages. To develop a matrix of distances, we subtract the concordances from the perfect agreement figure of 10 that each language has with itself. The subsequent assignments of distances are

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	0										
N	2	0									
Da	2	(1)	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
Sp	6	6	5	9	7	2	0				
I	6	6	5	9	7	(1)	(1)	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0

We first search for the minimum distance between pairs of languages (clusters). The minimum distance, 1, occurs between Danish and Norwegian, Italian and French, and Italian and Spanish. Numbering the languages in the order in which they appear across the top of the array, we have

$$d_{32} = 1 \qquad d_{86} = 1 \qquad \text{and } d_{87} = 1$$

Since $d_{76} = 2$, we can merge onl clusters 8 and 6 or clusters 8 and 7. We cannot merge clusters 6, 7, and 8 at level 1. We choose first to merge 6 and 8, and then to update the distance matrix and merge 2 and 3 to obtain the clusters (6 8) and (2 3). Subsequent computer calculations produce the dendrogram as shown below.



Single linkage dendrogram for distances between numbers in 11 languages

From the dendrogram, we see that Norwegian and Danish, and also French and Italian, cluster at the minimum distance (maximum similarity) level. When the allowable distance is increased, English is added to the Norwegian–Danish group, and Spanish merges with the French–Italian group. Notice that Hungarian and Finnish are more similar to each other than to the other clusters of languages. However, these two clusters (languages) do not merge until the distance between the nearest neighbors has increased substantially. Finally, all the clusters of languages are merged into a single cluster at the largest nearest neighbor distance, 9. □

Since single linkage joins clusters by the shortest link between them, the technique cannot discern poorly separated clusters. On the other hand, single linkage is one of he few clustering methods that can delineate nonellipsoidal clusters. The tendency of single linkage to pick out long stringlike clusters is known as *chainig*. Chaining can be

misleading if items at opposite ends of the chain are, in fact, quite dissimilar.

The clusters formed by the single linkage method will be unchanged by any assignment of distance (similarity) that gives the same relative orderings as the initial distances (similarities). In particular, any one of a set of similarity coefficients that are monotonic to one another will produce the same clustering.

Complete Linkage

Complete linkage clustering proceeds in much the same manner a single linking clusterings, with one important exception: At each stage, the distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, that are *most distant*. Thus, complete linkage ensures that all items in a cluster are within some maximum distance (or minimum similarity) of each other.

The general agglomerative algorithm again starts by finding the minimum entry in $\mathbf{D} = \{d_{ij}\}$ and merging the corresponding objects, such as U and V, to get cluster $(U\,V)$. For Step 3 of the general algorithm presented before, the distances between $(U\,V)$ and any other cluster W are computed by

$$d_{(U\,V)W} = \max\{d_{UW}, d_{VW}\}$$

Here d_{UW} and d_{VW} are the distances between the most distant members of clusters U and W and clusters V and W, respectively.

Example: Clustering using complete linkage Let us return to the distance matrix introduced in a previous example:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

At the first stage, objects 3 and 5 are merged, since they are most similar. This gives the cluster (3 5). At stage 2, we compute

$$\begin{aligned} d_{(3\,5)1} &= \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11 \\ d_{(3\,5)2} &= \max\{d_{32}, d_{52}\} = 10 \\ d_{(3\,5)4} &= \max\{d_{34}, d_{54}\} = 9 \end{aligned}$$

and the modified distance matrix becomes

$$\begin{array}{c} \\ (3\,5) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (3\,5) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & \textcircled{5} & 0 \end{bmatrix} \end{array}$$

The next merger occurs between the most similar groups, 2 and 4, to give the cluster (2 4). At stage 3, we have

$$d_{(2\,4)(3\,5)} = \max\{d_{2(3\,5)}, d_{4(3\,5)}\} = \max\{10, 9\} = 10$$
$$d_{(2\,4)1} = \max\{d_{21}, d_{41}\} = 9$$

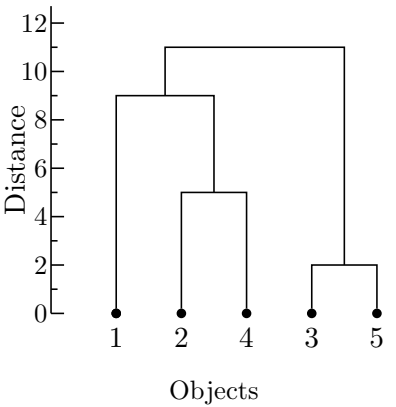
and the distance matrix

$$\begin{array}{c} \\ (3\,5) \\ (2\,4) \\ 1 \end{array} \begin{array}{c} \begin{array}{ccc} (3\,5) & (2\,4) & 1 \end{array} \\ \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & \textcircled{9} & 0 \end{bmatrix} \end{array}$$

The next merger produces the cluster (1 2 4). At the final stage, the groups (3 5) and (1 2 4) are merged as the single cluster (1 2 3 4 5) at level

$$d_{(1\,2\,4)(3\,5)} = \max\{d_{1(3\,5)}, d_{(2\,4)(3\,5)}\}$$
$$= \max\{11, 10\} = 11$$

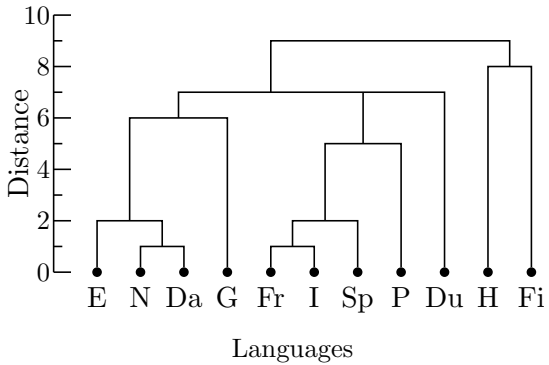
The dendrogram is given in the figure.



Complete linkage dendrogram for distances between five objects □

Comparing the two dendrograms for the five-objects examples, we see that the dendrograms for single linkage and complete linkage differ in the allocation of object 1 to previous groups.

Example: Complete linkage clustering of 11 languages In one of our previous examples, we presented a distance matrix for numbers in 11 languages. The complete linkage clustering algorithm applied to this distance matrix produces the dendrogram shown in the following figure:



Complete linkage dendrogram for distances between numbers in 11 languages

Comparing this complete linkage dedrogram to the previously rendered single linkage dendrogram, we see that both hierarchical methods yield the English–Norvegian–Danish and the French–Italian–Spanish language groups. Polish is merged with French–Italian–Spanish at an intermediate level. In addition, both methods merge Hungarian and Finnish only at the penultimate stage.

However, the two methods handle German and Dutch differently. Single linkage merges German and Dutch at an intermediate distance, and these two lnaguages remain a cluster until the final merger. Complete linkage merges German with the English–Norwegian–Danish group at an intermediate level. Dutch remains a cluster by itself

until it is merged with the English–Norwegian–Danish–German and French–Italian–Spanish–Polish groups at a higher distance level. The final complete linkage merger involves two clusters. The final merger in single linkage involves three clusters. □

Example: Clustering variables using complete linkage Data collected on 22 U.S. public utility companies for the year 1975 are listed in the table:

Public Utility Data (1975)

	Company	Variables							
		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	Arizona Public Service	1.06	9.2	151	54.4	1.6	9077	0.0	0.628
2	Boston Edison Co.	0.89	10.3	202	57.9	2.2	5088	25.3	1.555
3	Central Louisiana Electric Co.	1.43	15.4	113	53.0	3.4	9212	0.0	1.058
4	Commonwealth Edison Co.	1.02	11.2	168	56.0	0.3	6423	34.3	0.700
5	Consolidated Edison Co. (N.Y.)	1.49	8.8	192	51.2	1.0	3300	15.6	2.044
6	Florida Power & Light Co.	1.32	13.5	111	60.0	-2.2	11127	22.5	1.241
7	Hawaiian Electric Co.	1.22	12.2	175	67.6	2.2	7642	0.0	1.652
8	Idaho Power Co.	1.10	9.2	245	57.0	3.3	13082	0.0	0.309
9	Kentuchy Utilities Co.	1.34	13.0	168	60.4	7.2	8406	0.0	0.862
10	Madison Gas & Electric Co.	1.12	12.4	197	53.0	2.7	6455	39.2	0.623
11	Nevada Power Co.	0.75	7.5	173	51.5	6.5	17441	0.0	0.768
12	New England Electric Co.	1.13	10.9	178	62.0	3.7	6154	0.0	1.897
13	Northern States Power Co.	1.15	12.7	199	53.7	6.4	7179	50.2	0.527
14	Oklahoma Gas & Electric Co.	1.09	12.0	96	49.8	1.4	9673	0.0	0.588

15	Pacific Gas & Electric Co.	0.96	7.6	164	62.2	-0.1	6468	0.9	1.400
16	Puget Sound Power & Light Co.	1.16	9.9	252	56.0	9.2	15991	0.0	0.620
17	San Diego Gas & Electric Co.	0.76	6.4	136	61.9	9.0	5714	8.3	1.920
18	The Southern Co.	1.05	12.6	150	56.7	2.7	10140	0.0	1.108
19	Texas Utilities Co.	1.16	11.7	104	54.0	-2.1	13507	0.0	0.636
20	Wisconsin Electric Power Co.	1.20	11.8	148	59.9	3.5	7287	41.1	0.702
21	United Illuminating Co.	1.04	8.6	204	61.0	3.5	6650	0.0	2.116
22	Virginia Electric & Power Co.	1.07	9.3	174	54.3	5.9	10093	26.6	1.306

Keys to the Previous Table

Key: X_1: Fixed-charge coverage ratio (income/debt).
X_2: Rate of return on capital.
X_3: Cost of KW capacity in place.
X_4: Annual load factor.
X_5: Peak kWh demand growth from 1974 to 1975.
X_6: Sales (kWh use per year).
X_7: Percent nuclear.
X_8: Total fuel costs (cents per kWh).
Source: Data courtesy of H. E. Thompson.

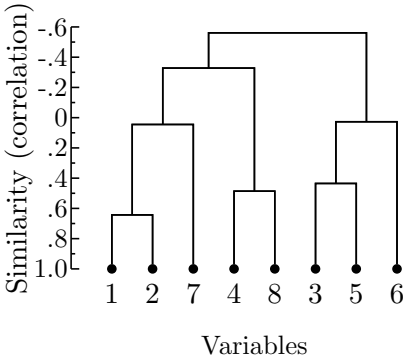
Although it is more interesting to group companies, we shall see here how the complete

linkage algorithm can be used to cluster vaiables. We measure the similarity between pairs of variables by the product-moment correlation coefficient. The correlation matrix is given in the next table.

Correlation Between Pairs of Variables (Public Utility Data)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1							
X_2	0.643	1						
X_3	−0.103	−0.348	1					
X_4	−0.082	−0.086	0.100	1				
X_5	−0.259	−0.260	0.435	0.034	1			
X_6	−0.152	−0.010	0.027	−0.288	0.176	1		
X_7	0.045	0.211	0.115	−0.164	−0.019	−0.374	1	
X_8	−0.013	−0.328	0.005	0.486	−0.007	−0.561	−0.185	1

When the sample correlations are used as similarity measures, variables with large negative correlations are regarded as very dissimilar; variables with large positive correlations are regarded as very similar. In this case, the "distance" between clusters is measured as the *smallest* similarity between members of the corresponding clusters. The complete linkage algorithm, applied to the foregoing similarity matrix, yields the following dendrogram.



Complete linkage dendrogram for similarities among eight utility company variables

We see that variables 1 and 2 (fixed size coverage ratio and rate of return on capital), variables 4 and 8 (annual load factor and total fuel costs), and variables 3 and 5 (cost per kilowatt capacity in place and peak kilowatthour demand growth) cluster at intermediate "similarity" levels. Variables 7 (percent nuclear) and 6 (sales) remain by themselves until the final stages. The final merger brings together the (1 2 4 7 8) group and (3 5 6) group. □

As in single linkage, a "new" assignment of distances (similarities) that have the same relative ordering as the initial distances will not change the configuration of the complete linkage clusters.

Average Linkage

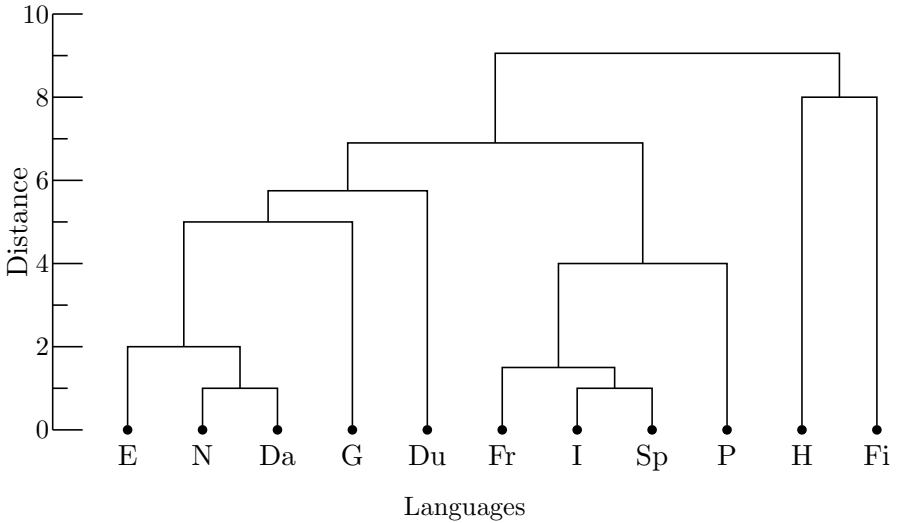
Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

Again, the input to the average linkage algorithm may be distances or similarities, and the method can be used to group objects or variables. The average linkage algorithm proceeds in the manner of the general algorithm We begin by searching the distance matrix $\mathbf{D} = \{d_{ik}\}$ to find the nearest (most similar) objects—for example, U and V. These objects are merged to form the cluster UV. For Step 3 of the general agglomerative algorithm, the distances between (UV) and the other cluster W are determined by

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$$

where d_{ik} is the distance between object i in the cluster (UV) and object k in the cluster W, and $N_{(UV)}$ and N_W are the number of items in clusters (UV) and W, respectively.

Example: Average linkage clustering of 11 languages The average linkage algorithm was applied to the "distances" between 11 languages given in one of the previous examples. The resulting dendrogram is displayed in the figure below:



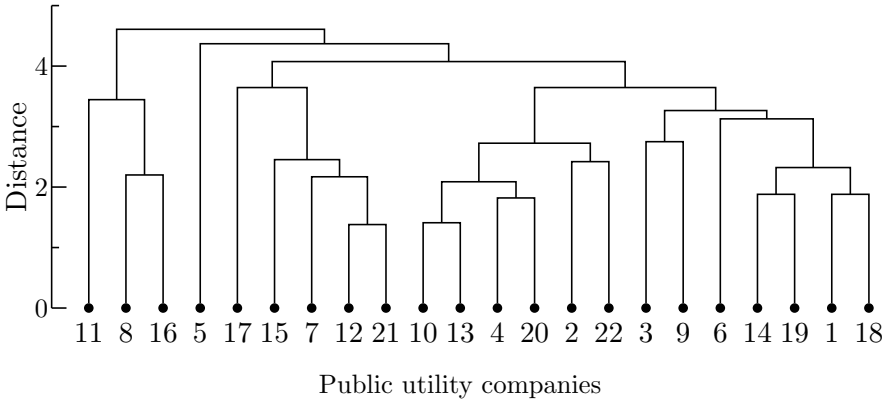
Average linkage dendrogram for distances between numbers in 11 languages

A comparison of this dendrogram with the corresponding single linkage dendrogram and complete linkage dendrogram indicates that average linkage yields a configuration very much like the complete linkage configuration. However, because distance is defined differently for each case, it is not surprising that mergers take place at different levels. □

Example: Average linkage clustering of public utilities An average linkage algorithm applied to the Euclidean distances between 22 public utilities (see the distances table) produced the dendrogram in the figure below.

Distances Between 22 Utilities

Firm No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	0.00																					
2	3.10	0.00																				
3	3.68	4.92	0.00																			
4	2.46	2.16	4.11	0.00																		
5	4.12	3.85	4.47	4.13	0.00																	
6	3.61	4.22	2.99	3.20	4.60	0.00																
7	3.90	3.45	4.22	3.97	4.60	3.35	0.00															
8	2.74	3.89	4.99	3.69	5.16	4.91	4.36	0.00														
9	3.25	3.96	2.75	3.75	4.49	3.73	2.80	3.59	0.00													
10	3.10	2.71	3.93	1.49	4.05	3.83	4.51	3.67	3.57	0.00												
11	3.49	4.79	5.90	4.86	6.46	6.00	6.00	3.46	5.18	5.08	0.00											
12	3.22	2.43	4.03	3.50	3.60	3.74	1.66	4.06	2.74	3.94	5.21	0.00										
13	3.96	3.43	4.39	2.58	4.76	4.55	5.01	4.14	3.66	1.41	5.31	4.50	0.00									
14	2.11	4.32	2.74	3.23	4.82	3.47	4.91	4.34	3.82	3.61	4.32	4.34	4.39	0.00								
15	2.59	2.50	5.16	3.19	4.26	4.07	2.93	3.85	4.11	4.26	4.74	2.33	5.10	4.24	0.00							
16	4.03	4.84	5.26	4.97	5.82	5.84	5.04	2.20	3.63	4.53	3.43	4.62	4.41	5.17	5.18	0.00						
17	4.40	3.62	6.36	4.89	5.63	6.10	4.58	5.43	4.90	5.48	4.75	3.50	5.61	5.56	3.40	5.56	0.00					
18	1.88	2.90	2.72	2.65	4.34	2.85	2.95	3.24	2.43	3.07	3.95	2.45	3.78	2.30	3.00	3.97	4.43	0.00				
19	2.41	4.63	3.18	3.46	5.13	2.58	4.52	4.11	4.11	4.13	4.52	4.41	5.01	1.88	4.03	5.23	6.09	2.47	0.00			
20	3.17	3.00	3.73	1.82	4.39	2.91	3.54	4.09	2.95	2.05	5.35	3.43	2.23	3.74	3.78	4.82	4.87	2.92	3.90	0.00		
21	3.45	2.32	5.09	3.88	3.64	4.63	2.68	3.98	3.74	4.36	4.88	1.38	4.94	4.93	2.10	4.57	3.10	3.19	4.97	4.15	0.00	
22	2.51	2.42	4.11	2.58	3.77	4.03	4.00	3.24	3.21	2.56	3.44	3.00	2.74	3.51	3.35	3.46	3.63	2.55	3.97	2.62	3.01	0.00



Average linkage dendrogram for distances between 22 public utility companies

Concentrating on the intermediate clusters, we see that the utility companies tend to group according to geographical location. The cluster that is easily and among the first recognized as such, both on the map and by looking at the dendrogram, consists of the firms numbered 10 (Madison Gas & Electric Co.), 13 (Northern States Power Co.), 4 (Commonwealth Edison Co.), and 20 (Wisconsin Electric Power Co.) They are all located east of the Great Lakes. Two firms that cluster with them are from the East coast: firms 2 (Boston Edison Co.) and 22 (Virginia Electric & Power Co.)—which fact shows us that, apart from the geographical location, there are also other factors that strongly influence similarites. Were it not so, a cluster containing 2 (Boston Edison

Co.), 21 (United Illuminating Co.), 5 (Consolidated Edison Co. (N.Y.)), 11 (Nevada Power Co.) and possibly 12 (New England Electric Co.) would have formed. One more cluster that was formed under a strong geographical influence, one we could name "ocean influence", consists of firms in coastal areas (East coast, West coast and in the Pacific): 21 (United Illuminating Co.),12 (New England Electric Co.), 7 (Hawaiian Electric Co.), 15 (Pacific Gas & Electric Co.), and 17 (Puget Sound Power & Light Co.). Another cluster that would not be immediately recognized just by looking at the map contains the firms 14 (Oklahoma Gas and Electric Company), 19 (Texas Utilities Company), 1 (Arizona Public Service), and 18 (The Southern Company—primarily Georgia and Alabama).

It is, perhaps, not surprising that utility firms with similar locations (or types of locations, as in our "ocean influence" above) cluster. One would expect regulated firms in the same area to use, basically, the same type of fuel(s) for power plants and face common markets. Consequently, types of generation, costs, growth rates, and so forth should be relatively homogeneous among these firms. This is apparently reflected in the hierarchical clustering. □

For average linkage clustering, changes in the assignment of distances (similarities) can affect the arrangement of the final configuration of clusters, even though the changes preserve relative orderings.

Ward's Hierarchical Clustering Method

Ward considered hierarchical clustering procedures based on minimizing the 'loss of information' from joining two groups. This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion, EES. First, for a given cluster k, let ESS_k be the sum of the squared deviations of every item in the cluster from the cluster mean (centroid). If there are currently K clusters, define ESS as the sum of the ESS_k or $\text{ESS} = \text{ESS}_1 + \text{ESS}_2 + \cdots + \text{ESS}_K$. At each step in the analysis, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS (minimum loss of information) are joined. Initally, each cluster consists of a single item, and, if there are N items, $\text{ESS}_k = 0$, $k = 1, 2, \ldots, N$, so $\text{ESS} = 0$. At the other extreme, when all the clusters are combined in a single group of N items, the value of ESS is given by

$$\text{ESS} = \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}})$$

where $\mathbf{x}_j$ is the multivariate measurement associated with the jth item and $\bar{\mathbf{x}}$ is the mean of all the items.

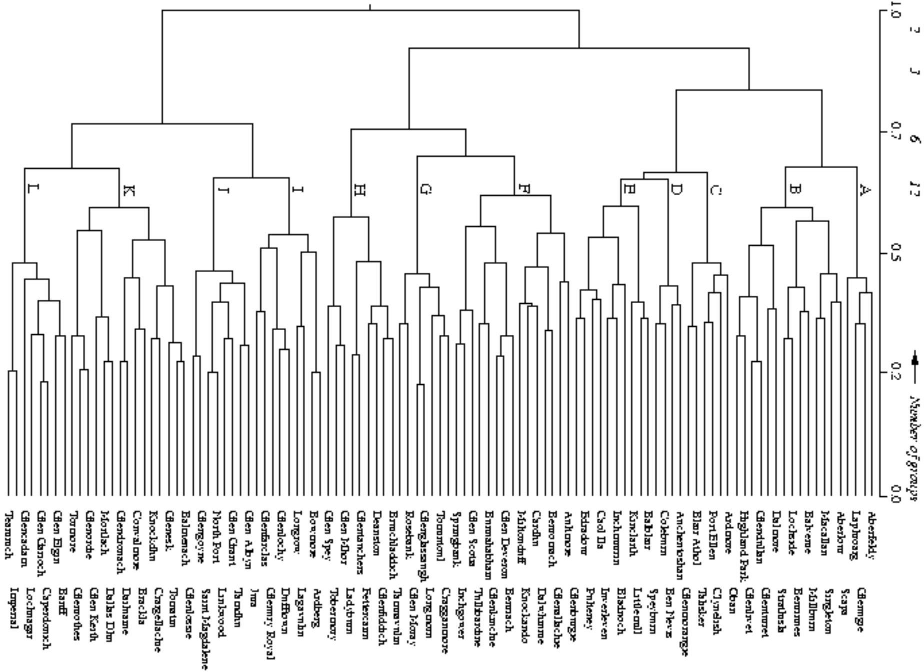
The results of Ward's method can be displayed as a dendrogram. The vertical axis gives the values of ESS at which the mergers occur.

Ward's method is based on the notion that the clusters of multivariate observations

are expected to be roughly elliptically shaped. It is a hierarchical precursor to non-hierarchical clustering methods that optimize some criterion for dividing data into a *given* number of elliptical groups. We will discuss nonhierarcical clustering later on.

Example: Clustering pure malt scotch whiskies Virtually all the world's pure malt Scotch whiskies are produced in Scotland. In one study, 68 binary variables were created measuring characteristics of Scotch whiskey that can be broadly classified as color, nose, body, palate and finish. For example, there were 14 color characteristics (descriptions), including white wine, yellow, very pale, pale, bronze, full amber, red, and so forth. LaPointe and Legendre clustered 109 pure malt Scotch whiskies, each from a different distillery. The investigators were interested in determining the major types of single-malt whiskies, their chief characteristics, and the best representative. In addition, they wanted to know whether the groups produced by the hierarchical clustering procedure corresponded to different geographical regions, since it is known that whiskies are affected by local soil, temperature, and water conditions.

Weighted similarity coefficients $\{s_{ik}\}$ were created from binary variables representing the presence or absence of characteristics. The resulting "distances," defined as $\{d_{ik} = 1 - s_{ik}\}$, were used with Ward's method to group the 10 pure (single-) malt Scotch whiskies. The resulting dendrogram is shown in the next slide. (An average linkage procedure applied to a similarity matrix produced almost exactly the same classification.)



A dendrogram for similarities between 109 pure malt Scotch whiskies.

The groups labbelled A–L in the figure are the 12 groups of similar Scotches identified by the investigators. A follow-up analysis suggested that these 12 groups have a large geographic component in the sense that Scotches with similar characteristics tend to be produced by distilleries that are located reasonably close to one another. Consequently, the investigators concluded, "The relationship with geographic features was demonstrated, supporting the hypothesis that whiskies are affected not only by distillery secrets and traditions but also by factors dependant on region, such as water, soil, microclimate, temperature and even air quality." □

Final Comments—Hierarchical Procedures

There are many agglomerative hierarchical clustering procedures besides single linkage, complete linkage, and average linkage. However, all the agglomerative procedures follow the basic algorithm.

As with most clustering methods, sources of error and variation are not formally considered in hierarchical procedures. This means that a clustering method will be sensitive to outliers, or "noise points."

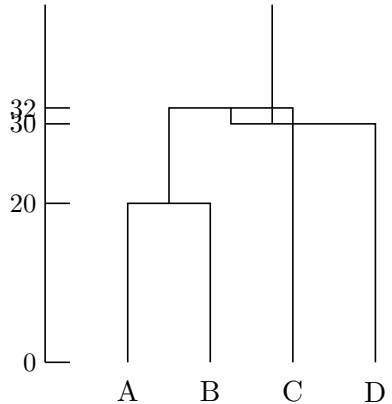
In hierarchical clustering, there is no provision for a reallocation of objects that may have been "incorrectly" grouped at an early stage. Consequently, the final configuration of clusters should always be carefully examined to see whether it is sensible.

For a particular problem, it is a good idea to try several clustering methods and, within a given method, a couple different ways of assigning distances (similarities). If the outcomes from the several methods are (roughly) consistent with one another, perhaps a case for "natural" groupings can be advanced.

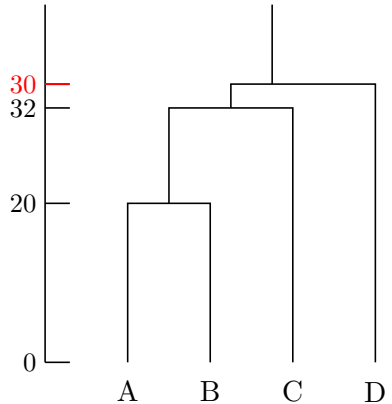
The *stability* of a hierarchical solution can sometimes be checked by applying the clustering algorithm before and after *small* errors (perturbations) have been added to the data units. If the groups are fairly well distinguished, the clusterings before perturbation and after perturbation should agree.

Common values (ties) in the similarity or distance matrix can produce multiple solutions to a hierarchical clustering problem. That is, the dendrograms corresponding to different treatments of the tied similarities (distances) can be different, particularly at the lower levels. This is not an inherent problem of any method; rather, multiple solutions occur for certain kinds of data. Multiple solutions are not necessarily bad, but the user needs to know of their existence so that the groupings (dendrograms) can be properly interpreted and different groupings (dendrograms) compared to assess their overlap.

Some data sets and hierarchical clustering methods can produce *inversions.* An inversion occurs when an object joins an exixting cluster at a smaller distance (greater similarity) than that of a previous consoidation. An inversion is represented two different ways in the following diagram:



(i)



(ii)

In this example, the clustering method joins A and B at distance 20. At the next step, C is added to the group $(A\,B)$ at distance 32. Because of the nature of the clustering algorithm, D is added to group $(A\,B\,C)$ at distance 30, a smaller distance than the distance at which C joined $(A\,B)$. In **(i)** the inversion is indicated by a dendrogram with crossover. In **(ii)**, the inversion is indicated by a dendrogram with a nonmonotonic scale.

Inversions can occur when there is no clear cluster structure and are generally associated with two hierarchical clustering algorithms known as the centroid method and

the median method. The hierarchical procedures discussed so far are not prone to inversions.

Nonhierarchical Clustering Methods

Nonhierarchical clustering techniques are designed to group *items*, rather than *variables*, into a collection of K clusters. The number of clusters, K, may either be specified in advance or determined as part of the clustering procedure. Because a matrix of distances (similarities) does not have to be determined, and the basic data do not have to be stored during the computer run, nonhierarchical methods can be applied to much larger data sets than can hierarchical techniques.

Nonhierarchical methods start from either (1) an initial partition of items into groups or (2) an initial set of seed points, which will form the nuclei of clusters. Good choices for starting configurations should be free of overt biases. One way to start is to randomly select seed points from among the items or to randomly partition the items into initial groups.

Now we will discuss one of the more popular nonhierarchical procedures, the K-means method.

K-means Method

MacQueen suggests the term *K-means* for describing an algorithm of his that assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of these three steps:

1. Partition the items into K initial clusters
2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
3. Repeat step 2 until no more reassignments take place.

Rather than starting with a partition of all items into K preliminary groups in Step 1, we could specify K initial centroids (seed points) and then proceed to Step 2.

The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points. Experience suggests that most major changes in assignment occur with the first reallocation step.

Example: Clustering using the *K*-means method Suppose we measure two variables X_1 and X_2 for each of four items $A, B, C,$ and D. The data are given in the following table:

Item	Observations	
	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

The objective is to divide these into $K = 2$ clusters such that the items within a cluster are closer to one another than they are to the items in different clusters. To implement the $K = 2$-means method, we *arbitrarily* partition the items into two clusters, such as $(A\,B)$ and $(C\,D)$, and compute the coordinaes $(\bar{x_1}, \bar{x_2})$ of the cluster centroid (mean).

Thus, at Step 1, we have

Cluster	Coordinates of centroid $\bar{x_1}$	$\bar{x_2}$
$(A\,B)$	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
$(C\,D)$	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

At step 2, we complete the Euclidean distance of each item from the group centroids and reassign each item to the nearest group. If an item is moved from the initial configuration,the cluster centroids (means) must be updated proceeding. The ith coordinate, $i = 1, 2, ...p$, of the centroid is easily updated using the formulas:

$$\bar{x}_{i,new} = \frac{n\bar{x}_i + x_{ji}}{n+1} \qquad \text{if the } j\text{th item is } \textit{added} \text{ to a group}$$

$$\bar{x}_{i,new} = \frac{n\bar{x}_i - x_{ji}}{n-1} \qquad \text{if the } j\text{th item is } \textit{removed} \text{ from a group}$$

Here n is the number of items in the "old" group with centroid $\bar{x}' = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)$.

Consider the initial clusters $(A\,B)$ and $(C\,D)$. The **coordinates** of the centroids are $(2, 2)$ and $(-1, -2)$ respectively. Suppose item A with coordinates (5,3) is moved to

the $(C\,D)$ group. The new groups are (B) and $(A\,C\,D)$ with updated centroids:

$$\begin{array}{rll} \text{Group } (B) & \bar{x}_{1,new} = \frac{2(2)-5}{2-1} = -2 & \bar{x}_{2,new} = \frac{2(2)-3}{2-1} = 1, \text{ the } \textbf{coordinates} \text{ of B} \\ \text{Group } (A\,C\,D) & \bar{x}_{1,new} = \frac{2(-1)+5}{2+1} = 1 & \bar{x}_{2,new} = \frac{2(-2)+3}{2+1} = -0.33 \end{array}$$

Returning to the initial groupings in Step 1, we compute the squared distances

$$\begin{array}{lll} d^2(A,(A\,B)) & = (5-2)^2 + (3-2)^2 & = 10 \\ d^2(A,(C\,D)) & = (5+1)^2 + (3+2)^2 & = 61 \\ d^2(A,(B)) & = (5+1)^2 + (3-1)^2 & = 40 \\ d^2(A,(A\,C\,D) & = (5-1)^2 + (3+.33)^2 & = 27.09 \end{array}$$

if A is not moved

if A is moved to CD group

Since A is closer to the center of $(A\,B)$ than it is to the center of $(A\,C\,D)$, it is not reassigned.

Continuing, we consider reassigning B. We get

$$\begin{array}{lll} d^2(B,(A\,B)) & = (-1-2)^2 + (1-2)^2 & = 10 \\ d^2(B,(C\,D)) & = (-1+1)^2 + (1+2)^2 & = 9 \\ d^2(B,(A)) & = (-1-5)^2 + (1-3)^2 & = 40 \\ d^2(B,(B\,C\,D) & = (-1+1)^2 + (1+1)^2 & = 4 \end{array}$$

if B is not moved

if B is moved to CD group

Since B is closer to the center of $(B\,C\,D)$ than it is to the center of $(A\,B)$, B is reassigned to the $(C\,D)$ group. We now have the clusters (A) and $(B\,C\,D)$ with centroid coordinates $(5,3)$ and $(-1,-1)$ respectively.

We check C for reassignment.

$$
\begin{array}{lll}
d^2(C,(A)) & = (1-5)^2 + (-2-3)^2 & = 41 \\
d^2(C,(B\,C\,D)) & = (1+1)^2 + (-2+1)^2 & = 5 \\
d^2(C,(A\,C)) & = (1-3)^2 + (-2-.5)^2 & = 10.25 \\
d^2(C,(B\,D) & = (1+2)^2 + (-2+.5)^2 & = 11.35
\end{array}
$$

if C is not moved

if C is moved to A group

Since C is closer to the center of $(B\,C\,D)$ than it is to the center of the $(A\,C)$ group, C is not moved. Continuing in this way, we find that no more reassignments take place and the final $K = 2$ clusters are (A) and $(B\,C\,D)$.

For the final clusters, we have

	Squared distances to group centroids			
	Item			
Cluster	A	B	C	D
(A)	0	40	41	89
$(B\,C\,D)$	52	4	5	5

The within cluster sum of squares (sum of squared distances to centroid) are

$$
\begin{array}{ll}
\text{Cluster } (A): & 0 \\
\text{Cluster } (B\,C\,D): & 4+5+5 = \quad 14
\end{array}
$$

Equivalently, we can determine the $K = 2$ clusters by using the criterion

$$\min E = \sum d^2_{i,c(i)}$$

where the minimum is over the number of $K = 2$ clusters and $d^2_{i,c(i)}$ is the squared distance of case i from the centroid (mean) of the assigned cluster.

In this example, there are seven possibilities for $K = 2$ clusters:

$$\begin{gathered} A, (B\,C\,D) \\ B, (A\,C\,D) \\ C, (A\,B\,D) \\ D, (A\,B\,C) \\ (A\,B), (C\,D) \\ (A\,C), (B\,D) \\ (A\,D), (B\,C) \end{gathered}$$

For the $A, (B\,C\,D)$ pair:

$$\begin{array}{rl} A & d^2_{A,c(A)} = 0 \\ (B\,C\,D) & d^2_{B,c(B)} + d^2_{C,c(C)} + d^2_{D,c(D)} = 4 + 5 + 5 = 14 \end{array}$$

Consequently, $\sum d^2_{i,c(i)} = 0 + 14 = 14$ For the remaining pairs, you may verify that

$$\begin{aligned} B, (A\,C\,D) \qquad & d^2_{i,c(i)} = 48.7 \\ C, (A\,B\,D) \qquad & d^2_{i,c(i)} = 27.7 \\ D, (A\,B\,C) \qquad & d^2_{i,c(i)} = 31.3 \\ (A\,B), (C\,D) \qquad & d^2_{i,c(i)} = 28 \\ (A\,C), (B\,D) \qquad & d^2_{i,c(i)} = 27 \\ (A\,D), (B\,C) \qquad & d^2_{i,c(i)} = 51.3 \end{aligned}$$

Since the smallest $\sum d^2_{i,c(i)}$ occurs for the pair of clusters (A) and $(B\,C\,D)$, this is the final partition. □

To check the stability of the clustering, it is desirable to rerun the algorithm with a new initial partition. Once clusters are determined, intuitions concerning their interpretations are aided by rearranging the list of items so that those in the first cluster appear first, those in the second cluster appear next, and so forth. A table of the cluster centroids (means) and within-cluster variances also helps to delineate group differences.

Example: *K*-means clustering of public utilities Let us return to the problem of clustering public utilities using the table containing eight variables per company. The

K-means algorithm for several choices of K was run. We present a summary of the results for $K = 4$ and $K = 5$. In general, the choice of a particular K is not clear cut and depends upon subject-matter knowledge, as well as data-based appraisals. (Data-based appraisals might include choosing K so as to maximize the between-cluster variability relative to the within-cluster variability. Relevant measures might include $\frac{|\mathbf{W}|}{|\mathbf{B}+\mathbf{W}|}$ and $\text{tr}(\mathbf{W}^{-1}\mathbf{B})$.) The summary is as follows:

$K = 4$

Cluster	Number of firms	Firms
1	5	Idaho Power Co, (8), Nevada Power Co. (11), Puget Sound Power & Light Co. (16), Virginia Electric & Power Co. (22), Kentucky Utilities Co. (9).
2	6	Central Louisiana Electric Co. (3), Oklahoma Gas & Electric Co. (14), The Southern Co. (18), Texas Utilities Co. (19), Arizona Public Service (1), Florida Power & Light Co. (6).
3	5	New England Electric Co. (12), Pacific Gas & Electric Co. (15), San Diego Gas & Electric Co. (17), United Illuminating Co. (21), Hawaiian Electric Co. (7).
4	6	Consolidated Edison Co. (N.Y.) (5), Boston Edison Co. (2), Madison Gas & Electric Co. (10), Northern States Power Co. (13), Wisconsin Electric Power Co. (20), Commonwealth Edison Co. (4).

Distances between Cluster Centres:

	1	2	3	4
1	0			
2	3.08	0		
3	3.29	3.56	0	
4	3.05	2.84	3.18	0

$K = 5$

Cluster	Number of firms	Firms
1	5	Nevada Power Co. (11), Puget Sound Power & Light Co. (16), Idaho Power Co. (8), Virginia Electric & Power Co. (22), Kentucky Utilities Co. (9).
2	6	Central Louisiana Electric Co. (3), Texas Utilities Co. (19), Oklahoma Gas & Electric Co. (14), The Southern Co. (18), Arizona Public Service (1), Florida Power & Light Co. (6).
3	5	New England Electric Co. (12), Pacific Gas & Electric Co. (15), San Diego Gas & Electric Co. (17), United Illuminating Co. (21), Hawaiian Electric Co. (7).
4	2	Consolidated Edison Co. (N.Y.) (5), Boston Edison Co. (2).
5	4	Commonwealth Edison Co. (4), Madison Gas & Electric Co. (10), Northern States Power Co. (13), Wisconsin Electric Power Co. (20).

Distances between Cluster Centres:

	1	2	3	4	5
1	0				
2	3.08	0			
3	3.29	3.56	0		
4	3.63	3.46	2.63	0	
5	3.18	2.99	3.81	2.89	0

We have

$$F_{nuc} = \frac{\text{mean square percent nuclear between clusters}}{\text{mean square percent nuclear within clusters}}$$
$$= \frac{3.335}{.255} = 13.1$$

so firms within different clusters are widely separated with respect to percent nuclear, but firms within the same cluster show little percent nuclear variation. Fuel costs (FUELC) and annual sales (SALES) also seem to of some importance in distinguishing the clusters.

Reviewing the firms in the five clusters, it is apparent that the K-means method gives results generally consistent with the average linkage hierarchical method. Firms with common or compatible geographical location cluster. Also, the firms in a given cluster weem to be roughly the same in terms of percent nuclear. □

We must caution, as we have throughout the book, that the importance of *individual* variables in clustering must be judged from a multivariate perspective. *All* of the variables (multivariate observations) determine the cluster means and the reassignment of items. In addition, the values of the descriptive statistics measuring the importance of individual variables are functions of the number of clusters and the final configuration of the clusters. On the other hand, descriptive measures can be helpful, after the fact, in assessing the "success" of clustering procedure.

Final Comments—Nonhierarchical Procedures

There are strong arguments for not fixing the number of clusters, K, in advance, including the following:

1. If two or more seed points inadvertently lie within a single cluster, their resulting clusters will be poorly differentiated.
2. The existence of an outlier might produce at least one group with very disperse items
3. Even if the population is known to consist of K groups, the sampling method may be such that data from the rarest group do not appear in the sample. Forcing the data into K groups would lead to nonsensical clusters.

In cases in which a signle run of the algorithm requires the user to specify K, it is always a good idea to rerun the algorithm for several choices.

Clustering Based on Statistical Models

The popular clustering methods discussed earlier in this chapter, including single linkage, complete linkage, average linkage, Ward's method and K-means clustering, are

intuitively reasonable procedures but that is as much as we can say without having a model to explain how the observations were produced. Major advances in clustering methods have been made through the introduction of statistical models that indicate how the collection of $(p \times 1)$ measuremaents $\mathbf{x}_j$, from the N objects, was generated. The most common model is one where cluster k has expected proportion p_k of the objects and the corresponding measurements are generated by a probability density function $f_k(\mathbf{x})$. Then, if there are K clusters, the observation vector for a single object is modeled as arising from the *mixing distribution*

$$f_{Mix}(\mathbf{x}) = \sum_{k-1}^{K} p_k f_k(\mathbf{x})$$

where each $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. This distribution $f_{\text{Mix}}(\mathbf{x})$ is called a mixture of the K distributions $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_K(\mathbf{x})$ because the observation is generated from the component distribution $f_k(\mathbf{x})$ with probability p_k. The collection of N objservation vectors generated from this distribution will be a mixture of observations from the component distributions.

The most common mixture model is a mixture of multivariate normal distributions where the k-th component $f_k(\mathbf{x})$ is the $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ density function.

The normal mixture model for one observation $\mathbf{x}$ is

$$f_{\text{Mix}}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) = \sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

Clusters generated by this model are ellipsoidal in shape with the heaviest concentration of observations near the center.

Inferences are based on the likelihood, which for N objects and a fixed number of clusters K, is

$$\begin{aligned} L(p_1, \ldots, p_k, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \prod_{j-1}^{N} f_{\text{Mix}}(\mathbf{x}_j|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) \\ &= \prod_{j=1}^{N}\left(\sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_j-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)\right) \end{aligned}$$

where the proportions $p_1, \ldots, p_k$, the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, and the covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k$ are unknown. The measurements for different objects are treated as independent and identically distributed observations from the mixture distribution.

There are typically far too many unknown parameters for parameters for making inferences when the number of objects to be clustered is at least moderate. However, certain conclusions can be made regarding situations where a heuristic clustering method

should work well. In particular, the likelihood-based procedure under the normal mixture model with all $\mathbf{\Sigma}_k$ the same multiple of the identity matrix, $\eta\mathbf{I}$, is approximately the same as K-means clustering and Ward's method. To date, no statistical models have been advanced for which the cluster formation procedure is approximately the same as single linkage, complete linkage or average linkage.

Most importantly, under the sequence of mixture models for different K, the problems of choosing the number of clusters and choosing an appropriate clustering method has been reduced to the problem of selecting an appropriate statistical model. This is a major advance.

A good approach to selecting a model is to first obtain the maximum likelihood estimates $\widehat{p}_1, \ldots, \widehat{p}_K, \widehat{\boldsymbol{\mu}}_{\mathbf{1}}, \widehat{\mathbf{\Sigma}}_{\mathbf{1}}, \ldots, \widehat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \widehat{\mathbf{\Sigma}}_{\boldsymbol{K}}$ for a fixed number of clusters K. These estimates must be obtained numerically, using special purpose software. The resulting value of the maimum of the likelihood

$$L_{\max} = L(\widehat{p}_1, \ldots, \widehat{p}_K, \widehat{\boldsymbol{\mu}}_{\mathbf{1}}, \widehat{\mathbf{\Sigma}}_{\mathbf{1}}, \ldots, \widehat{\boldsymbol{\mu}}_{\boldsymbol{k}}, \widehat{\mathbf{\Sigma}}_{\boldsymbol{K}})$$

provides the basis for model selection. How do we decide on a reasonable value for the number of clusters K? In order to compare models with different numbers of parameters, a penalty is subtracted from twice the maximized value of the log-likelihood to give

$$-2\ln L_{\max} - Penalty$$

where the penalty depends on the number of parameters estimated and the number of observations N. Since the probabilities p_k sum to 1, there are only $K-1$ probabilities that must be estimated, $K\times p$ means and $K\times p(p+1)/2$ variances and covariances. For the Akaike information criterion (AIC), the penalty is $2N\times$ (number of parameters) so

$$\text{AIC} = 2\ln L_{\max} - 2N\left(K\frac{1}{2}(p+1)(p+2)-1\right)$$

The Bayesian information criterion (BIC) is similar but uses the logarithm of the number of parameters in the penalty function

$$\text{BIC} = 2\ln L_{\max} - 2\ln(N)\left(K\frac{1}{2}(p+1)(p+2)-1\right)$$

There is still occasional difficulty with too many parameters in the mixture model so simple structures are assumed for the $\boldsymbol{\Sigma_k}$. In particular, progressively more complicated structures are allowed as indicated in the following table

Assumed form for $\boldsymbol{\Sigma}_k$	Total number of parameters	BIC
$\boldsymbol{\Sigma_k} = \eta\mathbf{I}$	$K(p+1)$	$\ln L_{\max} - 2\ln(N)K(p+1)$
$\boldsymbol{\Sigma_k} = \eta_k\mathbf{I}$	$K(p+2)-1$	$\ln L_{\max} - 2\ln(N)(K(p+2)-1)$
$\boldsymbol{\Sigma_k} = \eta_k Diag(\lambda_1,\lambda_2,\ldots,\lambda_p)$	$K(p+2)+p-1$	$\ln L_{\max} - 2\ln(N)(K(p+2)+p-1)$

Even for a fixed number of clusters, the estimation of a mixture model is complicated. One current software package, MCLUST, available in the R software library, combines hierarchical clustering, the EM algorithm and the BIC criterion to develop an appropriate model for clustering. In ghe 'E'-step of the EM algorithm, a $(N \times K)$ matrix is created whose jth row contains estimates of the conditional (on the current parameter estimates) probabilities that observation $\mathbf{x}_j$ belongs to cluster $1, 2, \ldots, K$. So, at convergence, the jth observation (object) is assigned to the cluster k for which the conditional probability

$$p(k|\mathbf{x}_j) = \frac{\widehat{p}_j f(\mathbf{x}_j|k)}{\sum_{i=1}^{K} \widehat{p}_i f(\mathbf{x}_i|k)}$$

of membership is the largest.

Example: A model based clustering of the iris data Consider Iris data in the following table:

Data on Irises

π_1: *Iris Setosa*				π_2: *Iris Versicolor*				π_3: *Iris Virginica*			
Sepal length x_1	Sepal width x_2	Petal length x_3	Petal width x_4	Sepal length x_1	Sepal width x_2	Petal length x_3	Petal width x_4	Sepal length x_1	Sepal width x_2	Petal length x_3	Petal width x_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1

4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4

5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Using *MCLUST* and specifically the *me* function, we first fit the $p = 4$ dimensional normal mixture model restricting the covariance matrices to satisfy $\sum_k = \eta_k \mathbf{I}$, $k = 1, 2, 3$.

Using the BIC criterion, the software chooses $K = 3$ clusters with estimated centers

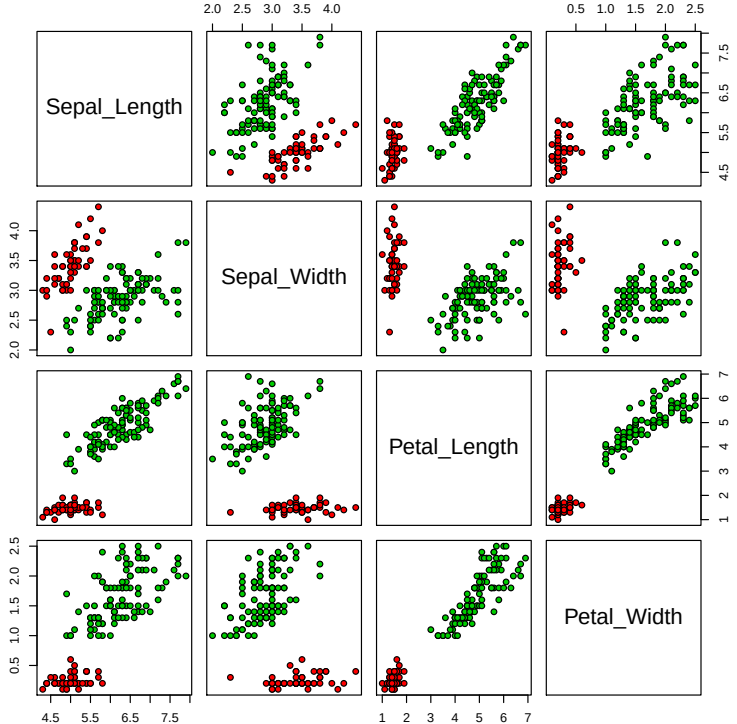
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{bmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} 5.90 \\ 2.75 \\ 4.40 \\ 1.43 \end{bmatrix}, \qquad \boldsymbol{\mu}_3 = \begin{bmatrix} 6.85 \\ 3.07 \\ 5.73 \\ 2.07 \end{bmatrix},$$

and estimated variance-covariance scale factors $\widehat{\eta}_1 = 0.76$, $\widehat{\eta}_2 = 0.163$ and $\widehat{\eta}_3 = 0.2534$. For this solution, BIC $= -853.8$. A matrix plot of the clusters for pairs of variables is shown in the next figure.

Once we have an estimated mixture model. a new object $\mathbf{x}_j$ will be assigned to the cluster for which the conditional probability of membership is the largest.

Assuming the $\boldsymbol{\Sigma_k} = \eta_k \mathbf{I}$ covariance structure and allowing up to $K = 7$ clusters, the BIC can be increased to BIC $= -705.1$.

Finally, using the BIC criterion with up to $K = 9$ groups and several different covariance structures, the best choice is a two-group mixture model with unconstrained covariances. The estimated mixing probabilities are $\widehat{p}_1 = 0.3333$ and $\widehat{p}_2 = 0.6667$. The



Multiple scatter plots of $K = 2$ clusters for Iris data

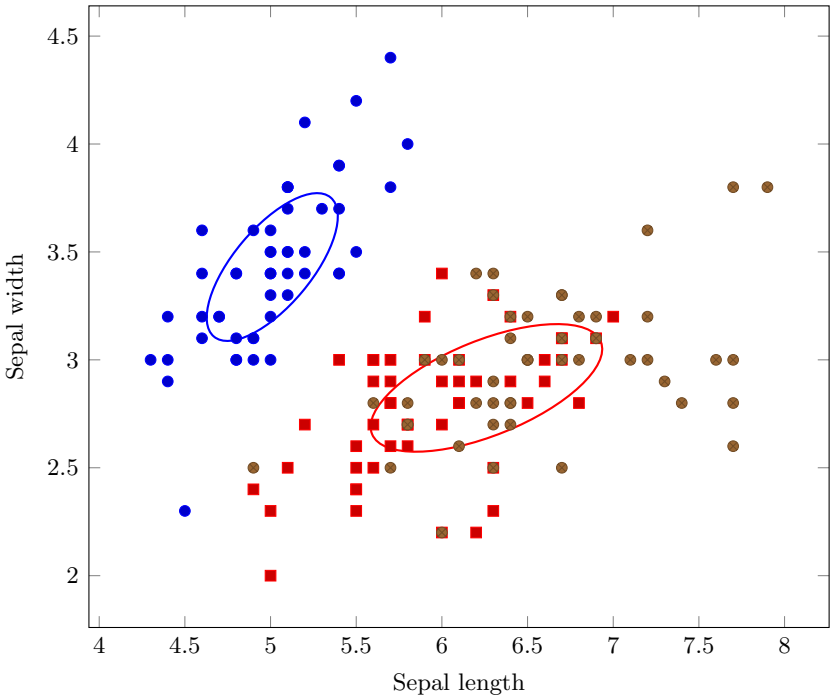
estimated group centers are

$$\boldsymbol{\mu_1} = \begin{bmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{bmatrix}, \qquad \boldsymbol{\mu_2} = \begin{bmatrix} 6.26 \\ 2.87 \\ 4.91 \\ 1.68 \end{bmatrix}$$

and the two estimated covariance matrices are

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}} = \begin{bmatrix} .1218 & .0972 & .0160 & .0101 \\ .0872 & .1408 & .0115 & .0091 \\ .0160 & .0115 & .0296 & .0059 \\ .0101 & .0091 & .0059 & .0109 \end{bmatrix} \qquad \widehat{\boldsymbol{\Sigma}}_{\mathbf{2}} = \begin{bmatrix} .4530 & .1209 & .4489 & .1655 \\ .1209 & .1096 & .1414 & .0792 \\ .4489 & .1414 & .6748 & .2858 \\ .1655 & .0792 & .2858 & .1786 \end{bmatrix}$$

Essentially, two species of Iris have been put in the same cluster as the projected view of the scatter plot of the sepal measurements in the next figure shows. □



Scatter plot of sepal measurements for best model.

Multidimensional Scaling

We will now discuss methods for displaying (transformed) multivariate data in low-dimensional space. We have already considered this issue when we discussed plotting scores on, say, the first two principal components or the scores on the first two linear discriminants. The methods we are about to discuss differ from these procedures in the sense that their *primary* objective is to "fit" the original data into a low-dimensional coordinate system such that any distortion caused by a reduction in dimensionality is minimized. Distortion generally refers to the similarities or dissimilarities (distances) among the original data points. Although Euclidean distance may be used to measure the closeness of points in the final low-dimensional configuration, the notion of similarity or dissimilarity depends upon the underlying technique for its definition. A low-dimensional plot of the kind we are alluding to is called an *ordination* of the data.

Multidimensional scaling techniques deal with the following problem: For a set of observed similarities (or distances) between every pair of N items, find a representation of the items in few dimensions such that the interitem proximities "nearly match" the original similarities (or distances).

It may not be possible to match exactly the ordering of the original similarities (distances). Consequently, scaling techniques attempt to find configurations in $q \leq N-1$ dimensions such that the match is as close as possible. The numerical measure of closeness is called the *stress*.

It is possible to arrange the N items in a low-dimensional coordinate system using only the *rank orders* of the $N(N-1)/2$ original similarities (distances), and not their magnitudes. When only this ordinal information is used to obtain a geometric representation, the process is called *nonmetric multidimensional scaling.* In the actual magnitudes of the original similarities (distances) are used to obtain a geometric representation in q dimensions, the process is called *metric dmultidimensional scaling.* Metric multidimensional scaling is also known as *principal coordinate analysis.*

Scaling tecnhiques were developed by Shepard Multidimensional scaling invariably requires the use of a computer, and several good computer programs are now available for the purpose.

The Basic Algorithm

For N items, there are $M = N(N-1)/2$ similarities (distances) between pairs of different items. These similarities constitute the basic data. (In cases where the similarities cannot be easily quantified as, for example, the similarity between two colors, the rank orders of similarities are the basic data.)

Assuming no ties, the similarities can be arranged in a strictly ascending order as

$$s_{i_1k_1} < s_{i_2k_2} < s_{i_Mk_M}$$

Here $s_{i_1k_1}$ is the smallest of the M similarities. The subscript i_1k_1 indicates the pair of

items that are least similar—that is, the items with rank 1 in the similarity ordering. Other subscripts are interpreted in the same manner. We want to find a q-dimensional configuration of the N items such that the distances, $d_{ik}^{(q)}$, between pairs of items match the ordering in the previous equation. If the distances are laid out in a manner corresponding to that ordering, a perfect match occurs when

$$d_{i_1k_1}^{(q)} > d_{i_2k_2}^{(q)} > \cdots > d_{i_Mk_M}^{(q)}$$

That is, the descending ordering of the distances in q dimensions is exactly analogous to the ascending ordering of the initial similarities. As long as the order defined above is preserved, the magnitudes of the distances are unimportant.

For a given value of q, it may not be possible to find a configuration of points whose pairwise distances are monotonically related to the original similarities. Kruskal proposed a measure of the extent to which a geometrical representation falls short of a perfect match. This measure, the stress, is defined as

$$\text{Stress}(q) = \left\{ \frac{\sum\sum_{i<k} (d_{ik}^{(q)} - \widehat{d}_{ik}^{(q)})^2}{\sum\sum_{i<k} \left[d_{ik}^{(q)}\right]^2} \right\}^{1/2}$$

The $\widehat{d}_{ik}^{(q)}$'s in the stress formula are numbers known to satisfy the descending ordering of the distances; that is, they are monotonically related to the similarities. The $\widehat{d}_{ik}^{(q)}$'s

are *not* distances in the sense that they satisfy the usual distance properties. They are merely reference numbers used to judge the nonmonotonicity of the observed $d_{ik}^{(q)}$'s.

The idea is to find a representation of the items as points in q-dimensions such that the stress is as small as possible. Kruskal suggests the stress be informally interpreted according to the following guidelines:

Stress [%]	*Goodness of fit*
20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

Goodness of fit refers to the monotonic relationship between the similarities and the final distances.

A second measure of discrepancy, introduced by Takane, is becoming the preferred criterion. For a given dimension q, this measure, denoted by SStress, replaces the d_{ik}'s

and $\widehat{d}_{ik}$'s in the previous equation by their squares and is given by

$$\text{SStress}(q) = \left[\frac{\sum\sum_{i<k} (d_{ik}^2 - \widehat{d}_{ik}^2)^2}{\sum\sum_{i<k} d_{ik}^4} \right]^{1/2}$$

The value of SStress is always between 0 and 1. Any value less than .1 is typically taken to mean that there is a good representation of the objects by the points in the given configuration.

Once items are located in q dimensions, their $q \times 1$ vectors of coordinates can be treated as multivariate observations. For display purposes, it is convenient to represent this q-dimensional scatter plot in terms of its principal component axes.

We have written the stress measure as a function of q, the number of dimensions for the geometrical representation. For each q, the configuration leading to the minimum stress can be obtained. As q increases, minimum stress will, within rounding error, decrease and will be zero for $q = N - 1$. Beginning with $q = 1$, a plot of these stress (q) numbers versus q can be constructed. The value of q for which this plot begins to level off may be selected as the "best" choice of the dimensionality. That is, we look for an "elbow" in the stress-dimensionality plot.

The entire multidimensional scaling algorithm is summarized in these steps:

1. For N items, obtain the $M = N(N-1)/2$ similarities (distances) between distinct pairs of items. Order the similarities from largest to smallest. If similarities (distances) cannot be computed, the rank orders must be specified.

2. Using a trial configuration in q dimensions, determine the interitem distances $d_{ik}^{(q)}$ and numbers $\hat{d}_{ik}^{(q)}$, where the later satisfy the descending ordering and minimize the Stress(q) or SStress. (The $\hat{d}_{ik}^{(q)}$ are frequently determined within scaling computer programs using regression methods designed to produce monotonic "fitted" distances.)

3. Using the $\hat{d}_{ik}^{(q)}$'s, move the points around to obtain an improved configuration. (For q fixed, an improved configuration is determined by a general function minimization procedure applied to the stress. In this context, the stress is regarded as a function of the $N \times q$ coordinates of the N items.) A new configuration will have new $d_{ik}^{(q)}$'s, new $\hat{d}_{ik}^{(q)}$'s, and smaller stress. The process is repeated until the best (minimum stress) representation is obtained.

4. Plot minimum stress (q) versus q and choose the best number of dimensions, q^*, from an examination of this plot.

We have assumed that the initial similarity values are symmetric ($s_{ik} = s_{ki}$), that there are no ties, and that there are no missing observations. Kruskal has suggested meth-

ods for handling asymmetries, ties, and missing observations. In addition, there are now multidimensional scaling computer programs that will handle not only Euclidean distance, but any distance of the Minkowski type.

The next three examples illustrate multidimensional scaling with distances as the initial (dis)similarity measures.

Example: Multidimensional scaling of U.S. cities The following table displays the airline distances between pairs of selected U.S. cities.

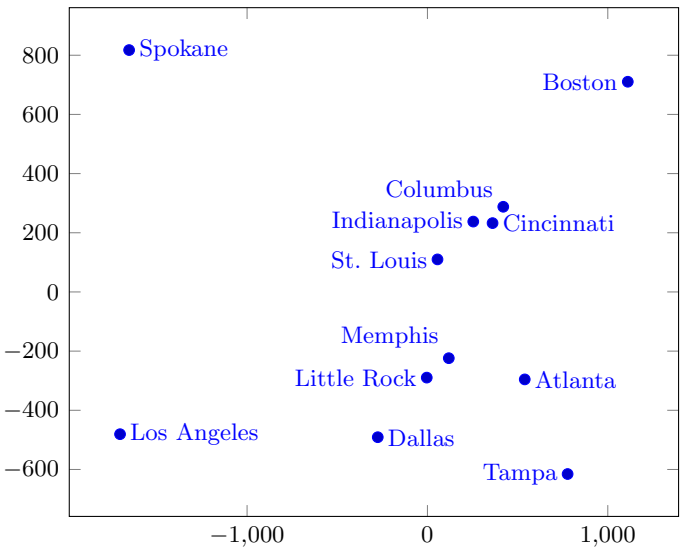
Airline-Distance Data

	Atlanta (1)	Boston (2)	Cincinnati (3)	Columbus (4)	Dallas (5)	Indianapolis (6)	Little Rock (7)	Los Angeles (8)	Memphis (9)	St. Louis (10)	Spokane (11)	Tampa (12)
(1)	0											
(2)	1068	0										
(3)	461	867	0									
(4)	549	769	107	0								
(5)	805	1819	943	1050	0							
(6)	508	941	108	172	882	0						
(7)	505	1494	618	725	325	562	0					
(8)	2197	3052	2186	2245	1403	2080	1701	0				
(9)	366	1355	502	586	464	436	137	1831	0			
(10)	558	1178	338	409	645	234	353	1848	294	0		
(11)	2467	2747	2067	2131	1891	1959	1988	1227	2042	1820	0	
(12)	467	1379	928	985	1077	975	912	2480	779	1016	2821	0

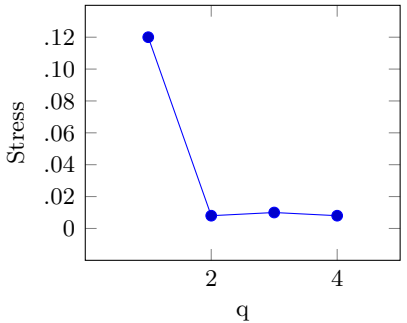
Since the cities naturally lie in a two-dimensional space (a nearly level part of the curved surface of the earth), it is not surprising that multidimensional scaling with

$q = 2$ will locate these items about as they occur on a map. Note that if the distances in the table are ordered from largest to smallest—that is,from a least similar to most similar—the first position is occupied by $d_{\text{Boston, L.A.}} = 3052$.

A multidimensional scaling plot for $q = 2$ dimensions is shown in the next figure. The axes lie along the sample principal components of the scatter plot.



A geometrical representation of cities produced by multidimensional scaling.

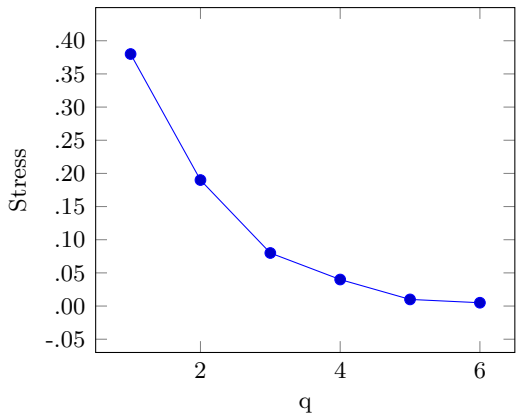


Stress function for airline distances

A plot of stress (q) versus q is shown in the figure on the left. Since stress (1) $\times$ 100% = 12%, a representation of the cities in one dimension (along a single axis) is not unreasonable. The "elbow" of the stress function occurs at $q = 2$. Here stress (2) $\times$ 100% = 0.8%, and the "fit" is almost perfect.

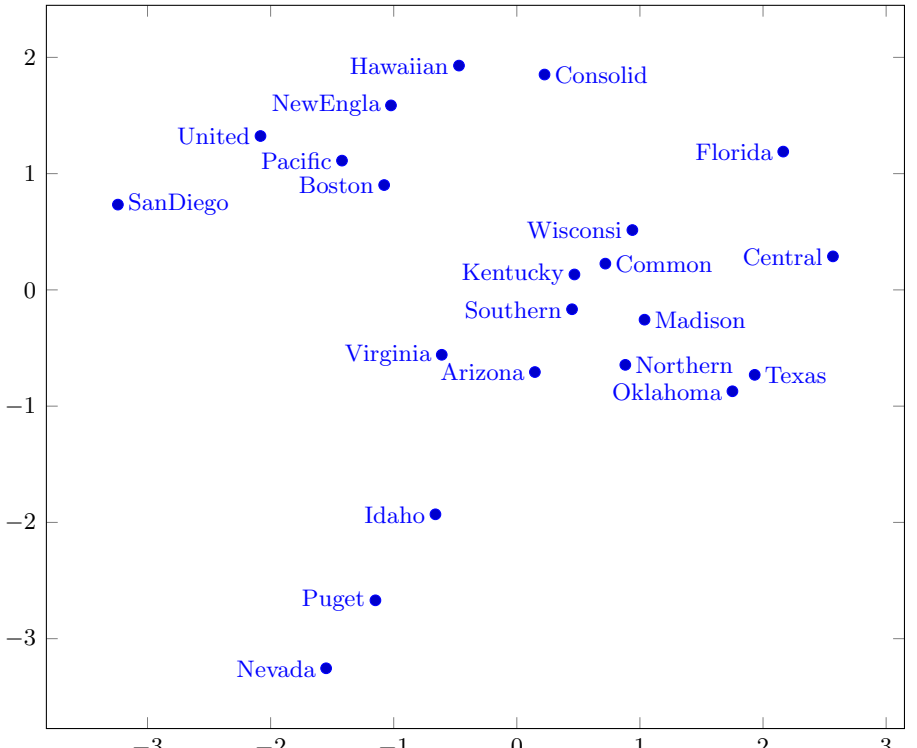
The plot indicates that $q = 2$ is the best choice for the dimension of the dimension of the final configuration. Note that the stress increases for $q = 3$. This anomaly can occur for extremely small values of stress because difficulties with the numerical search procedures used to locate the minimum stress. □

Example: Multidimensional scaling of public utilities Let us try to represent the 22 public utility firms discussed in one of previous examples as points in a low-dimensional space. The measures of (dis)similarities between pairs of firms are Euclidean distances listed in the table named 'Distances between 22 utilites'. Multidimensional scaling in $q = 1, 2, \ldots, 6$ dimensions produced the stress function shown in the figure below.



Stress function for distances between utilities.

The stress function in the figure has no sharp elbow. The plot appears to level out at "good" values of stress (less than or equal to 5%) in the neighborhood of $q = 4$. A good four-dimensional representation of the utilites is achievable, but difficult to display. We show a plot of the utility configuration obtained in $q = 2$ dimensions in the next figure. The axes lie along the sample principal components of the final scatter.



A geometrical representation of utilities produced by multidimensional scaling.

Although the stress for two dimensions is rather high (stress (2) $\times$ 100% = 19%), the distances between firms in the figure above are not wildly inconsistent with the clustering results presented earlier in this chapter. For example, the midwest utilites—Commonwealth Edison, Wisconsin Electric Power (WEPCO), madison Gas and Electric (MG & E), and Northern States Power (NSP)—are close together (similar). Texas Utilities and Oklahoma Gas and Electric (Ok. G & E) are also very close together (similar). Other utilities tend to group according to geographical locations or similar environments.

The utilities cannot be positioned in two dimensions such that the interutility distances $d_{ik}^{(2)}$ are entirely consistent with the original distances.More flexibility for positioning the points is required, and this can only be obtained by introducing additional dimensions. □

Example: Multidimensional scaling of universities Data related to 25 U.S. universities are given in the following table:

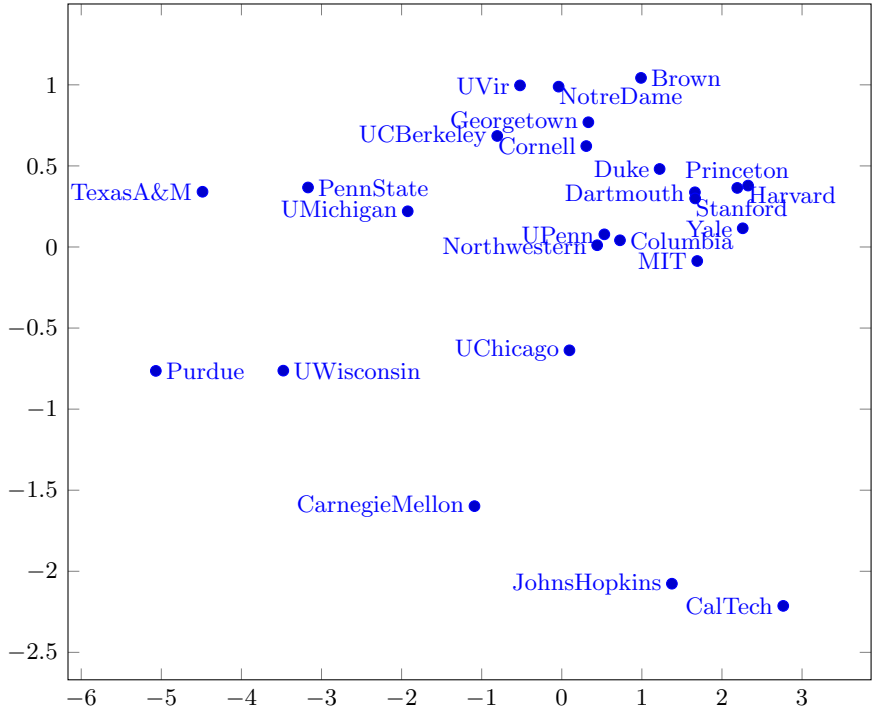
Data on Universities						
University	SAT	Top10	Accept	SFRatio	Expenses	Grad
Harvard	14.00	91	14	11	39.525	97
Princeton	13.75	91	14	8	30.220	95
Yale	13.75	95	19	11	43.514	96
Stanford	13.60	90	20	12	36.450	93
MIT	13.80	94	30	10	34.870	91
Duke	13.15	90	30	12	31.585	95
CalTech	14.15	100	25	6	63.575	81
Dartmouth	13.40	89	23	10	32.162	95
Brown	13.10	89	22	13	22.704	94
JohnsHopkins	13.05	75	44	7	58.691	87
UChicago	12.90	75	50	13	38.380	87
UPenn	12.85	80	36	11	27.553	90
Cornell	12.80	83	33	13	21.864	90
Northwestern	12.60	85	39	11	28.052	89
Columbia	13.10	76	24	12	31.510	88
NotreDame	12.55	81	42	13	15.122	94

UVirginia	12.25	77	44	14	13.349	92
Georgetown	12.55	74	24	12	20.126	92
CarnegieMellon	12.60	62	59	9	25.026	72
UMichigan	11.80	65	68	16	15.470	85
UCBerkeley	12.40	95	40	17	15.140	78
UWisconsin	10.85	40	69	15	11.857	71
PennState	10.81	38	54	18	10.185	80
Purdue	10.05	28	90	19	9.066	69
TexasA&M	10.75	49	67	25	8.704	67

Source: *U.S. News & World Report*, Sept. 18, 1995, p. 126

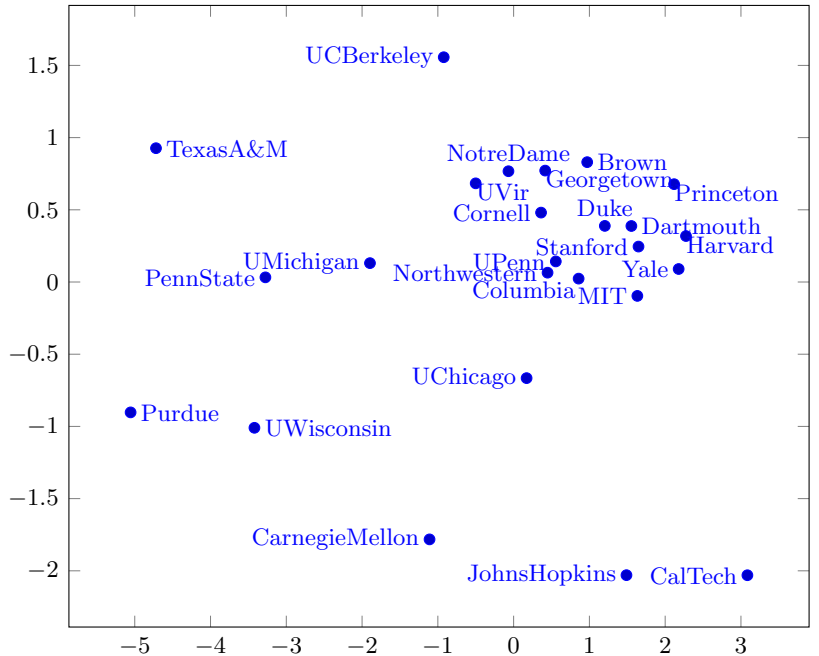
These data give the average SAT score of entering freshmen, percent of freshmen in top 10% of high school class, percent of applicants accepted, student–faculty ratio, estimated annual expense, and graduation rate (%). A metric multidimensional scaling algorithm applied to the standardized university data gives the two-dimensional representation shown in the figure below.

Notice how the private universities cluster on the right of the plot while the large public universities are, generally, on the left.



A two-dimensional representation of universities produced by metric multidimensional scaling.

A nonmetric multidimensional scaling two-dimensional configuration is shown next.



A two-dimensional representation of universities produced by nonmetric multidimensional scaling.

For this example, the metric and nonmetric scaling representations are very similar, with the two-dimensional stress value being approximately 10% for both scalings. □

Classical metric scaling, or principal coordinate analysis, is equivalent to plotting the principal components. Different software programs choose the signs of the appropriate eigenvectors differently, so at the first sight, two solutions may appear to be different. However, the solutions will coincide with a reflection of one or more of the axes.

To summarize, the key objective of multidimensional scaling procedures is a low-dimensional picture. Whenever multivariate data can be presented graphically in two or three dimensions, visual inspection can greatly aid interpretations.

When the multivariate observations are naturally numerical, and Euclidean distances in p-dimensions, $d_{ik}^{(p)}$, can be computed, we can seek a $q < p$-dimensional representation by minimizing

$$E = \left[\sum\sum_{i<k} \frac{\left(d_{ik}^{(p)} - d_{ik}^{(q)} \right)^2}{d_{ik}^{(p)}} \right] \left[\sum\sum_{i<k} d_{ik}^{(p)} \right]^{-1}$$

In this alternative approach, the Euclidean distances in p and q dimensions are compared directly. Techniques for obtaining low-dimensional representations by minimizing E are called *nonlinear mappings*.

The final goodness of fit of any low-dimensional representation can be depicted graphically by *minimal spanning trees*.

Correspondence Analysis

Developed by the French, correspondence analysis is a graphical procedure for representing associations in a table of frequencies or counts. We will concentrate on a two-way table of frequencies or *contingency table*. If the contingency table has I rows and J columns, the plot produced by correspondence analysis contains two sets of points: A set of I points corresponding to the rows and a set of J points corresponding to the columns. The positions of the points reflect associations.

Row points that are close together indicate rows that have similar profiles (conditional distributions) across the columns. Column points that are close together indicate columns with similar profiles (conditional distributions) down the rows. Finally, row points that are close to column points represent combinations that occur more frequently than would be expected from an independence model—that is, a model in which the row categories are unrelated to the column categories.

The usual output from a correspondence analysis includes the "best" two-dimensional representation of the data, along with the coordinates of the plotted points, and a measure (called the *inertia*) of the amount of information retained in each dimension.

Before briefly discussing the algebraic development of correspondence analysis, it is helpful to illustrate the ideas we have introduced with an example.

Example: Correspondence analysis of archaeological data

Frequencies of Types of Pottery

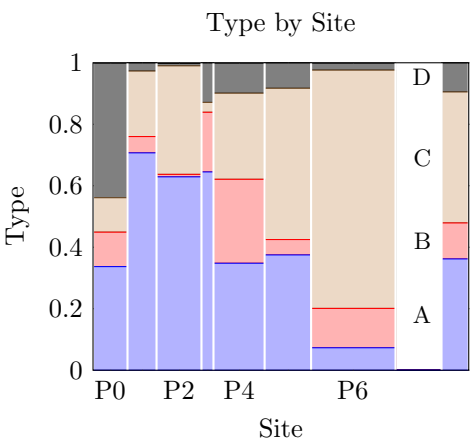
Site	Type				Total
	A	B	C	D	
P0	30	10	10	39	89
P1	53	4	16	2	75
P2	73	1	41	1	116
P3	20	6	1	4	31
P4	46	36	37	13	132
P5	45	6	59	10	120
P6	16	28	169	5	218
Total	283	91	333	74	781

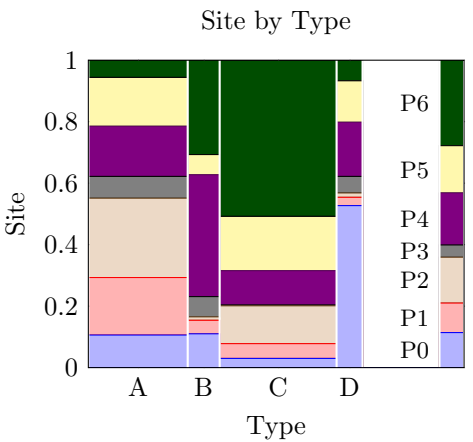
Source: Data courtesy of M. J. Tretter.

The table in this example contains the frequencies (counts) of $J = 4$ different types

of pottery (called potsherds) found at $I = 7$ archaeological sites in an area of the American Southwest. If we divide the frequencies in each row (archaeological site) by the corresponding row total, we obtain a profile of types of pottery. The profiles for the different sites (rows) are shown in a bar graph in the left figure below. The widths of the bars are proportional to the total row frequencies. In general, the profiles are different; however, the profiles for sites P1 and P2 are similar, as are the profiles for sites P4 and P5.

The archaeological site profile for different types of pottery (columns) are shown in a bar graph below, in the figure on the right.





The site profiles are constructed using the column totals. The bars in the figure appear to be quite different from one another. This suggests that the various types of pottery are not distributed over the archaeological sites in the same way.

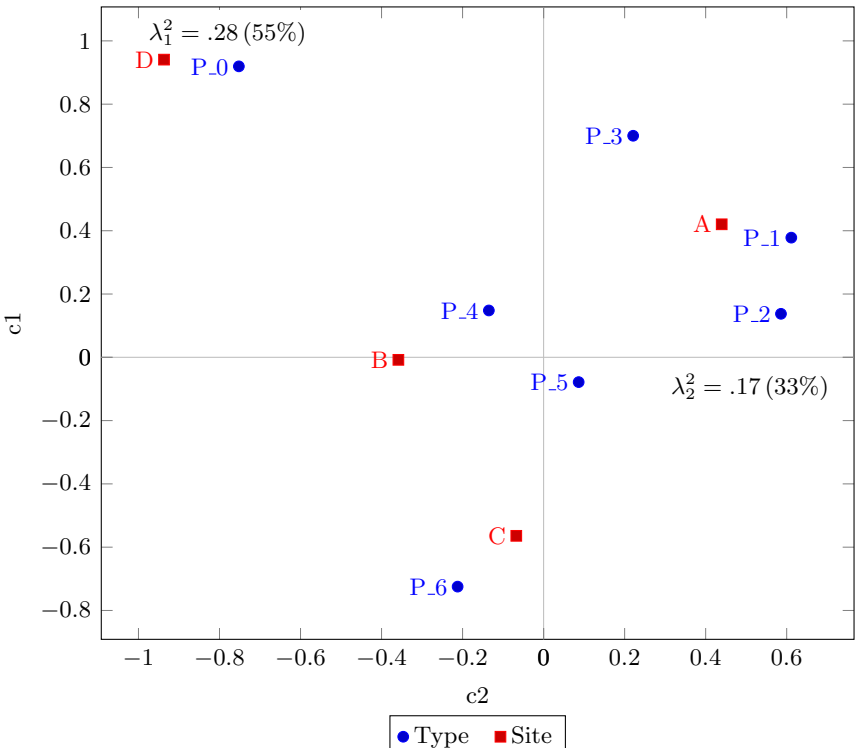
The two-dimensional plot from a correspondence analysis[1] of the pottery type–site data is shown in the next figure.

The plot indicates, for example, that sites P1 and P2 have similar pottery type profiles (the two points are close together), and sites P0 and P6 have very different profiles (the points are far apart). The individual points representing the types of pottery are spread out, indicating that their archaeological site profiles are quite different. These findings are consistent with the profiles pictured previously by the bar charts.

Notice that the points P0 and D are quite close together and separated from the remaining points. This indicates that pottery type D tends to be associated, almost exclusively, with the site P0. Similarly, pottery type A tends to be associated with site P1 and, to lesser degrees, with sites P2 and P3. Pottery type B is associated with sites P4 and P5, and pottery type C tends to be associated, again, almost exclusively, with site P6. Since the archaeological sites represent different periods, these associations are of considerable interest to archaeologists.

[1] the JMP software was used for a correspondence analysis of the data in the frequencies of types of pottery table

A correspondence analysis plot of the pottery type–site data.



The number $\lambda_1^2 = .28$ at the end of the first coordinate axis in the two-dimensional plots is the *inertia* associated with the first dimension. This inertia is 55% of the total inertia. The inertia associated with the second dimension $\lambda_2^2 = .17$, and the second dimension accounts for 33% of the total inertia. Together, the two dimensions account for $55\% + 33\% = 88\%$ of the total inertia. Since, in this case, the data could be exactly represented in three dimensions, relatively little information (variation) is lost by representing the data in two dimensions in the correspondence analysis plot. Equivalently, we may regard this plot as the best two-dimensional representation of the multidimensional scatter of row points and the multidimensional scatter of column points. The combined inertia of 88% suggests that the representation "fits" the data well.

In this example, the graphical output from a correspondence analysis shows the nature of the associations in the contingency table quite clearly. □

Algebraic Development of Correspondence Analysis

To begin, let $\mathbf{X}$, with elements x_{ij} be an $I \times J$ two-way table of unscaled frequencies or counts. In our discussion we take $I > J$ and assume that $\mathbf{X}$ is of full column rank J. The rows and columns of the contingency table bX correspond to different categories of two different characteristics. As an example, the array of frequencies of different pottery types at different archaeological sites, shown in the pottery table, with $I = y$

archaeological sites and $J = 4$ pottery types

If n is the total of the frequencies in the data matrix $\mathbf{X}$, we first construct a matrix of proportions $\mathbf{P} = \{p_{ij}\}$ by dividing each element of $\mathbf{X}$ by n. Hence:

$$p_{ij} = \frac{x_{ij}}{n}, \quad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \text{ or } \underset{(I\times J)}{\mathbf{P}} = \frac{1}{n} \underset{(I\times J)}{\mathbf{X}}$$

The matrix $\mathbf{P}$ is called the *correspondence matrix*.

Next define the vectors of row and column sums $\mathbf{r}$ and $\mathbf{c}$ respectively, and the diagonal matrices $\mathbf{D}_r$ and $\mathbf{D}_c$ with the elements of $\mathbf{r}$ and $\mathbf{c}$ on the diagonals. Thus

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, \qquad i = 1, 2, \ldots, I, \text{ or } \underset{(I\times 1)}{\mathbf{r}} = \underset{(I\times J)}{\mathbf{P}} \underset{(J\times 1)}{\mathbf{1}_J}$$

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, \qquad j = 1, 2, \ldots, J, \text{ or } \underset{(J\times 1)}{\mathbf{c}} = \underset{(I\times J)}{\mathbf{P}'} \underset{(J\times 1)}{\mathbf{1}_I}$$

where $\mathbf{1}_J$ is a $J \times 1$ and 1_I is a $I \times 1$ vector of 1's and

$$\mathbf{D}_r = \text{diag}(r_1, r_2, \ldots, r_I) \text{ and } \mathbf{D}_c = \text{diag}(c_1, c_2, \ldots, c_J)$$

We define the square root matrices

$$\mathbf{D}_r^{1/2} = \text{diag}\left(\sqrt{r_1}, \ldots, \sqrt{r_I}\right) \qquad \mathbf{D}_r^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{r_1}}, \ldots, \frac{1}{\sqrt{r_I}}\right)$$

$$\mathbf{D}_c^{1/2} = \text{diag}\left(\sqrt{c_1}, \ldots, \sqrt{c_J}\right) \qquad \mathbf{D}_c^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{c_1}}, \ldots, \frac{1}{\sqrt{c_J}}\right)$$

for scaling purposes. Correspondence analysis can be formulated as the weighted least squares problem to select $\widehat{\mathbf{P}} = \{\widehat{p}_{ij}\}$, a matrix of specified reduced rank, to minimize

$$\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - \widehat{p}_{ij})^2}{r_i c_j} = \text{tr}\left[(\mathbf{D}_r^{-1/2}(\mathbf{P} - \widehat{\mathbf{P}})\mathbf{D}_c^{-1/2})(\mathbf{D}_r^{-1/2}(\mathbf{P} - \widehat{\mathbf{P}})\mathbf{D}_c^{-1/2})'\right]$$

since $\frac{p_{ij} - \widehat{p}_{ij}}{\sqrt{r_i c_j}}$ is the (i, j) element of $\mathbf{D}_r^{-1/2}(\mathbf{P} - \widehat{\mathbf{P}})\mathbf{D}_c^{-1/2}$.

As the following Result demonstrates, the term $\mathbf{rc}'$ is common to the approximation $\widehat{\mathbf{P}}$ whatever the $I \times J$ correspondence matrix $\mathbf{P}$. The matrix $\widehat{\mathbf{P}} = \mathbf{rc}'$ can be shown to be the best rank 1 approximation to $\mathbf{P}$.

Result. The term $\mathbf{rc}'$ is common to the approximation $\widehat{\mathbf{P}}$ whatever the $I \times J$ correspondence matrix $\mathbf{P}$.

The reduced rank s approximation to $\mathbf{P}$, which minimizes the sum of squares, is given

by

$$\mathbf{P} \doteq \sum_{k=1}^{s} \widetilde{\lambda}_k (\mathbf{D}_r^{1/2} \widetilde{\mathbf{u}}_k)(\mathbf{D}_c^{1/2} \widetilde{\mathbf{v}}_k)' = \mathbf{r}\mathbf{c}' + \sum_{k=2}^{s} \widetilde{\lambda}_k (\mathbf{D}_r^{1/2} \widetilde{\mathbf{u}}_k)(\mathbf{D}_c^{1/2} \widetilde{\mathbf{v}}_k)'$$

where the $\widetilde{\lambda}_k$ are the singular values and the $I \times 1$ vectors $\widetilde{\mathbf{u}}_k$ and the $J \times 1$ vectors $\widetilde{\mathbf{v}}_k$ are the corresponding singular vectors of the $I \times J$ matrix $\mathbf{D}_r^{-1/2} \mathbf{P} \mathbf{D}_c^{-1/2}$. The minimum values of is $\sum_{k=s+1}^{J} \widetilde{\lambda}_k^2$.

The reduced rank $K > 1$ approximation to $\mathbf{P} - \mathbf{r}\mathbf{c}'$ is

$$\mathbf{P} - \mathbf{r}\mathbf{c}' \doteq \sum_{k=1}^{K} \lambda_k (\mathbf{D}_r^{1/2} \mathbf{u}_k)(\mathbf{D}_c^{1/2} \mathbf{v}_k)'$$

where the λ_k are the singular values and the $I \times 1$ vectors $\mathbf{u}_k$ and the $J \times 1$ vectors $\mathbf{v}_k$ are the corresponding singular vectors of the $I \times J$ matrix $\mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r}\mathbf{c}') \mathbf{D}_c^{-1/2}$. Here $\lambda_k = \widetilde{\lambda}_{k+1}$, $\mathbf{u}_k = \widetilde{\mathbf{u}}_{k+1}$, and $\mathbf{v}_k = \widetilde{\mathbf{v}}_{k+1}$ for $k = 1, \ldots, J-1$.

Proof. We first consider a scaled version $\mathbf{B} = \mathbf{D}_r^{-1/2} \mathbf{P} \mathbf{D}_c^{-1/2}$ of the correspondence matrix $\mathbf{P}$. According to the best low rank $= s$ approximation to $\widehat{\mathbf{B}}$ to $\mathbf{D}_r^{-1/2} \mathbf{P} \mathbf{D}_c^{-1/2}$ is given by the first s terms in the singular-value decomposition

$$\mathbf{D}_r^{-1/2} \mathbf{P} \mathbf{D}_c^{-1/2} = \sum_{k=1}^{J} \widetilde{\lambda}_k \widetilde{\mathbf{u}}_k \widetilde{\mathbf{v}}_k'$$

where

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\widetilde{\mathbf{v}}_k = \widetilde{\lambda}_k\widetilde{\mathbf{u}}_k \qquad \widetilde{\mathbf{u}}_k'\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \widetilde{\lambda}_K\widetilde{\mathbf{v}}_k'$$

and

$$|(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2})(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2})' - \widetilde{\lambda}_k^2\mathbf{I}| = 0 \quad \text{for} \quad k = 1, \ldots, J$$

The approximation to $\mathbf{P}$ is then given by

$$\widehat{\mathbf{P}} = \mathbf{D}_r^{1/2}\widehat{\mathbf{B}}\mathbf{D}_c^{1/2} \doteq \sum_{k=1}^{s}\widetilde{\lambda}_k(\mathbf{D}_r^{1/2}\widetilde{\mathbf{u}}_k)(\mathbf{D}_c^{1/2}\widetilde{\mathbf{v}}_k)'$$

and the error of approximation is $\sum_{k=s+1}^{J}\widetilde{\lambda}_k^2$.

Whatever the correspondence matrix $\mathbf{P}$, the term $\mathbf{rc}'$ always provides a (the best) rank one approximation. This corresponds to the assumption of independence of the rows and columns. To see this, let $\widetilde{\mathbf{u}}_1 = \mathbf{D}_r^{1/2}\mathbf{1}_I$ and $\widetilde{\mathbf{v}}_1 = \mathbf{D}_c^{1/2}\mathbf{1}_J$, where $\mathbf{1}_I$ is a $I \times 1$ and $\mathbf{1}_J$ a $J \times 1$ vector of 1's.

$$\begin{aligned}\widetilde{\mathbf{u}}_\mathbf{1}'\left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right) &= \left(\mathbf{D}_r^{1/2}\mathbf{1}_I\right)'\left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)\\ &= \mathbf{1}_I'\mathbf{P}\mathbf{D}_c^{-1/2} = \mathbf{c}'\mathbf{D}_c^{-1/2}\\ &= [\sqrt{c_1}, \ldots, \sqrt{c_J}] = \left(\mathbf{D}_c^{1/2}\mathbf{1}_J\right)' = \widetilde{\mathbf{v}}_\mathbf{1}'\end{aligned}$$

and

$$\begin{aligned}\left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)\widetilde{\mathbf{v}}_{\mathbf{1}} &= \left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)\left(\mathbf{D}_c^{1/2}\mathbf{1}_J\right) \\ &= \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{1}_J = \mathbf{D}_r^{-1/2}\mathbf{r} \\ &= \begin{bmatrix}\sqrt{r_1} \\ \vdots \\ \sqrt{r_I}\end{bmatrix} = \mathbf{D}_r^{1/2}\mathbf{1}_I = \widetilde{\mathbf{u}_{\mathbf{1}}}\end{aligned}$$

That is,

$$(\widetilde{\mathbf{u}}_{\mathbf{1}}, \widetilde{\mathbf{v}}_{\mathbf{1}}) = (\mathbf{D}_r^{1/2}\mathbf{1}_I, \mathbf{D}_c^{1/2}\mathbf{1}_J)$$

are singular vectors associated with singular value $\widetilde{\lambda}_1 = 1$. For any correspondence matrix, $\mathbf{P}$, the common term in every expansion is

$$\mathbf{D}_r^{1/2}\mathbf{u}_1\mathbf{v}_1'\mathbf{D}_c^{1/2} = \mathbf{D}_r\mathbf{1}_I\mathbf{1}_J'\mathbf{D}_c = \mathbf{r}\mathbf{c}'$$

Therefore, we have established the first approximation and can write

$$\mathbf{P} = \mathbf{r}\mathbf{c}' + \sum_{k=2}^{J}\widetilde{\lambda}_k(\mathbf{D}_r^{1/2}\widetilde{\mathbf{u}}_{\mathbf{k}})(\mathbf{D}_c^{1/2}\widetilde{\mathbf{v}}_{\mathbf{k}})'$$

Because of the common term, the problem can be rephrased in terms of $\mathbf{P} - \mathbf{r}\mathbf{c}'$ and its scaled version $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$. By orthogonality of the singular vectors of

$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$, we have $\widetilde{\mathbf{u}}_\mathbf{k}(\mathbf{D}_r^{1/2}\mathbf{1}_I) = 0$ and $\widetilde{\mathbf{v}}'_\mathbf{k}(\mathbf{D}_c^{1/2}\mathbf{1}_J) = 0$, for $k > 1$, so

$$\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \sum_{k=2}^{J} \widetilde{\lambda}_k \widetilde{\mathbf{u}}_\mathbf{k}\widetilde{\mathbf{v}}'_\mathbf{k}$$

is the singular-value decomposition of $\mathbf{D}_r^{1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$ in terms of the singular values and vectors obtained from $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$. Converting to singular values and vectors $\lambda_k, \mathbf{u}_\mathbf{k}$ and $\mathbf{v}_\mathbf{k}$ from $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$ only amounts to changing k to $k-1$ so $\lambda_k = \lambda_{k+1}, \mathbf{u}_\mathbf{k} = \widetilde{\mathbf{u}}_\mathbf{k+1}$, and $\mathbf{v}_\mathbf{k} = \widetilde{\mathbf{v}}_\mathbf{k+1}$ for $k = 1, \ldots, J-1$.

In terms of the singular value decomposition for $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$, the expansion for $\mathbf{P} - \mathbf{r}\mathbf{c}'$ takes the form

$$\mathbf{P} - \mathbf{r}\mathbf{c}' = \sum_{k=1}^{J-1} \lambda_k (\mathbf{D}_r^{1/2}\mathbf{u}_k)(\mathbf{D}_c^{1/2}\mathbf{v}_k)'$$

The best rank K approximation to $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$ is given by $\sum_{k=1}^{K} \lambda_k \mathbf{u}_k \mathbf{v}'_k$. Then, the best approximation to $\mathbf{P} - \mathbf{r}\mathbf{c}'$ is

$$\mathbf{P} - \mathbf{r}\mathbf{c}' \doteq \sum_{k=1}^{K} \lambda_k (\mathbf{D}_r^{1/2}\mathbf{u}_k)(\mathbf{D}_c^{1/2}\mathbf{v}_k)' \qquad \square$$

Remark. Note that the vectors $\mathbf{D}_r^{1/2}\mathbf{u}_k$ and $\mathbf{D}_c^{1/2}\mathbf{v}_k$ in the above expansion $\mathbf{P} - \mathbf{r}\mathbf{c}'$ need not have length 1 but satisfy the sclaing

$$(\mathbf{D}_r^{1/2}\mathbf{u}_k)'\mathbf{D}_r^{-1}(\mathbf{D}_r^{1/2}\mathbf{u}_k) = \mathbf{u}_k'\mathbf{u}_k = 1$$
$$(\mathbf{D}_c^{1/2}\mathbf{v}_k)'\mathbf{D}_c^{-1}(\mathbf{D}_c^{1/2}\mathbf{v}_k) = \mathbf{v}_k'\mathbf{v}_k = 1$$

Because of this scaling, the expansions in the last Result have been called a *generalized singular-value decomposition.*

Let $\mathbf{\Lambda}, \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_I]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_J]$ be the matrices of singular values and vectors obtained from $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$. It is usual in correspondence analysis to plot the first two or three columns of $\mathbf{F} = \mathbf{D}_r^{-1}(\mathbf{D}_r^{1/2}\mathbf{U})\mathbf{\Lambda}$ and $\mathbf{G} = \mathbf{D}_c^{-1}(\mathbf{D}_c^{1/2}\mathbf{V})\mathbf{\Lambda}$ or $\lambda_k\mathbf{D}_r^{-1/2}\mathbf{u}_k$ and $\lambda_k\mathbf{D}_c^{-1/2}\mathbf{v}_k$ for $k = 1, 2,$ and maybe 3.

The joint plot of the coordinates $\mathbf{F}$ and $\mathbf{G}$ is called a *symmetric map* since the points representing the rows and columns have the same normalization, or scaling, along the dimensions of the solution. That is, the geometry for the row points is identical to the geometry for the column points.

Example: Calculations for correspondence analysis Consider the 3×2 contingency table

	B1	B2	Total
A1	24	12	36
A2	16	38	64
A3	60	40	100
	100	100	200

The correspondence matrix is

$$\mathbf{P} = \begin{bmatrix} .12 & .06 \\ .08 & .24 \\ .30 & .20 \end{bmatrix}$$

with marginal totals $\mathbf{c}' = [.5, .5]$ and $\mathbf{r}' = [.18, .32, .50]$. The negative square root matrices are

$$\mathbf{D}_r^{-1/2} = \text{diag}(\frac{\sqrt{2}}{.6}, \frac{\sqrt{2}}{.8}, \sqrt{2}) \qquad \mathbf{D}_c^{-1/2} = \text{diag}(\sqrt{2}, \sqrt{2})$$

Then

$$\mathbf{P} - \mathbf{r}\mathbf{c}' = \begin{bmatrix} .12 & .06 \\ .08 & .24 \\ .30 & .20 \end{bmatrix} - \begin{bmatrix} .18 \\ .32 \\ .50 \end{bmatrix} \begin{bmatrix} .5 & .5 \end{bmatrix} = \begin{bmatrix} .03 & -.03 \\ -.08 & .08 \\ .5 & -.05 \end{bmatrix}$$

The scaled version of this matrix is

$$\mathbf{A} = \mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2} = \begin{bmatrix} \frac{\sqrt{2}}{.6} & 0 & 0 \\ 0 & \frac{\sqrt{2}}{.8} & 0 \\ 0 & 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} .03 & -.03 \\ -.08 & .08 \\ .05 & -.05 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix}$$

$$= \begin{bmatrix} 0.1 & -0.1 \\ -0.2 & 0.2 \\ 0.1 & -0.1 \end{bmatrix}$$

Since $I > J$, the square of the singular values and the $\mathbf{v}_i$ are determined from

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} .1 & -.2 & .1 \\ -.1 & .2 & -.1 \end{bmatrix} \begin{bmatrix} .1 & -.1 \\ -.2 & .2 \\ .1 & -.1 \end{bmatrix} = \begin{bmatrix} .06 & -.06 \\ -.06 & .06 \end{bmatrix}$$

It is easily checked that $\lambda_1^2 = .12, \lambda_2^2 = 0$, since $J - 1 = 1$, and that

$$\mathbf{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$$

Further,

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} .1 & -.1 \\ -.2 & .2 \\ .1 & -.1 \end{bmatrix} \begin{bmatrix} .1 & -.2 & .1 \\ -.1 & .2 & -.1 \end{bmatrix} = \begin{bmatrix} .02 & -.04 & .02 \\ -.04 & .08 & -.04 \\ .02 & -.04 & .02 \end{bmatrix}$$

A computer calculation confirms that the single nonzero eigenvalue is $\lambda_1^2 = .12$, so that the singular value has absolute value $\lambda_1 = .2\sqrt{3}$ and, as you can easily check,

$$\mathbf{u}_1 = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$$

The expansion of $\mathbf{P} - \mathbf{rc}'$ is then the single term

$$\lambda_1(\mathbf{D}_r^{1/2}\mathbf{u}_1)(\mathbf{D}_c^{1/2}\mathbf{v}_1)' = \sqrt{.12}\begin{bmatrix} \frac{.6}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{.8}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}$$

$$= \sqrt{.12}\begin{bmatrix} \frac{.3}{\sqrt{3}} \\ -\frac{.8}{\sqrt{3}} \\ \frac{.5}{\sqrt{3}} \end{bmatrix}\begin{bmatrix} \frac{1}{2} & \frac{-1}{2} \end{bmatrix} = \begin{bmatrix} .03 & -.03 \\ -.08 & .08 \\ .05 & -.05 \end{bmatrix} \quad \text{check}$$

There is only one pair of vectors to plot

$$\lambda_1 \mathbf{D}_r^{1/2}\mathbf{u}_1 = \sqrt{.12}\begin{bmatrix} \frac{.6}{\sqrt{2}} & 0 & 0 \\ 0 & \frac{.8}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} = \sqrt{.12}\begin{bmatrix} \frac{.3}{\sqrt{3}} \\ -\frac{.8}{\sqrt{3}} \\ \frac{.5}{\sqrt{3}} \end{bmatrix}$$

and

$$\lambda_1 \mathbf{D}_c^{1/2}\mathbf{v}_1 = \sqrt{.12}\begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}$$ □

There is a second way to define contingency analysis. Following Greencare, we call the preceding approach the *matrix approximation method* and the approach to follow the *profile approximation method.* We illustrate the profile approximation method using the row profiles; however, an analogous solution results if we were to begin with the column profiles.

Algebraically, the row profiles are the rows of the matrix $\mathbf{D}_r^{-1}\mathbf{P}$, and contingency analysis can be defined as the approximation of the row profiles by points in a low-dimensional space. Consider approximating the row profiles by the matrix $\mathbf{P^*}$. Using the square-root matrices $\mathbf{D}_r^{1/2}$ and $\mathbf{D})c^{1/2}$, we can write

$$(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P^*})\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2}(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P^*})\mathbf{D}_c^{-1/2}$$

and the least squares criterion can be written, with $p^*_{ij} = \frac{\widehat{p}_{ij}}{r_i}$, as

$$\begin{aligned}
\sum_i \sum_j \frac{\left(p_{ij} - \widehat{p}_{ij}\right)^2}{r_i c_j} &= \sum_i r_i \sum_j \frac{\left(\frac{p_{ij}}{r_i} - p^*_{ij}\right)^2}{c_j} \\
&= \operatorname{tr}\left[\mathbf{D}_r^{1/2}\mathbf{D}_r^{1/2}\left(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P^*}\right)\mathbf{D}_c^{-1/2}\mathbf{D}_c^{-1/2}\left(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P^*}\right)'\right] \\
&= \operatorname{tr}\left[\mathbf{D}_r^{1/2}\left(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P^*}\right)\mathbf{D}_c^{-1/2}\mathbf{D}_c^{-1/2}\left(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P^*}\right)'\mathbf{D}_r^{-1/2}\right] \\
&= \operatorname{tr}\left[\left[\left(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P^*}\right)\mathbf{D}_c^{-1/2}\right]\left[\left(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P^*}\right)\mathbf{D}_c^{-1/2}\right]'\right]
\end{aligned}$$

Minimizing the last expression for the trace the last row is precisely the first minimization problem treated in the Proof of the last result. By the singular-value decomposition, $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$ has the singular-value decomposition

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \sum_{k=1}^{J} \widetilde{\lambda}_k \widetilde{\mathbf{u}}_k \widetilde{\mathbf{v}}'_k$$

The best rank K approximation is obtained by using the first K terms of this expansion. Since we have $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{1/2}$ approximated by $\mathbf{D}_r^{1/2}\mathbf{P^*}\mathbf{D}_c^{-1/2}$, we left multiply by $\mathbf{D}_r^{-1/2}$ and right multiply by $\mathbf{D}_c^{1/2}$ to obtain the generalized singular-value decompo-

sition

$$\mathbf{D}_r^{-1}\mathbf{P} = \sum_{k=1}^{J} \widetilde{\lambda}_k \mathbf{D}_r^{-1/2} \widetilde{\mathbf{u}}_\mathbf{k} \left(\mathbf{D}_c^{1/2} \widetilde{\mathbf{v}}_\mathbf{k} \right)'$$

where, knowing that $(\widetilde{\mathbf{u}}_\mathbf{1}, \widetilde{\mathbf{v}}_\mathbf{1}) = (\mathbf{D}_r^{1/2}\mathbf{1}_I, \mathbf{D}_c^{1/2}\mathbf{1}_J)$ are singular vectors associated with singular value $\lambda_1 = 1$. Since $\mathbf{D}_r^{-1/2}(\mathbf{D}_r^{1/2}\mathbf{1}_I) = \mathbf{1}_I$ and $(\mathbf{D}_c^{1/2}\mathbf{1}_J)' = \mathbf{c}'$, the leading term in dcomposition above is $\mathbf{1}_I\mathbf{c}'$.

Consequently, in terms of the singular values and vectors from $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$, the reduced rank $K < J$ approximation to the row prifiles $\mathbf{D}_r^{-1}\mathbf{P}$ is

$$\mathbf{P^*} \doteq \mathbf{1}_I\mathbf{c}' + \sum_{k=2}^{K} \widetilde{\lambda}_k \mathbf{D}_r^{-1/2} \widetilde{\mathbf{u}}_\mathbf{k} (\mathbf{D}_c^{1/2} \widetilde{\mathbf{v}}_\mathbf{k})'$$

In terms of the singular values and vectors $\lambda_k, \mathbf{u}_k$ and $\mathbf{v}_k$ obtainded from $\mathbf{D}_r^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$, we can write

$$\mathbf{P^*} - \mathbf{1}_I\mathbf{c}' \doteq \sum_{k=1}^{K-1} \lambda_k \mathbf{D}_r^{-1/2} \mathbf{u}_\mathbf{k} (\mathbf{D}_c^{1/2} \mathbf{v}_\mathbf{k})'$$

(Row profiles for the archaeological data were shown in the "Type by Site" histogram above.)

Inertia

Total inertia is a measure of the variation in the count data and is defined as the weighted sum of squares

$$\operatorname{tr}\left[\mathbf{D}_r^{-1/2}\left(\mathbf{P}-\mathbf{r}\mathbf{c}'\right)\mathbf{D}_c^{-1/2}\left(\mathbf{D}_r^{-1/2}\left(\mathbf{P}-\mathbf{r}\mathbf{c}'\right)\mathbf{D}_c^{-1/2}\right)'\right] = \sum_i\sum_j \frac{(p_{ij}-r_ic_j)^2}{r_ic_j} = \sum_{k=1}^{J-1}\lambda_k^2$$

where the λ_k are the singular values obtained from the singular-value decomposlition of $\mathbf{D}_r^{-1/2}(\mathbf{P}-\mathbf{r}\mathbf{c}')\mathbf{D}_c^{-1/2}$ (see the Proof of the previous Result).[2]

[2]Total inertia is related to the chi-square measure of association in a two-way contingency table, $\chi^2 = \sum \frac{(O_{ij}-E_{ij}^2}{E_{ij}}$. Here $O_{ij} = x_{ij}$ is the observed frequency and E_{ij} is the expected frequency for the ij-th cell. In our context if the row variable is independent of (unrelated to) the column variable, $E_{ij} = nr_ic_j$, and

$$\text{Total inertia} = \sum_{i-1}^{I}\sum_{j=1}^{J}\frac{(p_{ij}-r_ic_j)^2}{r_ic_j} = \frac{\chi^2}{n}$$

Interpretation in Two Dimensions

Since the inertia is a measure of the data table's total variation, how do we interpret a large value for the proportion $\frac{\lambda_1^2 + \lambda_2^2}{\sum_{k=1}^{I-1} \lambda_k^2}$? Geometrically, we say that the associations in the centered data are well represented by points in a plane, and this best approximating plane accounts for nearly all the variation in the data beyond that accounted for by the rank 1 solution (independence model). Algebraically, we say that the approximation

$$\mathbf{P} - \mathbf{r}\mathbf{c}' \doteq \lambda_1 \mathbf{u}_1 \mathbf{v}_1' + \lambda_2 \mathbf{u}_2 \mathbf{v}_2'$$

is very good or, equivalently, that

$$\mathbf{P} \doteq \mathbf{r}\mathbf{c}' + \lambda_1 \mathbf{u}_1 \mathbf{v}_1' + \lambda_2 \mathbf{u}_2 \mathbf{v}_2'$$

Final Comments

Correspondence analysis is primarily a graphical technique designed to represent associations in a low-dimensional space. It can be regarded as a scaling method, and can be viewed as a complement to other methods such as multidimensional scaling and biplots. Correspondence analysis also has links to principal component analysis and canonical correlation analysis.

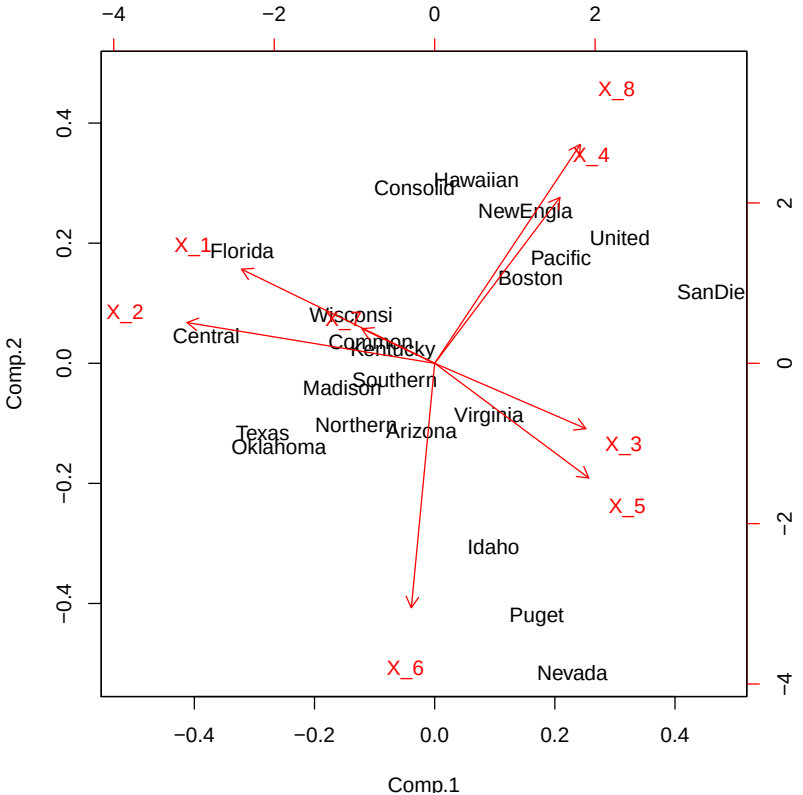
Biplots for Viewing Sampling Units and Variables

A *biplot* is a graphical representation of the information in an $n \times p$ data matrix. The *bi-* refers to two kinds of information contained in a data matrix. The information in the rows pertains to samples or sampling units and that in the columns pertains to variables.

When there are only two variables, scatter plots can represent the information on both the sampling units and the variables in a single diagram. This permits the visual inspection of the position of one sampling unit relative to another and the relative importance of each of the two variables to the position of any unit.

With several variables, one can construct a matrix array of scatter plots, but there is no one single plot of the sampling units. On the other hand, a two-dimensional plot of the sampling units can be obtained by graphing the first two principal components, as has been shown in previous lessons. The idea behind biplots is to add the information about the variables to the principal component graph.

The figure below gives an example of a biplot for the data from the Public Utility Data table, as cited in an example before.



A biplot of the data on public utilities.

You can see how the companies group together and which variables contribute to their positioning within this representation. For instance, X_1 = fixed-charge ratio and X_2 = rate of return on capital put the Florida and Louisiana companies together.

Constructing Biplots

The construction of a biplot proceeds from the sample principal components.

We already know, that the best two-dimensional approximation to the data matrix $\mathbf{X}$ approximates the jth observation $\mathbf{x}_j$ in terms of the sample values of the first two principal components In particular,

$$\mathbf{x}_j \doteq \bar{\mathbf{x}} + \widehat{y}_{j1}\widehat{\mathbf{e}}_{\mathbf{1}} + \widehat{y}_{j2}\widehat{\mathbf{e}}_{\mathbf{2}}$$

where $\widehat{\mathbf{e}}_{\mathbf{1}}$ and $\widehat{\mathbf{e}}_{\mathbf{2}}$ are the first two eigenvectors of $\mathbf{S}$ or, equivalently, of $\mathbf{X}_c{}'\mathbf{X}_c = (n-1)\mathbf{S}$. Here, $\mathbf{X}_c$ denotes the mean corrected data matrix with rows $(\mathbf{x}_j - \bar{\mathbf{x}})'$. The eigenvectors determine a plane, and the coordinates of the jth unit (row) are the pair of values of the principal components, $(\widehat{\lambda}_{j1}, \widehat{\lambda}_{j2})$.

To include the information on the variables in this plot, we consider the pair eigenvectors $(\widehat{\mathbf{e}}_{\mathbf{1}}, \widehat{\mathbf{e}}_{\mathbf{2}}$. These eigenvectors are the coefficient vectors for the first two sample principal components. Consequently, each row of the matrix $\widehat{\mathbf{E}} = [\widehat{\mathbf{e}}_{\mathbf{1}}, \widehat{\mathbf{e}}_{\mathbf{2}}]$ positions a variable in the graph, and the magnitudes of the coefficients (the coordinates of the

variable) show the weightings that variable has in each principal component. The positions of the variables in the plot are indicated by a vector. Usually, statistical computer programs include a multiplier so that the lengths of all of the vectors can be suitably adjusted and plotted on the same axes as the sampling units. Units that are close to a variable likely have high values on that variable. To interpret a new point $\mathbf{x}_0$, we plot its principal components $\widehat{\mathbf{E}}'(\mathbf{x}_0 - \bar{\mathbf{x}})$.

A direct approach to obtaining a biplot starts from the singular value decomposition which first expresses the $n \times p$ mean corrected matrix $\mathbf{X}_c$ as

$$\underset{(n\times p)}{\mathbf{X}_c} = \underset{(n\times p)}{\mathbf{U}} \; \underset{(p\times p)}{\Lambda} \; \underset{(p\times p)}{\mathbf{V}'}$$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\mathbf{V}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{X}_c{}'\mathbf{X}_c = (n-1)\mathbf{S}$. That is, $\mathbf{V} = \widehat{\mathbf{E}} = [\widehat{\mathbf{e}}_\mathbf{1}, \widehat{\mathbf{e}}_\mathbf{2}, \ldots, \widehat{\mathbf{e}}_\mathbf{p}]$. Multiplying the above equation

$$\mathbf{X}_c\widehat{\mathbf{E}} = \mathbf{U}\mathbf{\Lambda}$$

where the jth row of the left-hand side,

$$\left[(\mathbf{x}_j - \bar{\mathbf{x}})'\,\widehat{\mathbf{e}}_\mathbf{1}, (\mathbf{x}_j - \bar{\mathbf{x}})'\,\widehat{\mathbf{e}}_\mathbf{2}, \ldots, (\mathbf{x}_j - \bar{\mathbf{x}})'\,\widehat{\mathbf{e}}_\mathbf{p}\right] = [\widehat{y}_{j1}, \widehat{y}_{j2}, \ldots, \widehat{y}_{jp}]$$

is just the value of the principal components for the jth item. That is, $\mathbf{U}\mathbf{\Lambda}$ contains all of the values of the principal components, while $\mathbf{V} = \widehat{\mathbf{E}}$ contains the coefficients that define the principal components.

The best rank 2 approximation to $\mathbf{X}_c$ is obtained by replacing $\mathbf{\Lambda}$ by

$$\mathbf{\Lambda^*} = \text{diag}(\lambda_1, \lambda_2, 0, \ldots, 0).$$

This is called the *Eckart–Young theorem.* The approximation is then

$$\mathbf{X}_c \doteq \mathbf{U\Lambda^*\,V}' = [\widehat{\mathbf{y}}_1, \widehat{\mathbf{y}}_2] \begin{bmatrix} \widehat{\mathbf{e}}_1' \\ \widehat{\mathbf{e}}_2' \end{bmatrix}$$

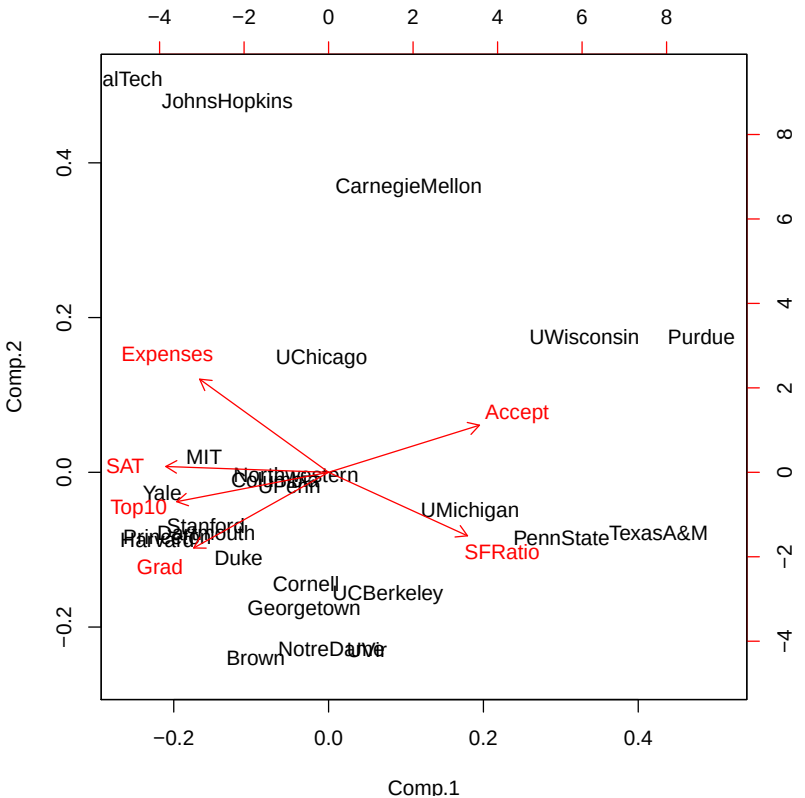
where $\widehat{\mathbf{y}}_1$ is the $n \times 1$ vector of values of the first principal component and $\widehat{\mathbf{y}}_2$ is the $n \times 1$ vector of values of the second principal component.

In the biplot, each *row* of the data matrix, or item, is represented by the point located by the pair of values of the principal components. The ith *column* of the data matrix, or variable, is represented as an arrow from the origin to the point with coordinates (e_{1i}, e_{2i}), the entries in the ith column of the second matrix $[\widehat{\mathbf{e}}_1, \widehat{\mathbf{e}}_2]'$ in the approximation above. This scale may not be compatible with that of the principal components, so an arbitrary multiplier can be introduced that adjusts all of the vectors by the same amount.

The idea of a biplot, to represent both units and variables in the same plot, extends to canonical correlation analysis, multidimensional scaling, and even more complicated nonlinear techniques.

Example: A biplot of universities and their characteristics The Data on universities table, as presented before, gives the data on some universities for certain variables used to compare or rank major universities. These variables indlude $X_1 =$ average SAT core of new freshmen, $X_2 =$ percentage of new freshmen in top 10% of high school class, $X_3 =$ percentage of applicants accepted, $X_4 =$ student–faculty ratio, $X_5 =$ estimated annual expenses and $X_6 =$ graduation rate (%).

Because the two of the variables, SAT and Expenses, are on a much different scale from that of the other variables, we standardize the data and base our biplot on the matrix of standardized observations $\mathbf{x}_i$. The biplot is given in figure below.



A biplot of the data on universities.

Notice how Cal Tech and Johns Hopkins are off by theselves; the variable Expense is mostly responsible for this positioning. The large state universities in our sample are to be left in the biplot, and most of our private schools are on the right. Large values for the variables SAT, Top10, and Grad are associated with the private schoool group. Northwestern lies in the middle of the biplot. □