Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

A Method for Comparing Two Hierarchical Clusterings

E. B. Fowlkes [a] & C. L. Mallows [a]
[a] Bell Laboratories, Murray Hill, NJ, 07974, USA

Version of record first published: 12 Mar 2012.

To cite this article: E. B. Fowlkes & C. L. Mallows (1983): A Method for Comparing Two Hierarchical Clusterings, Journal of the American Statistical Association, 78:383, 553-569

To link to this article: http://dx.doi.org/10.1080/01621459.1983.10478008

A Method for Comparing Two Hierarchical Clusterings

E. B. FOWLKES and C. L. MALLOWS*

This article concerns the derivation and use of a measure of similarity between two hierarchical clusterings. The measure, B_k, is derived from the matching matrix, $[m_{ij}]$, formed by cutting the two hierarchical trees and counting the number of matching entries in the k clusters in each tree. The mean and variance of B_k are determined under the assumption that the margins of $[m_{ij}]$ are fixed. Thus, B_k represents a collection of measures for $k = 2, \ldots, n - 1$. (k, B_k) plots are found to be useful in portraying the similarity of two clusterings. B_k is compared to other measures of similarity proposed respectively by Baker (1974) and Rand (1971). The use of (k, B_k) plots for studying clustering methods is explored by a series of Monte Carlo sampling experiments. An example of the use of (k, B_k) on real data is given.

KEY WORDS: Clustering; Measures of similarity; Statistical graphics.

1. INTRODUCTION

In carrying out a cluster analysis of a set p-dimensional multivariate data, one may wish to compare two or more hierarchical clusterings of the same set of objects. For example, one may wish to study the effect of using different metrics, or different clustering algorithms, or data from two different sources. Informally, one can inspect the clusterings to determine the important clusters, and the composition of the clusters from one tree can be compared with the composition of the clusters from a second tree. Matches indicate similarity between the two trees. This method of comparison of two clusterings can be extremely laborious and time-consuming and affords no measurement of strength of the comparisons. In this article we propose a method of comparing two hierarchical clusterings that gives a numerical measure for the degree of similarity. Not only does the method provide a comparison between two clusterings, it also is useful as a tool for studying hierarchical clustering in general. See Hartigan (1975) for a description of hierarchical clustering algorithms.

The article has five sections. Section 2 derives the measure of comparison, B_k, and establishes certain statistical properties. Section 3 discusses alternative methods of comparison developed by Baker (1974) and Rand (1971). Section 4 studies various properties of B_k via Monte Carlo sampling, and section 5 describes some uses of the measure, B_k, for a set of real data.

2. DERIVATION OF B_k

Suppose that we have two hierarchical clusterings of the same number of objects, n, which we may label A_1 and A_2. The hierarchical trees representing A_1 and A_2 are then cut to produce $k = 2, \ldots, n - 1$ clusters for each tree. The cutting of a hierarchical tree simply corresponds to setting a value of height or cluster strength and determining the cluster assignments of the tree at that strength. For each value of k we may label the clusters for A_1 and A_2 arbitrarily from one to k and form the matrix

$$M = [m_{ij}] \quad (i = 1, \ldots, k; j = 1, \ldots, k). \tag{2.1}$$

where the quantity m_{ij} is the number of objects in common between the ith cluster of A_1 and the jth cluster of A_2. Our measure of association is then defined to be

$$B_k = T_k / \sqrt{P_k Q_k}, \tag{2.2}$$

where

$$T_k = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}^2 - n, \tag{2.3}$$

$$m_{i\cdot} = \sum_{j=1}^{k} m_{ij}, \tag{2.4}$$

$$m_{\cdot j} = \sum_{i=1}^{k} m_{ij}, \tag{2.5}$$

$$m_{\cdot\cdot} = n = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}, \tag{2.6}$$

$$P_k = \sum_{i=1}^{k} m_{i\cdot}^2 - n, \tag{2.7}$$

$$Q_k = \sum_{j=1}^{k} m_{\cdot j}^2 - n. \tag{2.8}$$

B_k is calculated for every value of k, and a portrayal of the similarity of the two clusterings may be given by plotting B_k versus k. We have $0 \le B_k \le 1$ for each k. $B_k = 1$ when M has exactly k nonempty cells, which happens when the k clusters in each clustering correspond com-

* E.B. Fowlkes is a Member of Technical Staff and C.L. Mallows is Department Head, Bell Laboratories, Murray Hill, NJ 07974.

© Journal of the American Statistical Association
September 1983, Volume 78, Number 383
Invited Papers Section

pletely. Also $B_k = 0$ when each m_{ij} is 0 or 1, so that every pair of objects that appear in the same cluster in A_1 are assigned to different clusters in A_2. When $k = n$, M is a permutation matrix, and B_k is indeterminate. Figure 1 shows two hierarchical trees that have each been cut to produce 2 clusters and the corresponding matrix, m_{ij}, resulting from the cuts; B_k for $k = 2$ is .25.

B_k is also related in the following way to the sum over all pairs of objects of those pairs that have matching cluster assignments. For a given value of k, let $a_{uv} = 1$ if objects u and v are in the same cluster in both trees, and $a_{uv} = 0$ otherwise. Then

$$2 \sum_{u,v \in S} a_{uv} = 2 \sum_{i=1}^{k} \sum_{j=1}^{k} \binom{m_{ij}}{2} = \sum_{i=1}^{k} \sum_{j=1}^{k} m_{ij}^2 - n, \quad (2.9)$$

where S is the set of all pairs of objects. The last quantity in the expression is the numerator of B_k. Rand (1971) and Hartigan (1975) also considered similarity measures calculated from the matching matrix, $[m_{ij}]$. Rand's measure is considered in Sections 3 and 4. Hartigan proposed measures based on $[m_{ij}]$ but does not establish their properties or use them on real data.

Another interpretation of B_k is as a version of Daniels's generalized correlation coefficient (Daniels 1944). For objects u, v let $a_{uv} = sgn(u - v)$ if u, v are in the same cluster in A_1, and $a_{uv} = 0$ if not; similarly $b_{uv} = sgn(u - v)$ if u, v are in the same cluster in A_2, and $b_{uv} = 0$ if not. Then $B_k = \sum a_{uv} b_{uv} / (\sum a_{uv}^2 \sum b_{uv}^2)^{1/2}$, where all sums are over all u, v.

In defining B_k we have chosen to cut each tree to produce the same number of clusters, k. If there are exact ties in cluster strengths, one of the subtrees is selected at random for cutting in order to produce k clusters. Alternatively, we could have cut one tree to produce k clusters and the other to produce l clusters. The matrix $[m_{ij}]$ would then be k by l, and one could define a similarity measure B_{kl}. We have not considered B_{kl} in this article.

The mean and variance of B_k can be derived under the assumption that conditional on the margins of the matrix, $[m_{ij}]$, namely $(m_{i\cdot}, m_{\cdot j})$ being fixed, there is random allocation of the objects to the cells. This assumption is valid, for example, if the two clusterings are unrelated to each other. The Appendix derives the results

$$E(B_k) = \sqrt{P_k Q_k}/n(n-1),$$

$$\text{var}(B_k) = \frac{2}{n(n-1)} + \frac{4P_k' Q_k'}{n(n-1)(n-2)P_k Q_k} \quad (2.10)$$

$$+ \frac{(P_k - 2 - 4P_k'/P_k)(Q_k - 2 - 4Q_k'/Q_k)}{n(n-1)(n-2)(n-3)}$$

$$- \frac{P_k Q_k}{n^2(n-1)^2}, \quad (2.11)$$

where

$$P_k' = \sum_{i=1}^{k} m_{i\cdot}(m_{i\cdot} - 1)(m_{i\cdot} - 2), \quad (2.12)$$

$$Q_k' = \sum_{j=1}^{k} m_{\cdot j}(m_{\cdot j} - 1)(m_{\cdot j} - 2). \quad (2.13)$$

Plots of B_k versus k can be enhanced by the addition of $E(B_k)$, and the limits $E(B_k) \pm 2\,(\text{var}(B_k))^{1/2}$. If B_k falls outside the limits in a systematic fashion, similarity of the two clusters may be significant. We feel that the assumptions used to derive $E(B_k)$ are reasonable since values of B_k calculated for unrelated pairs of bivariate normal random samples of a moderate size ($n = 100$) nearly always fell within the above limits. This result can be seen in Section 4. Since successive values of B_k are correlated and the distribution of B_k is not normal, the defined limits give only an approximate indication of the significance of the similarity between two hierarchical clusterings.

In general, a measure of similarity between two hierarchical clusterings could depend on three things: the topologies of the two trees that represent the clusterings, the assignment of labels to the terminal nodes of these trees, and the heights (or "weights") of the internal nodes. Our measure B_k ignores the third of these components; indeed it is difficult to see how to use this information when the two clusterings are based on different metrics or different algorithms. However, our measure does depend on the first two components. Notice that the two trees in Figure 1 have identical topologies. Therefore B_k depends not only on the similarity of the topologies of the trees but also on the labeling of the nodes or objects in the trees. Figure 2(a) shows that with appropriate relabeling of the nodes in both trees, $B_k = 1.0$ for all values of k. Figure 2(b) gives a relabeling of the nodes for A_1 (Figure 2(a)), which increases B_2 in Figure 1 from .25 to .50. This poses an interesting question. Can B_k be decomposed into two parts, one measuring the similarity of the tree topologies, and one measuring the similarity of the node labeling given the topology? Clearly we can write $B_k = B_k' B_k''$, where B_k' is the maximum value of B_k over all possible relabelings of the two trees, holding their topologies fixed, and B_k'' is B_k / B_k'. Unfortunately, for large trees computation of B_k' does not appear to be feasible, and we have not developed this idea.

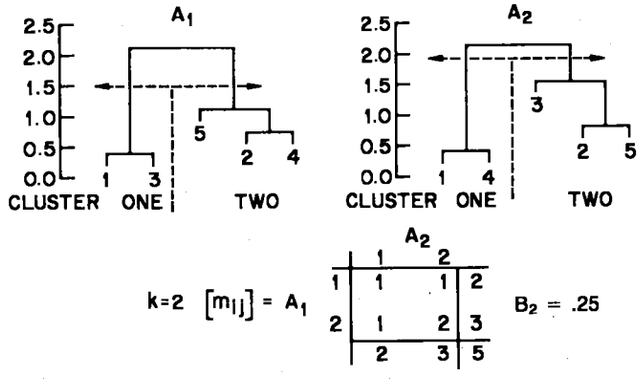

Figure 1. Dendrograms for two hierarchical clusterings of five objects and the formation of the $[m_{ij}]$ matrix for $k = 2$.

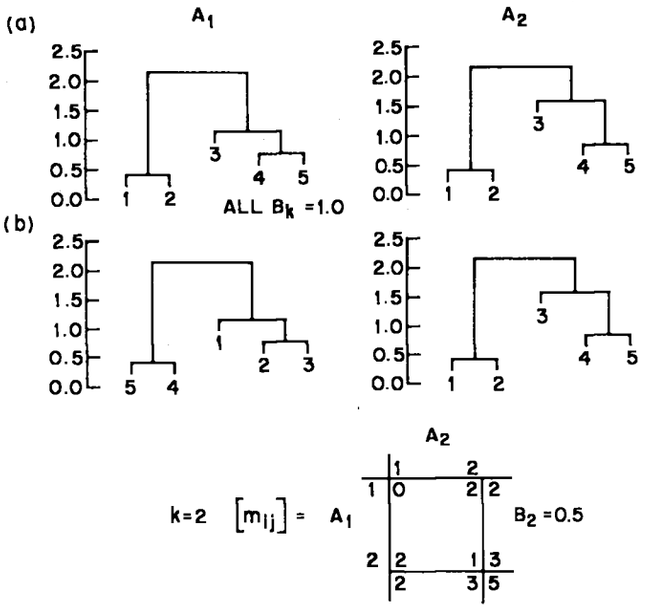

Figure 2. Dendrograms of two hierarchical clusterings of five objects showing the effect on B_k of relabeling of the objects.

3. OTHER MEASURES OF SIMILARITY

Several statistics have been developed to measure the similarity between hierarchical clusterings. Johnson (1968), Rand (1971), Anderberg (1973), and Hubert and Levin (1976) all proposed measures that are functions of the matrix, $[m_{ij}]$. Arabie and Boorman (1973) used multidimensional scaling to study the similarity between the measures themselves. They studied 12 measures, of which some were based on $[m_{ij}]$ while others used ideas from information theory (see Jardine and Sibson 1971 for example). Baker (1974) proposed a measure that uses rank correlation methods.

These measures have been used in a very different fashion from the use we propose for B_k. Either they use one number to summarize the similarity between two hierarchical clusterings or they compare the clusterings for some fixed number of partitions. None propose the use of a sequence of measures as the basis for a plotting procedure of the form that we have proposed for B_k. Nevertheless, we have chosen the measures of Rand (1971) and Baker (1974) for comparisons with B_k. These measures are presented and discussed in the following paragraphs. Certain comparisons of the Rand and Baker measures with B_k are given in Section 4 on Monte Carlo sampling experiments.

Rand's measure (which we henceforth refer to as R_k) is similar in construction to B_k since it is based on the matrix, $[m_{ij}]$. R_k may be defined as follows. Let n be the total number of objects to be clustered. Then, for a given number of clusters, k, R_k is defined as the ratio of the sum of the number of pairs of objects that cluster together in the two clusterings under comparison and the number of pairs of objects that fall in different clusters in both clusterings to the total number of pairs, $\binom{n}{2}$. Thus, R_k can be interpreted as the probability that two objects are treated alike in both clusterings. Furthermore, R_k can be shown to be

$$R_k = \left[T_k - \frac{1}{2}P_k - \frac{1}{2}Q_k + \binom{n}{2}\right] \Big/ \binom{n}{2} \quad (3.1)$$

and $0 \le R_k \le 1$. The statistics B_k and R_k may be thought of as resulting from two different methods of scaling T_k (see (2.3)) to lie in the unit interval. Rand did numerous Monte Carlo sampling experiments to determine properties of R_k and its utility in studying clustering methods. However, he derived few formal properties of R_k. Using the results (2.10) and (2.11) we may derive moments of R_k under the assumption of fixed margins, $m_{i\cdot}$ and $m_{\cdot j}$, and random allocation of matching counts of objects to $[m_{ij}]$. Thus the mean and variance of R_k can be shown to be

$$E(R_k) = 1 - (P_k + Q_k)/((n(n-1)) + (2P_kQ_k)/(n^2(n-1)^2), \quad (3.2)$$

$$\text{var}(R_k) = (4P_kQ_k)/(n^2(n-1)^2)\,\text{var}(B_k), \quad (3.3)$$

where $\text{var}(B_k)$ is defined by (2.7), (2.8), and (2.11), respectively.

It is easy to see that $E[R_k] \to 1.0$ as $k \to n$. In practice, as we show in Section 4, we have found that this limit is approached relatively quickly as k increases. This property is not seen in Rand's paper since he never chooses k to be large in relation to n (he has $k < 10$ for $n = 30$ for example). Also $\text{var}(R_k) \to 0$ as $k \to n$, and in practice we have found that the possible range of R_k is quite narrow for all values of k. The statistic B_k on the other hand can vary in a much wider range. (It should be pointed out, however, that neither R_k nor B_k can be zero for small values of k.)

For these reasons we feel that R_k is somewhat inappropriate for use in a plotting procedure analogous to that for B_k. In Section 4 (k, B_k) plots are shown to have interesting configurations for both null and nonnull cases of clustering. The configurations of (k, R_k) plots are largely dominated by the rapid approach of R_k to 1 and the narrow range of variation so that these configurations are masked.

To establish some formal properties of R_k, Rand derived results concerning clusters of a special type. Specifically, Rand derived values for comparing R_k for an original clustering (which had k clusters each with n points) with various simple and major modifications of this clustering. For example, one modification was to move one point from one cluster to another. Rand also determined the limiting behavior of R_k when either the number of clusters, k, or the number of observations, n, becomes infinite. We have reproduced Rand's results along with side-by-side comparisons of R_k with B_k in Table 1.

A third plotting procedure would use a centralized and

Table 1. Comparisons of Expressions for the Measures B_k and R_k Between Two Similar Clusterings, Given an Initial Clustering, Y, Which Has k Clusters of n Points Each

Modification of Y	R_k	$\lim R_k$ $n \to \infty$, k fixed	$\lim R_k$ $n \to \infty$, $k = \lambda n$	B_k	$\lim B_k$ $n \to \infty$, k fixed	$\lim B_k$ $n \to \infty$, $k = \lambda n$
	a. B_k and R_k Applied to the Clusterings Y and Y', Where Y' is a Simple Modification of an Original Clustering, Y					
Two Clusters Joined	$\frac{(k^2-2)n-k}{k^2n-k}$	$\frac{k^2-2}{k^2}$	1	$\left(\frac{k(n-1)}{k(n-1)+2n}\right)^{1/2}$	$\left(\frac{k}{k+1}\right)^{1/2}$	1
One Cluster Split Into Two Equal Parts (n even)	$\frac{(2k^2-1)n-2k}{2k^2n-2k}$	$\frac{2k^2-1}{2k^2}$	1	$\left(\frac{(k-\frac{1}{2})n-k}{k(n-1)}\right)^{1/2}$	$\left(\frac{k-\frac{1}{2}}{k}\right)^{1/2}$	1
One Cluster Split Into Single Point Clusters	$\frac{(k^2-1)n-k+1}{k^2n-k}$	$\frac{k^2-1}{k^2}$	1	$\left(\frac{(k-1)}{k}\right)^{1/2}$	$\left(\frac{k-1}{k}\right)^{1/2}$	1
One Point Taken From Each Cluster to Form a New Cluster of k Points	$\frac{kn^2-3n-k+3}{kn^2-n}$	1	1	$\frac{(n-2)(n-1)^{1/2}}{(n((n-1)^2-n+k}$	1	1
	b. B_k and R_k Applied to Clusterings Y' and Y'', Which are Similar Modifications of an Original Clustering, Y					
Movement of a Point to Different Clusters	$\frac{k^2n-k-4}{k^2n-k}$	1	1	$\frac{kn^2-kn-2n+2}{kn^2-n+2}$	1	$\frac{n-1}{n}$
Different Clusters Split Into Two Equal Parts	$\frac{(k^2-1)n-k}{k^2n-k}$	$\frac{k^2-1}{k^2}$	1	$\frac{kn-n-k}{kn\frac{n}{2}k}$	$\frac{k-1}{k\frac{1}{2}}$	1
Different Pairs of Clusters Joined	$\frac{(k^2-4)n-k}{k^2n-k}$	$\frac{k^2-4}{k^2}$	1	$\frac{k(n-1)}{kn+2n-k}$	$\frac{k}{k+2}$	1
	c. B_k and R_k Applied to Clusterings Y and Y', Where Y' is a Major Modification of an Original Clustering, Y					
All Clusters Joined Into One Large Cluster	$\frac{n-1}{kn-1}$	$\frac{1}{k}$	0	$\left(\frac{n-1}{kn-1}\right)^{1/2}$	$\frac{1}{\sqrt{k}}$	0
All Clusters Split Into Single Point Clusters	$\frac{(k-1)n}{kn-1}$	$\frac{k-1}{k}$	1	0	0	0
n Clusters are Formed With k Points in Each, One Point From Each Original Cluster	$\frac{(k-1)(n-1)}{kn-1}$	$\frac{k-1}{k}$	$\frac{n-1}{n}$	0	0	0

standardized version of T_k (see (2.3)), namely,

$$(k, (T_k - E(T_k))/(\text{var}(T_k))^{1/2}). \tag{3.4}$$

We have chosen not to do this since it conceals changes in $E(T_k)$ as k varies, that can give useful information concerning the margins $m_{i\cdot}$ and $m_{\cdot j}$.

Baker (1974) has derived a method of comparing two hierarchical trees that uses the gamma index, γ, of Goodman and Kruskal (1974), which is a measure of rank correlation. The measure is defined as the rank correlation between stages at which pairs of objects combine in the trees. Even though Baker's measure performed quite well in the studies reported in Section 4, we feel that it tends to mask certain comparisons between clusterings since it does summarize similarity by one number.

4. MONTE CARLO SAMPLING EXPERIMENTS

In this section we report on our study of various properties of B_k by means of a series of Monte Carlo experiments. First, we wished to determine the behavior of B_k for the null case where the clusterings being compared were unrelated. Second, we studied the effect of small perturbations of the coordinates of the objects being clustered; we calculated B_k between the original clustering and the perturbed clustering. We also studied one aspect

of the choice of metric, by examining the effect of scale changes of the coordinates, the effect of outliers, and the addition of variables containing no cluster structure to a set of variables having pronounced cluster structure. Last, we studied the sampling distribution of B_k. There are of course other questions one might raise, but we felt that these were crucial and interesting ones and that answering them would give us a deeper understanding of exactly what B_k was measuring.

We attempted to investigate these issues by generating bivariate normal samples of various types and calculating B_k for pairs of clusterings of these samples. The bivariate normal samples were generated in the following fashion. First, a pair of pseudorandom numbers U_1 and U_2 distributed uniformly on the interval (0, 1) were generated according to the method of Coveyou and MacPherson (1967). This method is congruential with modulus (2^{35}) and multiplier 273673163155 (base 8). U_1 and U_2 were transformed into pseudorandom unit normal deviates, z_1 and z_2, according to the method of Marsaglia (1962) via the "box-wedge-tail" algorithm. The deviates, z_1 and z_2, were transformed into a bivariate normal observation (y_1, y_2) with the covariance matrix

$$V = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}; \tag{4.1}$$

using the following transformation

$$\begin{aligned} y_1 &= \sigma_1 z_1 \\ y_2 &= \rho\sigma_2 z_1 + \sigma_2 (1 - \rho^2)^{1/2} z_2. \end{aligned} \tag{4.2}$$

4.1 Null Case

For the null case we generated 20 pairs of bivariate normal samples of size $n = 100$ and

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.0 \end{bmatrix}; \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}; \rho = .5. \tag{4.3}$$

We then performed hierarchical clusterings for each pair using the complete linkage method and Euclidean distance. In the complete linkage method, the distance between two clusters is the maximum of the distances between pairs of points, one in each cluster. B_k was calculated from the hierarchical trees for each pair. There was no reason to believe that the members of a pair would be related in any way. Figure 3(a) shows (k, B_k) plots for 2 of the 20 pairs. On each plot are superimposed $E(B_k)$ and $E(B_k) \pm 2\,(\text{var}(B_k))^{1/2}$. Notice that in general B_k falls inside the limits as one might expect for pairs of samples that are unrelated. (We shall soon see that the distribution of B_k is quite asymmetric for $k \geq 60$.) Figure 3(b) shows (k, R_k) plots for the same 2 samples that were considered in Figure 3(a). On each plot are superimposed $E(R_k) \pm 2(\text{var}(R_k))^{1/2}$. The plots show a general configuration that was anticipated by remarks in Section 3. The value of R_k rapidly approaches 1.0 for $k > 30$. We do not feel that this property is reasonable for samples that are unrelated. We shall see that when the two clusterings are related, our statistic B_k takes values that are usefully spread out over the interval (0, 1); in such cases the corresponding R_k values would be concentrated very near to 1. Notice that the 2 σ limits are very narrow indeed and that R_k agrees well with its expectation.

We were also interested in what the Baker measure, γ, would be for these 20 pairs of unrelated bivariate normal samples and how it would compare with B_k. To investigate this we made box plots of B_k for certain selected values of k and for γ. Box plots (Tukey 1977) are a tool for summarizing the distribution of a sample. The center line of the box is the sample median and the lower and upper extremities of the box are respectively the 25th and 75th percentiles. A line is drawn from the extremities of the box to the minimum and maximum of the sample respectively. Figure 4, shows box plots for B_k and γ when $k = 2, 10, 20, 30, 40, 50, 60, 70, 80, 90$. The box plot for γ shows that the median γ for the 20 samples is approximately zero. The distribution of γ is also shown to be quite symmetric with very small standard deviation (the interquartile range is about .001). Baker's γ, thus, performs quite well in the null case, where the clusters being compared are unrelated. We see that B_k has a very skewed distribution for $k \geq 60$ in this case.

4.2 Perturbation Experiments

4.2.1 Single Cluster Experiments. Figures 5 and 6 show comparisons using (k, B_k) plots for clusterings of original and perturbed data sets. Specifically, 20 samples of size 100 were generated from the bivariate normal distribution with parameters as in (4.3). For each sample, the coordinates (x_i, y_i), $(i = 1, \ldots, n)$ were perturbed by adding a pair of small random deviates ϵ_1 and ϵ_2, where (ϵ_1, ϵ_2) were normal with mean zero and standard deviation, σ_ϵ. Gnanadesikan, Kettenring, and Landwehr (1977) used a similar method of perturb or "shake" (in their terminology) a set of data. They studied the stability of a given clustering in the presence of perturbation. The (k, B_k) plots would thus be of use in studying the stability. Three sets of 20 original and perturbed samples were generated by taking $\sigma_\epsilon = .03, .06, .09$. Hierarchical clusterings of each sample were carried out using the complete linkage method and Euclidean distance. Since the original and perturbed data were not very different, one would expect that their clusterings would have marked similarity. Figure 5 shows (k, B_k) plots with lines for $E(B_k)$ and $E(B_k) \pm 2\,(\text{var}(B_k))^{1/2}$ superimposed for 2 individual samples for each of the cases $\sigma_\epsilon = .03, .06, .09$. Each plot shows significant similarity between the original and perturbed data sets since the points generally lie well beyond the limit, $E(B_k) + 2\,(\text{var}(B_k))^{1/2}$. The plots show also a decreasing similarity with increases in σ_ϵ. There appears also to be a tendency for B_k to remain relatively constant over long stretches of k, with a precipitous falloff at the very highest values of k. These patterns may be seen more

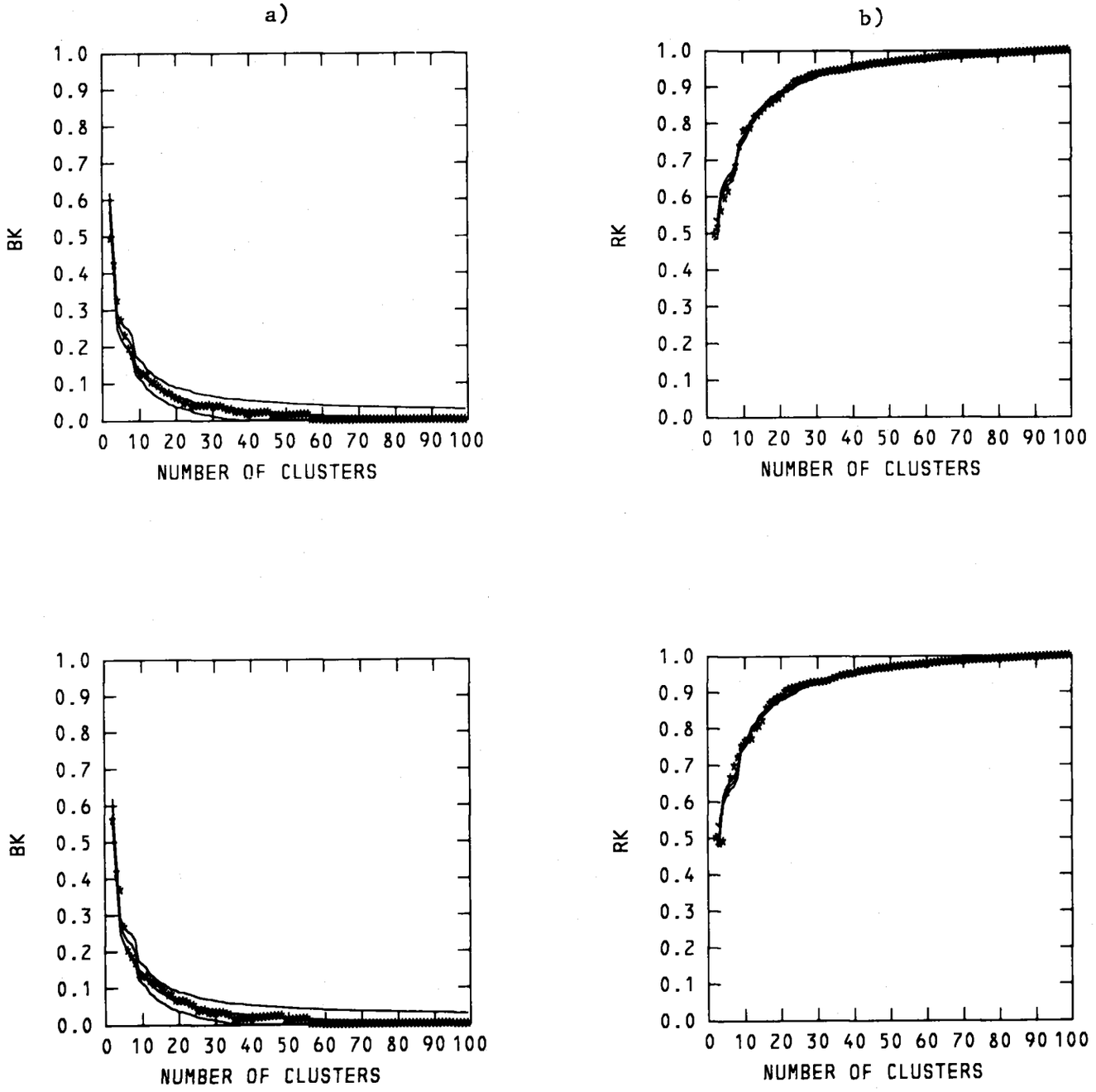

Figure 3. (k, B_k) and (k, R_k) plots for four pairs of unrelated bivariate samples, $n = 100$, $\mu_1 = .0$, $\mu_2 = .0$, $\sigma_1 = 1.0$, $\sigma_2 = 1.0$, $\rho = .5$.

clearly if an average configuration is calculated across the 20 samples for each value of σ_ϵ. Figure 6 shows $(k, \bar{B}_k)$ plots as well as (k, S_k) plots for $\sigma_\epsilon = .03, .06, .09$, where $\bar{B}_k$ is the .05 trimmed mean and S_k is the .05 trimmed standard deviation of the 20 values of B_k. Lines for $E(B_k)$ and $\pm 2\,(\text{var}(B_k))^{1/2}$ for the first of the 20 samples are included for reference. This practice is repeated in Figures 8 and 9. Figure 5 shows some very interesting effects. The tendency of B_k to remain relatively constant for long stretches may be seen more clearly here than in the individual plots of Figure 5. Notice, however, that these stretches grow shorter as σ_ϵ increases or, in other words, that B_k approaches zero faster as k approaches $n - 1$. The drop in level of B_k as σ_ϵ increases may also be seen. For $\sigma_\epsilon = .03$, $B_k \leq .8$ for $5 \leq k < 90$, for $\sigma_\epsilon = .06$, $B_k \leq .65$ for $5 \leq k < 70$, and for $\sigma_\epsilon = .09$, $B_k \leq .575$ for $5 \leq k < 50$. Figures 5 and 6 exhibit a very desirable property of B_k. Perturbation affects B_k least for small values of k and greatest for large values of k. One certainly would not want B_k for small values of k to be drastically affected by small perturbations in the data.

The (k, S_k) plots in Figure 6 also show an extremely

interesting configuration. The values of S_k remain somewhat constant at about .1 for most of the range of k. There is, however, a sharp increase in S_k if $k > 90$. Since $0 < B_k < 1$, one would certainly not expect B_k to be approximately normally distributed. Nevertheless, normal quantile-quantile plots of the B_k were made for the 20 sample values for each k. The extreme values of k show the most systematic departures from normality. There appears to be systematic curvature in the configurations for $2 < k < 10$ and pronounced discreteness in the distributions of B_k for $92 < k < 98$. The latter pattern is certainly not surprising since the matrix, M, for k in this range is quite sparse with a large number of its nonzero elements equal to one. The normal quantile plots for $10 < k < 90$ were remarkably straight. Thus, over wide ranges of k the distribution of B_k is approximately normal with somewhat constant mean and variance.

Both the Rand, R_k, and the Baker-Goodman-Kruskal, γ, similarity measures were calculated for each of the three sets of data referred to previously (σ_ϵ = .03, .06, .09). The value of R_k quickly approached one regardless of the magnitude of the perturbation; this property was anticipated by remarks in Section 3. Figure 7, shows box plots of the 20 sample values of γ for σ_ϵ = .09 along with box plots of B_k for the same 10 k values used in Section 4.1. The distribution of values of γ indicates significant similarity between the clusterings before and after perturbation. The median value of γ across the 20 sample values was approximately .60. The location of the distribution of γ is thus intermediate to the locations for the distributions of B_2 and B_{10}. The variance of the distribution of γ as estimated by the interquartile range appears, however, to be somewhat larger than the variances of either B_2 or B_{10}. The results are thus a bit mixed. That the median γ is somewhat high in the presence of a small perturbation of the sample is reasonable and desirable, but the large variance of the distribution of γ is cause for concern.

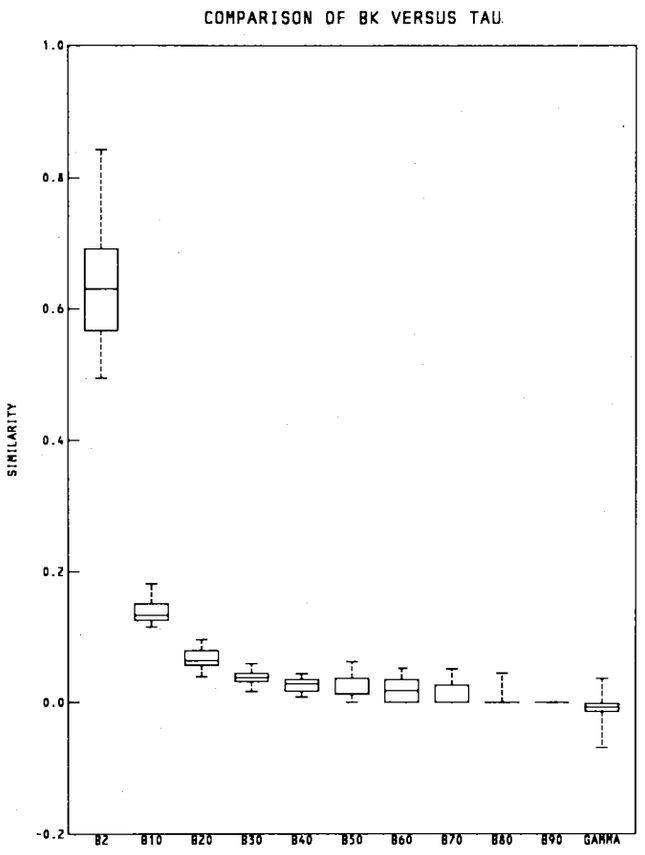

Figure 4. Box plots of B_k and γ for 20 samples of size n = 100 from each of two unrelated bivariate normal distributions, μ_1 = .0, μ_2 = .0, σ_1 = 1.0, σ_2 = 1.0, ρ = .5 (k = 2, 10, 20, 30, 40, 50, 70, 80, 90).

4.2.2 Multiple Cluster Experiments. Perturbation experiments were also carried out on bivariate normal data that had more than one cluster. Two experiments were run, one with two clusters and one with five. For the former case samples were generated in the following fashion. First, one sample of size 15 was generated from the bivariate normal distribution with parameters as in (4.3). Next, a sample of size 15 was generated from a bivariate normal distribution with the same parameters except that

$$\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} = \begin{bmatrix}2.5\\ 2.5\end{bmatrix}. \tag{4.4}$$

Then the two samples were combined to form an original sample of size 30, and the coordinates of each observation were perturbed in the same fashion used earlier in this section to produce a perturbed sample of size 30. Twenty such samples were generated for each of the perturbation standard deviations, σ_ϵ = .10, .20, .30, .40. The similarity measure, B_k, was calculated between the cluster trees for each original and perturbed sample, and Figure 8 shows $(k, \bar{B}_k)$ plot for all 4 σ_ϵ's, where $\bar{B}_k$ is again the .05 trimmed mean of the 20 sample values of B_k. The plots exhibit similar configurations to those in Figure 5 for one bivariate sample. The plots show significant similarity between the clusterings of the original and perturbed samples since most of the values of B_k lie beyond the limit $E(B_k) + 2(\text{var}(B_k))^{1/2}$. Also the plots show that values of B_k decrease with k more rapidly as σ_ϵ increases. For σ_ϵ = .10, there is the tendency for B_k to remain relatively constant across wide ranges of k. One would expect that since there are in fact two distinct clusters in the data, B_k would be large for k = 2, and indeed B_2 is greater than .80 in all panels of Figure 8; the perturbation does not thwart the recovery of the strong clustering in the data.

The five-cluster perturbation experiment was similar in design to that for two clusters. Each cluster was a sample of size 15 from a bivariate normal probability distribution with $\sigma_1 = \sigma_2 = 1.0$ and ρ = .5. The mean vectors for the five clusters were

$$\begin{bmatrix}0.0\\ 0.0\end{bmatrix}, \begin{bmatrix}10.0\\ 0.0\end{bmatrix}, \begin{bmatrix}5.0\\ 5.0\end{bmatrix}, \begin{bmatrix}0.0\\ 10.0\end{bmatrix}, \begin{bmatrix}10.0\\ 10.0\end{bmatrix}. \tag{4.5}$$

Twenty samples were generated according to this scheme and were perturbed using σ_ϵ = .10, .20, .40. This yielded

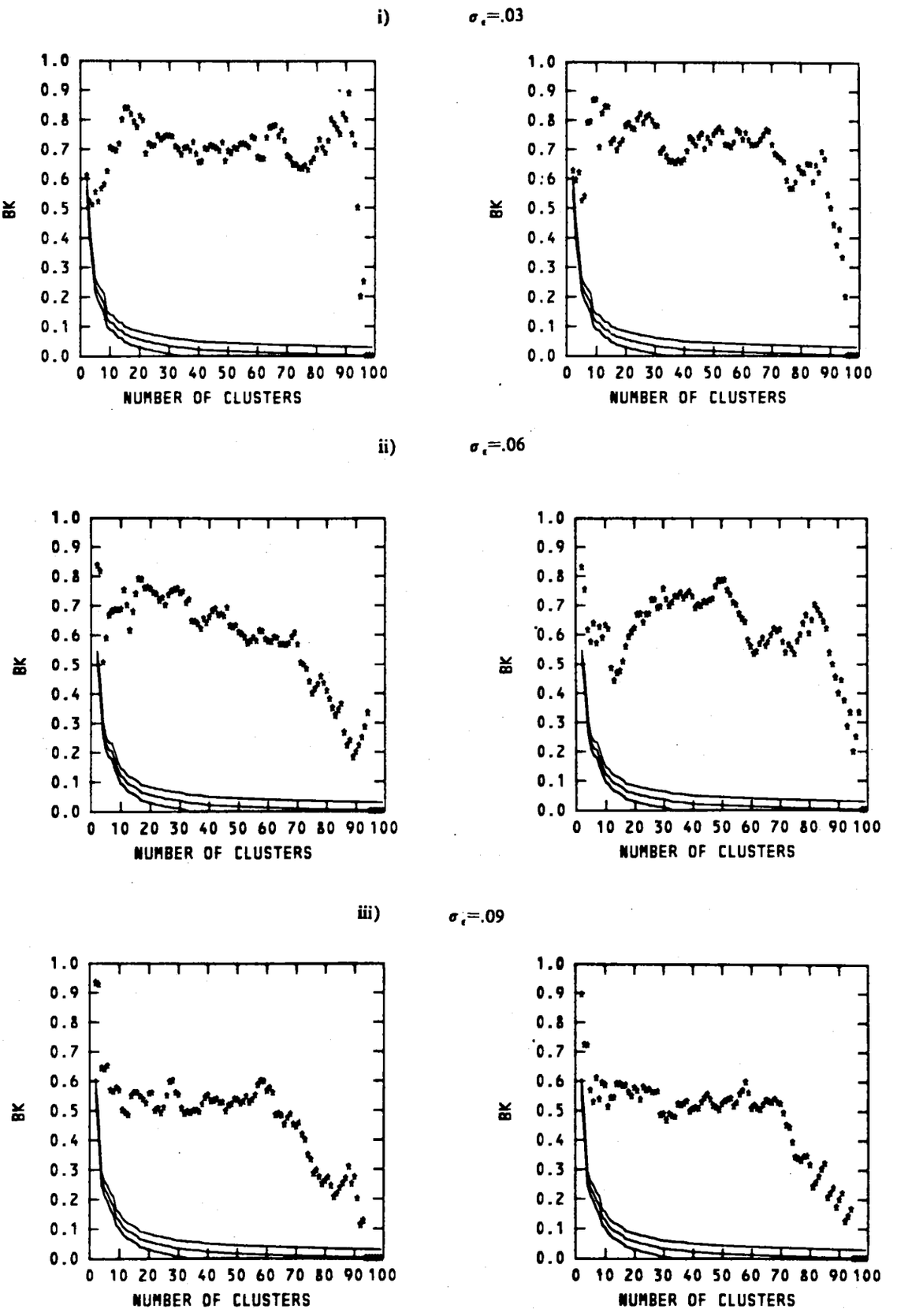

Figure 5. (k, B_k) plots for two particular bivariate normal samples of size $n = 100$ and their perturbations; $\mu_1 = .0$, $\mu_2 = .0$, $\sigma_1 = 1.0$, $\sigma_2 = 1.0$, $\rho = .5$.

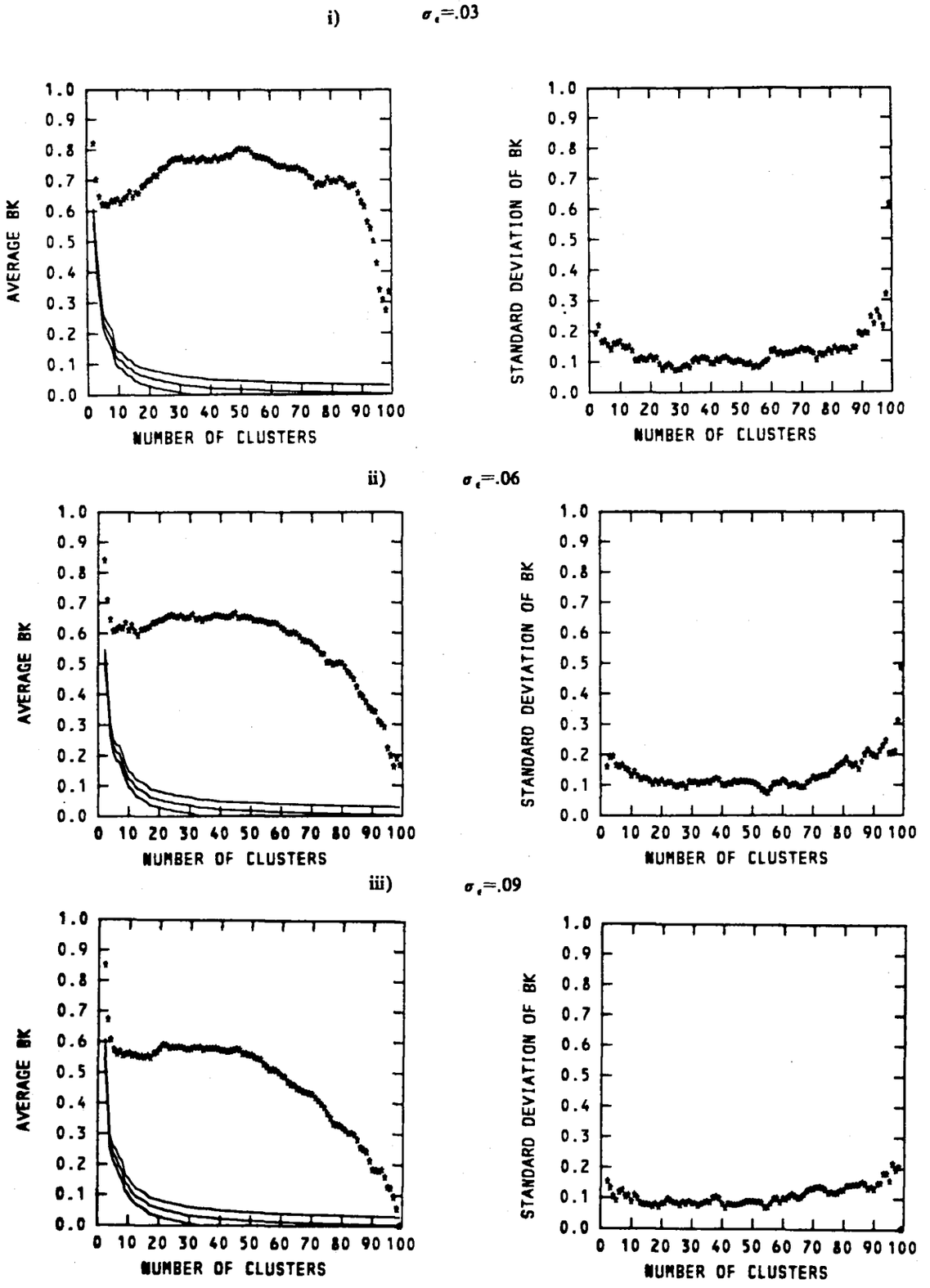

Figure 6. $(k, \bar{B}_k)$ and (k, S_k) plots for 20 bivariate normal samples of size $n = 100$ and their perturbations; $\mu_1 = .0$, $\mu_2 = .0$, $\sigma_1 = 1.0$, $\sigma_2 = 1.0$, $\rho = .5$.

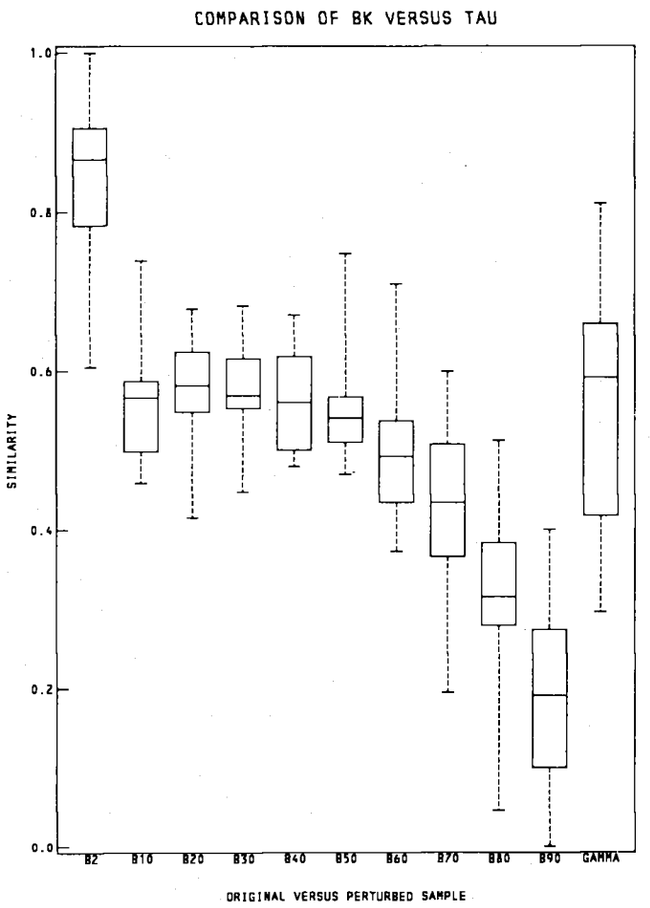

Figure 7. Box plots of B_k and γ for 20 bivariate normal samples of size n = 100 and their perturbations; $\mu_1 = .0$, $\mu_2 = .0$, $\sigma_2 = 1.0$, $\sigma_2 = 1.0$, $\rho = .5$; $k = 2, 10, 30, 40, 50, 60, 70, 80, 90$; $\sigma_\epsilon = .09$.

20 pairs of original and perturbed samples. Figure 9 shows $(k, \bar{B}_k)$ plots for the three values of σ_ϵ. As we would expect, B_k is large for $k = 5$ (actually $B_k > .95$ for all three cases), and the recovery of five clusters is not harmed by the perturbation. There is also the tendency for the configuration to drop off more quickly as σ increases.

A variation of the five-cluster perturbation experiment provided an interesting insight and showed how (k, B_k) plots may be used profitably to study clustering methods. First, a master sample of size 75 from the same mixture of 5 bivariate normals was generated. The 20 perturbations on this master sample were generated with $\sigma_\epsilon = .80$. Next (k, B_k) and $(k, \bar{B}_k)$ plots ($\bar{B}_k$ calculated across the 20 samples) were constructed for the comparison of the master sample and each of the 20 perturbed samples using both the complete linkage and the single linkage methods of hierarchical clustering. We have defined complete linkage previously, in the single linkage method the distance between two clusters is the minimum distance across all pairs of points, one in each cluster. Of interest was whether the (k, B_k), $(k, \bar{B}_k)$ plots could be used to see how each of the methods recovered the master structure. Figure 10 shows $(k, \bar{B}_k)$ plots for the two methods. Notice that the plot for the complete linkage method reaches a sharp maximum for $k = 5$ while the maximum $\bar{B}_k$ for the single linkage method is not sharp and does not occur at $k = 5$. This appears to indicate that the complete linkage method is doing rather better in recovering the structure of the master sample. Notice, however, that the $(k, \bar{B}_k)$ plot for the complete linkage method falls off much more rapidly than that for the single linkage method. To ascertain the reason for such behavior, we selected one of the 20 samples whose (k, B_k) configuration most nearly matched the $(k, \bar{B}_k)$ for both the complete linkage and the single linkage methods. Next, the four trees for the master and perturbed samples for both the complete linkage and the single linkage methods were cut at $k = 10$, where there was a large discrepancy between the B_k values (Euclidean distance was used in the clustering). Figure 11 shows scatter plots for the four cases with the points identified by cluster number (1 to 10). Notice that for the master sample the complete linkage method divides the (0., 0.), (10., 0.), (0., 10.), and (10., 10.) clusters roughly into two equal parts, while it trims one point (cluster number 10) from the (5., 5.) cluster. The single linkage method trims one point from the (5., 5.) cluster and divides the (10., 0.) and (10., 10.) clusters into three parts, but the (0., 0.) and (0., 10.) clusters are retrieved exactly. For the perturbed sample the complete linkage method still tends to split up each cluster (except the (0., 10.) cluster) into 2 roughly equally-sized groups, but the groups have quite different content than those for the master sample. It is easy to see that the similarity between the clusterings of the master and perturbed samples would be quite low in this case.

In contrast, the single-linkage method retrieves much of the same structure that it saw in the master sample. It trims the same point from the (5., 5.) cluster, recovers the (0., 10.) cluster exactly, and trims two points from the (0., 0.) cluster. These similarities alone account for a larger value of B_k for the single linkage method.

This experiment was repeated for a different master sample and different perturbed samples. Figure 12 shows $(k, \bar{B}_k)$ plots for the single linkage and complete linkage methods. The complete linkage method retrieves the 5 clusters better than the single linkage method does, but the latter again performs better for larger values of k.

4.3 Other Experiments

In this section we summarize the findings of three other experiments. In the first we used the 20 bivariate samples of size 100 described in Section 4.2.1. The first coordinate of each sample was multiplied by 2 and 4 ($km = 2, 4$), respectively, to generate two groups of 20 pairs of samples. Plots of (k, B_k) for each group showed that the changes of scale have large effects on the similarities between clusterings of the original sample and that of the sample whose first coordinate had been rescaled. For km = 2, B_k dropped to about .58 for $k = 5$ followed by a gradual increase to about .80 for $k = 90$. A similar con-

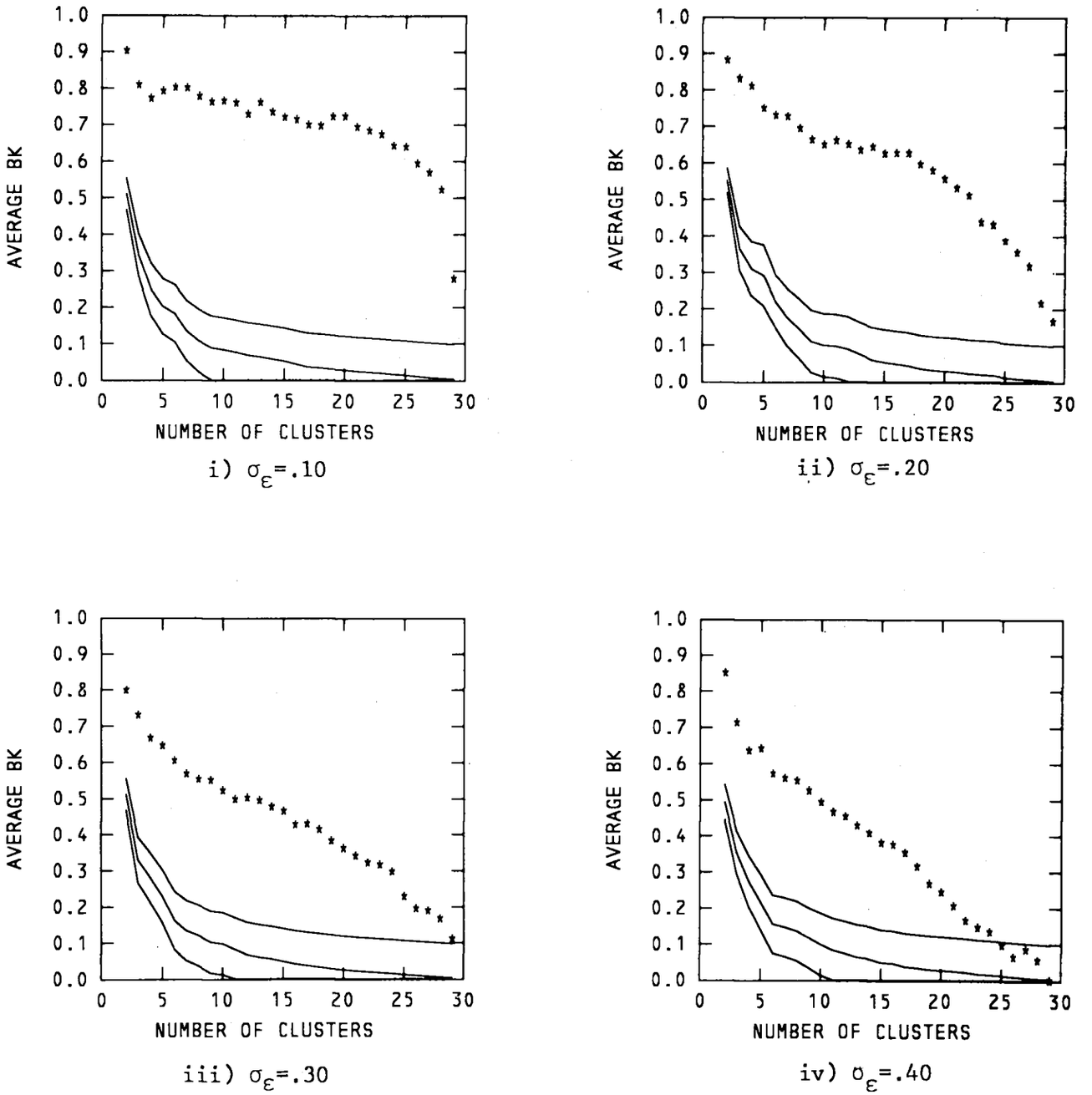

Figure 8. (k, $\bar{B}_k$) plots for 20 samples of size n = 30 from the mixture of two bivariate normals with mixing proportion .5, and their perturbations; μ_1 = (0., 0.), μ_2 = (2.5, 2.5); $\sigma_1 = \sigma_2$ = 1.0, $\rho_1 = \rho_2$ = .5 for each component of the mixture.

figuration was seen for km = 4, with a drop in B_k to about .47 for k = 5 with a gradual increase to about .65 for k = 90.

In the outlier experiments the 20 bivariate normal samples of size 100 were altered by adding + 10 to both x and y coordinates of a randomly selected observation from the 100. A plot of (k, B_k) was calculated across the 20 pairs of clusterings of original and altered samples. The plot showed a pronounced drop in similarity for small k and very high similarity for large k. For example, B_k fell to about .77 for k = 2 and was greater than .95 for k > 20.

For the noise variable experiments, independent univariate normal random variables whose variances were equal to the average of the observed variances for the individual coordinates were added as coordinates to the data from Section 4.2.2 containing five clusters (n = 75). Plots of (k, B_k) when three noise variables had been added, for example, showed a dramatic decrease in similarity between the clusterings for the original sample and the sample including the noise variables. For example, B_k was equal to what might be expected by chance alone for $2 \leq k \leq 7$, and only about .20 for $8 \leq k \leq 70$ for most of the 20 pairs of samples. This demonstrates how greatly

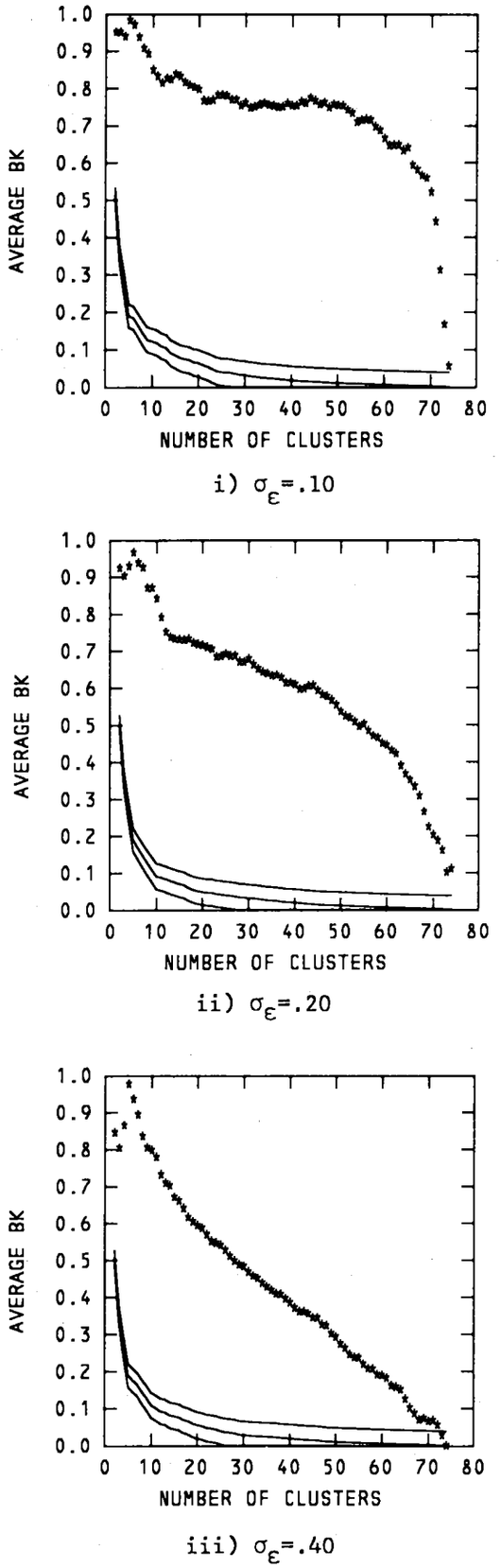

Figure 9. $(k, \bar{B}_k)$ plots for 20 samples of size $n = 75$ from the mixture of five bivariate normals with mixing proportions, .2, and their perturbations; $\mu_1 = (0, 0)$, $\mu_2 = (10, 0)$, $\mu_3 = (5, 5)$, $\mu_4 = (0, 10)$, $\mu_5 = (10, 10)$, $\sigma_1 = \cdots \sigma_5 = 1.0$, $\rho_1 = \cdots = \rho_5 = .5$.

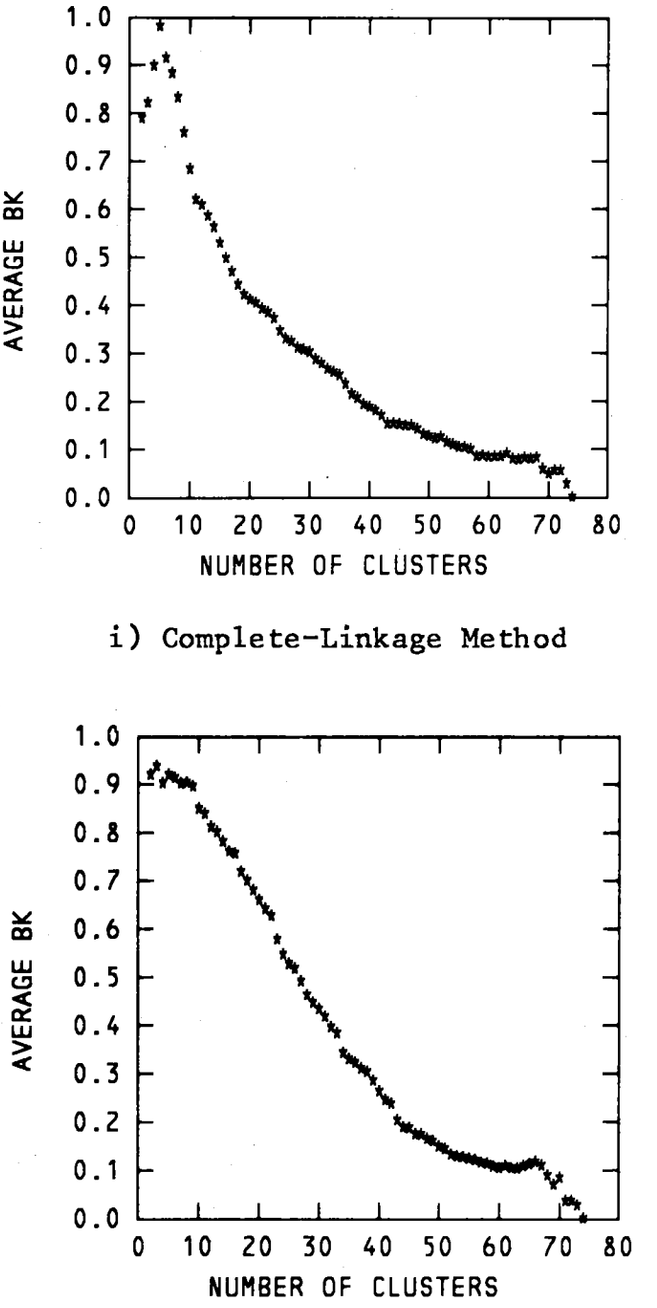

Figure 10. $(k, \bar{B}_k)$ plots for a master sample of size $n = 75$ and 20 perturbations from the mixture of five bivariate normals with mixing proportions, .2; $\mu_1 = (0.0)$, $\mu_2 = (10, 10)$ $\mu_1 = (0.0)$, $\mu_2 = (10, 0)$, $\mu_3 = (5, 5)$, $\mu_4 = (0, 10)$, $\mu_5 = (10, 10)$, $\sigma_1 = \cdots = \sigma_5 = 1.0$, $\rho_1 = \cdots = \rho_5 = .5$; $\sigma_\epsilon = .80$.

the recovery of cluster structure can be affected when only a subset of the variables contain the structure, and some of the variables contribute only noise.

4.4 Data Example

In this section we present an example of the use of the B_k statistic on real data. In the example B_k was used as

one of a battery of techniques for uncovering patterns and structure in the data.

The example concerned the authorship of 12 writings. The data were taken from a feasibility study by W.A. Larsen and R. McGill (1982) on the identification of authors using letter frequencies. Specifically, the frequency of each letter of the alphabet in selected passages of approximately 7,000 letters was recorded for two writings by each of the following six authors: Pearl Buck; James Michener; Arthur C. Clarke; Ernest Hemingway; William Faulkner; and Victoria Holt. Unusual words and proper nouns were excluded from the passages, since these would tend to be specific to individual texts and are not necessarily characteristic of the whole body of work of

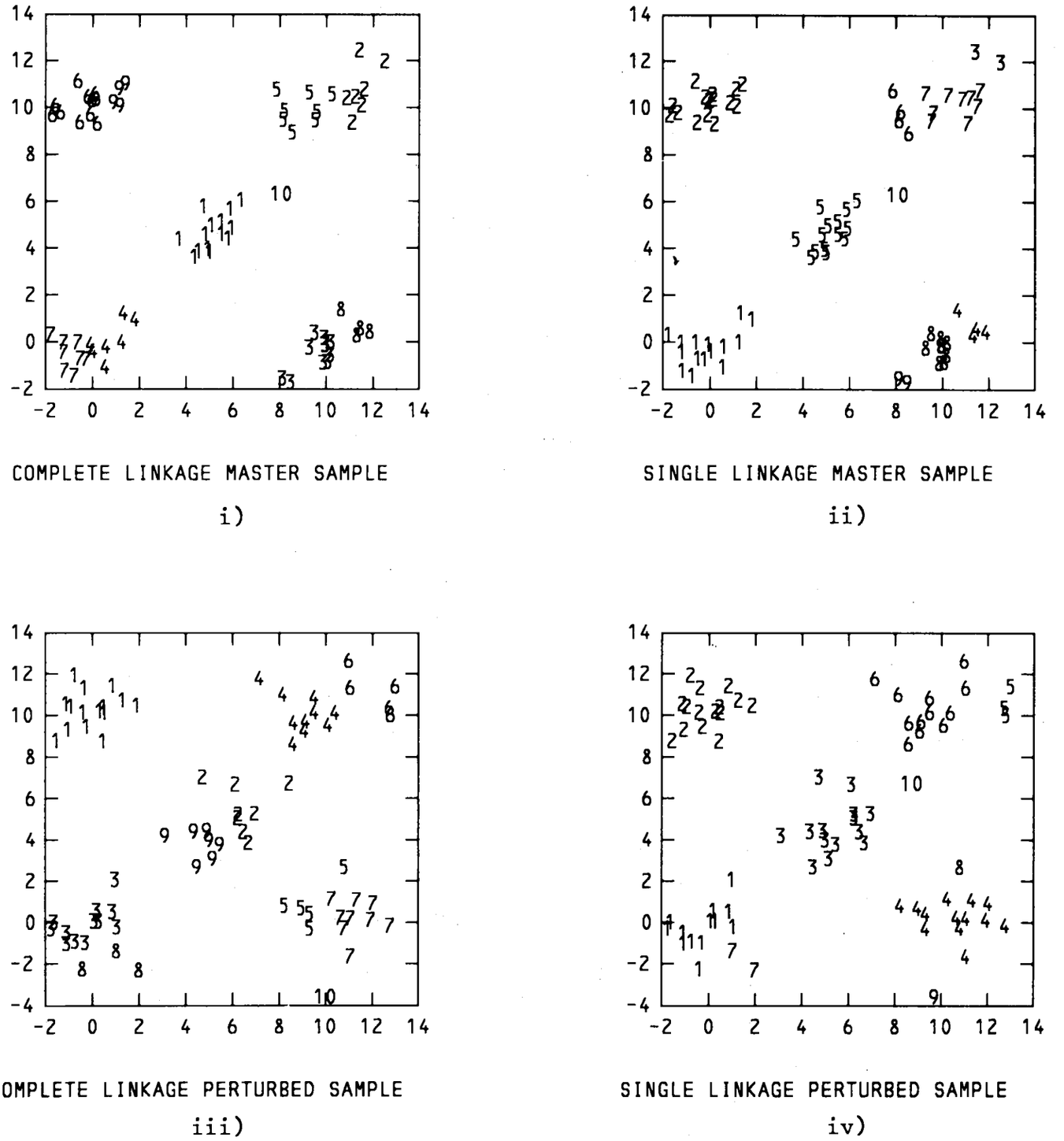

Figure 11. Scatter plots of master sample of size 75 and one perturbed sample from the mixture of five bivariate normals with mixing proportions, .2; μ_1 = (0, 0), μ_2(10, 0), μ_3 = (5, 5), μ_4 = (0, 10), μ_5 = (10, 10), $\sigma_1 = \cdots = \sigma_5$ = 1.0; $\rho_1 = \cdots = \rho_5$ = .5; σ_ϵ = .80.

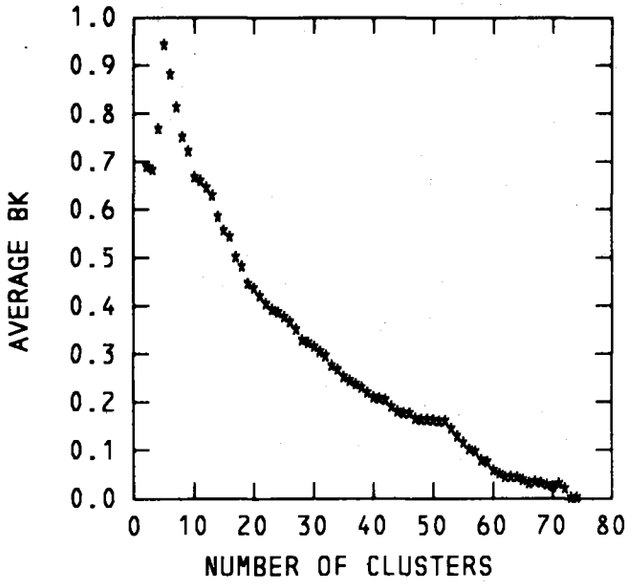

Figure 12. $(k, \bar{B}_k)$ plots for a master sample of size $n = 75$ and 20 perturbations from the mixture of five bivariate normals with mixing proportions, .2; $\mu_1 = (0.0)$, $\mu_2 = (10, 0)$, $\mu_3 = (5, 5)$, $\mu_4 = (0, 10)$, $\mu_5 = (10, 10)$, $\sigma_1 = \cdots = \sigma_5 = 1.0$; $\rho_1 = \cdots = \rho_5 = .5$; $\sigma_\epsilon = .80$.

an author. The specific list of writings is given in Table 2. The identifying numbers of the books are used as identifiers in the dendrograms shown in Figure 14.

The data thus comprised 26 variables on each of 12 observations. Before any analyses were carried out, the letter frequencies were converted to proportions by dividing by the passage lengths, also shown in Table 2. The letter frequencies and the total number of letters in each

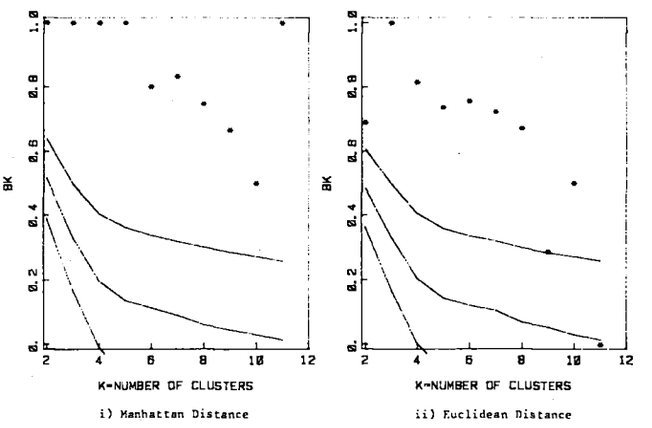

Figure 13. (k, B_k) plots for Larsen and McGill (1982) authorship data using scaled and unscaled variables.

passage are given in Table 3. Hierarchical clustering was carried out for the 12 observations using the complete linkage method with Manhattan and Euclidean distance and either scaled or unscaled variables. Manhattan distance is just the sum of the absolute differences for the coordinate values of a pair of objects. There are a number of choices for scaling the data. For example, proportions of letters across books could be used to estimate the appropriate binomial standard deviation. An estimate of a

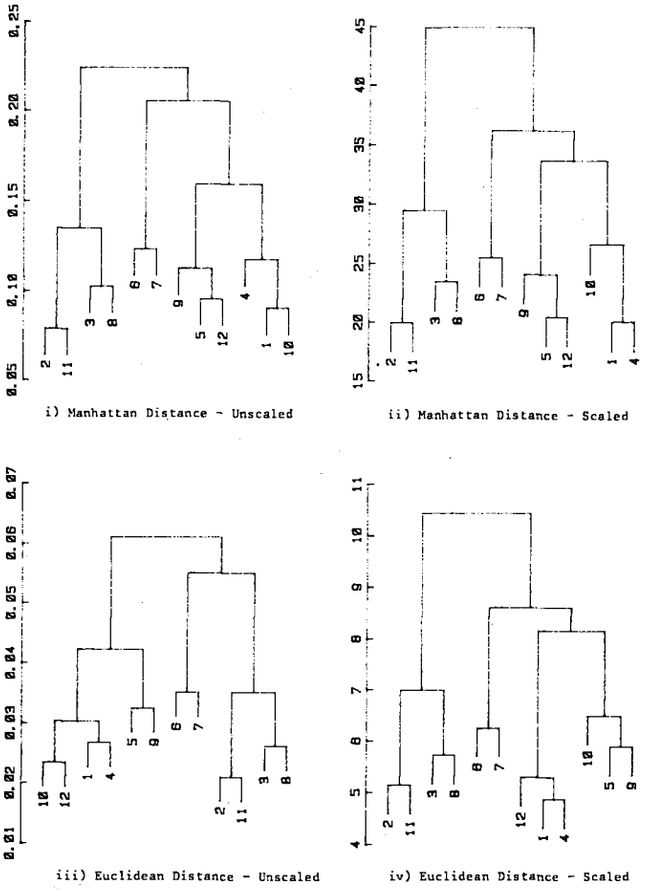

Figure 14. Dendrograms for hierarchical clustering (complete linkage method) of Larsen and McGill (1982) authorship data.

Table 2. Titles, Authors, and Total Letter Frequencies Used in Larsen and McGill Authorship Study

	Title	Author	Total Letter Frequency
1.	The Three Daughters of Madame Liang	Pearl Buck	7144
2.	The Drifters	James Michener	6669
3.	The Lost Worlds of 2001	Authur C. Clarke	7100
4.	East Wind, West Wind	Pearl Buck	7479
5.	A Farewell to Arms	Ernest Hemingway	6877
6.	The Sound and the Fury (Part 1)	William Faulkner	6885
7.	The Sound and the Fury (Part 2)	William Faulkner	6971
8.	Profiles of the Future	Authur C. Clarke	7505
9.	Islands In The Stream	Ernest Hemingway	6924
10.	Bride of Pendorric (Part 1)	Victoria Holt	6650
11.	The Voice of Asia	James Michener	6510
12.	Bride of Pendorric (Part 2)	Victoria Holt	6933

binomial standard deviation appropriate for each author could also be derived. For this example we chose instead to use the sample standard deviation of the proportions for each letter as the appropriate scale factor. A completely different approach would have been to use a variance stabilizing transformation like the arcsine-square root or even a logit transformation. Regardless of the choice of standardization procedure, it is an interesting question whether such standardization should be carried out at all since on the resulting scale, low frequency variables like x and z would have equal weight with high frequency variables like a and e. At any rate, in this example we simply want to illustrate how B_k can be used to study the effect of standardization or lack thereof.

The similarity between the hierarchical clusterings for the scaled and unscaled variables was calculated for both Manhattan and Euclidean distance. (k, B_k) plots for these two distances are shown respectively in Figure 13. The corresponding hierarchical clusterings are given in Figure 14. The (k, B_k) plot for Manhattan distance shows exact agreement in the clusterings for $k = 2, 3, 4, 5, 11$. It is interesting to note that the topologies for the two trees are, aside from height (or weight of the internal nodes) and labeling, identical. The same property is not true for the trees using Euclidean distance. This is reflected also in the fact that the (k, B_k) plot for Euclidean distance has as small or smaller values of B_k than the corresponding values for Manhattan distance. This also shows that the clusterings are much more affected by use of standardization for Euclidean distance than for Manhattan distance.

The flavor of the preceding discussion might imply that Manhattan distance is to be recommended for use with these data. Such is not really the case. It appears that the

Table 3. Data From Larsen and McGill (1982) Authorship Study

Letter Frequency	Passage 1	2	3	4	5	6	7	8	9	10	11	12
A	550	515	590	557	589	541	517	592	576	557	554	541
B	116	109	112	129	72	109	96	151	120	97	108	93
C	147	172	181	128	129	136	127	251	136	145	206	149
D	374	311	265	343	339	228	356	238	404	354	243	390
E	1015	827	940	996	866	763	771	985	873	909	797	887
F	131	167	137	158	108	126	115	168	122	97	164	133
G	131	136	119	129	159	129	189	152	156	121	100	154
H	493	376	419	571	449	401	478	381	593	479	328	463
I	442	432	514	555	472	520	558	544	406	431	471	518
J	2	8	6	4	7	5	6	7	3	10	4	4
K	52	61	46	76	59	72	80	39	90	94	34	65
L	302	280	335	291	264	280	322	416	281	240	293	265
M	159	146	176	247	158	209	163	236	142	154	149	194
N	534	470	403	479	504	471	483	526	516	417	482	484
O	516	561	505	509	542	589	617	524	488	477	532	545
P	115	140	147	92	95	84	82	107	91	100	145	70
Q	4	4	8	3	0	2	8	9	3	3	8	4
R	409	368	395	413	416	324	294	418	339	305	361	299
S	467	387	464	533	314	454	358	508	349	415	402	423
T	632	632	670	632	691	672	685	655	640	597	630	644
U	174	195	224	181	197	247	225	226	194	237	196	193
V	66	60	113	68	64	71	37	89	40	64	66	66
W	155	156	146	187	225	160	216	106	250	194	149	218
X	5	14	13	10	1	11	12	15	3	9	2	2
Y	150	137	162	184	155	280	171	142	104	140	80	127
Z	3	5	10	4	2	1	5	20	5	4	6	2
Total Frequency	7144	6669	7100	7479	6877	6885	6971	7505	6924	6650	6510	6933

use of Euclidean distance on the unscaled data reveals the most interesting structure of the four clusterings carried out. Notice, for this case, pairs of authors combine before combining with other authors and that the women authors (Buck and Holt) combine before combining with any of the men. Also the women are more similar than are any of the men. It is interesting too that Hemingway is closer to the women than to any of the men. All of these tidbits are not preserved for the Euclidean-scaled case or for either of the Manhattan distance cases. For each of these cases there is some confusion in the linking together of the two women and Hemingway. The authors Faulkner, Michener, and Clarke are always paired first in each of the four trees.

5. CONCLUSION

The main insight we have obtained from thinking about B_k and other measures of similarity of two clusterings is that similarity is not a one-dimensional concept. Clearly our plots can be helpful in selecting an appropriate number of clusters. They also show that two hierarchical clusterings can exhibit different degrees of similarity at different levels of cut.

APPENDIX: DERIVATION OF THE MEAN AND VARIANCE OF B_k

First, we write $a^{(p)} = a(a-1)\cdots(a-p+1)$. Then, since

$$B_k = T_k/(P_kQ_k)^{1/2}, \quad \text{where} \quad T_k = \sum_i \sum_j m_{ij}^{(2)}$$

and P_k, Q_k are defined in Section 2 and are fixed according to the following assumption, it is sufficient to derive the moments of T_k. This assumption is that the two clusterings are mutually independent, subject to the condition that the cluster sizes (in each clustering) are fixed at $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$, respectively. The a_i and b_j are the marginal totals of m_{ij}, namely $m_{i\cdot}$ and $m_{\cdot j}$, respectively. Given this assumption, the elements of the matrix $M = [m_{ij}]$ have a generalized hypergeometric distribution (Lancaster 1969, p. 214). Then, the pth factorial moment of m_{ij} is

$$E(m_{ij}^{(p)}) = (a_i^{(p)}b_j^{(p)})/n^{(p)},$$

and the cross factorial moments are

$$E(m_{ij}^{(p)}m_{ij}^{(q)}) = \frac{a_i^{(p+q)}b_j^{(p)}b_{j'}^{(q)}}{n^{(p+q)}}, \quad j \neq j',$$

and $$E(m_{ij}^{(p)}m_{i'j'}^{(q)}) = \frac{a_i^{(p)}a_{i'}^{(q)}b_j^{(p)}b_{j'}^{(q)}}{n^{(p+q)}}, \quad i \neq i', j \neq j'$$

Therefore

$$E(T_k) = \frac{\sum_i \sum_j q_i^{(2)}b_j^{(2)}}{n^{(2)}} = \frac{P_kQ_k}{n^{(2)}}.$$

In evaluating $E(T_k^2)$, we have to keep track of four different types of terms shown as follows:

$$T_k^2 = \sum_i \sum_j (m_{ij}^{(2)})^2 + \sum_i \sum_j \sum_{j' \neq j'} m_{ij}^{(2)}m_{ij'}^{(2)} + \sum_{i \neq i'} \sum_i \sum_j m_{ij}^{(2)}m_{i'j}^{(2)} + \sum_{i \neq i'} \sum_i \sum_{j \neq j'} \sum_j m_{ij}^{(2)}m_{i'j'}^{(2)}.$$

Now, $(m^{(2)})^2 = m^2(m-1)^2 = m^{(4)} + 4m^{(3)} + 2m^{(2)}$ so that

$$E(T_k^2) = \sum_i \sum_j \left\{ \frac{a_i^{(4)}b_j^{(4)}}{n^{(4)}} + \frac{4a_i^{(3)}b_j^{(3)}}{n^{(3)}} + \frac{2a_i^{(2)}b_j^{(2)}}{n^{(2)}} \right\} + \sum_i \sum_j \sum_{j \neq j'} \sum_{j'} \frac{a_i^{(4)}b_j^{(2)}b_{j'}^{(2)}}{n^{(4)}} + \sum_i \sum_{i \neq j} \sum_{i'} \sum_j \frac{a_i^{(2)}a_{i'}^{(2)}b_j^{(4)}}{n^{(4)}} + \sum_{i \neq i'} \sum_i \sum_{i'} \sum_{j \neq j'} \sum_j \sum_{j'} \frac{a_i^{(2)}a_{i'}^{(2)}b_j^{(2)}b_{j'}^{(2)}}{n^{(4)}}.$$

Collecting terms, this reduces to

$$E(T_k^2) = \frac{2}{n^{(2)}} P_kQ_k + \frac{4}{n^{(3)}} P_k'Q_k' + \frac{1}{n^{(4)}} (P_k^2 - 4P_k' - 2P_k)(Q_k^2 - 4Q_k' - 2Q_k),$$

where P_k' and Q_k' are defined by (2.12) and (2.13). The value of $\text{var}(B_k)$ follows immediately on subtracting $(P_kQ_k/n^{(2)})^2$ and dividing by P_kQ_k. Similarly the mean and variance of the Rand measure, R_k, may be determined using $E(T_k)$ and $E(T_k^2)$.

[*Received May 1981. Revised April 1982.*]

REFERENCES

ANDERBERG, M.R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

ARABIE, P., and BOORMAN, S.A. (1973), "Multidimensional Scaling of Measures of Distance Between Partitions," *Journal of Mathematical Psychology*, 10, 148–203.

BAKER, F.B. (1974), "Stability of Two Hierarchical Grouping Techniques Case I: Sensitivity to Data Errors," *Journal of the American Statistical Association*, 69, 440–445.

COVEYOU, R.R., and MAC PHERSON, R.D. (1967), "Fourier Analysis of Uniform Random Number Generators," *Journal of the Association of Computing Machinery*, 14, 100–119.

DANIELS, H.E. (1944), "The Relation Between Measures of Correlation in the Universe of Sample Permutations," *Biometrika*, 33, 129–135.

GNANADESIKAN, R., KETTENRING, J.R., and LANDWEHR, J.M. (1977), "Interpreting and Assessing the Results of Cluster Analysis," *Bulletin of the International Statistical Institute*, 47, 451–463.

GOODMAN, L.A., and KRUSKAL, W.H. (1974), "Measures of Association for Cross-Classifications," *Journal of the American Statistical Association*, 49, 732–764.

HARTIGAN, J.A. (1975), *Clustering Algorithms*, New York: John Wiley.

HUBERT, L.J., and LEVIN, J.R. (1976), "Evaluating Object Set Partitions: Free-Sort-Analysis and Some Generalizations," *Journal of Verbal Learning and Verbal Behavior*, 15, 549–470.

JARDINE, N., and SIBSON, R. (1971), *Mathematical Taxonomy*, London: John Wiley.

JOHNSON, S.C. (1968), "Metric Clustering," unpublished Bell Laboratories memorandum.

LANCASTER, H.O. (1969), *The Chi-Squared Distribution*, New York: John Wiley.
LARSEN, W.A., and McGILL, R. (1982), personal communication.
MARSAGLIA, G. (1962), "Expressing a Random Variable in Terms of Uniform Random Variables," *Annals of Mathematical Statistics*, 32, 874–898.
RAND, W.M. (1971), "Objective Criteria for Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 1971, 846–850.
TUKEY, J.W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley, 39–41.

Comment

DAVID L. WALLACE*

1. INTRODUCTION

Useful and interpretable methods for exploring and comparing the results of clustering algorithms are few, and developing inferential methods has proved difficult. Fowlkes and Mallows's contribution is most welcome. Their new measure, B_k, of association between two partitions of n objects into k clusters and the plots they propose for its examination can be readily computed and displayed and promote the exploration of clustering results.

The history of statistics is replete with measures of association and correlation. Just because a measure has plausibility at face value, ranges from 1 down to 0, and has a known sampling distribution under an extreme null hypothesis of total randomness is no guarantee of its usefulness or interpretability. What does .50 mean? And does .70 in one set of data really indicate greater association than .60 in another? For cross-classified data, the sequence of papers by Goodman and Kruskal (1954,1963) provided a model for the careful study of measures. They stressed the value of an interpretation of the measure for the population sampled or for the data generating process.

The multiple correlation coefficient illustrates how even when a clear meaning can be attached to a value—for example, the fraction of variance explained—that meaning may be misleading when compared across experiments. Mallows's C_p statistic has been valuable in part because it is cast as a measure for internal technical use and not as a test statistic or as a general measure of fit.

Topics such as clustering are tougher than two-way contingency tables or multiple regression. Even imperfect measures can be helpful when used with caution, and some starts must be made where few inferential concepts and tools are available. We want measures whose meaning and value are stable over unimportant changes in the situation and structure, while being sensitive to changes that do matter. With this background in mind, I consider the B_k measure and plots introduced by Fowlkes and Mallows (or FM, as I shall henceforth refer to them).

The two B_k plots in FM's Figure 12 illustrate the type of usage that can be made, and despite any warnings by the authors, will be made. Here each plot is a comparison of a clustering from one set of data with the clustering of a perturbed version of the data (strictly, the plots are averages of the measures for 20 perturbations). The two plots correspond to the same comparison based on two clustering algorithms. The comparisons within each such plot or between plots might be based on different distance measures entering a single algorithm, or for clusterings based on two sets of variables or two sets of objects, or whatever.

In each of the plots in Figure 12, B_k increases sharply and steadily as k, the number of clusters, decreases until k is very small. Furthermore, except when k is 5 or 6, the value of B_k is higher for the single linkage plot than for the complete linkage plot. The superficial conclusions are evident—the effect of perturbations on the clustering is greater for large k than for small k, and for any k, the effect is less for single linkage than for complete linkage. Are these conclusions valid, or could some or all of the increase be an artifact of the measure? What does B_k measure, and how stable is the meaning under incidental changes in the situation?

FM do not answer these questions, nor will I, but I hope to shed some light on them by examining the structure of B_k. I use their Larsen-McGill data on letter frequencies in 12 texts by 6 authors. This example is small and thereby not fully satisfactory, but real examples are a foil for judgments, comparisons, and conclusions that simulations and artificial examples cannot match.

2. CLUSTERING TREES AND PARTITIONS

FM have chosen to present their methodology under two unnecessary limitations. Removing these extends ap-

* David L. Wallace is Professor of Statistics, Department of Statistics, University of Chicago, Chicago, IL 60637.

© Journal of the American Statistical Association
September 1983, Volume 78, Number 383
Invited Papers Section