

Quantitative Evaluation of Clustering Results Using Computational Negative Controls *

Ronald K. Pearson, Tom Zylkin, James S. Schwaber, Gregory E. Gonye †

Abstract

Most partition-based cluster analysis methods (e.g., k -means) will partition any dataset \mathcal{D} into k subsets, regardless of the inherent appropriateness of such a partitioning. This paper presents a family of permutation-based procedures to determine both the number of clusters k best supported by the available data and the weight of evidence in support of this clustering. These procedures use one of 37 cluster quality measures to assess the influence of structure-destroying random permutations applied to the original dataset. Results are presented for a collection of simulated datasets for which the correct cluster structure is known unambiguously.

1 Introduction

There is considerable interest in the use of unsupervised clustering methods to discover structure in datasets [4, 5, 9, 11, 12, 15], with much recent interest in clustering gene expression data [2, 6, 8]. Partition-based clustering methods will generally partition *any* dataset \mathcal{D} into a specified number of disjoint subsets, regardless of how appropriate such a partitioning may be to the dataset. As a practical consequence, the following two questions are both important and difficult to answer:

- A. Does the dataset under consideration actually exhibit a natural cluster structure?
- B. If so, how many clusters k are present in the dataset?

In biology, *negative controls* are commonly used to assess the effects of an experimental treatment: samples that are known *a priori* not to respond to the treatment provide a basis for deciding how “significant” the observed responses are for the other samples. Here, we apply this idea to assessing the significance of clustering

results by creating a collection of *computational negative controls* from the original dataset. Specifically, we generate a collection of m randomized datasets from our original dataset, in which independent random permutations are applied to the individual components of the attribute vector associated with each object. We then cluster both the original dataset and the m randomizations using the same clustering procedure and compare the results. The idea is that large differences between the original results and the randomizations are indicative of significant cluster structure that has been destroyed by the randomization. Applying this procedure then provides a basis for addressing both of the practical questions raised above. Specifically, the difference between the original and randomized results provides the basis for a quantitative assessment of the significance of the clustering results obtained, and these significance results can then be used together with cluster quality results to decide how many clusters are present in the dataset when there is evidence in support of a cluster structure. The following example provides a simple illustration of this basic procedure, and subsequent sections of the paper consider various aspects of the procedure in more detail.

2 Illustrative example

The example considered here is based on the Ruspini dataset [15], consisting of 75 bivariate attribute vectors. This dataset was chosen because it represents a simple, well-known example that is commonly used as a benchmark problem in evaluating clustering methods and is widely available, incorporated as a built-in data object in both the *R* and *S-plus* statistics packages. The left hand plot in Fig. 1 is a scatterplot of these attribute vectors, which shows the existence of four reasonably well-defined clusters, although some have argued in favor of a five cluster description [5]. The right-hand plot in Fig. 1 shows the results of applying a random permutation to the second component of the attribute vector, largely destroying this cluster structure. The consequences of such permutations are discussed in detail in Sec. 2.3, using the basic computational negative control (CNC) procedure described next.

*The authors wish to acknowledge support for this work from the National Institutes of General Medical Sciences and Mental Health (MH64459-01) and the Defense Advanced Research Program Administration BioSpice initiative (F30602-01-0578).

†Daniel Baugh Institute for Functional Genomics and Computational Biology, Thomas Jefferson University, Philadelphia, PA. T. Zylkin is currently with Department of Mathematics, University of Pennsylvania.

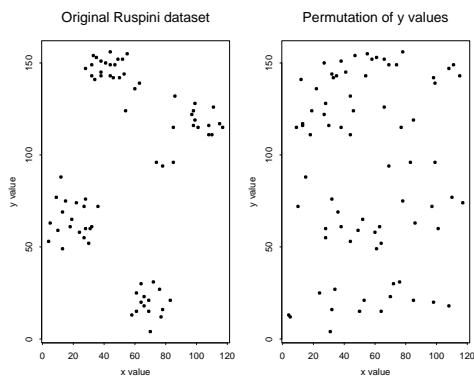


Figure 1: Plots of the original Ruspini dataset (left-hand side) and the result of a random permutation applied to the y variable (right-hand side)

2.1 Computational procedure The basic CNC procedure proposed here consists of the following steps:

0. Formulation: specify a dataset \mathcal{D} , the maximum number k^* of clusters to consider, a clustering method \mathcal{M} , a cluster quality measure $Q(\cdot)$, and a number m of negative control datasets to generate.
1. For each $k = 2, 3, \dots, k^*$, do:
 - a. Obtain the k -cluster partition \mathcal{P}_0^k of the dataset \mathcal{D} using method \mathcal{M} .
 - b. Compute the quality measure $Q(\mathcal{P}_0^k)$ for this clustering result.
 - c. For each i in 1 to m , do:
 - i. Generate a new structure-destroying permutation $\pi_i \mathcal{D}$ of the dataset \mathcal{D} .
 - ii. Obtain the k -cluster partition \mathcal{P}_i^k of the dataset $\pi_i \mathcal{D}$ using method \mathcal{M} .
 - iii. Compute the quality measure $Q(\mathcal{P}_i^k)$ for this clustering result.
2. Develop and examine both graphical and numerical summaries of the results:
 - a. Generate side-by-side boxplot summaries of the m $Q(\mathcal{P}_i^k)$ values for $k = 2, 3, \dots, k^*$.
 - b. Plot the values of $Q(\mathcal{P}_0^k)$ for $k = 2, 3, \dots, k^*$ on the same plot with the boxplots generated in (a). Note that if significant cluster structure in the original dataset has been destroyed by the random permutations, some of the values of $Q(\mathcal{P}_0^k)$ should fall well outside the range of variation seen in the $Q(\mathcal{P}_i^k)$ boxplots.
 - c. Compute p -values p_k for the null hypothesis that $Q(\mathcal{P}_0^k)$ has the same distribution as the permutation results $\{Q(\mathcal{P}_i^k)\}$.

3. Generate answers to Questions A and B above:

- A. If $\min\{p_k\} < \theta$ where θ is a specified significance threshold, the dataset \mathcal{D} exhibits evidence of significant structure.
- B. If significant structure is detected, select the “best” choice of k from these results.

This paper investigates the 37 cluster quality measures $Q(\cdot)$ described by Bolshakova and Azuaje [2], who used them with random subset selection to assess cluster stability (i.e., the degree to which “similar” datasets gave “similar” clustering results). Here, our objective is the opposite: we wish to destroy the structure present in the original dataset and look for large changes in the result, indicative of significant cluster structure in the original dataset. Because we consider so many different quality measures, this paper restricts consideration to a single clustering method \mathcal{M} , the *Partitioning Around Medoids* (PAM) procedure described by Kaufman and Rousseeuw [12], with the popular Euclidean dissimilarity measure. This procedure was chosen because it is available in both the *R* and *S-plus* software packages, it has been described in reasonable detail [12], and it overcomes a number of the known limitations of the more popular k -means clustering procedure (e.g., its outlier sensitivity and its dependence on the original ordering of the objects in the dataset). Clearly, the quality of the results obtained here can depend strongly on the method \mathcal{M} chosen, and follow-on studies are planned to examine this influence in detail. Indeed, one of our motivations for undertaking this work is to provide a practical, data-based platform for comparing the performance of different clustering algorithms.

2.2 Silhouette coefficients One of the 37 cluster quality measures considered here is the *silhouette coefficient* [11, 12], based on the idea that a “good” clustering should consist of *well-separated, cohesive* clusters, and defined as follows. Given a partitioning \mathcal{P} of a set of N objects into k clusters, consider any fixed object i and let C_i denote the set of indices for all objects clustered together with object i . As a cohesion measure for cluster C_i , take the average dissimilarity between object i and all other objects in C_i :

$$a(i) = \frac{1}{n_i} \sum_{j \in C_i} d_{ij},$$

where d_{ij} is the dissimilarity between objects i and j and n_i is the number of objects in cluster C_i . To characterize the separation between clusters, let K_ℓ denote the ℓ^{th} *neighboring cluster*, distinct from C_i , for $\ell = 1, 2, \dots, k - 1$. Define $b(i)$ as the average

dissimilarity between object i in cluster C_i and the objects in the closest neighboring cluster, i.e.

$$b(i) = \min_{\ell} \left\{ \frac{1}{n_{\ell}} \sum_{j \in K_{\ell}} d_{ij} \right\}.$$

The silhouette coefficient $s(i)$ for object i is defined as the normalized difference:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

For the silhouette coefficient $s(i)$ to be well-defined, \mathcal{P} must contain at least two clusters and every cluster in \mathcal{P} must contain at least two objects. Then, it is not difficult to show that the silhouette coefficient satisfies $-1 \leq s(i) \leq 1$ for all i and these limits have the following interpretations. If cluster C_i is “tight,” all of its objects exhibit small dissimilarities, implying that $a(i)$ is a small positive number. Similarly, if cluster C_i is well-separated from all of its neighbors, $b(i)$ is a much larger positive number so $\max\{a(i), b(i)\} = b(i)$ and $b(i) - a(i) \simeq b(i)$. Hence, for an object i in a tight cluster, well separated from its neighbors, $s(i) \simeq 1$. Conversely, suppose object i has been assigned to the wrong cluster by a clustering procedure. It then follows that $a(i)$ is a large positive number (reflecting its dissimilarity to the other objects it is clustered with) and $b(i)$ is a small positive number (i.e., its “nearest neighbor cluster” is really the cluster it belongs in). In this case, $\max\{a(i), b(i)\} = a(i)$ and $b(i) - a(i) \simeq -a(i)$, implying $s(i) \simeq -1$. Finally, if there is little or no inherent cluster structure, we would expect that $a(i)$ and $b(i)$ would be about the same (i.e., the “best cluster” for object i is essentially no better than the “second best cluster” for that object), implying $s(i) \simeq 0$. A useful overall quality measure for a given clustering is its average silhouette coefficient:

$$\bar{s} = \frac{1}{N} \sum_{i=1}^N s(i).$$

Based on their experience, Kaufman and Rousseeuw [12, p. 88] offer the following suggestions for the interpretation of \bar{s} as a measure of evidence in support of cluster structure:

- $0.70 < \bar{s} \leq 1.00 \Rightarrow$ strong evidence
- $0.50 < \bar{s} \leq 0.70 \Rightarrow$ reasonable evidence
- $0.25 < \bar{s} \leq 0.50 \Rightarrow$ weak evidence
- $\bar{s} \leq 0.25 \Rightarrow$ no significant evidence.

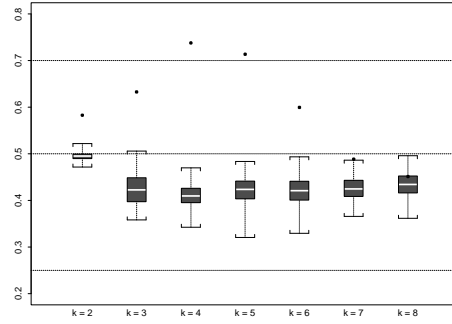


Figure 2: Boxplot summaries of the permutation-based average silhouette coefficient values, plotted against the number of clusters k

The advantage of the permutation results presented here is that they give additional support to these conclusions when they are valid, offering the possibility of attaching statistical significance to these statements. In particular, note that if the value of \bar{s} computed from the original dataset is larger than all but $q - 1$ of the m permutation values, the (one-sided) probability of observing a value this extreme in the permutation data is less than q/m . As a consequence, we can take the observation of such average silhouette coefficient values as evidence that structure in the original dataset has been destroyed at a significance level of q/m . This additional information is especially useful in cases where the silhouette coefficients are either somewhat marginal (e.g., less than 0.5) or misleading by themselves, a problem discussed further in Sec. 6.

2.3 Results and significance Fig. 2 shows a boxplot summary of the average silhouette coefficient values \bar{s} obtained after applying the PAM clustering procedure to each of 200 random permutations of the Ruspini dataset (specifically, these permutations were applied to the second component of each attribute vector; the first component was not modified). In addition, the average silhouette coefficient values computed from the original dataset are shown as solid circles in this plot, and the horizontal dashed lines represent the interpretation threshold values discussed above. Here, the clustering results for $k = 2$ through $k = 6$ all clearly support the structure hypothesis at the 0.5% level, since they all exceeded the corresponding 200 permutation values. Taking the maximum of these significant \bar{s} values as the most likely number of clusters present gives the correct result, $k = 4$. Also, note that the evidence in support of the secondary choice $k = 5$ appears almost as strong as that for $k = 4$, lending credence to the argument for

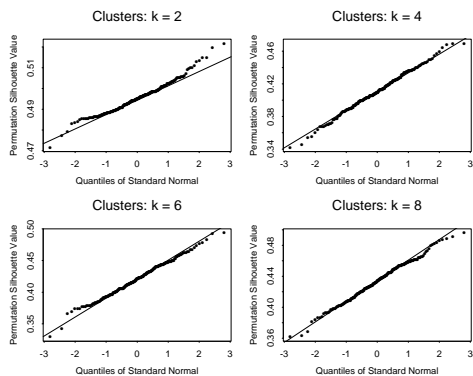


Figure 3: Normal Q-Q plots for the k cluster permutation results, $k = 2, 4, 6,$ and 8 .

this secondary choice. Support for other choices of k is substantially weaker, both in terms of the average silhouette values themselves and in terms of the distance these values lie from the 200 permutation results.

This last observation raises an important point. The arguments just presented do not depend on the distribution of the permutation values, only on the hypothesis that all permutations are equivalent (specifically, that the randomized datasets are exchangeable [10]). As a consequence, in the practically important case where the original result falls outside the range of the permutation results, strictly rank-based interpretations make no use of the *distance* between the original and the randomized results. Frequently, however, the randomized results exhibit an approximately normal distribution, permitting us to make stronger statements about significance that do account for this distance. Specifically, define the z score for each set of results as:

$$z = \frac{\bar{s}_0 - \bar{s}_p}{\sigma},$$

where \bar{s}_0 is the value of \bar{s} computed from the original dataset, \bar{s}_p is the average of the m permutation values for \bar{s} , and σ is the standard deviation of these m permutation \bar{s} values. Under the approximate normality assumption just described, the probability of observing a value of z larger than Z is $1 - \Phi(Z)$ where $\Phi(x)$ is the standard Gaussian cumulative distribution function.

The *normal quantile-quantile (Q-Q) plot* provides a useful informal test of the approximate normality hypothesis for a given dataset [3]. This plot shows the rank-ordered data values, sorted from smallest to largest, as a function of the rank order: if the approximate normality assumption is reasonable, this plot should appear as a straight line. D’Agostino and Stephens [3] argue against using Q-Q plots for datasets smaller than about 50 numbers, but these plots are quite

k	\bar{s}	z	p
2	0.58	11.66	0
3	0.63	6.47	5×10^{-11}
4	0.74	13.60	0
5	0.71	10.20	0
6	0.60	6.48	5×10^{-11}
7	0.49	2.55	0.005
8	0.45	0.71	0.239

Table 1: Summary of silhouette coefficient permutation results for the Ruspini dataset

useful for characterizing the 200 permutation results considered here. Fig. 3 shows normal Q-Q plots for these results for $k = 2, 4, 6,$ and 8 , from which it is clear that the approximate normality assumption is reasonable here. Consequently, the Gaussian p -values are computed and listed in the last column of Table 1. The extreme significance of these p values greatly strengthens the permutation-based conclusions presented above.

Finally, note that even if normality is not a reasonable approximation, so long as the permutation results exhibit a unimodal distribution, the probability of observing a particular result in the randomized data decreases monotonically (and usually rapidly) with increasing z scores. Hence, even under this much weaker distributional assumption, larger z scores can be taken as stronger evidence in support of a particular cluster structure hypothesis.

3 Other quality measures

In addition to the silhouette coefficients defined in Sec. 2.2, we also consider the 36 cluster validity measures introduced by Bolshakova and Azuaje [2], representing 18 variations of each of the cluster validity measures proposed by Dunn [9] and Davies and Bouldin [4].

3.1 Dunn’s index The index proposed by Dunn [9] is based on a partitioning \mathcal{P} of the dataset \mathcal{D} into k clusters $C_i, i = 1, 2, \dots, k$, a distance measure (intercluster distance) $\delta(C_i, C_j)$ between distinct clusters, and a cohesion measure (intracluster distance) $\Delta(C_i)$ for each cluster. These distances may be computed in various ways, discussed further in Sec. 3.3, but here it is enough to note that both of these quantities are assumed to be positive for all clusters. Given these quantities, Dunn’s

metric for the quality of a clustering is given by:

$$Q(\mathcal{P}) = \min_i \left\{ \min_{j \neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{\ell} \{\Delta(C_{\ell})\}} \right\} \right\} \\ = \frac{\min_i \{\min_{j \neq i} \{\delta(C_i, C_j)\}\}}{\max_{\ell} \{\Delta(C_{\ell})\}}.$$

Note that if all clusters are well-separated with respect to the intercluster distance measure $\delta(\cdot, \cdot)$ and “tight” with respect to the intracluster distance measure $\Delta(\cdot)$, Dunn’s index will assume a large positive value. Hence, one approach to selecting the number of partitions k in a dataset is to maximize $Q(\mathcal{P})$ with respect to k . Like the silhouette coefficients $s(i)$, Dunn’s metric is not applicable to unclustered datasets: it is not computable for the case $k = 1$.

3.2 The Davies-Bouldin index The index of Davies and Bouldin [4] represents a different combination of the same elements used to define Dunn’s index. Specifically, given a clustering \mathcal{P} and the quantities $\delta(C_i, C_j)$ and $\Delta(C_i)$ defined in Sec. 3.1, the Davies-Bouldin index is defined as

$$(3.1) \quad Q(\mathcal{P}) = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left\{ \frac{\Delta(C_i) + \Delta(C_j)}{\delta(C_i, C_j)} \right\}.$$

In contrast to Dunn’s index, the Davies-Bouldin index should exhibit small values for well-separated, compact clusters. Hence, the optimum number of clusters in a dataset may be determined by minimizing $Q(\mathcal{P})$ with respect to k . Also, like Dunn’s index, the Davies-Bouldin index is not defined for the unclustered case, $k = 1$. Finally, as a practical matter, we have found plots of the logarithm of the Davies-Bouldin index to be more useful than plots of the raw index values. Hence, in the following discussions, the term “Davies-Bouldin index” will refer to the logarithm of the expression defined in Eq. (3.1).

3.3 The Bolshakova-Azuaje family Evaluation of either of the cluster quality indices just described requires computable definitions of the intercluster distance $\delta(C_i, C_j)$ between two distinct clusters C_i and C_j and the intracluster distance $\Delta(C_i)$ for every individual cluster C_i . The following paragraphs describe the six choices of $\delta(C_i, C_j)$ and the three choices of $\Delta(C_i)$ proposed by Bolshakova and Azuaje [2] in their study of cluster stability.

Let \mathcal{S} and \mathcal{T} be two different clusters, of size $|\mathcal{S}|$ and $|\mathcal{T}|$. The *single linkage distance* $\delta_1(\mathcal{S}, \mathcal{T})$ is the closest distance between objects belonging to the distinct clusters \mathcal{S} and \mathcal{T} , given by:

$$\delta_1(\mathcal{S}, \mathcal{T}) = \min_{i \in \mathcal{S}, j \in \mathcal{T}} d_{ij}.$$

The *complete linkage distance* $\delta_2(\mathcal{S}, \mathcal{T})$ is the distance between the two most distant objects in the two clusters, defined by:

$$\delta_2(\mathcal{S}, \mathcal{T}) = \max_{i \in \mathcal{S}, j \in \mathcal{T}} d_{ij}.$$

The *average linkage distance* $\delta_3(\mathcal{S}, \mathcal{T})$ is the average distance between objects in the two clusters, given by:

$$\delta_3(\mathcal{S}, \mathcal{T}) = \frac{1}{|\mathcal{S}||\mathcal{T}|} \sum_{i \in \mathcal{S}, j \in \mathcal{T}} d_{ij}.$$

Note that all three of these intercluster distance measures are closely related to hierarchical clustering procedures [12], and they are computable from the object dissimilarity matrix alone. The next two intercluster distance measures require knowledge of the distance function itself, since they involve the *centroids* \mathbf{c}_s and \mathbf{c}_t of the clusters \mathcal{S} and \mathcal{T} , whose j^{th} components are defined by:

$$[\mathbf{c}_s]_j = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} x_j^i, \quad [\mathbf{c}_t]_j = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} x_j^i,$$

where x_j^i is the j^{th} component of the attribute vector \mathbf{x}^i for object i . Note that the centroids are generally *not* objects in the original dataset, so complete knowledge of the original dissimilarity matrix is not sufficient to compute either the centroids or the following two intercluster distance measures. Denoting the distance function between attribute vectors \mathbf{x} and \mathbf{y} by $d(\mathbf{x}, \mathbf{y})$, the *centroid linkage distance* $\delta_4(\mathcal{S}, \mathcal{T})$ is given by:

$$\delta_4(\mathcal{S}, \mathcal{T}) = d(\mathbf{c}_s, \mathbf{c}_t),$$

and the *average to centroids linkage distance* $\delta_5(\mathcal{S}, \mathcal{T})$ is given by:

$$\delta_5(\mathcal{S}, \mathcal{T}) = \frac{1}{|\mathcal{S}| + |\mathcal{T}|} \left(\sum_{i \in \mathcal{S}} d(\mathbf{x}^i, \mathbf{c}_t) + \sum_{i \in \mathcal{T}} d(\mathbf{x}^i, \mathbf{c}_s) \right).$$

Finally, the *Hausdorff distance* $\delta_6(\mathcal{S}, \mathcal{T})$ is defined as:

$$\delta_6(\mathcal{S}, \mathcal{T}) = \max\{h(\mathcal{S}, \mathcal{T}), h(\mathcal{T}, \mathcal{S})\} \\ h(\mathcal{X}, \mathcal{Y}) = \max_{i \in \mathcal{X}} \{\min_{j \in \mathcal{Y}} \{d_{ij}\}\}.$$

The three intracluster distances proposed by Bolshakova and Azuaje [2] are the *complete diameter*, defined as

$$\Delta_1(\mathcal{S}) = \max_{i, j \in \mathcal{S}} \{d_{ij}\},$$

the *average diameter*, defined as

$$\Delta_2(\mathcal{S}) = \frac{1}{|\mathcal{S}|(|\mathcal{S}| - 1)} \sum_{i, j \in \mathcal{S}, j \neq i} d_{ij},$$

Model	p	k	Description
1	10	1	Uniform on the unit cube in R^{10}
2	2	3	Well-separated Gaussian clusters
3	10	4	3 informative, 7 noise components
3 ⁰	3	4	Informative part of Model 3
5	3	2	Two elongated clusters
6	10	2	Model 5 plus 7 noise components
7	10	2	Overlapping, 9 noise components

Table 2: Summary of the Dudoit and Fridlyand models and variations considered here

and the *centroid diameter*, defined as

$$\Delta_3(\mathcal{S}) = \frac{2}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} d(\mathbf{x}^i, \mathbf{c}_s).$$

Note that like the intercluster distances $\delta_4(\mathcal{S}, \mathcal{T})$ and $\delta_5(\mathcal{S}, \mathcal{T})$, the centroid diameter $\Delta_3(\mathcal{S})$ requires knowledge of the dissimilarity function $d(\cdot, \cdot)$ and not just the pairwise dissimilarity matrix between the objects in the original dataset.

4 Case study

The following paragraphs present partial results from a case study based on the Ruspini dataset and a variant of the eight simulation-based datasets described by Dudoit and Fridlyand [8]. These models are summarized in Table 2 and they include a uniformly distributed random dataset (Model 1) that has been advocated as a structureless null model for cluster validation [11, p. 186], three models with clear structure like the Ruspini dataset (Models 2, 3⁰, and 5), extensions of two of these three models obtained by adding spurious noise components to the attribute vectors (Models 3 and 6), and a model with two overlapping one-dimensional clusters and nine spurious noise components (Model 7).

Fig. 4 shows the results obtained for the Ruspini dataset using the Dunn index based on the complete linkage intercluster distance $\delta_2(\mathcal{S}, \mathcal{T})$ and the complete diameter $\Delta_1(\mathcal{S})$. The boxplots in this figure summarize 50 random permutations of the dataset, giving one-sided significance values of $p = 0.02$ whenever the Dunn index computed from the original dataset exceeds all of the permutation values. Here, the only significant result is that for $k = 4$, which corresponds to the correct cluster structure. For comparison, Fig. 5 shows the results obtained using the average to centroids distance $\delta_5(\mathcal{S}, \mathcal{T})$ with $\Delta_1(\mathcal{S})$, which exhibit significant results

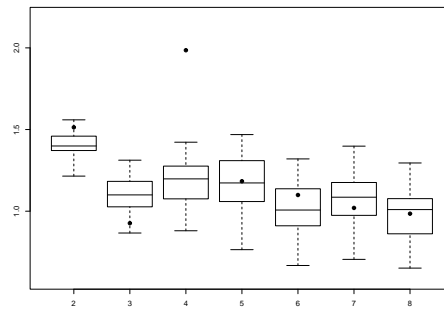


Figure 4: Permutation results for the Ruspini dataset with the Dunn(2,1) index

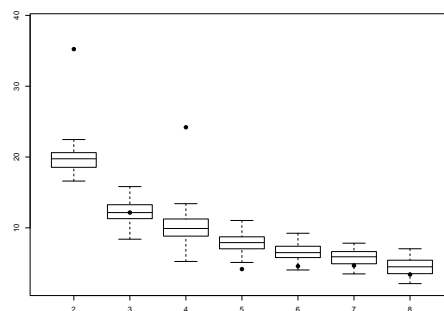


Figure 5: Permutation results for the Ruspini dataset with the Dunn(5,1) index

for both $k = 2$ and $k = 4$. In fact, both the Dunn index values ($Q(\mathcal{P}) \simeq 36$ vs. 25) and their associated z values (10.51 vs. 7.93) strongly favors $k = 2$ over the correct value $k = 4$. This result is considered further at the end of this section, but the key point is that the choice of the 18 options considered here for defining Dunn indices is quite important in practice.

Results obtained for the Ruspini dataset using the Davies-Bouldin indices with the same choices of intercluster and intracluster distances as in the first example are shown in Fig. 6. Recall that for the Davies-Bouldin index, small values are better than larger ones, so we are interested in those values of k for which the Davies-Bouldin index values computed from the original dataset lie below the smallest of the permutation values. Here, we have two such values: $k = 4$ and $k = 5$, in agreement with the results obtained from the silhouette coefficients. Both the fact that the Davies-Bouldin index value is smaller for $k = 4$ than for $k = 5$ and the fact that the z value is more negative for $k = 4$ than for $k = 5$ support the generally preferred interpretation of the Ruspini dataset as containing four clusters.

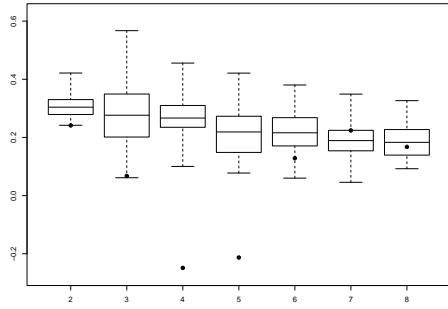


Figure 6: Permutation results for the Ruspini dataset with the DB(2,1) index

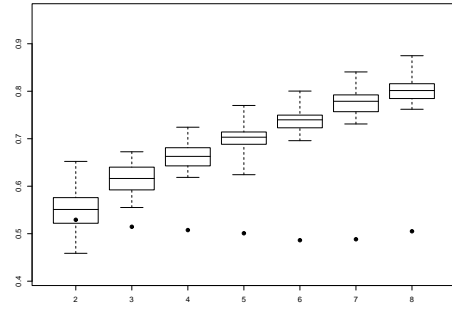


Figure 8: Permutation results for the Model 1 dataset with the DB(2,3) index

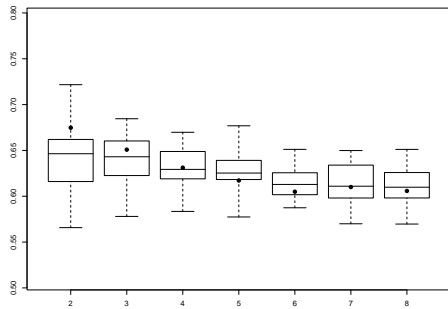


Figure 7: Permutation results for the Model 1 dataset with the DB(2,1) index

Results obtained from Model 1 with the DB(2,1) index are shown in Fig. 7. Here, as in all other cases involving attribute vectors of dimension $p > 2$, independent random permutations are applied to all but the first of the p components of this vector. In this particular example, no cluster structure is present, so none of the original cluster quality indices should be significant relative to the permutation results, and this is precisely the result seen in Fig. 7 for $k = 2$ through $k = 8$. In contrast, Fig. 8 shows the results obtained for this structureless dataset from the Davies-Bouldin index based on the same intercluster distance (complete linkage, $\delta_2(\mathcal{S}, \mathcal{T})$) but with the centroid diameter $\Delta_3(\mathcal{S})$ instead of the complete diameter, $\Delta_1(\mathcal{S})$. There, significant evidence appears in support of any number of clusters between $k = 3$ and $k = 8$, with strongest support for $k = 7$, which has both the smallest Davies-Bouldin index value and the most negative z score.

More generally, the behavior seen in Fig. 8 is characteristic of the results obtained using the centroid diameter $\Delta_3(\mathcal{S})$ with any choice of intercluster distance, for either the Dunn indices or the Davies-Bouldin in-

trices. That is, results based on $\Delta_3(\mathcal{S})$ almost always give high significance to a wide range of k values, frequently with a shallow maximum (for the Dunn indices) or minimum (for the Davies-Bouldin indices) at a fairly large value of k . Further, of the 37 cluster quality measures considered here, only those based on the centroid diameter provided evidence for cluster structure in the structureless Model 1 dataset.

Similarly, the average to centroids intercluster distance $\delta_5(\mathcal{S}, \mathcal{T})$ also appears to perform badly for many of the examples considered here. For example, this intercluster distance was the only one that gave strong support for the incorrect number of clusters in the Ruspini dataset, as seen in Fig. 5. Further, $\delta_5(\mathcal{S}, \mathcal{T})$ was the only intercluster distance that led to incorrect clustering results for Model 3⁰ using the Dunn indices. Also, the results obtained for the Davies-Bouldin indices were consistently incorrect for this model whenever they were based on $\delta_5(\mathcal{S}, \mathcal{T})$.

Comparing the Dunn and Davies-Bouldin indices across all of the datasets considered here, it appears that the Dunn indices generally yield better results. For example, the Dunn index based on the centroid linkage distance $\delta_4(\mathcal{S}, \mathcal{T})$ and the complete diameter $\Delta_1(\mathcal{S})$ found no structure in the structureless Model 1 dataset, but the Davies-Bouldin index based on the same pair of distances gave clear support for $k = 5$ clusters. More dramatic differences are seen for Model 3, where 7 of the 18 Dunn indices gave reasonably clear evidence for the correct cluster structure ($k = 4$), but *none* of the Davies-Bouldin indices gave clear support for this structure.

5 Extraneous attributes

This last result illustrates an important point: Models 3, 6, and 7 are difficult to cluster correctly, due to the presence of many extraneous noise components.

Noise Components	\bar{s} , $k = 2$	\bar{s} , $k = 4$	z , $k = 2$	z , $k = 4$
0	0.636	0.750	70.38	22.89
1	0.619	0.709	52.88	19.51
2	0.604	0.675	49.12	20.99
3	0.587	0.638	28.51	25.29
4	0.579	0.619	23.53	17.44
5	0.568	0.595	23.49	23.92
6	0.557	0.573	22.15	17.47
7	0.548	0.555	23.02	18.82

Table 3: Average silhouette coefficients and z values for $k = 2$ and $k = 4$ as a function of the number of noise components included in Model 3

This point has been noted previously [11, p. 24] as an argument against including possibly extraneous attributes in cluster analysis. Our results strongly support this cautionary argument. For example, results obtained for Model 3⁰ give clear evidence for the correct number of clusters for 14 of the 18 Dunn indices considered here, but we only obtain comparably clear evidence for the correct clustering for 7 of the 18 Dunn indices for the Model 3 data, which includes the three informative attributes from Model 3⁰ and seven unrelated noise components. This difference is even more pronounced for Models 5 and 6, which also differ by the presence of seven unrelated noise components. There, both the Dunn and the Davies-Bouldin indices give the correct results in all 18 cases for Model 5, but only two of the Dunn indices and one of the Davies-Bouldin indices find the correct structure for Model 6. The results for Model 7, which has overlapping univariate clusters and nine unrelated noise components, are even worse: none of the 37 cluster quality measures considered here gives clear evidence for the correct structure.

Table 3 shows how the results obtained for Model 3⁰ with the silhouette coefficient degrade as additional noise components are included in the attribute vectors used for clustering. In the absence of any extraneous noise components (i.e., for Model 3⁰), the average silhouette value for $k = 4$, the correct result, is significantly larger than that for $k = 2$, the case with the second-largest \bar{s} value. As noise components are added, two things happen. First, the average silhouette coefficient values all decrease in magnitude, suggesting poorer clustering, although they all remain highly significant relative to the permutation results, based on the very

large z values listed in Table 3. Second, the difference between the average silhouette coefficient for $k = 4$ and that for $k = 2$ decreases monotonically as more noise components are included in the attribute vectors. In particular, note that by the time seven noise components have been included (i.e., for Model 3), these average silhouette coefficients only differ by about 1%.

6 Overall assessment

The general CNC strategy described here starts with a set $\{\mathcal{P}_k\}$ of k -partition clusterings for a given dataset over a specified range of k values. Using structure-destroying permutations, we then assess the significance of the cluster structure found for each k with some measure of cluster quality, $Q(\mathcal{P}_k)$. Finally, we retain only those partitionings \mathcal{P}_k that show significance, either in terms of permutation statistics alone (i.e., the rank of the original data clustering relative to the m permutations) or using z scores as a more quantitative measure of significance. The best of these significant partitionings with respect to the quality measure $Q(\cdot)$ is then taken as the partitioning \mathcal{P}_{k^*} best supported by the data. In favorable cases, the clustering that maximizes the quality measure $Q(\cdot)$ will also exhibit the most significant z score, and this z score will be far larger than all of the others. This behavior is nicely illustrated by the Dunn(3,1) index results obtained for Model 3⁰. There, the correct clustering ($k = 4$) exhibits both the largest Q value (1.41) and the most significant z score (20.73); in contrast, the second-best result is obtained for $k = 2$, with a Q value of 1.05 and a z score of 8.78, and the third-best result is that for $k = 3$, with a Q value of 0.74 and a z score of 5.41. Also, these Q values are the only three that fall outside the range of the 50 randomizations considered for this example. Hence, in this case, both the z score and the Q value argue in favor of the correct result, $k = 4$.

Conversely, it frequently happens that the partitioning that maximizes the quality measure $Q(\cdot)$ is not significant and indeed, corresponds to the incorrect partitioning. A specific example where this occurs is the result obtained for the Dunn(2,3) index computed from Model 3. There, the largest Q value is 2.15, obtained for $k = 2$, but this Q value falls within the range of the randomized Q values, with a z score of only 0.91. In contrast, the largest *significant* Q value is 1.98 for the correct structure, $k = 4$, falling well outside the range of randomized Q values, with a z score of 7.36. It is also important to note that maximizing the z score does not give the correct partitioning in this case: the maximum z score seen in this example is 10.89, obtained for $k = 8$, but the Q value for this case is only 1.20. In general, we have found that the combined use of both Q values

for the original dataset and their associated z scores as described here gives better results than the use of either Q values or z scores alone.

Overall, the results obtained for the datasets considered here demonstrate that the proposed CNC approach to cluster validation is extremely effective in favorable cases. In particular, all of the 37 cluster quality measures we examined gave the correct results for the Model 5 dataset, and most gave the correct results for the Ruspini and Models 1, 2, and 3^0 datasets. Conversely, we have also seen that this approach fails in difficult cases like those containing many spurious noise components: we never obtained the correct results for Model 7, rarely for Model 6, and only occasionally for Model 3. Part of this difficulty may lie in the clustering algorithm used here (procedure PAM with Euclidean dissimilarities), although clustering is known to be difficult under these circumstances, as discussed in Sec. 5. We are exploring the question of method dependence further and will report the results elsewhere.

In general, our results suggest that silhouette coefficients represent the best of the 37 cluster quality measures considered here. Specifically, silhouette coefficients gave clear, correct results for the Ruspini dataset and for Models 1, 2, 3^0 , and 5. Further, silhouette coefficients gave marginally correct results for Model 6, for which almost none of the other 36 quality measures gave correct results. The poorest relative performance for the silhouette coefficients was obtained for Model 3, for which the results were very marginal and clearly inferior to those obtained with three of the Dunn indices (specifically, $\text{Dunn}(1,1)$, $\text{Dunn}(3,1)$ and $\text{Dunn}(6,2)$). As noted, none of the 37 measures considered here gave correct results for Model 7. In addition to their generally superior performance relative to the other cluster quality measures considered here, the silhouette coefficients also have the advantage of normalization, lying in the range $-1 \leq \bar{s} \leq 1$ and having the informal interpretation guidelines presented in Sec. 2.2. In contrast, the other 36 cluster quality measures considered here are only constrained to be positive and the observed ranges of variation depend strongly both on the choices of $\delta(\mathcal{S}, \mathcal{T})$ and $\Delta(\mathcal{S})$ and on the datasets considered.

Conversely, the results just noted also indicate that silhouette coefficients do sometimes fail to give the correct result in situations where some of the other cluster quality measures considered here can give the correct result. This observation motivates our continued interest in these alternatives, although it does appear that the list of 36 choices considered here can be pared down substantially. First, almost without exception in the examples considered here, the Dunn index based on any given combination of intercluster and intraclus-

i, j	$k^*, 3^0$	$Q, 3^0$	$z, 3^0$	$k^*, 3$	$Q, 3$	$z, 3$
1,1	4	0.70	28.95	4	0.65	20.79
1,2	4	1.68	27.26	4 [†]	1.15	17.04
1,3	4	1.18	47.51	4	0.83	24.88
2,1	4	2.10	14.49	2	1.64	4.90
2,2	4	5.00	15.35	2*	3.15	3.04
2,3	4	3.51	21.07	4	1.98	7.36
3,1	4	1.41	20.73	4	1.10	12.82
3,2	4	3.34	23.96	2 [†]	1.95	10.34
3,3	4	2.35	28.24	4	1.39	11.97
4,1	4	1.36	26.35	4 [†]	0.96	16.89
4,2	4	3.24	32.00	2	1.79	18.23
4,3	4	2.27	36.29	2 [†]	1.23	5.74
5,1	2	73.40	10.04	2	70.09	8.12
5,2	2	156.85	18.03	2	134.24	11.55
5,3	2*	102.13	2.50	2	91.67	3.45
6,1	4 [†]	1.45	10.68	2*	1.33	2.29
6,2	4	3.46	11.33	4	2.07	3.83
6,3	4	2.42	18.14	4	1.48	7.25

Table 4: Summary of the 18 $\text{Dunn}(i,j)$ index results for Models 3^0 and 3

ter distances gave better results than the corresponding Davies-Bouldin index. Also, it was noted that both Dunn and Davies-Bouldin indices based on the centroid diameter $\Delta_3(\mathcal{S})$ often gave poor results, as did these indices based on the average to centroids linkage distance, $\delta_5(\mathcal{S}, \mathcal{T})$. This point is illustrated clearly in Table 4, which summarizes the results obtained for all 18 of the Dunn indices considered here, for Models 3^0 and 3. Specifically, this table lists the number of clusters k^* identified in the dataset, the associated quality measure $Q = Q(\mathcal{P}_{k^*})$, and the associated z -score. Entries where k^* is marked with the symbol [†] represent ambiguous cases where the difference in Q values between the best and second-best k values were quite small. Note that all three of the $\text{Dunn}(5,j)$ indices gave the incorrect results for Model 3^0 , and these were the *only* cases where the Dunn indices failed to identify the correct structure, as noted in Sec. 4. Conversely, note that in this case if we had chosen k to maximize the z score, we would have obtained the correct result. This situation also occurs twice for Model 3, for the $\text{Dunn}(2,2)$ and $\text{Dunn}(6,1)$ indices; these cases are marked with * in Table 4. Also, as in the results presented in Sec. 5 for the silhouette coefficients, the degradation caused by the addition of unrelated noise components is clear for the Dunn indices as well. In particular, note that the Q

values obtained for Model 3, containing the 7 extraneous noise attributes, are always smaller than those for Model 3⁰, which contains only the 3 information-bearing attributes from Model 3. Similarly, in every case except Dunn(5,3), the z scores for the predicted clustering are smaller for Model 3 than for Model 3⁰. Conversely, it is also clear that the extent of this degradation is a strong function of the specific Dunn index considered. For example, it may be seen in Table 4 that the differences between the results for Model 3⁰ and Model 3 are fairly minor for Dunn(1,1): the Q value is about 7% smaller for Model 3, the z score is about 28% smaller but still quite significant, and the correct number of clusters is identified in both cases. In contrast, the differences are much more pronounced for Dunn(2,2): the Q value declines by 37%, the z score declines by 80%, and the correct number of clusters is not identified for Model 3.

7 Computational considerations

The results presented in this paper were obtained using the basic CNC procedure outlined in Sec. 2.1, but several reviewers raised concerns over the computational complexity, which is $m + 1$ times that of standard clustering method \mathcal{M} on which it is based. We acknowledge the importance of this concern and we are exploring several approaches to reduce this basic complexity. The following discussions briefly outline two preliminary ideas that are clearly useful in special cases and which may be useful in developing more general approaches.

7.1 Exploratory comparisons Although it does not provide the basis for computing statistical significance levels, applying the general procedure described here with a single randomization (i.e., with $m = 1$) does provide comparisons that may be extremely useful in an exploratory analysis context. In what follows, we suppose $Q(\cdot)$ is a cluster quality measure like Dunn's index or the silhouette coefficient, for which larger values correspond to better clustering results, although the idea extends immediately to indices like the Davies-Bouldin index for which smaller values correspond to better clustering results. Given a dataset \mathcal{D} and a range $k_{min} \leq k \leq k_{max}$ of candidate cluster numbers, first partition the unmodified dataset \mathcal{D} for this range of k values using any method \mathcal{M} to obtain the collection $\{\mathcal{P}_k\}$ of clusterings for $k = k_{min}$ through $k = k_{max}$. From these results, compute the corresponding quality measures $\{Q(\mathcal{P}_k)\}$. Next, apply independent random permutations to components 2 through p of the p -component attribute vector associated with each object to obtain a single randomized dataset \mathcal{D}^r . Apply clustering method \mathcal{M} to this new dataset to obtain the corresponding set of partitionings $\{\mathcal{P}_k^r\}$ for $k = k_{min}$

through $k = k_{max}$ and, from these, the corresponding set of quality measures $\{Q(\mathcal{P}_k^r)\}$. Since $Q(\mathcal{P}) > Q(\mathcal{P}')$ implies partitioning \mathcal{P} is better than partitioning \mathcal{P}' here, useful insights may be obtained by looking at the following differences:

$$(7.2) \quad \delta_k = Q(\mathcal{P}_k) - Q(\mathcal{P}_k^r).$$

In particular, note that if δ_k is a large positive number, this result provides preliminary evidence in support of k significant clusters. To determine what is "large," it may be desirable to consider a normalized quantity like

$$(7.3) \quad \Delta_k = \frac{Q(\mathcal{P}_k) - Q(\mathcal{P}_k^r)}{Q(\mathcal{P}_k) + Q(\mathcal{P}_k^r)},$$

which satisfies $|\Delta_k| \leq 1$ for Dunn's index or other inherently positive cluster quality measures. Conversely, for normalized measures like the silhouette coefficient, it is enough to examine δ_k directly since $|Q(\mathcal{P})| \leq 1$ implies $|\delta_k| \leq 2$. In any case, preliminary evaluation of δ_k or Δ_k values provides a basis for excluding values of k for which significant evidence of cluster structure in the dataset appears unlikely.

This idea extends directly to comparison of other clustering characteristics besides the number of clusters k . In particular, the same strategy could be used to compare different clustering methods $\{\mathcal{M}_i\}$, different dissimilarity measures, or any other method- or pretreatment-related computational options that might be of interest. The practical value of this idea lies in the fact that we can exclude analysis options that are unlikely to characterize whatever structure is present in the dataset at twice the standard clustering cost, rather than $m + 1$ times this cost for some large m .

7.2 Separable dissimilarities A significant contribution to the computational cost of clustering using general procedures like the PAM algorithm employed here is the expense of forming the $N \times N$ dissimilarity matrix \mathbf{D} , given a set of N attribute vectors of dimension p . In fact, since dissimilarities satisfy $d_{ii} = 0$ for all i and $d_{ji} = d_{ij}$ for all i and j [11, 12], it is enough to compute the $N(N-1)/2$ elements of this matrix, d_{ij} for $i > j$. If the cost of computing each element d_{ij} is C , the total cost associated with the basic CNC procedure described here is approximately $mN^2C/2$. This computational cost may be traded for storage space if we consider *separable dissimilarity measures* of the general form:

$$(7.4) \quad d_{ij} = \sum_{\ell=1}^p \delta(i, j, \ell),$$

where $\delta(i, j, \ell)$ is the contribution of component ℓ of the attribute vectors \mathbf{x}_i and \mathbf{x}_j to the total dissimilarity.

The best known separable dissimilarity measure is the Manhattan (L_1) distance, for which

$$(7.5) \quad \delta(i, j, \ell) = |x_i^\ell - x_j^\ell|.$$

Also, note that while the Euclidean (L_2) dissimilarity is not of this form, the squared Euclidean distance is separable, corresponding to

$$(7.6) \quad \delta_{ij\ell} = |x_i^\ell - x_j^\ell|^2,$$

an observation that leads to some possible extensions, discussed at the end of this section.

Now, suppose d_{ij} is a separable dissimilarity of the form (7.4) and consider the dissimilarity d_{ij}^r between the randomized attribute vectors \mathbf{x}_i^r and \mathbf{x}_j^r . Note that the effect of the randomization r on the ℓ component of the original attribute vector \mathbf{x}_i is to replace x_i^ℓ with $x_{\pi(r, \ell; i)}^\ell$, where $\pi(r, \ell; i)$ is the i^{th} element of the N -element permutation vector $\pi(r, \ell)$. Here, r corresponds to a randomization index, taking values from 1 to m , and ℓ corresponds to an attribute component index, taking values from 1 to p . The reason this permutation depends on both r and ℓ is that the permutations applied to each component of the attribute vector are assumed independent; otherwise, if the same random permutation were applied to all components of the attribute vector, π would not depend on ℓ , only on the randomization r . The dissimilarity d_{ij}^r for the randomized attributes is given by:

$$(7.7) \quad d_{ij}^r = \sum_{\ell=1}^p \delta(\pi(r, \ell; i), \pi(r, \ell; j), \ell).$$

The key observation here is that if we initially compute and store the $N(N-1)p/2$ component dissimilarities $\delta(i, j, \ell)$ for $1 \leq i < j \leq N$ and $1 \leq \ell \leq p$ that defines any separable dissimilarity, we can compute d_{ij}^r for any randomization r by simply indexing the appropriate components of this array and adding the results. While this approach is not practical for very large p due to increased storage requirements, it represents a considerable computational savings in cases where it is applicable.

Finally, note that this same idea is also applicable to Euclidean distances. In particular, if we store the array of component squared Euclidean distances, we can simply apply the result (7.7) to the separable distance measure defined in Eq. (7.6) and take the square root of the resulting distances d_{ij} to obtain the elements of the desired Euclidean dissimilarity matrix. In fact, an even better alternative exists, based on the *Pythagorean sum* algorithm introduced by Moler and Morrison [13] and refined by Dubrulle [7]. Specifically, given two real

numbers, x and y , their Pythagorean sum is defined as

$$(7.8) \quad x \oplus y = \sqrt{x^2 + y^2},$$

and Moler and Morrison describe a procedure for computing $x \oplus y$ directly from x and y , without either forming squares or taking square roots. In the context of the problem considered here, note that the Euclidean distance d_{ij} between attribute vectors \mathbf{x}_i and \mathbf{x}_j is

$$(7.9) \quad d_{ij} = \delta(i, j, 1) \oplus \delta(i, j, 2) \oplus \cdots \oplus \delta(i, j, p),$$

where $\delta(i, j, \ell)$ is the distance component defined in Eq. (7.5) for the Manhattan dissimilarity. Hence, although Euclidean dissimilarities are not separable with respect to ordinary addition, they are separable with respect to Pythagorean addition. In fact, this idea could be extended to other dissimilarity measures besides the Euclidean case by considering dissimilarities that are separable with respect to other variants of addition. More specifically, Aczel [1] discusses the general class of *associative binary operators* that share many characteristics of ordinary addition and which can be represented in the general form

$$(7.10) \quad x \circ y = \phi^{-1}(\phi(x) + \phi(y)),$$

where the function $\phi: R \rightarrow R$ is strictly monotonic and continuous. Note that the function $\phi(x) = x^2$ satisfies these conditions when restricted to positive arguments x and this choice yields the Pythagorean sum $x \oplus y$. More generally, any dissimilarity that can be expressed as an associative binary combination of individual component dissimilarities can be computed from these component dissimilarities by direct application of Eq. (7.10) as

$$(7.11) \quad d_{ij} = \phi^{-1} \left(\sum_{\ell=1}^p \phi(\delta(i, j, \ell)) \right).$$

In a very different context, this idea has been used to define a broad class of nonlinear discrete-time dynamic models that can exhibit strongly nonlinear qualitative behavior while still retaining certain characteristics of linear dynamic models [14].

8 Summary

This paper has described a detailed permutation-based procedure for detecting the presence of significant cluster structure in a dataset using unsupervised, partition-based clustering algorithms. The basic idea is to compute a cluster quality measure for a partitioning of the original dataset and compare it with the range of quality measures obtained from a collection of random permutations applied to the dataset to destroy any cluster

structure that may be present. Large differences in these results may be taken as evidence in support of cluster structure, and this idea may be used to determine the most probable number of clusters present in the dataset. To assess the practicality of this idea, we have compared results obtained for the 37 cluster quality measures described by Bolshakova and Azuaje [2], applied to eight datasets: the Ruspini dataset [12, 15], six of the eight simulation datasets considered by Dudoit and Fridlyand [8], and a variation on one of these datasets (Model 3⁰, described in Sec. 4).

Based on the results presented in Secs. 4, 5, and 6, it appears reasonable to restrict consideration to the silhouette coefficient and the Dunn indices, and to drop those indices based on the distance measures $\delta_5(\mathcal{S}, \mathcal{T})$ and $\Delta_3(\mathcal{S})$, reducing the list of alternative cluster quality measures considered from 37 to 11. Also, note that these two omitted distances represent two of the three centroid-based distances, all of which incur additional computational effort in constructing the required centroids. In particular, the other four intercluster distances—single linkage, complete linkage, average linkage and Hausdorff distances—can all be computed directly from the dissimilarity matrix required for clustering algorithms like the PAM procedure considered here. For this reason, we also drop the centroid linkage distance $\delta_4(\mathcal{S}, \mathcal{T})$, leaving nine quality measures: the silhouette coefficient and the Dunn(*i,j*) indices for $i = 1, 2, 3$, and 6 and $j = 1$ and 2.

Having established the utility of the permutation-based procedure proposed here and narrowing the set of cluster quality measures from 37 to 9, we are currently examining the influence of alternative dissimilarity measures (e.g., Manhattan distances or product-moment correlations) and alternative clustering algorithms (e.g., *k*-means). In addition, we are applying these results to biological datasets, starting with publicly available datasets like those considered by Dudoit and Fridlyand [8] and ultimately examining gene expression datasets that are currently being generated in our laboratory. Based on the preliminary results presented here, we expect these ongoing investigations to provide useful guidance in selecting clustering procedures and in evaluating the effectiveness of different microarray normalization and data pretreatment methods. In particular, note that one of the strengths of the computational negative control framework proposed here is that it can be used to assess the dependence of the weight of evidence in support of a final clustering result on changes in any of the data collection or analysis steps that contribute to that final clustering. For example, Warren and Liu [16] compare a range of normalization methods for microarray data analysis, presenting con-

vincing evidence that reported gene expression values (i.e., the attributes on which a gene clustering would be based) can depend strongly on the various method choices and tuning parameter values adopted (e.g., the smoothing bandwidth in the popular *lowess* nonparametric smoother). Procedures like the computational negative controls method proposed here should be extremely useful in guiding these selections in cases where a desired intermediate or final result is a functionally significant clustering of genes.

References

- [1] J. Aczel, *A Short Course on Functional Equations*, Reidel, Dordrecht, 1987.
- [2] N. Bolshakova and F. Azuaje, *Cluster validation techniques for genome expression data*, Signal Processing, 83 (2003), pp. 825–833.
- [3] R. D’Agostino and M. Stephens, *Goodness-of-fit Techniques*, Marcel Dekker, New York, 1986.
- [4] D. Davies and D. Bouldin, *A cluster separation measure*, IEEE Trans. Pattern Recognition Machine Intell., 1 (1979), pp. 224–227.
- [5] M. Delattre and P. Hansen, *Bicriterion cluster analysis*, IEEE Trans. Pattern Recognition Machine Intell., 2 (1980), pp. 277–291.
- [6] S. Draghici, *Data Analysis Tools for DNA Microarrays*, Chapman and Hall/CRC, New York, 2003.
- [7] A.A. Dubrulle, *A Class of Numerical Methods for the Computation of Pythagorean Sums*, IBM J. Res. Develop., 27 (1983), pp. 582–589.
- [8] S. Dudoit and J. Fridlyand, *A prediction-based resampling method for estimating the number of clusters in a dataset*, Genome Biology, 3 (2002), pp. research0036.1–research0036.21.
- [9] J. Dunn, *Well separated clusters and optimal fuzzy partitions*, J. Cybernet., 4 (1974), pp. 95–104.
- [10] P. Good, *Permutation Tests*, Springer Verlag, New York, 2000.
- [11] A. Gordon, *Classification*, Chapman and Hall, New York, 1999.
- [12] L. Kaufman and P. Rousseeuw, *Finding Groups in Data*, Wiley, New York, 1990.
- [13] C. Moler and D. Morrison, *Replacing Square Roots by Pythagorean Sums*, IBM J. Res. Develop., 27 (1983), pp. 557–581.
- [14] R.K. Pearson, Ü. Kotta, and S. Nömm, *Systems with Associative Dynamics*, Kybernetika, 38 (2002), pp. 585–600.
- [15] E. Ruspini, *Numerical methods for fuzzy clustering*, Inform. Sci., 2 (1970), pp. 319–350.
- [16] L. Warren and B. Liu, *Comparison of Normalization Methods for cDNA Microarrays*, ch. 7 in Methods of Microarray Data Analysis III, K.F. Johnson and S.M. Lin, eds., Kluwer, Boston, 2003.