Psychological Methods
2011, Vol. 16, No. 3, 285–297

© 2011 American Psychological Association
1082-989X/11/$12.00 DOI: 10.1037/a0023346

Choosing the Number of Clusters in *K*-Means Clustering

Douglas Steinley
University of Missouri

Michael J. Brusco
Florida State University

Steinley (2007) provided a lower bound for the sum-of-squares error criterion function used in *K*-means clustering. In this article, on the basis of the lower bound, the authors propose a method to distinguish between 1 cluster (i.e., a single distribution) versus more than 1 cluster. Additionally, conditional on indicating there are multiple clusters, the procedure is extended to determine the number of clusters. Through a series of simulations, the proposed methodology is shown to outperform several other commonly used procedures for determining both the presence of clusters and their number.

Keywords: cluster analysis, *K*-means clustering, choosing the number of clusters

One of the primary difficulties (if not the primary difficulty) in cluster analysis is determining whether the obtained solution actually represents the underlying structure in the data or is merely an artifact of the procedure used to obtain that solution. When expounding on this issue, some definitions are helpful. A *partition* consists of *K* nonempty, nonoverlapping, and exhaustive subsets of a set of objects (subjects, patients, items, etc.). A *partitioning procedure* is an exact or approximate algorithm that can be used to obtain a partition of the objects. It is well-known (Milligan, 1996) that partitioning procedures will always return a *K*-cluster partition of the objects, regardless of whether that partition provides an appropriate representation of the underlying structure of the data (see Figure 1 for an illustration). The left-hand panel of Figure 1 indicates a situation where there are two clusters that most reasonable clustering algorithms would be able to detect; contrarily, the right-hand panel of Figure 1 illustrates a single cluster that has been erroneously divided into two, showing the potential outcome of incorrectly applying a clustering procedure when there is only one cluster.

To avoid this pitfall, it is helpful to decide on a definition of *cluster*. Perhaps the most commonly used definition of a cluster was provided by Cormack (1971), who provided the now classic definition that clusters should be "internally cohesive" (p. 329) and "externally isolated" (p. 329), indicating that objects within a cluster should be more similar (geometrically closer by some measure of distance) than objects that are in different clusters. Mathematically, this can be represented by the general equation for a mixture of distributions

$$f(\mathrm{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathrm{x}), \qquad (1)$$

where K is the number of clusters, π_k is the probability of observing the kth cluster (e.g., sometimes referred to as the *base rate*), and f_k is the probability distribution function of that generates the data observed in the kth cluster. Whenever a clustering procedure is implemented, there is always a possibility of imposing an improper structure on the data—this is the problem on which we focus our attention. The implications (i.e., distorting theory, poor generalizability, etc.) of adopting a partition that is merely a product of the clustering algorithm has been well-documented (see Milligan, 1996; Steinley, 2003).

Milligan (1996) provided a conceptual framework for conducting a cluster analysis that consisted of the following steps:

- Determine the observations to be clustered.
- Select the variables to be included in the clustering procedure.
- Decide if and how to standardize the variables chosen in the previous step.
- Select a measure of association (i.e., a distance measure).
- Select a clustering algorithm.
- Determine the number of clusters.
- Conduct cluster validation via interpretation, testing, and replication.

Conceptually, the process of determining whether a clustering is "good" falls under the rubric of cluster validation. Thus, a responsible implementation of a cluster analysis must include a diagnostic component that tests the partition to determine whether it adequately represents a cluster structure or is an arbitrary division of the data set. The three steps of the validation process, as conceptualized by Milligan (1996), contain a mixture of subjective and objective assessment. Specifically, the interpretation and testing (e.g., also termed *external validation* and usually consisting of a multivariate analysis of variance using the cluster solution as a grouping variable and a set of covariates that were not included in the analysis as dependent variables) are subjective in the sense that they both require domain-specific knowledge: the former because the interpretation of individual cluster solutions will depend on the theoretical significance of the variables included in the clustering proce-

This article was published Online First July 4, 2011.

Douglas Steinley, Department of Psychological Sciences, University of Missouri; Michael J. Brusco, Department of Marketing, Florida State University.

This article was partially supported by Army Research Institute Grant W91WAW-08-C-0034 and National Institute of Health Grant 1-K25-A017456-01 to Douglas Steinley.

Correspondence concerning this article should be addressed to Douglas Steinley, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65203. E-mail: steinleyd@missouri.edu

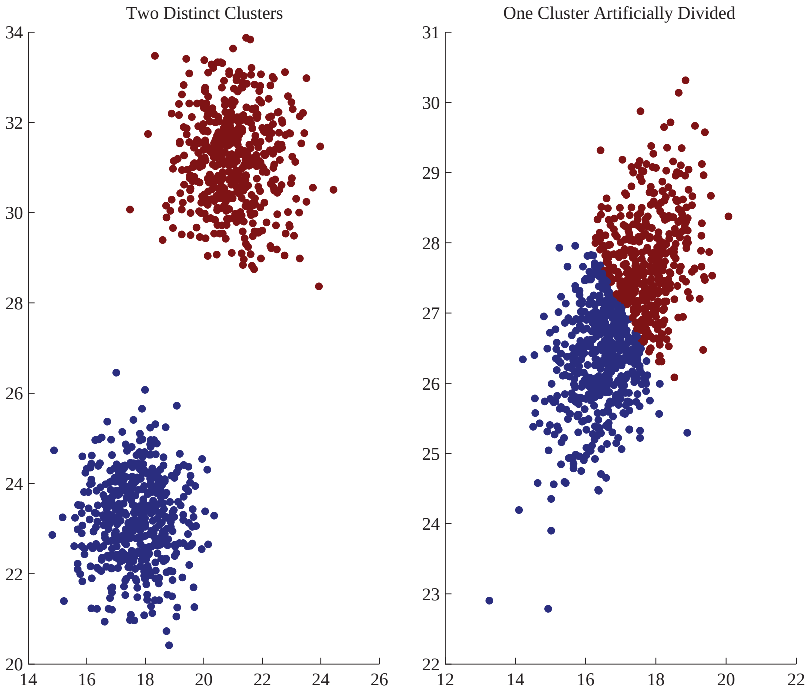

Figure 1. True versus fake clusters.

dures and the latter because domain experts will have to determine a set of theoretical justified covariates.

The final component of the cluster validation process is replication, and it is also considered to be a method for determining the validity and generalizability of the final cluster solution. In the present article, we discuss the most widely used type of replication, simply termed *replication analysis* (see Breckenridge, 1989, 2000), and show that replication analysis can succeed regardless of whether the other steps in the process were correct. Specifically, it is possible to get replicable solutions when an incorrect number of clusters are postulated. As such, we propose reformatting the steps outlined to combine choosing the number of clusters within the context of cluster validation. Concretely, this would appear as

- Cluster validation
  - ∘ Determine if there is more than one cluster.
  - ∘ If there is more than one cluster, determine the number of clusters.
  - ∘ Interpret, test, and validate.

So, the main change is mostly conceptual, in that choosing the correct number of clusters should be viewed as a part of the validation process. Additionally, and often overlooked, it should be verified that there is actually more than one cluster.

We address both the first and second steps of cluster validation within the context of a popular clustering procedure, *K*-means clustering (see Kogan, 2007; Steinley, 2006a), which, in a survey among experts in cluster analysis reported by Wu et al. (2008), has been named the most influential algorithm for unsupervised learning (i.e., cluster analysis). A testing procedure is developed that is based on the theoretical ratio between the within sum of squares and the total sum of squares one could expect when no cluster structure is present. The performance of the new techniques is compared with both replication analysis in the context of traditional clustering algorithms (see Breckenridge, 1989, 2000; McIntyre & Blashfield, 1980; Milligan, 1996) and the Bayesian information criterion (BIC) in the context of mixture models (see Banfield & Raftery, 1993).

After developing the initial procedure to screen for more than one cluster, we provide a technique for estimating the number of clusters in the data set. Although the seminal study of Milligan and Cooper (1985) examined nearly 30 procedures for determining the number of clusters, none of the procedures allowed for the possibility that only one cluster was present in the data. The proposed technique for determining the number of clusters is an extension of the procedure for assessing the presence of clusters. Once again, the performance of the procedure is compared with a range of other methods, including replication analysis and the top-performing method (e.g., the Calinski–Harabasz [CH] index), among others, identified by Milligan and Cooper (1985).

Cluster Validity and Choosing the Number of Clusters

In this section, we define the various methods that are compared for choosing the number of clusters. Generally, the methods can be divided into four types: traditional (e.g., formulaic procedures used in conjunction with classical clustering procedures), likelihood (e.g., BIC, Akaike information criterion [AIC]), replication analysis, and the proposed technique based on the lower bound of the sum-of-squares error in *K*-means clustering. Each technique is described in turn.

Traditional Measures

Milligan and Cooper (1985) conducted the first wide-scale investigation of several methods for determining the number of clusters, evaluating 30 different techniques. In this article, we examine some of the top methods identified in their article. Each of the following three measures assumes that there are at least two clusters.

CH index. The CH index (Calinski & Harabasz, 1974) is defined as

$$CH_K = \frac{SSB/(K-1)}{SSW/(N-K)}, \tag{2}$$

where N is the number of observations, K is the number of clusters, and SSB and SSW are the between- and within-cluster sums of squares, respectively. The correct number of clusters is chosen to be the value of K such that CH is a maximum.

Wilks's Λ. Wilks's Λ is traditionally used in multivariate analysis of variance to test whether groups are significantly different on a set of outcome variables. As an internal validation criterion for choosing the number of clusters in cluster analysis, Λ can be thought of as an indication of the separation of the clusters. Defined as

$$\Lambda_K = \frac{|\mathrm{W}|}{|\mathrm{T}|}, \tag{3}$$

$|\mathrm{W}|$ and $|\mathrm{T}|$ are the determinants of the within-cluster sum-of-squares cross-products and total sum-of-squares cross-products matrix, respectively. The number of clusters is chosen such that Λ is a minimum.

C index. The C index was proposed by Hubert and Levin (1976) and garnered support after a simulation study by Milligan (1981). The C index is computed as

$$C_K = \frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)}, \tag{4}$$

where d_w is the sum of within-cluster distances. The final value of K is chosen to correspond to the minimum value of C_K.

Mixture Modeling

Although the current article focuses on developing a procedure for determining the number of clusters in the context of K-means clustering, a related method of clustering observations is mixture model analysis (see McLachlan & Peel, 2000). Although a full-blown comparison between mixture model analysis and K-means clustering can be found in Steinley and Brusco (2011), we include the BIC in our comparison. The BIC was chosen over other criteria because it has consistently been recommended in the literature as the criterion of choice (see Martinez & Martinez, 2005, p. 185; McLachlan & Peel, 2000, p. 209). Using the most common assumption that the K clusters are multivariate normal, the population density of the clusters is provided by

$$f(\mathrm{x}; \alpha, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \sum_{k=1}^{K} \alpha_k \frac{\exp\left\{-\frac{1}{2}(\mathrm{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\mathrm{x}_i - \boldsymbol{\mu}_k)\right\}}{(2\pi)^{p/2} |\boldsymbol{\Sigma}_k|^{1/2}}, \tag{5}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the p-dimensional mean vector and the $p \times p$ covariance matrix for the kth cluster, respectively. The BIC is given as

$$BIC_c = 2L(\mathbf{X}, \hat{\theta}) - m\log(N),$$

where $\mathrm{L}(\mathbf{X}, \hat{\theta})$ is the log likelihood given the data ($\mathbf{X}$) and the estimated parameters $\hat{\theta}$, and m is the number of independent parameters.

Replication Analysis

McIntyre and Blashfield (1980) developed replication analysis for validating the solution provided by a cluster analysis. Replication analysis has seen frequent use as a method for validating cluster structures. For instance, Herzberg and Roth (2006) reported an application in personality research; Phinney, Dennis, and Gutierrez (2005) used replication analysis to identify profiles of Latino college freshmen; Zhang and Zhao (2000) used a type of replication analysis to develop consensus trees for gene clustering data; and Clatworthy, Buick, Hankins, Weinman, and Horne (2005) indicated that replication analysis is the most common method used to determine the stability of a cluster solution in psychology. Additionally, Henry, Tolan, and Gorman-Smith (2005) stated that replication analysis may soon be included in standard statistical software packages as a way to validate cluster analysis solutions. The procedure itself is a basic cross-validation procedure. The steps of the procedure are as follows:

1. Obtain two samples of data, S_1 and S_2 (usually the observed data set is randomly divided into two smaller data sets of the same size).

2. Cluster analyze S_1 and determine the set of K cluster centroids, denoted as $\bar{\mathrm{x}}_1$, from the obtained partition, P_1.

3. Determine the distance, D_2, between each observation in S_2 and $\bar{\mathrm{x}}_1$.

4. Assign data points in S_2 to the centroid in $\bar{\mathrm{x}}_1$ to which they are closest, obtaining the partition P_2.

5. Directly cluster analyze S_2 and obtain P_2^*.

6. Compare the P_2 and P_2^* by computing a measure of agreement. The most popular method for comparing two partitions is the adjusted Rand index (see Hubert & Arabie, 1985; Steinley, 2004), denoted as $\mathrm{ARI}(P_2, P_2^*)$, which has an upper bound of one when the two partitions are identical and a value of zero when they only agree at chance levels.

Step 1 is congruent with the common practice of obtaining a holdout sample to test the quality of a solution provided by a general analytic technique (see Cudeck & Browne, 1983, for cross-validation in the context of analyzing covariance structures; see Diana & Tommasi, 2002, as applied to principal component analysis; see Mosier, 1951, as applied to linear regression). Step 2 derives the reference cluster structure from the first half of the data, whereas Step 3 and Step 4 determine the distance from the second half of the data to the reference cluster and assign the observations to the cluster from the first sample to which they are closest. Step 5 determines the cluster structure from the second half of the data. The two partitions obtained from the holdout sample of the data (one partition derived from the means of the structure present in the first sample and one partition obtained by directly clustering the second sample) are compared in Step 6 to determine the degree to which the clusters are consistent across the two samples of data.

Choosing the number of clusters. Replication analysis has been advocated for choosing the number of clusters (see Milligan, 1996, for a review). A range of replication analyses are conducted for values of K ($K = 2, \ldots, K_{\text{max}}$), with the final value of K chosen to be the one that maximizes ARI(P_2, P_2^*). Breckenridge (1989, 2000) showed that the value obtained from replication analysis—that is, ARI(P_2, P_2^*)—was highly correlated with the true, underlying cluster structure. If we were to use the fact that the two measures are highly correlated, the method may be used to conclude that there is a true cluster structure.

However, it is crucial to realize that ARI(P_2, P_2^*) is unable to differentiate between $K = 1$ and $K = 2$. Extending replication analysis to the case of $K = 1$ is straightforward and is accomplished by adding a simple bootstrap procedure to estimate the variability of the point estimate ARI(P_2, P_2^*). Specifically, to determine the variability associated with the point estimate, Step 1 can be repeated several times (say, T times) where S_1 and S_2 are randomly constructed each time. For this example, the replication analysis was conducted 1,000 times using the K-means clustering procedure described in Steinley (2006b) and the different values for ARI(P_2, P_2^*) (abbreviated as ARI$_t$, $t = 1, \ldots, T$) observed on each replicate were collected in the vector $\mathbf{a} = [\text{ARI}_{(1)}, \ldots, \text{ARI}_{(T)}]$, where $\text{ARI}_{(1)} \leq \text{ARI}_{(2)} \leq \cdots \leq \text{ARI}_{(T)}$. In general, to conclude that there is no cluster structure, one would expect ARI(P_2, P_2^*) = 0, indicating no consistent structure inherent in the data set. Thus, if one were to adopt the standard significance level of $\alpha = .05$, then the test decision can be determined by comparing $\text{ARI}_{(.05 \times T)}$ with 0.[1] Specifically,

if $\text{ARI}_{(.05 \times T)} > 0$, conclude there is a cluster structure;

if $\text{ARI}_{(.05 \times T)} \leq 0$, conclude there is no definable cluster structure.

If one is to conclude that there is a cluster structure, 95% of the distribution of ARI(P_2, P_2^*) has to be greater than 0, providing strong evidence for a cluster structure. Then, after it is concluded that at least some type of cluster structure exists, the prior procedure for choosing the number of clusters can be implemented.

Validity versus consistency. On the surface, replication analysis appears to be testing the validity of a cluster structure actually present in the data; however, the procedure really is testing the ability of a given clustering routine to consistently find the same structure in the data. For instance, we generated 100 observations ($N = 100$) measured on four variables ($V = 4$), where each observation on each variable was distributed as a uniform distribution from zero to one. Next, the data were arbitrarily clustered into three groups, and we conducted the replication analysis above to determine whether there was a true cluster structure—the conclusion should be that there is no cluster structure because the data were generated uniformly from a four-dimensional box. In this instance, ARI(P_2, P_2^*) = .6905.

In this example, the estimated distribution is displayed in Figure 2 and the 95% cutoff is $\text{ARI}_{(50)} = .0896$, indicating that there is cluster structure, when, in fact, no cluster structure is present. Using this cutoff, we see that the replication analysis would actually indicate there is adequate recovery of an underlying cluster structure (perfect agreement was indicated five times). Thus, the observed agreement is likely due to the procedure being able to repeatedly partition the unstructured data in a consistent fashion in the absence of any type of true structure that is present in the data.

Lower Bound of SSE

An initial screening process is conducted prior to accepting that more than one cluster is present in the data. To develop the procedure, we look to the specifics of K-means clustering. In general, K-means clustering can be formulated as an alternating least squares problem that minimizes

$$SSE = \text{trace}[(\mathbf{X} - \mathbf{MC})'(\mathbf{X} - \mathbf{MC})], \qquad (6)$$

where $\mathbf{X} = \{x_{ij}\}$ is the $N \times V$ data matrix, $\mathbf{M} = \{m_{ik}\}$ is the $N \times K$ cluster membership matrix with $m_{ik} = 1$ if observation i is in the kth cluster and zero otherwise, and $\mathbf{C} = \{c_{kj}\}$ is the $K \times V$ matrix of cluster means with c_{kj} representing the mean of the kth cluster on the jth variable.[2] The K-means method alternates between estimating $\mathbf{M}$ and $\mathbf{C}$ until SSE can no longer be reduced.

Much like in analysis of variance, in K-means clustering, the variability of the data when divided into K clusters can be partitioned as

$$SST - SSB_K + SSE_K, \qquad (7)$$

where SST, SSB_K, and SSE_K (i.e., $SSE = SSW$), are the sum-of-squares total, between-clusters sum of squares, and within-cluster sum of squares (i.e., error) for a K-cluster solution, respectively (note that SST is not subscripted because SST is constant across all values of K). Furthermore, we offer the reminder that SSE_K is monotonically decreasing as K increases (i.e., $SSE_K \geq SSE_{K+1}$). Clearly, K-means clustering can be thought of as either minimizing the within-cluster variability or maximizing the between-cluster variability. Traditional procedures for determining the number of clusters when using classical clustering procedures, such as K-means clustering, assume that there are at least two clusters. However, the decision about $K = 1$ versus $K > 1$ can be evaluated by determining the properties of SSE and SST when $K = 1$, but the distribution is improperly divided into two. In other words, we could look at the relationship of SSE_2 (i.e., the within-cluster sum of squares for a two cluster solution) in relation to SST. To be as conservative as possible, the ratio

$$LBR = SSE_2/SST, \qquad (8)$$

should be less than the lower bound obtainable when there is only one distribution. Because SST is constant, the task turns to deter-

[1] It is possible, if not likely, that $.05 \times T$ will not be an integer value. In such instances, $.05 \times T$ is rounded to the nearest integer.

[2] The trace($\mathbf{X}$) = Σx_{ii}, or the sum of the diagonal elements of $\mathbf{X}$. Additionally, prime indicates the transpose of a matrix $\mathbf{X}$, such that $x_{ij} = x'_{ji}$.

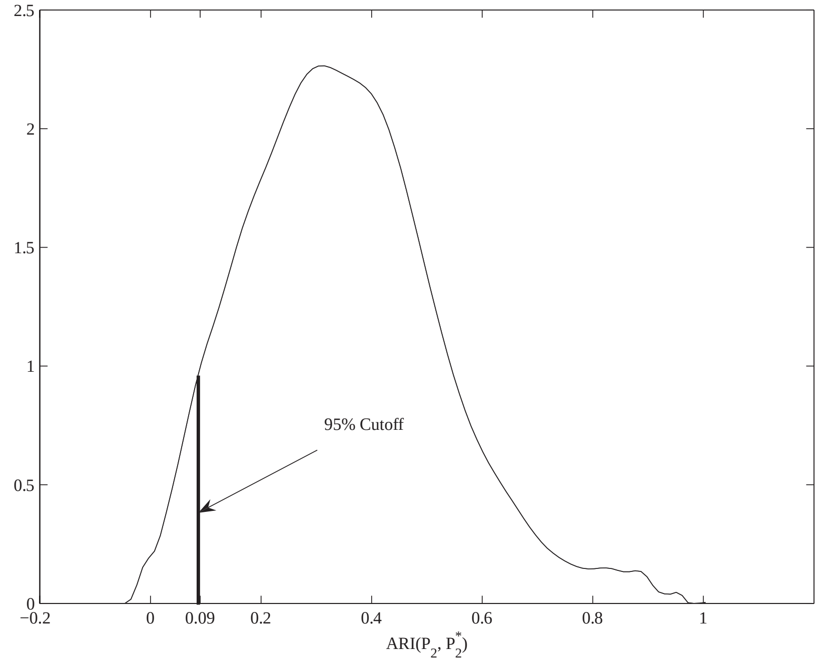

Figure 2. Estimated distribution and cutoff for the adjusted Rand index (ARI) in replication analysis.

mining the minimum of SSE_2 when dividing one distribution into two parts. Clearly, the minimum will occur when the distribution is divided exactly in half, with 50% of the observations being below the division and 50% of the observations being above the division (see Steinley & Brusco, 2008a, for a detailed discussion of this finding).

Although proofs are provided in Appendix A and Appendix B, the lower bound of the ratio (denoted as *LBR* in Equation 8) observable of SSE_2/SST for dividing a uniform (or multivariate uniform) distribution in half is $LBR^{(u)} = .25$ (see Appendix A); additionally, the analogous value for dividing a normal (or multivariate normal) distribution in half is $LBR^{(n)} = 1 - 2/\pi \approx .36$ (see Appendix B), where the superscripts of *u* and *n* indicate the uniform and normal distributions, respectively. Thus, to conclude that $K > 1$, the value of SSE_2/SST would have to be below the reference cutoff (either .25 or .36, depending on the choice of the reference distribution). The uniform distribution has been shown to be the distribution that is most likely to lead to the erroneous conclusion that clusters exist when they truly do not (see Steinley, 2007; Steinley & Brusco, 2008a), whereas the normal distribution is ubiquitous in the social and behavioral sciences. If the value in SSE_2/SST is less than the prescribed cutoff, then it has been determined that at least two clusters exist and the task of determining exactly how many clusters there are can be undertaken.

Choosing the Number of Clusters

Steinley (2007) derived the lower bound of the *SSE* for partitioning $\mathbf{X}_{N \times V}$ into *K* groups as

$$SSE_{\min}^{(K)} = \text{trace}(\mathbf{X}'\mathbf{X}) - \sum_{i=1}^{K} \lambda_i^{(\mathbf{XX}')} \tag{9}$$

where $\lambda^{(\mathbf{XX})} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ are the *N* eigenvalues of $\mathbf{XX}'$. The lower bound is trivial if $V \leq K$ (i.e., $SSE_{\min} = 0$); however, as long as $K < V$, the lower bound is meaningful.[3]

On the basis of this decomposition, a normalized index can be created that measures the closeness of the observed value of *SSE* to the minimum value of *SSE*, $SSE_{\min}^{(K)}$. Specifically, we define the lower bound technique (LBT) as

$$LBT_K = \frac{SSE_K - SSE_{\min}^{(K)}}{SST}. \tag{10}$$

When LBT_K is close to zero, the observed partitioning of the data is more compact and closer to its optimal value; conversely, as LBT_K approaches unity, the observed clustering is farther from its optimal value. The LBT_K is adapted for finding the number of clusters within the data set. Specifically, for a given data set, $\mathbf{X}_{N \times V}$, the data are partitioned into $K = 2, \ldots, K_{\max}$ (where $K_{\max} < V$) and LBT_K is computed for each partition, respectively. Then, the true number of clusters, K_{true}, is chosen such that

$$K_{\text{true}} = K^* \text{ such that } LBT_{K^*} = \min_K LBT_K. \tag{11}$$

[3] When $V \leq K$, the lower bound of *SSE* is zero because the number and sum of the nonzero eigenvalues of $\mathbf{XX}'$ and $\mathbf{X}'\mathbf{X}$ are equivalent.

Numerical Example of LBT: Fisher Iris Data

This section illustrates the proposed method by applying it to the classic Iris data set (Fisher, 1936). The iris data contain four variables (sepal length, sepal width, petal length, and petal width) measured in millimeters on 50 iris specimens from each of three species, *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Thus, the final data matrix for the iris data contains 150 observations measured on four variables. This data set provides a nice real-world example where the number of clusters (i.e., different species of iris) is generally assumed to be known prior to the analysis. Thus, when conducting a cluster analysis, we generally assume that we have 150 flowers that belong to three different species, and the goal of the current techniques will be to correctly identify that there are indeed three clusters.

First, to determine if there is more than cluster present in the data, the initial screening process is computed by partitioning the iris data into two clusters and comparing the ratio of SSE_2/SST with either $LBR^{(u)}$ or $LBR^{(n)}$. In this case, the value of $SSE_2/SST = 152.3/681.3 = .22$, and because .22 is less than both .36 (the cutoff for the normal distribution) and .25 (the cutoff for the uniform distribution), it is safe to assume that more than one cluster is present.

Table 1 provides the essential information needed to compute LBT for two and three clusters (recall that we only go up to $K = 3$ because the measure is meaningless when the number of clusters is equal to or more than the number of variables). Computing LBT for $K = 2$ results in

$$LBT_2 = \frac{152.3 - 15.5}{681.3} = .2008\ ;$$

likewise, for $K = 3$,

$$LBT_3 = \frac{78.9 - 3.6}{681.3} = .1105,$$

indicating that the value of K is three because $LBT_3 < LBT_2$.

Simulations

Simulation I: No Cluster Structure

It is necessary to conduct simulation studies to determine the Type I error rate of the proposed test. In other words, if there is no cluster structure present in the data, what is the probability that the LBR will make an incorrect decision and lead to the conclusion that there is a cluster structure when there is not (i.e., incorrectly conclude that $K > 1$ when in fact $K = 1$). To evaluate the LBR, we chose to generate data from multivariate standard normal distribution so the effects of different levels of correlation between the observed variables could be tested.

As Steinley (2006b) and Steinley and Brusco (2007) found that there is not a significant effect for the sample size on the recovery capabilities of K-means clustering, all data sets were generated to contain 200 observations. Additionally, to vary the dimensionality of the data, multivariate ellipses were generated in $V = 3, 4, \ldots,$ 15 dimensions (this type of generation satisfies the condition that the number of clusters has to be less than the number of variables). For example, to compute the lower bound for data sets in which SSE_2/SST is below the critical threshold, the minimum value of V has to be three. Fifteen is chosen as an upper limit on the number of variables to correspond with the parameters of the next simulation study. Finally, for each level of V, five population correlation structures followed the generation procedure outlined in Steinley and Brusco (2008b). The first was to use the identity matrix ($\mathbf{\Sigma} = \mathbf{I}$), while the other four correlation structures took the form of $\mathbf{\Sigma} = r \times \mathbf{J} + (1 - r)\mathbf{I}$, where $\mathbf{J}$ is a matrix of ones that is the appropriate size and the value of r is assumed to be .2, .4, .6, and .8, respectively. This resulted in a balanced design with $13 \times 5 = 65$ different conditions. For each condition, 2,500 replications were conducted, resulting in 162,500 different data sets. For all conditions, the LBR partitioned the data set into two clusters while the replication analysis (RA) used 1,000 bootstrapped samples to determine whether the partition arose from chance (to make the comparison comparable with the lower bound screening processes, the decision was always between one group or two groups). Finally, the BIC also was used to choose between either one or two groups, where no constraints were placed on the covariance structures in the mixture models.

For the LBR, none the 162,500 data sets ever exhibited a value of SSE_2/SST less than the cutoff of .36, indicating that the screening process would have never resulted in mistakenly partitioning a single distribution into multiple clusters. Figure 3 displays the kernel density plots of the ratio in SSE_2/SST by each level of correlation. As the correlation increases, the proposed procedure moves closer to concluding that there is more than one cluster; however, it never crosses the critical threshold. Finally, kernel density plots by the number of variables did not vary in a systematic or meaningful manner, indicating that the number of variables had little (if any) influence on the distribution of the SSE_2/SST values.

The Type I error rate for the BIC was .16, indicating that 16% of the time, it incorrectly indicated that more than one cluster was present. As seen in Table 2, the highest error rates for BIC are at low to moderate levels of correlation (for a detailed comparison of mixture models and K-means clustering, including the BIC and CH indices, see Steinley & Brusco, 2011). RA had the highest error rate (59%) and shows an increasing error rate as the correlation between variables increases. This is not too surprising as it becomes easier to consistently assign the same points to the same clusters when a distribution is elongated versus when it is spherical.

Table 1
Components of the Lower Bound Technique for the Iris Data

K	SSE	$SSE_{\min}$	SST
2	152.3	15.5	681.3
3	78.9	3.6	681.3

Note. SSE = within-cluster sum of squares; SST = sum-of-squares total.

Simulation II: Cluster Structure

Simulation parameters. Much like Milligan and Cooper's (1985) groundbreaking study, it is necessary to generate clusters that follow Cormack's (1971) definition—internally cohesive and

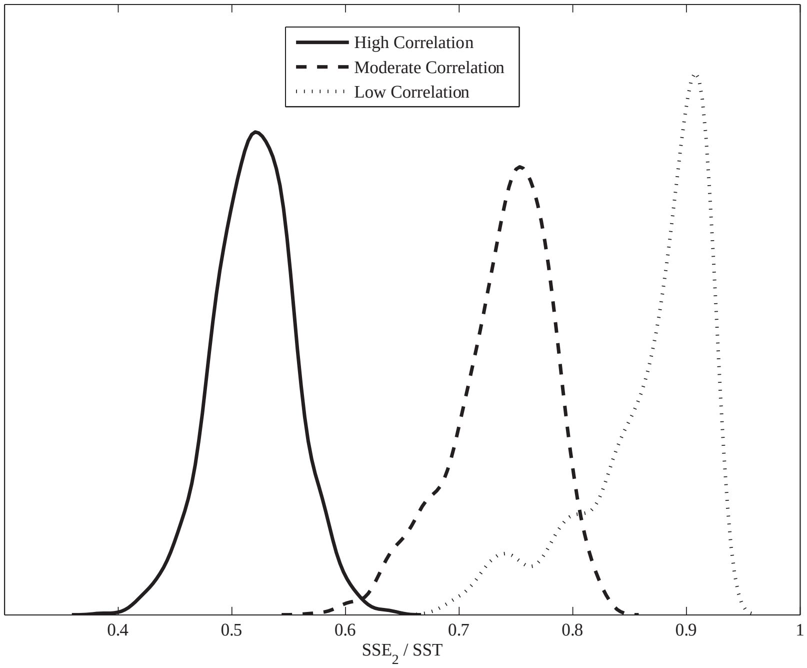

Figure 3. The effect of within-cluster correlation on the lower bound rule. SSE = within-cluster sum of squares; SST = sum-of-squares total.

externally isolated. An additional similarity to Milligan and Cooper's work is the manner in which clusters are generated. Specifically, the clusters are not allowed to overlap in the full-dimensional space; however, on any one marginal slice of the joint space, a set of two or more clusters may overlap.

Data sets were generated that corresponded to cluster structures consisting of $K = 2, 3, \ldots, 10$ clusters. Further echoing Milligan and Cooper (1985) and Steinley (2003, 2006b), the relative cluster density assumed three levels: (a) All clusters had the same number of observations, (b) one cluster had 60% of the observations while the remaining observations were evenly divided among the remaining clusters, and (c) one cluster had 10% of the observations while the remaining observations were evenly divided among the remaining clusters. Because of the relatively benign influence of the number of variables in the first simulation study, the number of variables (i.e., the dimensionality of the data set) was set equal to 15. Additionally, a third factor was the number of noise dimensions. Defined by Milligan (1985) as variables with no cluster structure, the noise dimensions were drawn from standard normal distributions. This factor also assumed three levels: no noise, one additional noise variable, and two additional noise variables. Although seemingly a small number of noise dimensions, Steinley and Brusco (2008a) showed that the addition of as little as one extraneous variable can have extremely detrimental effects on the ability of the *K*-means clustering procedure to perform effectively.

This design resulted in 81 different cells (nine levels for the number of clusters, three levels for the relative cluster density, and three levels for the number of noise dimensions). For each cell, 10 data sets were generated—many more than the standard three data sets per cell found in previous cluster analysis studies (Brusco, 2004; Milligan, 1980; Milligan & Cooper, 1985, 1988; Steinley, 2003).

Finally, following Milligan's (1985) generation method, each cluster was generated from truncated multivariate normal distributions. Furthermore, the generation process results in clusters that exhibit a mild level of pairwise correlations (see Appendix C for an example of one such correlation matrix). It has been shown that within-cluster correlations had much less of an influence on the performance of *K*-means clustering than did unequal within-cluster variances (see Steinley, 2006b; Steinley & Brusco, 2008a, 2011).

Table 2
Type I Error Rate for Determining One-Cluster Solutions

	Correlation					
Method	0	.20	.40	.60	.80	*M*
BIC	.03	.78	.13	.00	.00	.16
RA	.06	.68	.73	.75	.75	.59
LBR	.00	.00	.00	.00	.00	.00

Note. BIC = Bayesian information criterion; RA = replication analysis; LBR = lower bound of the ratio.

Table 3
Accuracy in Choosing the Number of Clusters

	True number of clusters									
Number chosen	2	3	4	5	6	7	8	9	10	Overall
Lower bound technique ($\Phi = 321$, $A = 91\%$)										
2 or fewer			3	0	0	0	2	0	0	5
1 too few		10	7	1	0	0	0	0	0	18
Correct	90	80	80	88	90	89	78	76	69	740
1 too many	0	0	0	1	0	1	5	3	3	13
2 or more	0	0	0	0	0	0	5	11	18	34
Calinski–Harabasz index ($\Phi = 412$, $A = 79\%$)										
2 or fewer			16	7	1	0	0	0	0	24
1 too few		45	4	1	0	0	0	0	0	50
Correct	90	45	70	82	76	70	70	70	69	642
1 too many	0	0	0	0	13	17	11	10	8	59
2 or more	0	0	0	0	0	3	9	10	13	35
C index ($\Phi = 9{,}143$, $A = 34\%$)										
2 or fewer			37	40	45	42	50	46	50	290
1 too few		33	24	24	6	8	1	5	0	101
Correct	45	25	29	26	29	30	29	29	30	272
1 too many	29	22	0	0	0	0	0	0	0	51
2 or more	16	10	0	0	0	0	0	0	0	26
Replication analysis ($\Phi = 15{,}591$, $A = 13\%$)										
2 or fewer			79	86	86	88	87	88	87	601
1 too few		82	9	3	3	1	3	1	1	103
Correct	90	8	2	1	1	1	0	1	2	106
1 too many	0	0	0	0	0	0	0	0	0	0
2 or more	0	0	0	0	0	0	0	0	0	810
Wilks's Λ ($\Phi = 34{,}560$, $A = 0\%$)										
2 or fewer			0	0	0	0	0	0	0	0
1 too few		0	0	0	0	0	0	0	0	0
Correct	0	0	0	0	0	0	0	0	0	0
1 too many	0	0	0	0	0	0	0	0	0	0
2 or more	90	90	90	90	90	90	90	90	90	810

Accordingly, we repeated the above design and then multiplied each within-cluster variable by $\sqrt{c}$, where $c \sim$ Uniform(1, 10), resulting in highly elongated, nonspherical clusters.

Analysis and results. All of the mentioned procedures were tested on every data set. There are two overall statistics of interest, accuracy and precision. *Accuracy* was defined as the proportion of times the procedure correctly predicted the true number of clusters. Here, *precision* is defined as

$$\Phi = \sum_{i=1}^{\Delta} (t_i - e_i)^2, \tag{12}$$

where i indexes each data set, Δ is the total number of data sets, t_i is the target number of clusters, and e_i is the estimated number of clusters.

Simulation IIa: Standard clusters. The LBT had the highest accuracy (91%; see Table 3), whereas the only other procedure with a remotely similar accuracy was the CH index (79%). The other three methods for choosing the number of clusters performed dismally, with RA being correct only 34% of the time, and the C index and Wilks's Λ being correct only 13% and 0% of the time, respectively. The poor performance of Wilks's Λ is not surprising because it performed the worst in Milligan and Cooper's (1985) study. It is unclear why the C index performs so poorly in this study while performing relatively well in the Milligan and Cooper (1985) study. The main difference is the type of algorithm used for the clustering: In the current article, we use *K*-means clustering, whereas the Milligan and Cooper study used hierarchical methods. Additionally, the C index was originally designed to be used in conjunction with hierarchical clustering, so the decrease in performance may be due to changing the fundamental way in which the clusters are formed.[4] Finally, the poor performance of RA can likely be attributed to the difficulty of consistently classifying

[4] The clusters are nested in hierarchical clustering, but that restriction is lifted in the context of *K*-means and other partitioning procedures. It is possible that the good performance of the C index in prior studies may be linked to the nested structure of the clusters.

observations into the same cluster when the number of clusters increases; thus, the ARI will be higher when there are a smaller number of observations, leading RA to choose the number of clusters to be two most of the time. Not only was the LBT most accurate, it was the most precise technique as well (31% more precise than the CH index and orders of magnitude more precise than the other three methods). By all accounts and measures, the LBT outperforms all of the other techniques for choosing the number of clusters. In the case of the CH index, the better performance of LBT is moderate; however, when compared with the other three techniques, the difference is vast.

Simulation IIb: Highly elongated clusters. Table 4 provides the results of the measures when the clusters were highly elongated (i.e., some axes of the clusters were potentially as much as 10 times longer than others). Because it has been shown that *K*-means clustering performs best when the data are more spherical, it was suspected that there would be a degradation of performance when examining highly nonspherical data. For the most part, this was observed for all techniques The LBT's accuracy dropped from 91% to 82%; however, it was still the most accurate and precise of all the techniques. In fact, the rank ordering of performance did not change when moving from slightly nonspherical clusters in the first part of this simulation to the highly elongated, nonspherical clusters in the second part.

Conclusion

LBRs of the within-cluster sum of squares to the total sum of squares were developed for determining whether clusters, according to Cormack (1971), exist in a data set. The LBT serves as a checkpoint for proceeding to assume that there are clusters and trying to determine how many clusters are present.

It was readily apparent from our analysis that the replication analysis supports cluster structure whether or not there is one present. This is because the data can be repeatedly and consistently partitioned into *K* sections whether or not the partitioning is consistent with a cluster structure. This finding lends empirical support to the cautionary statement by Krieger and Green (1999) regarding the use of an internal criterion to validate cluster solutions. Furthermore, the proposed test can be readily used in any

Table 4
Accuracy in Choosing the Number of Clusters: Highly Elongated Clusters

	True number of clusters									
Number chosen	2	3	4	5	6	7	8	9	10	Overall
Lower bound technique (Φ = 382, *A* = 82%)										
2 or fewer			7	3	5	3	3	0	1	22
1 too few		27	13	5	1	0	0	1	0	47
Correct	90	63	70	81	79	79	71	71	58	662
1 too many	0	0	0	1	2	6	7	5	13	34
2 or more	0	0	0	0	3	2	9	13	18	45
Calinski–Harabasz index (Φ = 399, *A* = 77%)										
2 or fewer	—	—	8	5	1	1	0	0	0	15
1 too few	—	51	11	2	0	0	0	0	0	64
Correct	83	39	71	81	77	74	70	69	62	626
1 too many	6	0	0	2	11	13	17	14	18	81
2 or more	1	0	0	0	1	2	3	7	10	24
C index (Φ = 8,738, *A* = 32%)										
2 or fewer	—	—	45	46	60	51	60	53	57	372
1 too few	—	32	18	16	3	10	1	7	7	94
Correct	50	29	26	28	27	29	29	30	26	274
1 too many	25	22	0	0	0	0	0	0	0	47
2 or more	15	7	1	0	0	0	0	0	0	23
Replication analysis (Φ = 14,821, *A* = 13%)										
2 or fewer	—	—	70	84	82	86	83	84	85	587
1 too few	—	81	15	5	4	2	3	4	2	116
Correct	77	8	5	1	4	2	3	2	3	105
1 too many	0	1	0	0	0	0	1	0	0	2
2 or more	0	0	0	0	0	0	0	0	0	0
Wilks's Λ (Φ = 34, 560, *A* = 0%)										
2 or fewer	—	—	0	0	0	0	0	0	0	0
1 too few	—	0	0	0	0	0	0	0	0	0
Correct	0	0	0	0	0	0	0	0	0	0
1 too many	0	0	0	0	0	0	0	0	0	0
2 or more	90	90	90	90	90	90	90	90	90	810

statistical software package that is able to compute eigenvalues of a square matrix.[5] Although only the *K*-means procedure is discussed herein, the LBT and associated techniques are appropriate to use in conjunction with any clustering procedure that is attempting to minimize *SSE* (e.g., Ward's hierarchical clustering).

A limitation of the simulations concerns the manner in which the clusters were generated. Simulation I generated data from the multivariate normal distribution, whereas Simulation II generated data from truncated normal distribution. The advantage of using multivariate normal distributions is the ease in which correlation between variables within each cluster can be manipulated. However, it is possible that, under other distributional assumptions, the procedures would perform differently. Steinley (2006a) indicated that *K*-means clustering performed best when the variables within clusters had equal variance; this was followed closely by the uniform distribution and then the triangular distribution. The worst condition examined was the normal distribution, where the variables within each cluster had different variances, performing about 18% worse. Given the robust performance of the LBT method in the presence of highly unequal variances, we do not expect that other distributions would differentially affect the LBT procedure as compared with the other methods for choosing the number of clusters.

The only apparent limitation to the proposed technique itself is the requirement that the number of variables exceed the number of clusters, leaving open the recommendation of what to do when this is not the case (i.e., the number of clusters exceeds the number of variables). In such a case, we would continue to recommend the CH index, as it performed second only to the LBT in the simulation studies. However, it is common that researchers usually have more variables than the number of clusters they are positing are present in the data set. For this approach to work, the researcher must choose variables on a theoretical basis of what is believed to contribute to the cluster structure (or use appropriate variable selection techniques); otherwise, irrelevant variables will be included and the natural clustering can be obscured (see Steinley & Brusco, 2008a, 2008b). Although this requires more thoroughness on the part of the investigator, the only alternative is to use a procedure that indicates a cluster structure is present on every data set. Prudence at the onset of a cluster analysis is clearly preferable to a coin flip at the end of a cluster analysis.

[5] The procedure is available as m files within the MatLab environment or as *R* code from Douglas Steinley on request.

References

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics, 49,* 803–821.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research, 24,* 147–162.

Breckenridge, J. N. (2000). Validating cluster analysis: Consistent replication and symmetry. *Multivariate Behavioral Research, 35,* 261–285.

Brusco, M. J. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods, 9,* 510–523.

Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3,* 1–27.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., & Horne, R. (2005). The use and reporting of cluster analysis in health psychology: A review. *British Journal of Health Psychology, 10,* 329–358.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society, Series A, 134,* 321–367.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Diana, G., & Tommasi, C. (2002). Cross-validation methods in principal component analysis: A comparison. *Statistical Methods & Applications, 11,* 71–82.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188

Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology, 19,* 121–132.

Herzberg, P. Y., & Roth, M. (2006). Beyond resilients, undercontrollers, and overcontrollers? An extension of personality prototype research. *European Journal of Personality, 20,* 5–28.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin, 83,* 1072–1080.

Kogan, J. (2007). *Introduction to clustering large and high-dimensional data.* New York, NY: Cambridge University Press.

Krieger, A. M., & Green, P. E. (1999). A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika, 64,* 341–353.

Martinez, W. L., & Martinez, A. R. (2005). *Exploratory data analysis with MATLAB.* Boca Raton, FL: Chapman & Hall/CRC.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research, 15,* 225–238.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York, NY: Wiley.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45,* 325–342.

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika, 46,* 187–199.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. *Psychometrika, 50,* 123–127.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analysis. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 341–375). River Edge, NJ: World Scientific.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50,* 159–179.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification, 5,* 181–204.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement, 11,* 5–11.

Phinney, J. S., Dennis, J. M., & Gutierrez, D. M. (2005). College orientation profiles of Latino students from low socioeconomic backgrounds: A cluster analytic approach. *Hispanic Journal of Behavioral Sciences, 72,* 387–408.

Steinley, D. (2003). Local optima in *K*-means clustering: What you don't know may hurt you. *Psychological Methods, 8,* 294–304.

Steinley, D. (2004). Properties of the Hubert–Arabie adjusted Rand index. *Psychological Methods, 9,* 386–396.

Steinley, D. (2006a). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59,* 1–34.

Steinley, D. (2006b). Profiling local optima in *K*-means clustering: Developing a diagnostic technique. *Psychological Methods, 11,* 178–192.

Steinley, D. (2007). Validating clusters with the lower bound for sum of squares error. *Psychometrika, 72,* 93–106.

Steinley, D., & Brusco, M. J. (2007). Initializing *K*-means batch clustering: A critical evaluation of several techniques. *Journal of Classification, 24,* 99–121.

Steinley, D., & Brusco, M. J. (2008a). A new variable weighting and selection procedure for *K*-means cluster analysis. *Multivariate Behavioral Research, 43,* 77–108.

Steinley, D., & Brusco, M. J. (2008b). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika, 73,* 125–144.

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: Recommendations and cautions. *Psychological Methods, 16,* 63–79. doi:10.1037/a0022673

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., . . . Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14,* 1–37.

Zhang, K., & Zhao, H. (2000). Assessing reliability of gene clusters from gene expression data. *Functional & Integrative Genomics, 1,* 156–173.

Appendix A

Proof for Uniform Distribution

Theorem: The minimum ratio of the within-cluster sum of squares to the corrected total sum of squares for a uniform distribution partitioned into two groups is 1/4.

Proof: To begin, we define the corrected total sum of squares for a univariate variable, *x*, to be $\Sigma_{i=1}^{N}(x_i - \bar{x})^2$. Likewise, the within-cluster sum of squares for the two clusters is provided by $\sum_{i\in \mathbf{C}_k}^{n_k}(x_i - \bar{x}_k)^2$ for $k = (1, 2)$; where C_k denotes the *k*th cluster, n_k denotes the number of objects in the *k*th cluster, $\bar{x}_k$ denotes the mean of the *k*th cluster, $x_1 \cup x_2 = x$, and $n_1 + n_2 = n$. Then, the quotient to be evaluated is

$$\frac{\sum_{i\in C1}(x_i - \bar{x}_1)^2 + \sum_{i\in C2}(x_i - \bar{x}_2)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{A1}$$

If each of the three terms in Equation A1 (e.g., the two terms in the numerator and the one term in the denominator) are multiplied and divided by their respective sample sizes, we arrive at a function of variances, given as

$$\frac{(n_1 - 1)\mathrm{Var}(C_1) + (n_2 - 1)\mathrm{Var}(C_2)}{(n-1)\mathrm{Var}(x)}. \tag{A2}$$

Given that the variance for the uniform distribution is $(b - a)^2/12$ (i.e., the length squared divided by 12), where *b* is the upper bound and *a* is the lower bound, and recognizing that each half of the uniform distribution will be half of the length of the entire variable, Equation A2 can be rewritten as

$$\frac{(n_1 - 1)\dfrac{\left(\dfrac{b-a}{2}\right)^2}{12} + (n_2 - 1)\dfrac{\left(\dfrac{b-a}{2}\right)^2}{12}}{(n-1)\dfrac{(b-a)^2}{12}} = \frac{(n-2)\dfrac{(b-a)^2}{48}}{(n-1)\dfrac{(b-a)^2}{12}} = \frac{(n-2)(b-a)^2 12}{(n-1)(b-a)^2 48} \tag{A3}$$

After the appropriate cancellations, this formula asymptotically converges to 1/4.

Furthermore, the generalization to a multivariate space with *V* variables is straightforward. The appropriate ratio becomes

$$\frac{(n-2)\left[\sum_{k=1}^{2}\sum_{i\in C_k}(x_i - \bar{x}_{1k})^2 + \sum_{k=1}^{2}\sum_{i\in C_k}(x_i - \bar{x}_{2k})^2 + \ldots + \sum_{k=1}^{2}\sum_{i\in C_k}(x_i - \bar{x}_{Vk})^2\right]}{(n-1)\left[\sum_{i=1}^{N}(x_i - \bar{x}_1)^2 + \sum_{i=1}^{N}(x_i - \bar{x}_2)^2 + \ldots + \sum_{i=1}^{N}(x_i - \bar{x}_V)^2\right]}, \tag{A4}$$

(*Appendices continue*)

and, from the univariate result, the minimum for the within sum of squares of the *v*th variable is

$$\frac{\sum_{k=1}^{2}\sum_{\in C_k}(x_i - \bar{x}_{vk})^2}{\sum_{i=1}^{N}(x_i - \bar{x}_v)^2} = \frac{1}{4} \Rightarrow \frac{(n-2)[1+1+\ldots+1]}{(n-1)[4+4+\ldots+4]} \Rightarrow \frac{(n-2)(V)(1)}{(n-1)(V)(4)}. \tag{A5}$$

As such, Equation A5 can be rewritten as $(n - 2)/(n - 1) \times 1/4$, which also asymptotically converges to 1/4.

Appendix B

Proof for Normal Distribution

Theorem: The minimum ratio of the within-cluster sum of squares to the corrected total sum of squares for a standard normal distribution partitioned into two groups is $1 - 2/\pi$.

Proof: Representing the standard normal distribution as *z*, the integral of the probability density function is given by

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-z^2/2}dz. \tag{B1}$$

As with the uniform distribution, the numerator of the ratio we are concerned with is minimized by splitting the original normal distribution in half, leaving

$$\frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} e^{-z^2/2}dz. \tag{B2}$$

Now the goal will be to find the mean and variance of Equation B2; however, it should be noted that Equation B2 must be multiplied by 2 to provide a proper probability density function that integrates to unity. Then, the formulas for the mean and variance are

$$\mathbf{E}(z) = \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} ze^{-z^2/2}dz \tag{B3}$$

and

$$\mathrm{Var}(z) = \frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} z^2e^{-z^2/2}dz - [\mathbf{E}(z)]^2, \tag{B4}$$

respectively. The integral in Equation B3 is solved by substituting $y = -\frac{1}{2}z^2$ and $dy = -zdz$, leading to

(*Appendices continue*)

$$-\frac{2}{\sqrt{2\pi}}\int_{0}^{\infty} e^{y}dy \Rightarrow -e^{y} = -\frac{2}{\sqrt{2\pi}}e^{-z^{2}/2}\Big|_{0}^{\infty} \Rightarrow 0 + \frac{2}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}. \tag{B5}$$

Focusing on the first term in Equation B4, it helps to recall the Weibull distribution, $W(\gamma, \beta) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^{\gamma}/\beta}$. Now it becomes clear that the first term in Equation B4 is the integral for finding the mean of a Weibull where $\gamma = \beta = 2$. Using the well-known formula for the mean of the Weibull, $\beta^{1/\gamma}\Gamma(1 + 1/\gamma)$, leads us to solve Equation B4 as

$$\frac{2}{2\pi}\sqrt{2}\Gamma\left(1 + \frac{1}{2}\right) - \left(\frac{2}{\sqrt{2\pi}}\right)^{2} = 1 - \frac{2}{\pi} \approx .3634. \tag{B5}$$

The extension of this result to a multivariate situation follows the exact logic as presented for the uniform distribution.

Appendix C

Example Generated Within-Cluster Correlation Matrix

The following correlation matrix, **R**, is an example of the relative degrees of within-cluster correlation observed. The maximum within-cluster pairwise correlation observed across all simulations was .63, while the minimum within-cluster pairwise correlation was −.57.

$$\mathbf{R} = \begin{bmatrix}
1.00 & 0.03 & -0.11 & 0.05 & 0.00 & 0.06 & -0.20 & -0.07 & 0.03 & 0.07 & 0.25 & 0.14 & 0.02 & -0.26 & 0.09 \\
0.03 & 1.00 & -0.09 & 0.18 & -0.01 & -0.42 & -0.17 & -0.04 & -0.09 & 0.00 & -0.13 & -0.01 & 0.20 & 0.05 & 0.17 \\
-0.11 & -0.09 & 1.00 & -0.07 & -0.13 & 0.09 & -0.09 & 0.07 & 0.03 & 0.01 & 0.25 & 0.16 & -0.07 & 0.12 & -0.19 \\
0.05 & 0.18 & -0.07 & 1.00 & -0.06 & -0.07 & -0.02 & 0.20 & 0.05 & 0.07 & -0.01 & 0.14 & 0.03 & -0.18 & 0.08 \\
0.00 & -0.01 & -0.13 & -0.06 & 1.00 & -0.15 & 0.06 & 0.22 & -0.09 & 0.03 & -0.20 & 0.06 & -0.11 & -0.17 & 0.11 \\
0.06 & -0.42 & 0.09 & -0.07 & -0.15 & 1.00 & -0.09 & -0.10 & 0.00 & -0.16 & 0.16 & -0.21 & 0.02 & 0.19 & 0.05 \\
-0.20 & -0.17 & -0.09 & -0.02 & 0.06 & -0.09 & 1.00 & -0.09 & 0.02 & -0.08 & -0.21 & 0.04 & -0.13 & -0.05 & -0.07 \\
-0.07 & -0.04 & 0.07 & 0.20 & 0.22 & -0.10 & -0.09 & 1.00 & 0.08 & -0.13 & -0.15 & 0.35 & -0.01 & -0.07 & -0.17 \\
0.03 & 0.09 & 0.03 & 0.05 & -0.09 & 0.00 & 0.02 & 0.08 & 1.00 & -0.10 & -0.25 & 0.08 & 0.06 & -0.10 & -0.14 \\
0.07 & 0.00 & 0.01 & 0.07 & 0.03 & -0.16 & -0.08 & -0.13 & -0.10 & 1.00 & 0.10 & -0.18 & -0.05 & -0.14 & -0.03 \\
0.25 & -0.13 & 0.25 & -0.01 & -0.20 & 0.16 & -0.21 & -0.15 & -0.25 & 0.10 & 1.00 & -0.13 & -0.04 & -0.05 & 0.02 \\
0.14 & -0.01 & 0.16 & 0.14 & 0.06 & -0.21 & 0.04 & 0.35 & 0.08 & -0.18 & -0.13 & 1.00 & -0.04 & -0.19 & -0.13 \\
0.02 & 0.20 & -0.07 & 0.03 & -0.11 & 0.02 & -0.13 & -0.01 & 0.06 & -0.05 & -0.04 & -0.04 & 1.00 & 0.07 & -0.07 \\
-0.26 & 0.05 & 0.12 & -0.18 & -0.17 & 0.19 & -0.05 & -0.07 & -0.10 & -0.14 & -0.05 & -0.19 & 0.07 & 1.00 & 0.23 \\
0.09 & 0.17 & -0.19 & 0.08 & 0.11 & 0.05 & -0.07 & -0.17 & -0.14 & -0.03 & 0.02 & -0.13 & -0.07 & 0.23 & 1.00
\end{bmatrix}$$

Received October 17, 2008
Revision received September 14, 2010
Accepted October 4, 2010 ■