Journal of Classification 22:221-250 (2005)
DOI: 10.1007/s00357-005-0015-6

OCLUS: An Analytic Method for Generating Clusters with Known Overlap

Douglas Steinley

University of Missouri, Columbia

Robert Henson

University of North Carolina, Greensboro

Abstract: The primary method for validating cluster analysis techniques is through Monte Carlo simulations that rely on generating data with known cluster structure (e.g., Milligan 1996). This paper defines two kinds of data generation mechanisms with cluster overlap, marginal and joint; current cluster generation methods are framed within these definitions. An algorithm generating overlapping clusters based on shared densities from several different multivariate distributions is proposed and shown to lead to an easily understandable notion of cluster overlap. Besides outlining the advantages of generating clusters within this framework, a discussion is given of how the proposed data generation technique can be used to augment research into current classification techniques such as finite mixture modeling, classification algorithm robustness, and latent profile analysis.

Keywords: Cluster generation; Overlapping clusters.

We would like to thank Lawrence Hubert for comments and suggestions on earlier versions of this paper. The first author was partially supported by Office of Naval Research Grant #000014-02-1-0877.

Author’s Address: Douglas Steinley, Department of Psychological Sciences, University of Missouri-Columbia, 210 McAlester Hall, Columbia, Missouri 65211, USA, e-mail: steinleyd@missouri.edu

1. Introduction

According to Milligan (1996), the validation of a clustering technique requires the generation of artificial data sets and testing via Monte Carlo simulation so the researcher has prior knowledge of the exact structure of the data. After generation and the subsequent clustering technique application, the resulting clusters are compared with the known structure. Frequently, different kinds of error are added to the known structure before the clustering algorithms are implemented, attempting to assess their resilience. Milligan (1996) notes, however, that results provided by these methods are only generalizable to the extent allowed by the data generation.

This paper will focus on the initial data generation step of validating clusters. For a detailed overview of the remaining steps of cluster validation, see Milligan (1996). A brief critique of current data generation techniques is provided, followed by a the proposal of new cluster generation method that has been implemented in several cluster validation studies (Steinley 2003; 2004). We end with a discussion of applicability for the proposed technique to several research areas.

2. Critique of Existing Techniques

1. Milligan (1985): Although a few simulation studies were already in the literature at the time, Milligan (1985) pioneered the extensive use of a Monte Carlo approach in cluster validation by developing an easily implemented algorithm, that used well-separated clusters from truncated (slightly) multivariate normal distributions. Standard normal error or outliers were added to the clusters to simulate measurement error and "messy" data, with both additions increasing the variance within clusters.

   Results:

   Standard normal error. Let x be normally distributed with mean μ and variance σ^2, and represented as $x \sim N(\mu, \sigma^2)$. If $e \sim N(0, 1)$ and x and e are independent, it follows from elementary statistics that $(x + e) \sim N(\mu, \sigma^2 + 1)$. Because $\sigma^2 + 1 > \sigma^2$, the variance has been increased by adding standard normal error, a result generalizable to the MVN distribution.

   Outliers. Let $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}'$ be an $n \times 1$ vector of observations. The x_i arranged in order from smallest to largest are the order statistics, denoted by $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ (Bickel & Doksum 2001; David 1981). For the order statistics, the mean is denoted by $\bar{x}$, the minimum by $x_{(1)}$, and the maximum by $x_{(n)}$. Let $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}'$ be an $m \times 1$ vector of outliers where each y_i is farther from $\bar{x}$ than any element of $\mathbf{x}$, so when $\mathbf{y}$ is added to $\mathbf{x}$ and the resulting order statistics are examined, there are three distinct possibilities:

(1) All m outliers are greater than $x_{(n)}$, resulting in the order statistics of the combined set being $x_{(1)}, \ldots, x_{(n)}, y_{(1)}, \ldots, y_{(m)}$;
(2) all m outliers are less than $x_{(1)}$, resulting in the combined order statistics $y_{(1)}, \ldots, y_{(m)}, x_{(1)}, \ldots, x_{(n)}$;
(3) b of the m outliers are less than $x_{(1)}$ and $m - b$ outliers are greater than $x_{(n)}$, resulting in the combined order statistics $y_{(1)}, \ldots, y_{(b)}, x_{(1)}, \ldots, x_{(n)}, y_{(b+1)}, \ldots, y_{(m)}$.

If $\mathbf{x}$ is viewed as a set of fixed data, robust estimation theory (Rousseeuw and Leroy 1987; Hampel, Ronchetti, Rousseeuw, and Stahel 1986) shows the breakdown point (the smallest number of arbitrary data points that needs to be added to the observed data to change the estimate) of the population variance is zero. Thus, if any additional data points (outliers) are added to $\mathbf{x}$ outside the range, $x_{(n)} - x_{(1)}$, the variance of the sample will increase. McIntyre and Blashfield (1980) note that increasing the variance of the clusters increases the degree of overlap between clusters. Moreover, the initial separation between clusters, the parameters of the MVN distributions forming the clusters, and the underlying distribution for the outliers are all random and internal (i.e., not user-specified) within Milligan's (1985) program. Because of these properties, the addition of outliers and error will have differing, unpredictable effects on the underlying clusters. Atlas and Overall (1994) note that Milligan's (1985) generation method creates an unrealistic degree of separation between clusters, and concrete statements about how much the clusters overlap cannot be made; however, this generation method is widely used in the literature (e.g., see Milligan 1980; Milligan, Soon, and Sokal 1983; Milligan and Cooper 1986; Milligan and Cooper 1988; Helsen and Green 1991; Balakrishnan, Cooper, Jacob, and Lewis 1994; Waller, Kaiser, Illian, and Manry 1998; Carmone, Kara, and Maxwell 1999; Brusco and Cradit 2001).

2. Kuiper and Fisher (1975): Kuiper and Fisher (1975) generated clusters from a sample of differing MVN populations with either identity or diagonal covariance matrices and different means; however, they never provide insight into the amount of overlap between the clusters used in the Monte Carlo study. One can infer that overlap changed as the means and covariance matrix changed, but it is impossible to quantify the degree of change.

3. Gold and Hoffman (1976): Gold and Hoffman (1976) sampled a primary population from a standard MVN distribution (i.e., with covariance matrix equal to the identity). They created sub-populations by adding random variables with differing expectations to data from the primary population, but failed to note the distribution of the added random variables, making it impossible to determine the degree of overlap between generated distributions.

4. Blashfield (1976): Blashfield (1976) structured the group covariance matrices to allow for correlations between the populations. In addition, after the data were sampled from the specified populations, measurement error sampled from a random uniform was added, causing the populations to be more mixed (i.e., overlap was increased). As in Milligan (1985), the degree of population overlap is impossible to determine because so many of the parameters are randomly chosen.

5. McIntyre and Blashfield (1980): McIntyre & Blashfield (1980) altered the overlap of the populations by increasing (or decreasing) the standard deviations of the various mixtures. But once again, no precise notion is available of how much the populations overlapped.

6. Price (1993): Price's method empirically creates overlapping clusters by "scooting" the means of the different distributions back and forth until the desired amount of overlap is achieved. Price (1993) looked at three levels of overlap between clusters (2%, 20%, and 40%). Because this method is iterative in nature, and depending on the number of clusters and the number of dimensions, it can be *very* time consuming. Also, because of the empirical nature of calculating the overlap, not all possible values of overlap are obtainable as a result of a mathematically impossibility due to sample size restrictions or confoundings from multiple dimensions. This severely limits the generalizability of Price's (1993) method.

7. Atlas and Overall (1994): Atlas and Overall rely on the manipulation of the intra-class correlation to control cluster overlap but note that the intraclass correlation "does not provide a perceptually meaningful description of population overlap" (p. 583).

8. Waller et al. (1999): These authors use what they refer to as indicator validities and compactness to control cluster overlap, and note that "when the indicator validities account for a large percentage of the variance, the clusters are well separated and easily discerned by visual inspection (p. 129)." However, visual inspection only allows comparisons of relativeness, such as, "These clusters overlap more than those clusters", as well

as restricting comparisons to three or fewer dimensions. Although the Waller et al. (1999) method is intended to generate Plasmodes (clusters that are based on real data) and is able to qualitatively relate how much clusters overlap, the overlap cannot be described in a quantitative manner, which is necessary when generating high-dimensional data sets.

3. OCLUS

Beauchaine and Beauchaine (2002) caution that although some success has been achieved, it is unrealistic to develop methods based solely on non-overlapping distributions. The objective of this paper is develop a procedure generating multivariate data from known distributions, and with a known amount of overlap between clusters. Within the literature, this has been a difficult task; for example, Atlas and Overall (1994) state:

> Although it is easy to generate artificial data representing random samples from underlying populations with different degrees of overlap in their multivariate distributions, it is not easy to display or otherwise communicate the extent of the population overlap in such a manner that a reader can readily appreciate its significance (p. 583).

The proposed data generation procedure, OCLUS (overlapping clusters), makes the concept of overlap understandable by approaching cluster overlap as the percentage of shared density between clusters. OCLUS was programmed in MATLAB 7 and exists as a collection of m-files (available by contacting the first author) and is able to capitalize on the strengths of many previous clustering procedures but avoids the weakness of not being able to assess cluster overlap.

3.1 Notation

The following notation is required in to describe OCLUS:

V : the number of dimensions (i.e., the number of variables);

K : the number of clusters desired, where C_k represents the k^{th} cluster, $1 \leq k \leq K$;

$\mathbf{n}$:= $\{n_1, \ldots, n_k\}$, the $k \times 1$ vector of the number of objects within each cluster where $N = \sum_{k=1}^{K} n_k$ is the total number of observations, For example, $\mathbf{n} = [50, 50, 50]'$ represents three clusters with 50 observations each;

$\boldsymbol{\eta}$:= $\{\eta_1, \ldots, \eta_k\}$ is a $k \times 1$ vector of mixing proportions, indicating the probability of observing an observation from the k^{th} cluster and providing an alternative method for sampling objects from specific clusters subject to the constraint $\sum_{k=1}^{K} \eta_k = 1$. For the example above using sample sizes, the corresponding vector of mixing proportions is $\eta = [0.3\bar{3}, 0.3\bar{3}, 0.3\bar{3}]$ and $N = 150$;

$\mathbf{\Sigma}_k$:= The desired covariance matrix for the variables in the k^{th} cluster;

$\mathbf{R}_k$:= The desired correlation matrix for the variables in the k^{th} cluster;

$p_{kk^*}^{(v)},\ p^{(v)},\ P$: $p_{kk^*}^{(v)}$ is the overlap between the two clusters C_k and C_{k^*} on dimension v, $p^{(v)}$ is the total amount of overlap on dimension v, and P is the average amount of overlap in $\mathbf{R}^{\mathbf{V}}$;

$\mathbf{D}$:= $\{d_{kk^*}\}$, where $1 \leq k, k^* \leq K$. Let the $K \times K$ identity matrix, $\mathbf{I_{K\times K}}$, represent K clusters with no overlap; in general, for $k = 1, \ldots, K$ and $k^* = 1, \ldots, K$, if $d_{(k,k^*)} = 1$ then the clusters C_k and C_{k^*} overlap, and if $d_{(k,k^*)} = 0$ then clusters C_k and C_{k^*} do not;

$\mathbf{X}$: the $N \times V$ data matrix;

x_{kvm} : the m^{th} observation on the v^{th} dimension from the k^{th} cluster;

$f_{kv}(x, \theta_{kv})$: the probability density function for the v^{th} dimension of the k^{th} cluster; θ_{kv} represents the vector of parameters relevant to $f_{kv}(x)$;

l_{kv}, u_{kv} : the lower and upper bounds, respectively, for x_{kv};

s : the separation parameter denoting how disjoint the *non-overlapping* clusters will be and represents the number of standard deviations the non-overlapping clusters should be from each other. A higher value indicates more separation between clusters;

$dist_v$: the different family of distributions from which clusters can be generated. The choice of distributions and their parameters are defined in Table 1;

z : for two clusters, C_k and C_{k^*}, z is the value such that $f_{kv}(z, \theta_{kv}) = f_{k^*v}(z, \theta_{k^*v})$, for positive values of $f_{kv}(z, \theta_{kv})$ and $f_{k^*v}(z, \theta_{k^*v})$.

3.2 OCLUS Algorithm

The OCLUS algorithm operates in the following manner:

1. Assumptions: all dimensions are independent and all clusters are independent.

Table 1. Distributions available in OCLUS

Distribution	Notation	Range of x	Parameter Definition
Uniform	$U(a,b)$	$a \leq x \leq b$	a : lower bound b : upper bound
Normal	$N(\mu,\sigma)$	$-\infty \leq x \leq \infty$	μ : mean σ : standard deviation
Gamma	$\gamma(\alpha,\beta)$	$0 \leq x \leq \infty$	α : shape parameter β : scale parameter
Triangular	$T(a,b,c)$	$a \leq x \leq b$	a : lower bound b : upper bound c : shape parameter, $a \leq c \leq b$

2. The user provides $\mathbf{D}$, P, $\mathbf{n}$ (or $\boldsymbol{\eta}$ and N), V, s, K, $\mathbf{\Sigma}_k$ (or $\mathbf{R}_k$) for each cluster, and $dist$. For $dist$, the distribution can be specified by choosing a different family of distributions for each dimension or specifying that all dimensions are generated from the same family of distributions.

3. A matrix, $\mathbf{O}$, denoting the order of clusters on each dimension is computed from $\mathbf{D}$.

4. Each row of $\mathbf{O}$ is randomized so the ordering is not the same on every dimension.

5. Compute θ_{kv} for dimension j, $j = 1$.

6. Let $j = j + 1$. Repeat step 5 until $j = V$ (once for each dimension).

7. Generate data from computed distributions. The computation of the parameters proceeds in a sequential fashion. First, the parameters for the first cluster is established and based on the specified overlap in $\mathbf{D}$ and by P the parameters for the second cluster are computed. Then, based on those parameters (and the overlap considerations), the parameters for the third cluster are computed. This process continues until the parameters for all K clusters have been computed. (*Note*: Most distributions we use can be generated from commands built into MATLAB. If not available, however, see Evans, Hastings, and Peacock 2000, for a guide to generating data from various distributions).

3.2.1 Assumptions

Krzanowski and Marriott (1994) note "…directly generating samples from an arbitrary high-dimensional joint distribution may not be possible" (p.

154). Assuming cluster and dimension independence allows the clusters to be "built" from the marginals. By taking the product of the marginals across the dimensions, the known joint distribution of each cluster can be formed, with the inter-cluster independence allowing for direct computation of overlap. The direct implications of the assumptions will be clearly seen in the section on computing the distributional overlap.

3.2.2 User-defined Options

All user-defined options are explained in the notation section; however, some restrictions are imposed on $\mathbf{D}$:

1. For any number of K, a given row of $\mathbf{D}$ cannot indicate cluster overlap between more than three clusters.
2. The maximum number of rows that can indicate overlap with three clusters is $K-2$.
3. The maximum sum of the off-diagonals of $\mathbf{D}$ is $2K-2$.

These constraints are arrived at by considering a set of clusters on in a unidimensional setting. Assuming the clusters differ in terms of their means, there will always be two clusters on opposite ends of the continuum that have one neighboring cluster; whereas, the clusters between the two extreme clusters will have two neighboring clusters. For clarification, $\mathbf{D}$ will be further illustrated through an example design matrix, $\hat{\mathbf{D}}$. $\hat{\mathbf{D}}$ is a symmetric design matrix indicating the number of clusters and their overlap. For example,

$$\hat{\mathbf{D}} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

indicates that cluster 1 overlaps with clusters 2 and 3, and clusters 4 and 5 overlap. On inspection, $\hat{\mathbf{D}}$ follows all three restrictions. In $\hat{\mathbf{D}}$ there are two sets, S_1 and S_2, of overlapping clusters where $S_1 = \{k_1, k_2, k_3\}$ and $S_2 = \{k_4, k_5\}$. S_1 and S_2 are separated by the user defined value s.

3.2.3 User-defined Value s

The value of s determines the separation of clusters on each dimensions. For distributions with bounded domains (uniform and triangular) on a given dimension, the default separation between non-overlapping clusters is just the value that ensures the lower-bound of one cluster does not overlap with the upper-bound of a neighboring cluster (or vice-versa). For distributions with unbounded domains (normal and gamma), the default separation between non-overlapping clusters is just the distance between the means that ensures that there is less than a 0.01 probability of two clusters overlapping. The user-defined value of s is an additional value that is added to the default values to increase the degree of separation. Thus, for two non-overlapping discrete clusters, if $s = 0$, the clusters won't overlap but they will "bump" up against each other in V dimensional space; however, as the value of s increases "empty" space will be created between the clusters (see Figures 2 and 4 in the examples section for an illustration of empty space).

3.2.4 Computing and Randomizing $\mathbf{O}$

From $\mathbf{D}$, OCLUS calculates the matrix $\mathbf{O}$ by locating the diagonal blocks in $\mathbf{D}$ to determine which clusters are always going to be overlapping with each other and which clusters will never overlap with each other (i.e., identifying the subsets S_1 and S_2 from above). Thus, defining the order clusters are generated for each dimension. For $\hat{\mathbf{D}}$, let $V = 3$ and the initial computation of $\hat{\mathbf{O}}$ be

$$\hat{\mathbf{O}} = \begin{bmatrix} 2 & 1 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \end{bmatrix}.$$

After $\hat{\mathbf{O}}$ is computed, it is randomized within row so cluster generation will be random on each dimension, thereby allowing different clusters to have different relative magnitudes for values across the dimensions. For example, it might be that cluster two exhibits the lowest values on the first dimension while cluster five exhibits the lowest values on the third dimension.

Furthermore, the randomization scheme is quite simple. First, a subset of variables is selected at random without replacement. Then, knowing which clusters must overlap in the variable subset (from $\mathbf{D}$), the order of the clusters is randomly chosen to be the original order within the subset provided in the initial computation of $\mathbf{O}$ or the reverse order. One possible randomization of the example is

$$\hat{\mathbf{O}} = \begin{bmatrix} 2 & 1 & 3 & 4 & 5 \\ 3 & 1 & 2 & 5 & 4 \\ 5 & 4 & 2 & 1 & 3 \end{bmatrix},$$

where on the first dimensions OCLUS would generate the clusters in the order specified by the first row of $\hat{\mathbf{O}}$; the order of generation on the second and third dimensions would be determined by the respective rows of $\hat{\mathbf{O}}$. This randomization allows for numerous multidimensional configurations to arise from the same underlying structure, a feature found desirable and a key component for generating random and clustered data (Milligan 1996; Waller et al. 1999).

3.2.5 Determining Overlap Among Different Clusters

To find overlap among the different clusters, the joint distribution of each dimension is determined by simple transformations on the joint distribution of the clusters. By the independence-of-dimensions assumption, for $f_{kv}(x, \theta_{kv})$ and $l_{kv} \leq x_{kvm} \leq u_{kv}$, the joint distribution of cluster C_k is

$$f_{k1}(x_{k1}, \theta_{k1}) \ldots f_{kV}(x, \theta_{kV}). \tag{1}$$

The expression in (1) can be rewritten as

$$\prod_{v=1}^{V} f_{kv}(x, \theta_{kv}). \tag{2}$$

By determining (2) for each cluster, the distribution of each cluster can be written as the matrix (each row represents a cluster)

$$\mathbf{dist_K} = \begin{pmatrix} \prod_{v=1}^{V} f_{1v}(x, \theta_{1v}) \\ \prod_{v=1}^{V} f_{2v}(x, \theta_{2v}) \\ \vdots \\ \prod_{v=1}^{V} f_{Kv}(x, \theta_{Kv}) \end{pmatrix}.$$

Now, the overlap component can be calculated: the overlap between two clusters, C_k and C_{k^*}, on dimension v is

$$\begin{aligned} p_{kk^*}^{(v)} &= \min\Bigg[\Big(\int_{l_{k^*v}}^{z} f_{k^*v}(x, \theta_{k^*v})dx + \int_{z}^{u_{kv}} f_{kv}(x, \theta_{kv})dx\Big), \\ &\quad \Big(\int_{l_{kv}}^{z} f_{kv}(x, \theta_{kv})dx + \int_{z}^{u_{k^*v}} f_{k^*v}(x, \theta_{k^*v})dx\Big)\Bigg], \end{aligned} \tag{3}$$

given positive values of

$$f_{k^*v}(z, \theta_{k^*v}) \text{ and } f_{kv}(z, \theta_{kv}); \tag{4}$$

z must exist and (4) must be satisfied for (3) to hold. If the values in (4) are zero, then C_k and C_{k^*} do not overlap and the function in (3) will also equal zero. If z does exist and (4) does not hold, the two clusters will overlap but the overlap will not equal $p_{kk^*}^{(v)}$ (i.e., (3) is defined, but the desired value of $p_{kk^*}^{(v)}$ is not achieved). When (3) is defined and (4) is true, the two clusters will overlap by the desired amount and data from the respective distributions can be generated. In deciding how to generate overlapping clusters, two types of overlap called marginal and joint are considered.

Marginal Overlap. Marginal overlap is defined by separately establishing the overlap for all V marginals. Thus, the goal of the data generation procedure is to establish either an equal and fixed amount of overlap for all V dimensions or to establish a different (but fixed) amount of overlap for each.

Joint Overlap. Joint overlap between two clusters, p^{kk^*}, is defined by establishing an overall and fixed amount of overlap for each of the V margins. By (3) and the independence-of-dimensions assumption, the joint overlap is computed by the product of the marginal overlaps:

$$p^{kk^*} = \prod_{v=1}^{V} p_{kk^*}^{(v)} \ . \tag{5}$$

Results. The definitions of marginal overlap and joint overlap directly lead to two asymptotic results.

Result 1. If the amount of overlap for each marginal is fixed, joint overlap converges to zero as the number of dimensions increases.

Proof. Each cluster, C_k, exists in $\mathbf{R^V}$. For every v, $p_{kk^*}^{(v)}$ can be calculated $K-1$ times (the maximum number of overlapping regions imposed by the restrictions on $\mathbf{D}$). Recalling that for each $p_{kk^*}^{(v)}$, if z does not exist $p_{kk^*}^{(v)}$ is zero, and the marginal overlap for dimension v is

$$p^{(v)} = \sum_{C_k \neq C_{k^*}} p_{kk^*}^{(v)}/(K-1) \quad . \tag{6}$$

(6) is computed V times to calculate the marginal overlap for each dimension. Given $0 \leq p^{(v)} < 1$ for all $p^{(v)}$, by (2), as $V \rightarrow \infty$, the joint overlap is

$$P = \prod_{v=1}^{\infty} p^{(v)} = 0 \quad , \tag{7}$$

due to an infinite product of fractions less than unity. Thus, as the number of dimensions increases and the joint overlap converges to zero, the clusters should become more discernable. Result 1 is a strong indication that Milligan's (1985) method, which shows increased cluster recovery as the number of variables increase (Milligan 1980; 1996), manipulates marginal overlap by adding error and outliers to the data.

> **Result 2**. If the amount of joint overlap is fixed, overlap for each marginal distribution converges to 1 as the number of dimensions increases.
>
> *Proof*. For simplification, assume that the $p^{(v)}$ are equal for all v. To obtain a total overlap of P between clusters C_k and C_{k^*}, the overlap on each dimension v must be $P^{\frac{1}{V}}$ (since by (2), joint overlap is the product of the dimensions, $(P^{\frac{1}{V}})^V = P$). Then,
>
> $$\lim_{V\to\infty} P^{\frac{1}{V}} = 1 \quad . \tag{8}$$

Thus, for small or moderate joint overlap in a highly dimensional space, the marginal distributions will have a very high degree of overlap. Joint overlap should be considered and studied by cluster analysts because it indicates that two clusters actually occupy the same region of $\mathbf{R^V}$ space.

3.2.6 Distributions

This section will show how to determine, when P is given, which specific distributions to use from a family of distributions. The derivations will only provide the analytical results for two clusters on one dimension. Similar results for more than two clusters and one dimension would require a large amount of space. Nonetheless, these results are easily derived for multiple clusters overlapping within a data set across multiple dimensions by using the results in (1)—(4). Regardless, OCLUS implements derivations for any value of K and V for each distribution. Because all examples illustrate the generation process for one dimension, the dimensionality superscript will be dropped. (*Note*: To calculate marginal overlap, set $V = 1$). Additionally, the complete derivation is only provided for the normal with equal variances; whereas, the results for the other distributions are provided but the derivation is omitted.

The normal–equal variances. Letting $x_1 \sim N(\mu_1, \sigma^2)$, $x_2 \sim N(\mu_2, \sigma^2)$, where μ_1, σ, and P are known, μ_2 unknown, and Φ is the cumulative distribution function of the standard normal distribution, we obtain (Lehman & Casella, 1998, p. 93)

$$P(x \leq u) = \Phi(\frac{u-\mu}{\sigma}), \tag{9}$$

where $x \sim N(\mu, \sigma^2)$. The integral evaluated is

$$\begin{aligned} &\int_{-\infty}^{z} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\{\frac{-(x_2-\mu_2)^2}{2\sigma^2}\} dx_2 \\ + &\int_{z}^{\infty} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\{\frac{-(x_1-\mu_1)^2}{2\sigma^2}\} dx_1 = P^{\frac{1}{V}} \quad . \end{aligned} \tag{10}$$

Solving for μ_2,

$$\begin{aligned}\int_{-\infty}^{z} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\exp\{\frac{-(x_2-\mu_2)^2}{2\sigma^2}\}dx_2 & \\ +\int_{z}^{\infty} \frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\exp\{\frac{-(x_1-\mu_1)^2}{2\sigma^2}\}dx_1 &= P^{\frac{1}{V}} \Rightarrow \\ P(x_2 \leq z) + 1 - P(x_1 \leq z) &= P^{\frac{1}{V}} \Rightarrow\end{aligned}$$

invoking (9) results in two equations for z,

$$\begin{aligned}z &= \mu_1 + \sigma\Phi^{-1}(1-\frac{P^{\frac{1}{V}}}{2}) \\ z &= \mu_2 - \sigma\Phi^{-1}(1-\frac{P^{\frac{1}{V}}}{2}) \Rightarrow\end{aligned}$$

solving these gives

$$\begin{aligned}\mu_1 + \sigma\Phi^{-1}(1-\frac{P^{\frac{1}{V}}}{2}) &= \mu_2 - \sigma\Phi^{-1}(1-\frac{P^{\frac{1}{V}}}{2}) \Rightarrow \\ \mu_2 &= \mu_1 + 2\sigma(\Phi^{-1}(1-\frac{P^{\frac{1}{V}}}{2})) \qquad (11)\end{aligned}$$

OCLUS generates overlapping clusters from this family of distributions by the following steps:

1. Choose μ_1 from a $U(0, 10)$ distribution.
2. Set the variances equal to one (or choose randomly and set equal).
3. Find z by using a “built–in” cumulative distribution function.
4. Solve for μ_2.

For example, let $\mu_1 = 0$, $P = .05$, and $V = 1$. The first distribution is known to be $N(0, 1)$ and $z = \Phi^{-1}(1 - \frac{.05}{2}) = 1.96$. Thus, the second distribution has to have a mean of $2(1.96) = 3.92$. By generating data for the first cluster from a distribution of $N(0, 1)$ and data for the second cluster from $N(3.92, 1)$, the overlap between the two clusters will be $.05$. Figure 1 plots the pdfs of these latter two distributions.

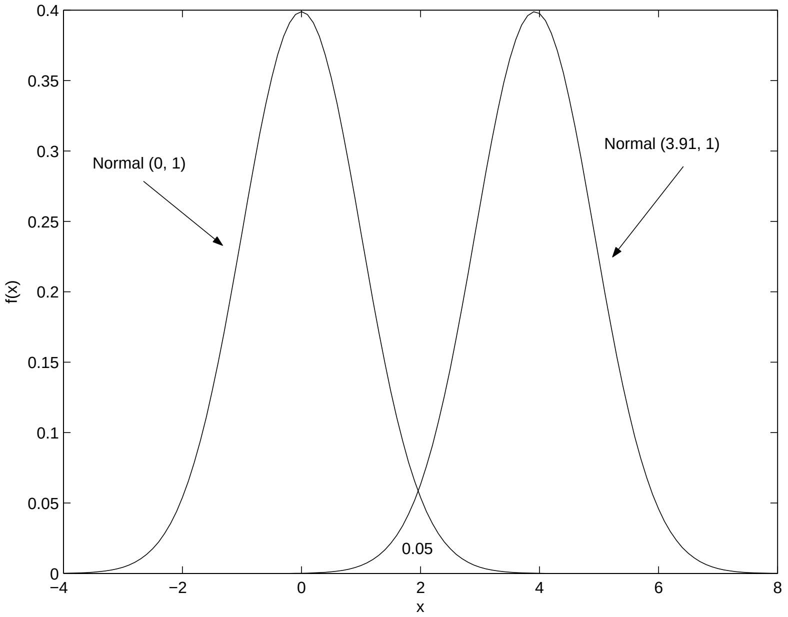

Figure 1. Example of two overlapping normal distributions with equal variance

The uniform. Let $x_1 \sim U(a_1, b_1)$, $x_2 \sim U(a_2, b_2)$, where $b_1 \leq b_2$ (refer to Table 1 for an explanation of the parameters). For dimension v, let a_1, b_1, and P be known while a_2 and b_2 are unknown. OCLUS generates overlapping clusters from the uniform distribution by the following steps:

1. Choose a_1 from a $U(0, L)$ distribution.
2. Set $b_1 = a_1 + L$.
3. Solve $a_2 = b_1 - L(P^{\frac{1}{V}})$.
4. Set $b_2 = a_2 + L$.

where L is the length (i.e., a function of the variability) of the uniform distributions on the v^{th} dimension. If the two clusters are generated from $U(a_1, b_1)$ and $U(a_2, b_2)$, respectively, then they will have an overlap of $P^{\frac{1}{V}}$ on dimension v. The same procedure can be repeated for any number of dimensions and any two distributions.

The normal–unequal variances. Let $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$ (refer to Table 1 for an explanation of the parameters). Let p_1 and p_2 equal $Prob[x_1 > z]$ and $Prob[x_2 < z]$, respectively, $p_1 + p_2 = p^{(v)}$, and $\mu_1 < \mu_2$. (Note: If $p_1 = p_2$, then $\sigma_1 = \sigma_2$). In addition, let μ_2, σ_2, P, and p_2 be known. Two restrictions are required so $f_1(x_1)$ and $f_2(x_2)$ intersect only once: $\mu_2 > \mu_1$; $\sigma_2 < \sigma_1$. The integral of interest is

$$\int_{-\infty}^{z} \frac{1}{\sigma_2(2\pi)^{\frac{1}{2}}}\text{exp}\{\frac{(x_2-\mu_2)^2}{-2\sigma_2^2}\}dx_2+\int_{z}^{\infty} \frac{1}{\sigma_1(2\pi)^{\frac{1}{2}}}\text{exp}\{\frac{(x_1-\mu_1)^2}{-2\sigma_1^2}\}dx_1 = P^{\frac{1}{V}} \quad , \tag{12}$$

and z can be calculated by

$$z = \mu_2 - \sigma_2\Phi^{-1}(1-p_2) \quad , \tag{13}$$

and, in turn, use (13) to obtain

$$f_2(z,\mu_2,\sigma_2) = \frac{1}{\sigma_2(2\pi)^{\frac{1}{2}}}\text{exp}\{\frac{(z-\mu_2)^2}{-2\sigma_2^2}\} \quad . \tag{14}$$

OCLUS generates overlapping clusters from normal distributions with unequal variances by the following steps:

1. Choose μ_2 from a $U(0,20)$ distribution.
2. Choose σ_2 from a $U(1,5)$ distribution.
3. Set $\sigma_1 = \frac{\text{exp}\{\frac{(\Phi^{-1}(1-p_1))^2}{-2}\}}{f_2(z,\mu_2,\sigma_2)(2\pi)^{\frac{1}{2}}}$
4. Set $\mu_1 = z - \frac{\text{exp}\{\frac{(\Phi^{-1}(1-p_1))^2}{-2}\}}{f_2(z,\mu_2,\sigma_2)(2\pi)^{\frac{1}{2}}}\Phi^{-1}(1-p_1)$

The gamma. Let $x_1 \sim \gamma(\alpha_1,\beta_1)$, $x_2 \sim \gamma(\alpha_2,\beta_2)$ (refer to Table 1 for an explanation of the parameters). z (the point where $p_1 = Prob[x_1 > z]$ and $p_2 = Prob[x_2 < z]$), α_1, and β_1 are known and β_2 is unknown. The pdf of the gamma is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\text{exp}\{\frac{-x}{\beta}\}, \tag{15}$$

where

$$\Gamma(\alpha) = \int_0^\infty (\frac{x}{\beta})^{\alpha-1}\text{exp}\{\frac{-x}{\beta}\}(\frac{1}{\beta})dx \tag{16}$$

reducing to

$$\Gamma(\alpha) = (\alpha-1)! \tag{17}$$

when α is an integer (Hogg & Craig, 1995). The integral of interest is

$$\int_0^z \frac{1}{\Gamma(\alpha_2)\beta_2^{\alpha_2}}x_2^{\alpha_2-1}\text{exp}\{\frac{-x_2}{\beta_2}\} + \int_z^\infty \frac{1}{\Gamma(\alpha_1)\beta_1^{\alpha_1}}x_1^{\alpha_1-1}\text{exp}\{\frac{-x_1}{\beta_1}\} = P^{\frac{1}{V}} \quad . \tag{18}$$

Integrating the two terms on the left-hand-side of (18) requires integration by parts $\alpha - 2$ and $\alpha - 1$ times, respectively. As the normal with unequal variance, when z, p_1, and p_2 are known, there is a unique combination of α_2 and β_2 that will lead to the overall desired level of P. Calculate z by

$$z = \gamma^{-1}(1 - p_1, \alpha_1, \beta_1) \quad . \tag{19}$$

Use (15) and (19) to obtain,

$$f_1(z, \alpha_1, \beta_1) = \frac{1}{\Gamma(\alpha_1)\beta_1^{\alpha_1}} z^{\alpha_1 - 1} \text{exp}\{\frac{-z}{\beta_1}\} \quad . \tag{20}$$

Setting (20) equal to the pdf of x_2, yields two equations with two unknowns, but solving this system of equations requires a search because integrating the second term on the left hand side of (18) results in a continued fraction (Weisstein, 2003). To find the appropriate α_2 and β_2, a grid search method (GAMSEARCH) is used, creating a vector, $\vec{\beta_2}$, of possible β_2's ranging from 1 to 10 in steps of 0.1, and a vector, $\vec{\alpha_2}$, of α_2's ranging from $\beta_1(\alpha_1 - 1)$ to $\beta_1(\alpha_1 - 1) + 10$. All pairs of values from $\vec{\beta_2}$ and $\vec{\alpha_2}$ are evaluated, and the unique solution are those values satisfying (3) and (4). Through empirical trial, the aforementioned range is usually suitable for finding the unique solution, but can be widened if (3) and (4) are unsatisfied in the initial search.

The exponential and chi-square distributions. The exponential and chi-square distributions are each special cases of the gamma. The exponential is a $\gamma(1, \beta)$ and the chi-square distribution is a $\gamma(\alpha, 2)$ (where α is the degrees of freedom of the chi-square) (Evans *et al.*, 2000). For all exponential distributions, α is fixed at 1; for all chi-square distributions, β is fixed at 2. Under these conditions, (3) may be obtained but (4) violated, causing for the generation of $K > 2$ populations, multiple points of intersection resulting in a convolution of the desired amount of overlap. Thus, OCLUS does not generate data from these two distributions.

The triangular distribution. The pdf of the triangular distribution (Evans *et al.* 2000, pp. 187–188) is

$$\begin{aligned} f(x) &= \frac{2(x-a)}{[(b-a)(c-a)]} \text{ if } a \leq x < c \,; \\ &= \frac{2(b-x)}{[(b-a)(b-c)]} \text{ if } c \leq x < b. \end{aligned} \tag{21}$$

Let $x_1 \sim T(a_1, b_1, c_1)$, $x_2 \sim T(a_2, b_2, c_2)$ (refer to Table 1 for explanation of parameters), where a_1, b_1, and c_1 are known. Let z be the point where $P[x_1 >$

$z]$ and $P[x_2 < z]$ are known and equal to p_1 and p_2. The integral evaluated is

$$\int_z^{b_1} \frac{2(b_1 - x)}{(b_1 - a_1)(b_1 - c_1)} dx + \int_{a_2}^{z} \frac{2(x - a_2)}{(b_2 - a_2)(c_2 - a_2)} dx = P^{\frac{1}{V}}. \tag{22}$$

Choosing p_1 and finding z by solving the first term on the left-hand-side of (22) and applying the quadratic formula, gives

$$z = \frac{2b_1 \pm (4b_1^2 - 4(b_1^2 - p_1(b_1 - c_1)(b_1 - a_1)))^{\frac{1}{2}}}{2} \tag{23}$$

The result from (23) can be used in conjunction with p_2 to find the parameters for the second distribution.

To find overlapping clusters from the triangular distribution, OCLUS follows these steps:

1. Choose a_1 from a $U(0, 20)$ distribution.
2. Choose b_1 from a $U(a_1, 20)$ distribution.
3. Let $c_1 = \frac{a_1+b_1}{2}$ (for other alternatives, see the section below regarding skewed data).
4. Set $a_2 = z - \frac{2p_2}{f(z)}$.
5. Choose $f(c_2)$ from a $U(f(z), f(c_1))$, then $c_2 = a_2 + \frac{f(c_2)f(z-a_2)}{f(z)}$.
6. Solve for b_2 by $b_2 = \frac{2}{f(c_2)} + a_2$.

3.2.7 Skewed Data

Waller et al. (1999) generate skewed data to simulate “real world” data. After normal data are generated, skewed data can be created by a simple transformation from the non-normal distribution with desired skewness and kurtosis, Fleishman (1978) provides the formula

$$\mathbf{X_s = a + bX + cX^2 + dX^3}. \tag{24}$$

For a table of the three constants (b, c, and d) controlling the skewness and kurtosis, see Fleishman (1978, pp. 524–525). An indepth discussion of generating non-normal skewed data is given in Fleishman (1978), Tadikamalla (1980), and Vale and Maurelli (1983), but it should be remembered that causing normally distributed data to be skewed will change the original amount of overlap.

OCLUS provides two natural methods of generating skewed data while still knowing the overlap. First, for the gamma distribution the mean is $\alpha\beta$ and the mode is $\beta(\alpha-1)$. The relationship, $\alpha\beta > \beta(\alpha-1)$, indicates data generated from the gamma distribution are naturally skewed. Second, skewed data with known overlap can be generated from the triangular distribution by noting the following relationship:

1. A symmetric triangular distribution is found by letting $c = \frac{b+a}{2}$;
2. a left-skewed triangular distribution can be formed by letting $c = \frac{b+3a}{4}$;
3. a right-skewed triangular distribution can be formed by setting $c = \frac{3b+a}{4}$.

These two distributions allow cluster validation studies to include skewness as a factor without altering the amount of overlap present in the generated clusters. However, OCLUS does not include a method (beyond the generation of variables with different variances) to directly control for the degree of kurtosis.

3.2.8 Correlated Data

Assume that $\mathbf{X}$ is a data matrix of n observations from a d-dimensional distribution that contains K clusters. Thus, under a model where the correlation structures of the groups are fully unrestricted, the first and second moments of C_k can be represented by a $V \times 1$ mean vector, $\boldsymbol{\mu}_k$, and a $V \times V$ covariance matrix, $\boldsymbol{\Sigma}_k$, respectively. Then, the squared statistical distance (the basis of computing overlap between clusters), from an arbitrary data point, $\mathbf{x}^*_{V\times 1}$, to the center of C_k is

$$d_x^2 = (\mathbf{x}^* - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^* - \boldsymbol{\mu}_k)\ . \tag{25}$$

Now we define a rotational transformation $\mathbf{L}_{V\times V}$ such that $\mathbf{Y} = \mathbf{XL}$. This transformations results in transformed group means and covariance matrices for the transformed observations, and therefore the group means, for $k = 1, \ldots, K$, of the transformed variable Y are then $\boldsymbol{\mu}_k^Y = \boldsymbol{\mu}_k\mathbf{L}$ and the covariance matrices for each group are $\boldsymbol{\Sigma}_k^Y = \mathbf{L}'\boldsymbol{\Sigma}_k\mathbf{L}$. Given the new variable the statistical distance, d_Y^2, of any point $\mathbf{y}^*$ and a transformed group mean $\boldsymbol{\mu}_k^Y$ is:

$$d_y^2 = (\mathbf{y}^* - \boldsymbol{\mu}_k^Y)(\boldsymbol{\Sigma}_k^Y)^{-1}(\mathbf{y}^* - \boldsymbol{\mu}_k^Y)'. \tag{26}$$

Through substitution (26) is:

$$\begin{aligned}
d_y^2 &= (\mathbf{x}^*\mathbf{L} - \boldsymbol{\mu}_k\mathbf{L})(\mathbf{L}'\boldsymbol{\Sigma}_k\mathbf{L})^{-1}(\mathbf{x}^*\mathbf{L} - \boldsymbol{\mu}_k\mathbf{L})' \\
d_y^2 &= (\mathbf{x}^* - \boldsymbol{\mu}_k)\mathbf{L}(\mathbf{L}^{-1}\boldsymbol{\Sigma}_k^{-1}\mathbf{L}'^{-1})((\mathbf{x}^* - \boldsymbol{\mu}_k)\mathbf{L})' \\
d_y^2 &= (\mathbf{x}^* - \boldsymbol{\mu}_k)\mathbf{L}(\mathbf{L}^{-1}\boldsymbol{\Sigma}_k^{-1}\mathbf{L}'^{-1})\mathbf{L}'(\mathbf{x}^* - \boldsymbol{\mu}_k)' \\
d_y^2 &= (\mathbf{x}^* - \boldsymbol{\mu}_k)(\mathbf{L}\mathbf{L}^{-1})\boldsymbol{\Sigma}_k^{-1}(\mathbf{L}'^{-1}\mathbf{L}')(\mathbf{x}^* - \boldsymbol{\mu}_k)'.
\end{aligned} \tag{27}$$

Since $\mathbf{L}\mathbf{L}^{-1} = \mathbf{L}'^{-1}\mathbf{L}' = \mathbf{I}$ then:

$$d_y^2 = (\mathbf{x}^* - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}^* - \boldsymbol{\mu}_k)' \ , \tag{28}$$

resulting in $d_y^2 = d_x^2$. Since the definition of overlap is based on statistical distance, any linear combination $\mathbf{L}$ will not change the proportion of overlap that was defined in the original data set, $\mathbf{X}$.

Generally, using the OCLUS algorithm as described above, $\mathbf{X}^*$ can be generated from normal distributions with univariate variance, resulting in the k^{th} cluster $\mathbf{X}_k^*$, being distributed as

$$\mathbf{X}_k^* \sim MVN(\boldsymbol{\mu}_k, \mathbf{I}).$$

where $\boldsymbol{\mu}_k$ represents the mean vector for the C_k. Correlation between the variables for the k^{th} group, defined either by the covariance matrix $\boldsymbol{\Sigma}_k$ or the correlation matrix $\mathbf{R}_k$, can be incorporated into the data by setting $\mathbf{L} = \boldsymbol{\Sigma}_k^{\frac{1}{2}}$ or $\mathbf{L} = \mathbf{R}_k^{\frac{1}{2}}$, respectively. For clusters that are overlapping, $\mathbf{L}$ must be the same for all observations within the overlapping clusters (i.e., common within-cluster covariance matrices are assumed, $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$). If clusters or groups of clusters (see the section on sub-clusters below) are well-separated, different values of $\mathbf{L}$ (or within-cluster covariance matrices may be unique) for the well-separated clusters (or groups of clusters) may be used with caution (i.e., the user must check the resulting transformations to determine if unintended overlap was introduced into the system). An example of correlated cluster structure is provided below in Example 5. Additionally, this allows for normal distributions with unequal variances to be arrived at via transformations of normal distributions with equal variances.(*Note:* The method for generating data with known overlap and known correlation matrix via transformations of the original data was arrived at through helpful comments provided by Reviewer 3).

4. Practical Concerns of OCLUS

4.1 Sub-clusters

Sub-clusters are defined as two or more sets of non-overlapping clusters containing either a single cluster or a group of overlapping clusters. $\hat{\mathbf{D}}$ from the illustrative example given above indicates a design matrix containing two sub-clusters. The inclusion of sub-clusters in a particular cluster generation design alters the way overlap is operationalized. Because marginal overlap is considered the average of overlapping regions within a dimension, the overlap in the sub-clusters must be adjusted for the fixed value of the denominator in

(6). For example, imagine three clusters embedded in one dimension with a desired marginal overlap of $p^{(v)} = .10$. If the clusters are generated as a "string", cluster 1 will overlap with cluster 2 which also overlaps with cluster 3. By (6), the overlap between each pair must be 0.10, but the same set of clusters could be embedded in one dimensional space with two sub-clusters, indicating two overlapping clusters comprise one sub-cluster while a singleton cluster comprises the remaining sub-cluster. For (6) to remain equal to 0.10, the overlap in the first sub-cluster must then be set to 0.20.

Even if overlap is equal, d-dimensional spaces with sub-clusters cannot be regarded the same as d-dimensional spaces with "strings" of clusters, allowing for the structure of the overlap to become a factor in cluster validation studies. Depending on the overlap structure of the data, various clustering algorithms may perform differently in the presence of sub-clusters.

4.2 Three Overlapping Groups

All of the previous illustrations of data generation focus on generating two clusters, but (3) and (4) can calculated for any pair of clusters. It is possible that due to the value of $p^{(v)}$, a set of three or more clusters will occupy the same bounded sub-space. If this possibility were ignored, the generated clusters would result in a greater degree of overlap then intended. OCLUS considers additional overlap caused by other clusters and adjusts the various θ_{kv}'s to achieve the desired value of $p^{(v)}$. This adjustment made by OCLUS is carried out by slightly moving the means of the clusters and recomputing their overlap in an iterative fashion until the desired overlap is achieved. For practical purposes, any two clusters with $P < .01$ are not considered overlapping (if more separation is required between non-overlapping clusters, greater values of s can be chosen–see above in the discussion of s). Additionally, the maximum value of P allowed by OCLUS is 0.50, a limit placed on P because it makes little sense to search for clusters when the joint overlap is greater than 50% (*Note*. This does not restrict the marginal overlap from being more than 0.50).

5. Examples

When discussing different distributions from which OCLUS is able to generate data, examples of pdfs were provided so overlap could be visualized as shared densities. This section applies the above methods to generate artificial data with known structure. For each example, scatter plots of the generated data points are displayed.

Example 1: Non-overlapping Normals, $K = 5$, $V = 2$

Example 1 provides a depiction of data that adheres to Cormack's (1971) definition of internally cohesive and externally isolated clusters. For non-

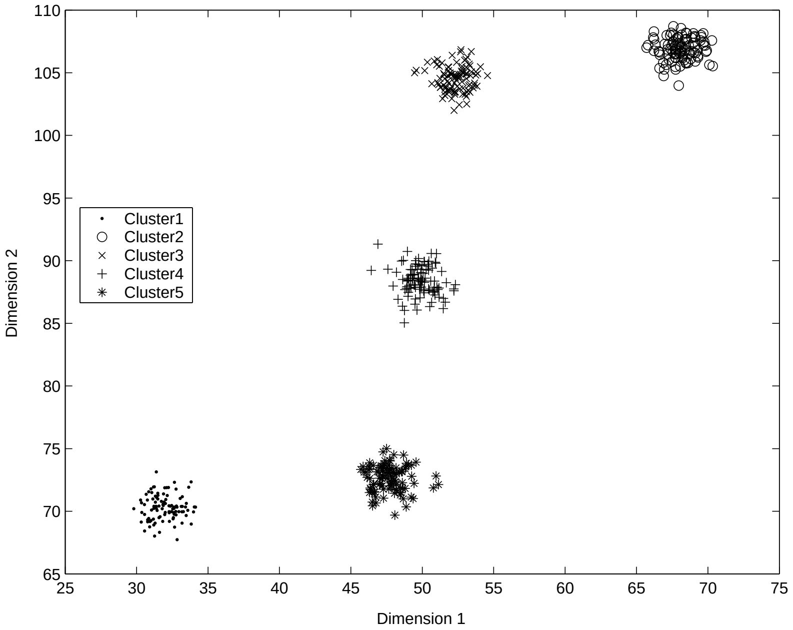

Figure 2. Example of non-overlapping clusters in 2 dimensions

overlapping clusters, $\mathbf{D} = \mathbf{I_{K\times K}}$. Figure 2 provides a scatter plot for the data generated from these distributions, where $\mathbf{n} = [100, 100, 100, 100, 100]'$.

Example 2: Mixture of Normal and Uniform Dimensions, $K = 3$**,** $V = 2$

This example shows the ability of OCLUS to generate different mixtures for each dimension on which a cluster is measured. This is a direct advantage of assuming independence across dimensions. For this example, $P = .10$. Thus, the total overlap on each dimension must be $\sqrt{.10} = .3162$. Let the design matrix be

$$\mathbf{D} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

indicating that cluster 1 overlaps with cluster 2, which overlaps with cluster 3. Figure 3 provides a scatter plot for the data generated from these distributions, for $\mathbf{n} = [100, 100, 100]'$. The advantages of mixing distributional clusters can be seen, from Figure 3. The generated clusters are neither rectangles as clusters generated from a uniform distribution or spheres as clusters generated from a

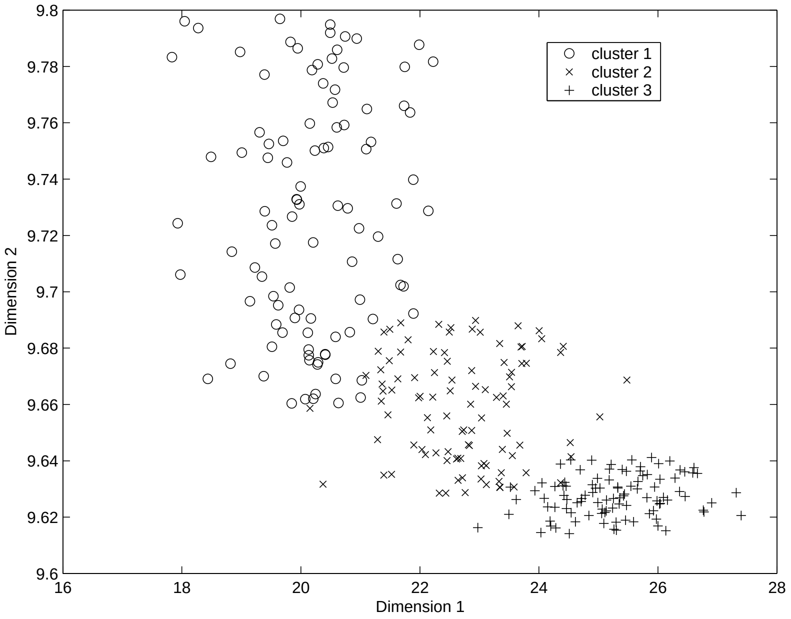

Figure 3. Example of overlapping clusters in 2 dimensions

normal distribution. Instead, they seem almost as rectangles with rounded corners and edges. This mixing allows for the overlap to stay the same while altering the shape of the generated clusters. In turn, this enables testing the effects of overlap on procedures that favor clusters of a particular shape.

Example 3: Overlapping Uniform Clusters, $K = 5$, $V = 3$

For this example, design matrix from the *user-defined options* section, $\hat{\mathbf{D}}$, is used, with overlap chosen to be $P = 0.20$ in three dimensions. Figure 4 provides a scatter plot of data generated from these distributions. As indicated by $\hat{\mathbf{D}}$, there are two groups of overlapping clusters. Group 1, containing the first three clusters in the lower right hand corner of Figure 4, is well separated from group 2, in the upper right hand corner of Figure 4. This flexibility of the design matrix allows for the testing of several different scenarios and orientations of clusters, and for evaluating procedures that consider smaller amounts of observations located away from the majority of objects to be outliers (Wishart, 1969).

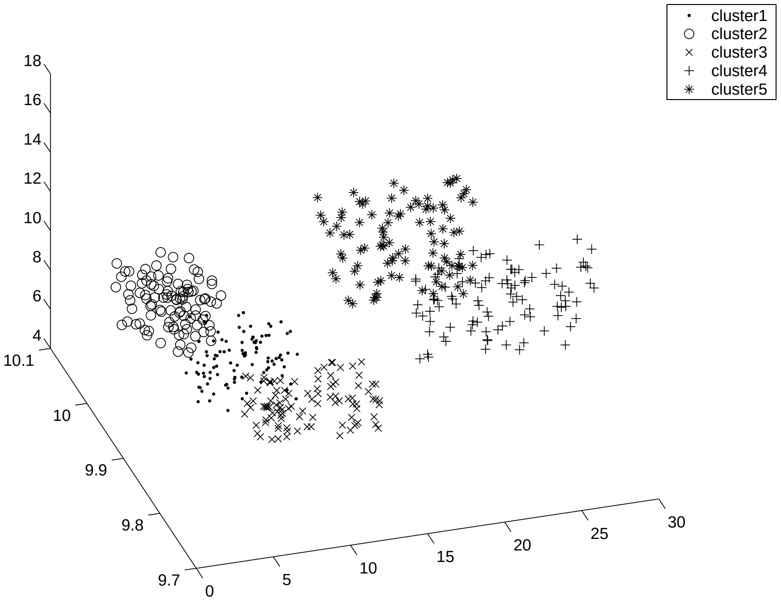

Figure 4. Example of overlapping clusters in 3 dimensions

Example 4: Introducing Correlated Variables

This example illustrates the introduction of correlation into the cluster structure in a bivariate space where there are only two clusters. First, assume that we want to generate two clusters that overlap with probability 0.20 in a two dimensional space where the variables are not correlated, illustrated in Figure 5.

The means and correlations for the two groups are

$$\boldsymbol{\mu}_1 = [16.02, 13.62] \quad \boldsymbol{\mu}_2 = [17.91, 11.73]$$

$$\mathbf{R}_1 = \begin{bmatrix} 1.00 & 0.04 \\ 0.04 & 1.00 \end{bmatrix} \quad \mathbf{R}_2 = \begin{bmatrix} 1.00 & 0.02 \\ 0.02 & 1.00 \end{bmatrix},$$

where the correlation between the two variables in both groups is entirely due to sample variation. Now suppose that we wanted the variables in each cluster to have a correlation of 0.40, then we would multiply both $\mathbf{X}_1$ and $\mathbf{X}_2$ by

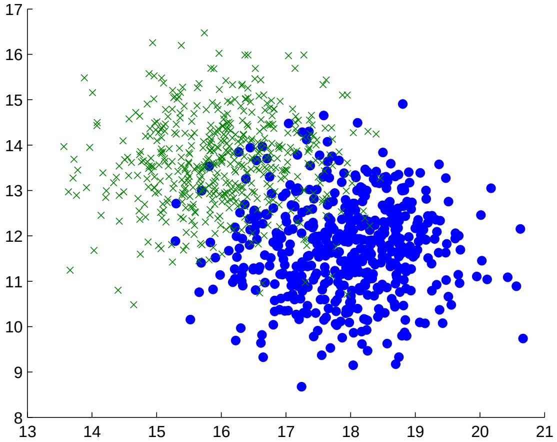

Figure 5. Example of uncorrelated variables in 2 dimensions

$$\mathbf{L}^{1/2} = \begin{bmatrix} 0.98 & 0.20 \\ 0.20 & 0.98 \end{bmatrix},$$

where $\mathbf{L}$ is the desired correlation matrix. The resulting data is depicted in Figure 6.

Now the means and covariances of the transformed clusters are

$$\boldsymbol{\mu}_1^* = [18.47, 16.61] \quad \boldsymbol{\mu}_2^* = [19.93, 15.14]$$

$$\mathbf{R}_1^* = \begin{bmatrix} 1.00 & 0.40 \\ 0.40 & 1.00 \end{bmatrix} \quad \mathbf{R}_2^* = \begin{bmatrix} 1.00 & 0.40 \\ 0.40 & 1.00 \end{bmatrix},$$

where the means have slightly shifted due to the oblique rotation of the data. However, the desired correlation between the variables has been achieved and the theoretical probability of overlap between the clusters is still 0.20.

6. Discussion

6.1 Advantages of OCLUS

The most attractive feature of OCLUS is its versatility in creating data with a known amount of overlap. In addition to its ability to create the well-separated clusters to satisfy Cormack's (1971) definition, OCLUS can generate well-separated groups of overlapping clusters (see $\hat{\mathbf{D}}$ above). Instead of manipulating the structure of the covariance matrices and creating uninterpretable

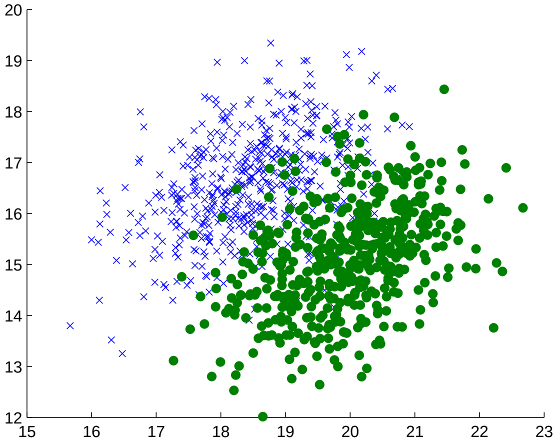

Figure 6. Example correlated variables in 2 dimensions

clusters to obtain overlap, OCLUS achieves the same goal analytically, yielding readily understandable overlap in terms of shared density. Not only does this retain the original interpretability of the clusters, it avoids the iterative method employed by Price (1993), giving a quantification of overlap while saving computing time. Another novel feature of OCLUS is the ability to generate data with known overlap from several different distributions, considered important for advancing the understanding of different clustering algorithms (Milligan, 1996).

6.2 Limitations

A limitation of OCLUS is the inability to sample directly from the joint distribution of the clusters, restricting the generation of the joint distribution clusters to be the product of the marginal distributions. A Markov Chain Monte Carlo simulation might be a way to sample directly from the joint distribution, but the trade-off will be a substantial increase in the computing time. Additionally, the ability to introduce correlation into the system while preserving group overlap may make generating clusters from the joint distribution unnecessary.

Similarly, another limitation of the OCLUS procedure is the inability to generate clusters with known skew and kurtosis. As noted above when discussing skewed data, the triangular distribution is used to generate data with skewed features; however, the exact degree of skewness is not known. Thus, in situations where researchers desire to have the most control over the degree of skewness and kurtosis in the generated clusters, we encourage the use of the Waller et al. (1999) procedure; on the other hand, if cluster overlap is of primary focus, we recommend the OCLUS procedure.

6.3 Future Applications of OCLUS

OCLUS should aid the advancement of the field of cluster analysis and classification by helping to study the robustness of both traditional (single-link, complete-link, k-means, etc.) and untraditional strategies (ADCLUS, pyramid clustering, fuzzy clustering, etc.), and the effects of various distributions and mixtures of distributions on the performance of these algorithms while phrasing the results in terms of cluster overlap.

OCLUS may play a useful role in investigating techniques of variable selection (Brusco & Cradit, 2001; Fowlkes, Gnanadesikan, & Kettering, 1987; Fowlkes, Gnanadesikan, & Kettering, 1988; Carmone, Kara, & Maxwell, 1999) and variable weighting (De Soete, DeSarbo, & Carroll, 1985; De Soete, 1986) by directly manipulating the dimensions independently. Additionally, data can be generated to investigate the robustness of various methods used for determining the number of clusters, both for those that are classical (Calinski & Harabasz, 1974; Duda & Hart, 1973; Hubert & Levin, 1976; Baker & Hubert, 1975; Beale, 1969; Atlas & Overall, 1994) and those of more recent vintage based in finite mixture modeling and model selection (Bozdogan & Sclove, 1984; Banfield & Raftery, 1993; Windham & Cutler, 1992; Bozdogan, 1993). OCLUS can also be used to test the sensitivity of methods attempting to determine the number of modes in a data set (Hartigan, 1988; Hartigan, 2000; Hartigan & Hartigan, 1985; Hartigan & Mohanty, 1992). Overall, OCLUS provides an interpretable mechanism for evaluation of technique robustness as developed over several different areas of cluster analysis, finite mixture modeling, and latent profile analysis.

References

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

ATLAS, R. S., and OVERALL, J. E. (1994). “Comparative Evaluation of Two Superior Stopping Rules for Hierarchical Cluster Analysis,” *Psychometrika, 59*, 581–591.

BAKER, F. B., and HUBERT, L. J. (1975). “Measuring the Power of Hierarchical Cluster Analysis,” *Journal of the American Statistical Association, 70*, 31–38.

BALAKRISHNAN, P. V., COOPER, M. C., JACOB, V. S., and LEWIS, P. A. (1994). “A Study of Classification Capabilities of Neural Networks Using Unsupervised Learning: A Comparison With K-means Clustering,” *Psychometrika, 59*, 509–525.

BALL, G. H., and HALL, D. J. (1967). “A Clustering Technique for Summarizing Multivariate Data,” *Behavioral Science, 12*, 153–155.

BANFIELD, J. D., and RAFTERY, A. E. (1993). “Model-based Gaussian and Non-Gaussian Clustering,” *Biometrics, 49*, 803–821.

BAYNE, C. K., BEAUCHAMP, J. J., BEGOVICH, C. L., and KANE, V. E. (1980). “Monte Carlo Comparisons of Selected Clustering Procedures,” *Pattern Recognition, 12*, 51–62.

BEAUCHAINE, T. P., and BEAUCHAINE, R. J., III. (2002). “A Comparison of Maximum Covariance and K-means Cluster Analysis in Classifying Cases into Known Taxon Groups,” *Psychological Methods, 7*, 245–261.

BEALE, E. M. L. (1969). *Cluster Analysis*, London: Scientific Control Systems.

BICKEL, P. J., and DOKSUM, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics* (2nd ed.), Upper Saddle River, NJ: Prentice Hall.

BLASHFIELD, R. K. (1976). “Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods,” *Psychological Bulletin, 83*, 377–388.

BOZDOGAN, H. (1993). “Choosing the Number of Component Clusters in the Mixture-Model Using a New Informational Complexity Criterion of the Inverse-Fisher Information Matrix,” *Information and Classification*, Eds., O. Opitz, B. Lausen, and R. Klar, Heidelberg: Springer, 40–54.

BOZDOGAN, H., and SCLOVE, S. L. (1984). “Multi-Sample Cluster Analysis Using Akaike’s Information Criterion,” *Annals of The Institute of Statistical Mathematics, 36*, 163–180.

BRUSCO, M. J., and CRADIT, J. D. (2001). “A Variable Selection Heuristic for K-means Clustering,” *Psychometrika, 66*, 249–270.

CALINSKI, R. B., and HARABASZ, J. (1974). “A Dendrite Method for Cluster Analysis,” *Communications in Statistics, 3*, 1–27.

CARMONE, F. J., KARA, A., and MAXWELL, S. (1999). “HINoV: A New Model to Improve Market Segment Definition by Identifying Noisy Variables,” *Journal of Marketing Research, 36*, 501–509.

CARROLL, J. D., and ARABIE, P. (1983). “INDCLUS: An Individual Differences Generalization of the ADCLUS Model and the MAPCLUS Algorithm,” *Psychometrika, 48*, 157–169.

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

DAVID, H. A. (1981). *Order statistics* (2nd ed.), New York: Wiley.

DE SOETE, G. (1986). “Optimal Variable Weighting for Ultrametric and Additive Tree Clustering,” *Quality and Quantity, 20*, 169–180.

DE SOETE, G., DESARBO, W. S., and CARROLL, J. D. (1985). Optimal Variable Weighting for Hierarchical Clustering: An Alternating Least-Squares Algorithm,” *Journal of Classification, 2*, 173–192.

DESARBO, W. S. (1982). “GENNCLUS: New Models for General Nonhierarchical Clustering Analysis,” *Psychometrika, 47*, 449–475.

DESARBO, W. S. (1984). “Constrained Classification: The Use of a Priori Information in Cluster Analysis,” *Psychometrika, 49*, 187–215.

DIDAY, E. (1986). “Orders and Overlapping Clusters by Pyramids,” *Multidimensional Data Analysis*, Eds., J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley, Leiden: DSWO, 201–234.

DIDAY, E., and BERTRAND, P. (1986). “An Extension of Hierarchical Clustering: The Pyramidal Presentation,” *Pattern recognition in practice II*, Eds., E. S. Gelsema and L. N. Kanal, Amsterdam: North-Holland, 411–424.

DUDA, R. O., and HART, P. E. (1973). *Pattern Classification and Scene Analysis*, New York: Wiley.

EDELBROOK, C. (1979). "Comparing the Accuracy of Hierarchical Clustering Algorithms: The Problem of Classifying Everybody," *Multivariate Behavioral Research, 14*, 367–384.

EVANS, M., HASTINGS, N., and PEACOCK, B. (2000). *Statistical distributions* (3rd ed.), New York: Wiley.

FLEISHMAN, A. J. (1978). "A Method for Simulating Non-Normal Distributions," *Psychometrika, 43*, 521–532.

GOLD, E. M., and HOFFMAN, P. J. (1976). "Flange Detection Cluster Analysis," *Multivariate Behavioral Research, 11*, 217–235.

GORDON, A. D. (1987). "A Review of Hierarchical Classification," *Journal of the Royal Statistical Society, A, 150*, 119–137.

FOWLKES, E. B., GNANADESIKAN, R., and KETTERING, J. R. (1987). "Variable Selection in Clustering and Other Contexts," *Design, Data, and Analysis*, Ed., C. L. Mallows, New York: Wiley, 13–34

FOWLKES, E. B., GNANADESIKAN, R., and KETTERING, J. R. (1988). "A Method for Comparing Two Hierarchical Clusterings (with Comments and Rejoinder)," *Journal of the American Statistical Association, 78*, 553–584.

GORDON, A. D. (1996). "Hierarchical classification," *Clustering and Classification*, Eds., P. Arabie, L. J. Hubert, and G. De Soete, River Edge, NJ: World Scientific, 65–121.

GORDON, A. D. (1999). *Classification* (2nd ed.), New York: Chapman and Hall/CRC.

HAMPBEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J., and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, New York: Wiley.

HANSEN, P., JAUMARD, B., and SANLAVILLE, E. (1994). "Partitioning Problems in Cluster Analysis: A Review of Mathematical Programming Approaches," *New Approaches in Classification and Data Analysis*, Eds., E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtchy, Springer: Berlin, 228–240.

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

HARTIGAN, J. A. (1988). "The Span Test of Unimodality, *Classification and Related Methods*, Ed., H. H. Bock, Amsterdam: North Holland, 229–236.

HARTIGAN, J. A. (2000). "Testing for Antimodes," *Data Analysis: Scientific Modelling and Practical Application*, Eds., W. Gaul, O. Opitz, and M. Schader, Berlin: Springer, 385–392.

HARTIGAN, J. A., and HARTIGAN, P. M. (1985). "The Dip Test of Multimodality," *Annals of Statistics, 13*, 70–84.

HARTIGAN, J. A., and MOHANTY, S. (1992). "The Runt Test for Multimodality," *Journal of Classification, 9*, 63–70.

HARTIGAN, J. A., and WONG, M. A. (1979). "Algorithm AS 136. A K-means Clustering Algorithm," *Applied Statistics, 28*, 100–108.

HELSEN, K., and GREEN, P. E. (1991). "A Computational Study of Replicated Clustering with an Application to Market Segmentation," *Decision Sciences, 22*, 1124–1141.

HOGG, R. V., and CRAIG, A. T. (1995). *Introduction to Mathematical Statistics* (5th ed.), Upper Saddle River, NJ: Prentice Hall.

HUBERT, L. J., and LEVIN, J. R. (1976). "A General Statistical Framework for Assessing Categorical Clustering in Free Recall," *Psychological Bulletin, 83*, 1072–1080.

JARDINE, N., and SIBSON, R. (1968). "The Construction of Hierarchic and Non-Hierarchic Classifications," *The Computer Journal, 11*, 177–184.

KRZANOWSKI, W. J., and MARRIOTT, F. H. C. (1994). *Multivariate Analysis Part I: Distributions, Ordination, and Inference*, New York: Wiley.

KUIPER, F. K., and FISHER, L. (1975). "A Monte Carlo Comparison for Six Clustering Procedures," *Biometrics, 31*, 777-784.

LANCE, G. N., and WILLIAMS, W. T. (1966). "A Generalised Sorting Strategy for Computer Classifications," *Nature, 212*, 218.

LANCE, G. N., and WILLIAMS, W. T. (1967). "A General Theory of Classificatory Strategies I. Hierarchical Systems," *The Computer Journal, 9*, 373–380.

LEHMAN, E. L., and CASELLA, G. (1998). *Theory of Point Estimation*, New York: Springer.

MACQUEEN, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Eds., L. M. Le Cam and J. Neyman, Berkeley: University of California Press, 281–297.

MCINTYRE, R. M., and BLASHFIELD, R. K. (1980). "A Nearest-Centroid Technique for Evaluating the Minimum Variance Clustering Procedure," *Multivariate Behavioral Research, 15*, 225–238.

MCQUITTY, L. L. (1960). "Hierarchical Linkage Analysis for the Isolation of Types. " *Educational and Psychological Measurement, 20*, 55–67.

MILLIGAN, G. W. (1980). "The Validation of Four Ultrametric Clustering Algorithms," *Pattern Recognition, 12*, 41–50.

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 Analyses," Eds., P. Arabie, L. J. Hubert, and G. De Soete, River Edge, NJ: World Scientific, 341–375.

MILLIGAN, G. W., and COOPER, M. C. (1986). "A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis," *Multivariate Behavioral Research, 21*, 441–458.

MILLIGAN, G. W., and COOPER, M. C. (1988). "A Study of Standardization of Variables in Cluster Analysis," *Journal of Classification, 5*, 181–204.

MILLIGAN, G. W., SOON, S. C., and SOKAL, L. M. (1983). "The Effect of Cluster Size, Dimensionality, and the Number of Clusters on Recovery of True Cluster Structure," *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*, 40–47.

MOJENA, R. (1977). "Hierarchical Grouping Methods and Stopping Rules-An Evaluation," *The Computer Journal, 20*, 359–363.

PRICE, L. J. (1993). "Identifying Cluster Overlap with NORMIX Population Membership Probabilities," *Multivariate Behavioral Research, 28,* 235–262.

ROUSSEEUW, P. J., and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*, New York: Wiley.

SHEPARD, R. N., and ARABIE, P. (1979). "Additive Clustering: Representation of Similarities as Combinations of Discrete Overlapping Properties," *Psychological Review, 86*, 87–123.

SNEATH, P. H. A. (1957). "The Application of Computers in Taxonomy," *Journal of General Microbiology, 17*, 201–226.

SOKAL, R. R., and MICHENER, C. D. (1958). "A Statistical Method for Evaluating Systematic Relationships," *University of Kansas Science Bulletin, 38*, 1409–1438.

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). “Standardizing Variables in K-means Clustering,” *Classification, Clustering, and Data Mining Applications*, Eds., D. Banks, L. House, F. R. McMorris, P. Arabie, and W. Gaul, New York: Springer, 53–60.

TADKIKAMALLA, P. R. (1980). “On Simulating Non-Normal Distributions,” *Psychometrika, 45*, 273–279.

VALE, C. D., and MAURELLI, V. A. (1983). “Simulating Multivariate Nonnormal Distributions,” *Psychometrika, 48*, 465–471.

WALLER, N. G., KAISER, H. A., ILLIAN, J. B., and MANRY, M. (1998). “A Comparison of the Classification Capabilities of the 1-dimensional Kohonen Neural Network with Two Partitioning and Three Hierarchical Cluster Analysis Algorithms,” *Psychometrika, 63*, 5–22.

WALLER, N. G., UNDERHILL, J. M., and KAISER, H. A. (1999). “A Method for Generating Simulated Plasmodes and Artificial Test Clusters with User-Defined Shape, Size, and Orientation,” *Multivariate Behavioral Research, 34*, 123–142.

WARD, Jr., J. H. (1963). “Hierarchical Grouping to Optimize an Objective Function,” *Journal of the American Statistical Association, 58*, 236–244.

WEISSTEIN, E. W. (2003). *CRC Concise Encyclopedia of Mathematics* (2nd ed.), London: Chapman and Hall/CRC.

WINDHAM, M. P., and CUTLER, A. (1992). “Information Ratios for Validating Cluster Analyses,” *Journal of the American Statistical Association, 87*, 1188–1192.