A Comparison of Clustering Methods when Group Sizes are Unequal, Outliers are Present, and in the Presence of Noise Variables

W. Holmes Finch
Ball State University

Cluster analysis is a widely used statistical tool for assisting researchers in identification of subgroups within the population based upon a set of variables. Research has shown that many clustering algorithms have difficulty in correctly grouping individuals within a sample when the subgroups within the population are of very different sizes. In addition, cluster algorithms can also yield inaccurate clustering results when outliers are present in the data, as well as if some of the variables used in the clustering are not actually associated with subgroup separation. A variety of clustering algorithms have been developed to deal with these problems, including approaches based on the popular Lasso regularization estimator, a trimmed estimator, density based clustering, the use of medoids rather than centroids, and a robust approach. Prior research has examined several of these methods with some, but not all of the challenging data scenarios outlined above. The purpose of this simulation study, therefore, was to compare a number of these clustering algorithms when group sizes were unequal, outliers were present, and some of the variables represented noise, rather than actually contributing to cluster separation. Results of the study showed that the robust, trimmed, and density based clustering methods yielded the most accurate clustering results when all of these issues were present in the data. When none were present, all of the methods examined here performed similarly and well. An empirical example is also demonstrated, and implications of the simulation study are discussed.

Cluster analysis is used by researchers to identify subgroups within a population (Prokasky, Rudasill, Molfese, Putnam, Gartstein, & Rothbart, 2017; Lewandoeski, Sperry, Cohen, & Ongur, 2014; Henry, Tolan, & Gorman-Smith, 2005). Typically, these subgroups will not have been previously labeled, and therefore may present researchers with a new or different understanding of the population under study. For example, researchers may be interested in understanding the number and profile of subgroups of students in a school district based upon their responses to a set of items regarding school climate. There exist two broad paradigms of cluster analysis: (1) hierarchical clustering, in which individuals are placed together in an iterative fashion until all are placed in a single cluster, and (2) an a priori supposition regarding the number of clusters (k) present based upon hypotheses derived from previous research is made and individuals in the sample are placed into one of the k clusters. With respect to hierarchical clustering, the key issue is the determination of the number of clusters to retain, which is based on a variety of statistical and conceptual considerations (Hahs-Vaughn, 2017). For the second clustering approach, determining the optimal number of clusters to retain is also a key issue, which is typically based on a well-developed theory supporting the a priori hypotheses regarding the number of subgroups present in the population.

The focus of this study is on the second application of cluster analysis, and therefore the remainder of this manuscript will be devoted to it. There exist a wide array of tools for clustering individuals into a finite number of groups. These methods vary in terms of the algorithms used to place individuals into one of the *k* clusters. Perhaps the most common such method, Kmeans clustering, has been shown to work well in a variety of situations. However, prior research has also demonstrated that it can have some difficulty in correctly identifying the correct clusters when outliers are present in the data (Aggarwal, 2016), and when some of the variables used to identify which individuals belong in which cluster are not in fact associated with cluster membership in the population (Raftery & Dean, 2006). This latter case is commonly referred to as the problem of noise variables. Several clustering algorithms have been investigated with respect to these problematic conditions, and been shown to work well (Galimberti, Manisi, & Soffritti, 2018). However, this prior work has not examined the accuracy of these methods for identifying correct cluster membership when the subgroups within the population are of very different sizes, a situation that has been shown to compromise commonly used approaches (Holden, Finch, & Kelley, 2011; Candel & van Breukelen, 2010). Thus, the goal of this study was to examine the ability of several cluster analysis methods

to correctly recover the underlying subgroup structure when these groups are of different sizes in the population, in conjunction with the presence of outliers and noise variables. The clustering methods to be examined in this study are described below, and include commonly used approaches, as well as those that have been demonstrated to be robust to outliers and/or noise variables in prior research.

Methods of Cluster Analysis

Kmeans

Kmeans is one of the most popular approaches to clustering a set of *n* observations using *p* observed variables (MacQueen, 1967). The optimal solution for a given sample and a prespecified number of *k* clusters is determined by minimizing the within cluster sum of squares:

$$SS_{wpk} = \sum_{r=1}^{K} \sum_{i \in K_r} \left(x_{ij} - \bar{x}_{jr}\right)^2 \quad (1)$$

where: x_{ij} =Value of variable *j* for individual *i* and $\bar{x}_{jr}$ =Mean of variable *j* for cluster *r*. Thus, for each cluster *r* within the full set of *k* clusters, the sum of squares is calculated for each of the *p* variables. These individual variable sums of squares are then summed to obtain a total sum of squares for the entire cluster solution:

$$SS_w = \sum_{j=1}^{p} SS_{wpk} \quad . \quad (2)$$

To begin the Kmeans algorithm, a set of *k* initial cluster centeroids are selected randomly from the sample, where *k* is defined by the data analyst. Each member of the sample is then placed with the cluster for which its value from equation (2) is minimized. The cluster centroids are then updated, the sums of squares in equations (1) and (2) are recalculated, and individuals in the sample are once again placed in the cluster leading to the overall smallest sum of squares. This series of steps is repeated until no switching of cluster membership lowers the SS_w, at which point the optimal Kmeans cluster solution is achieved.

Although Kmeans has been shown to be an effective tool in many situations, it also has proven to be problematic in some conditions. For example, when the underlying subgroups in the population are of different sizes, Kmeans has a tendency to place most members of the sample into the cluster associated with the largest group, leading to low sensitivity rates for the small group (Holden, Finch, & Kelley, 2011). Such a result is particularly problematic in practice when the small group represents a rare but important subgroup, such as individuals at-risk for attempting suicide, or children with a learning disability. In addition, because Kmeans relies on cluster means to calculate SS_w, it is potentially sensitive to the presence of outliers (Brodinova, Filzmoser, Ortner, Breiteneder, & Rohm, 2019; Garcia-Esudero & Gordaliza, 1999). When a sample does contain outliers, the correct cluster solution may not be identified (Brodinova, et al.). Finally, like many prediction algorithms (supervised and unsupervised), the performance of Kmeans, can be deleteriously impacted by the presence of noise variables; i.e., variables that are used in the cluster analysis but which are not actually related to cluster separation (Hajnal & Loosveldt, 1998). Given these potential weaknesses, researchers have developed alternative cluster analysis methods that are designed to work better when outliers are present, group sizes are very unequal, and/or noise variables are present in the data.

Kmedoids

Given the negative impact of outliers on the performance of Kmeans clustering, one alternative that has been proposed involves the use of an alternative to the mean as a measure of the cluster center (Kaufman & Rousseeuw, 1987). The Kmedoids algorithm begins identically to Kmeans, with the random selection of initial cluster centroids, and the calculation of SS_w. However, rather than then using the cluster means in calculating SS_w in the next iteration step, the medoid of each cluster is identified instead. The medoid is the individual member of the cluster that minimizes SS_w when its values are used in lieu of the cluster means. Thus, if using the variable values of individual A in cluster 1 leads to a minimization of SS_w, these scores will be used, rather than the means of the individual variables. In the second step of the Kmedoids algorithm, the quantities in equations (1) and (2) are once again calculated, based on the medoid points for each cluster, and the steps used with Kmeans are followed until SS_w can be minimized no further. Kmedoids clustering has been suggested as an alternative to Kmeans for use in situations where outliers are present, given the susceptibility of the mean to the influence of outliers (Kaufman & Rousseeuw, 1990).

Trimmed Kmeans

One general approach to the presence of outliers that is commonly used in the robust methods literature is that of trimming (Neykov, et al., 2007). With trimming, a predetermined proportion (α) of the observations at either extreme of the distribution is removed from the sample, and then the mean (or whichever statistic is of interest) is calculated. In the context of cluster analysis, trimming involves identification and removal of the α individuals that are at the extremes on both ends of the distribution of distance from their cluster centroid. Thus, if α is set to be 0.1, then individuals with distance values (e.g., Euclidean) above the 90th or below the 10th percentiles are removed from the analysis and the Kmeans clustering algorithm is carried out in the same fashion as described above. Euclidean distance for a pair of points, *i*, *l*, is calculated as:

$$d_{il} = \sqrt{\sum_{i=1}^{p}\left(x_{ip} - x_{lp}\right)^2} \quad (3)$$

where, x_{ip} =Value of variable *p* for individual *i* and x_{lp} =Value of variable *p* for individual *l*. Larger values of d_{il} indicate a greater difference in scores on the measured variables for the two individuals.

Model Based Clustering

An alternative framework to Kmeans clustering and its associated methods (Kmedoids and trimmed Kmeans) comes in the form of model based clustering (Fraley & Raftery, 2002; Mclust). Mclust is based on the assumption that each member of the sample comes from a finite mixture of *G* probability distributions. These *G* distributions correspond conceptually to the clusters in the methods described above. This mixture model is written as:

$$p(y_i) = \sum_{g=1}^{G} \tau_g f_g\left(y_i|\theta_g\right) \quad (4)$$

where, y_i =Measured variables for subject *i*; assumed to be multivariate normal; τ_g =Probability that the observation belongs in group *g*; and $f_g\left(\cdot\,|\theta_g\right)$ =Density of the *g*th component given its parameters, θ_g. In the context of identifying subgroups within the population, the parameters typically associated with θ_g are the means of the observed variables, along with their covariance matrix. Estimation of the model parameters in equation (3) is typically carried out using the expectation maximization (EM) algorithm for the log-likelihood function of the model:

$$\sum_{i=1}^{n} log\left[\sum_{g=1}^{G} \tau_g f_g\left(y_i|\theta_g\right)\right] . \quad (5)$$

The combination of parameters in equation (4) that maximizes the function in equation (5) are retained and the resulting clusters for the sample are assumed to represent subgroups within the population.

The distribution of the *p* observed variables is assumed to be multivariate normal, and thus serves as an important assumption underlying the EM algorithm. Research has shown that when the number of observed variables is large relative to the sample size, the covariance matrix is unstable or, in some cases, impossible to estimate (Witten & Tibshirani, 2010). In addition to issues associated with high dimensional data, the presence of outliers can also prove to be problematic for Mclust. Rehm, Klawonn, and Kruse (2007) showed that the method has a tendency to incorrectly identify outliers as single observation clusters, rather than correctly placing them in their actual group. Evans, Love, and Thurston (2015) reported a similar result, and also found that this tendency to place outliers in their own cluster was a particular issue with spherical as opposed to ellipsoid population clusters.

Sparse Penalized Cluster Analysis

As noted above, Kmeans clustering can have difficulty finding the correct solution when the number of variables, *p*, is large and/or many of the variables used in the analysis are not relevant to the problem of differentiating the subgroups in the population (Witten & Tibshirani, 2010). Such a scenario could occur, for example, when the researcher is unsure of which variables will best differentiate the underlying subgroups, and so includes a large number of them in the cluster analysis. The inclusion of these variables introduces noise with respect to cluster membership, which can reduce the accuracy of traditional clustering methods. In order to address this issue, Witten and Tibshirani described a variant of Kmeans clustering that relies on the well known Lasso (Tibshirani, 1996) approach to variable selection for high dimensional data problems. The Lasso penalizes models such that the inclusion of variables comes with a cost. Thus, for a variable to be worth inclusion in the model, it must provide improved fit above some predefined threshold.

The higher this threshold, the fewer variables will be included in the final model. The magnitude of this threshold is typically determined through an iterative process whereby the fit of the model for a set of data is determined for a variety of values, and the one yielding optimal fit as measured by, for example, an information index, is retained. The Lasso has been proven useful for a wide variety of models.

In the context of cluster analysis, the Lasso approach (SPARCL) is designed to maximize the following function:

$$\sum_{j=1}^{p} w_j f_j\left(x_j \Theta\right) \tag{6}$$

where, x_j =Variable j; w_j =Weight indicating the contribution of variable j to the cluster solution; Θ =Cluster partition consisting of number of clusters and individual cluster membership; and f_j =Function indicating degree of cluster separation; e.g., between cluster sum of squares. Larger values of w_j indicate that the variable contributes relatively more to the cluster solution, whereas when $w_j = 0$ the variable does not contribute to the solution at all. The value of w_j is determined in part by the sparsity parameter *s*, which the researcher must select based on relative model fit, as described above. Readers interest in a more complete discussion of the Lasso clustering algorithm are encouraged to read Witten and Tibshirani (2010).

Robust Clustering

Brodinova, et al. (2019) described a robust clustering algorithm that was designed to both down weight outlying observations, and select informative (as opposed to noise) variables for the purposes of clustering. This algorithm consists of three phases, each of which includes multiple steps. In the first phase, each observation in the dataset is given a weight with the goal of reducing the impact of possible outliers on the derivation of a cluster solution. This selection of observation weights begins with the identification of a solution for *k* clusters, much as with Kmeans. The weights are then calculated for each observation based on the application of the biweight function (Rocke, 1996) to the local outlier factor (LOF; Breunig, Kriegel, Ng, & Sander, 2000). A value of 0 is assigned to an observation that is determined to be an outlier, whereas a value of 1 is assigned to observations that are definitely not outliers. Observations for which this determination is uncertain are assigned a weight between 0 and 1 based on the LOF, where values closer to 0 indicate that the individual is more likely to be an outlier. This value of LOF serves as a weight such that observations with lower values play less of a role in the cluster solution than do those with higher values. The between cluster sum of squares is calculated using the weights to identify an initial cluster solution, much as is the case with Kmeans and the within cluster sum of squares value. The robust algorithm then reforms the clusters in an attempt to lower the between cluster sum of squares, and the between cluster sum of squares is once again calculated applying the individual weights, and compared to the value from the previous iteration. This process continues until convergence is reached; i.e., observations no longer switch clusters. It is important to note that with this robust clustering approach, observations can be identified as outliers and not assigned to any cluster.

The goal of the second phase of the robust clustering algorithm is to identify only those variables that are salient with regard to the identification of clusters, and down weight all of the others. In this respect, the robust clustering method is very similar in spirit to the sparse clustering method described above. A primary difference between the two approaches is that a variable's weight reflects not only its importance in terms of correctly identifying cluster number and membership, but also which variables clearly separate outliers from clusters. Thus, a variable could have a relatively high weight in part (or in full) because it accurately identifies outlying observations. As with Lasso clustering (Witten & Tibshirani, 2010) a sparsity parameter is applied to the robust algorithm, and each variable is assigned a weight between 0 and 1, with larger values indicating that the variable is informative in terms of cluster separation. The determination of these weights is carried out in an iterative fashion much as with the weights for individuals in the sample.

The final phase of the robust clustering algorithm involves the assignment of each individual to their closest cluster using only those variables with weights greater than 0. The variable weights are applied in this step so that those with larger values are more important in terms of determining cluster membership than those with lower values. In summary, the robust clustering algorithm involves the identification of outlying observations in phase 1, followed by the identification of salient variables in phase 2. This latter determination is made by considering how well individual variables accurately identify outliers, as well as their ability to differentiate among clusters. In the final step, those variables that have non-zero weights are used to place individuals within clusters using a weighted algorithm.

DBSCAN

The final clustering algorithm to be examined in this study was Density-Based Spatial Clustering and Application with Noise (DBSCAN; Ester, et al., 1996). This approach to clustering observations is designed to identify density regions associated with each subgroup within the population and then associate individuals within these regions to the corresponding cluster. In this context, density simply refers to the number of observations within the neighborhood of a given point, based on some measure of distance, such as d_{il}. When a large number of observations are near point *t*, we conclude that the density is high. The size of the neighborhood around the point is denoted as ε, with larger values indicating a larger desired neighborhood. The other important parameter in DBSCAN is the minimum number of acceptable points within the radius defined by ε. That is, a viable region must contain a minimal number of observations (m) in order to be acceptable for the purposes of clustering. The DBSCAN algorithm begins by identifying all of the core points in the dataset, where core points are those with at least *m* points in its neighborhood. Border points are defined as those that are not core points, but do belong to a neighborhood associated with a core point. Outliers are defined as those observations that are neither core nor border points.

In addition to placing individual points into one of these three categories, relationships among the points must also be assigned. Two points are considered to be density reachable if there are a set of core points connecting them to one another. For example, observation *u* is density reachable to observation *v* if *u* is in the neighborhood of core point *z*, which is in the neighborhood of core point *w*, and observation *v* is also in the neighborhood of *w*. In turn, points are density connected if they are both density reachable to a common core point. Thus, if observations *u* and *a* are both density reachable to core point *d*, then we conclude that they are density connected.

The DBSCAN algorithm begins with the calculation of the distance between each observation and all of the others in the sample. Next, using *m* and ε, core points are identified based on the heuristic described above. These core points are the seeds for clusters. If two core points are within ε of one another then they are assigned to the same cluster. After this initial assignment of cluster centers, all points that are density connected to a center are placed in its cluster. This sequence of cluster assignments is repeated until all data points are either in a cluster or not. Points not assigned to a cluster are identified as outliers. DBSCAN has been shown to be effective at dealing with datasets containing outliers (Ester, Kriegel, Sander, & Xu, 1996), and for clusters that are not spherical in form (Viswanath & Babu, 2009).

Study Goals

The purpose of this study was to compare the performance of several methods of cluster analysis in the case where outliers, and noise variables are present, and population subgroups are of different sizes. Prior work has examined the performance of several of these approaches with one another in the presence of outliers and/or noise variables (e.g., Brodinova, et al., 2019). However, these studies have not compared all of the methods included here with one another, nor have they examined the performance of these methods when both outliers and unequally sized groups are present. Given prior literature showing that standard approaches to clustering (e.g., Kmeans, Mclust) have difficulty when outliers are present (Witten & Tibshirani, 2010) and separately when population subgroups are of different sizes (Holden, Finch, & Kelley, 2011), it is important to know the impact of both conditions simultaneously on their ability to correctly identify existing clusters. In addition, prior work has not fully explored the performance of robust methods such as robust clustering, Lasso clustering, and DBSCAN when population clusters are of different sizes. Thus, the current work adds to the literature by examining both of these situations simultaneously, and when noise variables are present as well. Based on prior research (Brodinova, et al., 2019; Ester, Kriegel, Sander, & Xu, 1996), it is hypothesized that robust clustering and DBSCAN will yield the most accurate clustering results when groups are of different sizes and outliers are present, and Kmeans and Mclust will be the least accurate. In addition, it is hypothesized that when there are not outliers present, but some of the variables are unrelated to clustering (i.e., noise), Lasso and robust clustering will be the most accurate at assigning individuals to clusters, with Kmeans, Kmedoids, and Mclust being the least accurate.

Methods

In order to address the research goals outlined above, a Monte Carlo simulation study was conducted. Across all combinations of the manipulated study conditions outlined below, 1000 replications were generated and analyzed. All data generation and analysis was conducted using the R software package,

version 3.6 (R Development Core Team, 2018). Data were generated using the `SimData` function contained in the `wrsk` library. For all conditions, there were 3 population clusters and a total sample size of 120. These settings were selected in order to be representative of data encountered in actual practice, with a relatively modest overall sample size. Indicator variables were simulated from the multivariate standard normal distribution, with a mean of 0 and standard deviation of 1. The following conditions were manipulated in the study.

Number of Predictors

The number of variables was 10, 50, 100, 200, 400, or 800. These values were selected so as to represent a wide array of dimensionality conditions, from relatively low (10) to extremely high (800). Given that a primary goal of the current study was to investigate the performance of these methods in the high dimensional case, it was important to include a variety of such conditions.

Ratio of noise to informative variables

In order to examine the impact of the noise to informative variable (N/I) ratio on the performance of the clustering methods, two conditions were examined: (1) No noise variables were included in the sample (0 noise/1 informative) and (2) the number of noise variables equaled the number of informative variables (1/1). In this latter case, therefore, half of the variables were noise and did not contribute to differentiating the population subgroups. In the former case, all of the variables contributed to group separation.

Group Size Ratio

As noted in the study goals section, a primary purpose of this study was to compare the performance of clustering methods when the number of individuals in each group was substantially different. To this end, two group size ratio conditions were included in this study: 1/1/1 and 1/1/2. In the former condition, all groups were of the same size with 40 individuals in each. In the latter condition, two groups had 30 individuals each, whereas the third group included a sample size of 60. This set of conditions allowed for comparison of the methods in a best case (equal cluster sizes) as well as a more challenging situation (one large and two small clusters).

Proportion of Outlying Observations

The proportion of outlying observations was 0, 0.1, 0.2, 0.3, and 0.4. All outliers were on the informative variables only.

Cluster Separation

Data were generated so that clusters were separated on the informative variables by 1, 2, or 3 standard deviations. As an example, in the 1 standard deviation case, the means for each of the informative variables were separated by 1 standard deviation among the groups. Thus, group 1 had a mean of 0 on each informative variable, whereas group 2 had a mean of 1 on each informative variable, and group 3 had a mean of 2. For the noninformative variables, the groups' means did not differ from one another. The magnitude and direction of the mean differences was consistent across all of the informative variables.

Clustering Methods

Each of the clustering algorithms described above were included in the current study. These methods were Kmeans, Kmedoids, trimmed Kmeans, SPARCL, robust clustering, Mclust, and DBSCAN. With regard to DBSCAN, the default of 5 minimum points and $\varepsilon = 0.2$ were used. For SPARCL, 100 values of the regularization parameter were used and results compared against one another using the BIC. The optimal such parameter was that which resulted in the smallest BIC. For all methods, except DBSCAN the correct number of clusters was assumed. Therefore, Kmeans for example, was used to cluster members of the sample into 3 groups. The exception to this was DBSCAN, which determines the number of clusters itself and cannot be force to a particular solution. Finally, for Kmeans, Kmedoids, and trimmed Kmeans, 10 random starts were used.

Study Outcomes

The study outcomes were overall classification accuracy, and sensitivity (percent of true positives that were classified as such). Given the focus of this study on the ability of the clustering algorithms to correctly identify individuals in the smaller clusters, sensitivity for these specific groups, as well the overall

classification accuracy served as the study outcomes. In order to identify which of the manipulated variables and their interactions contributed to these outcome variables, analysis of variance (ANOVA) was used, in conjunction with the η^2 effect size. Only those manipulated factors and their interactions are described in the results below.

Results

Error Rate

The ANOVA for error rate indicated that the interactions of clustering method by sample size ratio ($F_{7,281}$ = 37.961, p <0.001, η^2 = 0.486), clustering method by N/I ratio by percent of outliers ($F_{28,1136}$ = 22.286, p <0.001, η^2 = 0.355), and clustering method by N/I ratio by cluster separation ($F_{14,564}$ = 1.834, p<0.001, η^2 = 0.277) were all significantly related to the overall error rate. Table 1 displays the error rates by sample size ratio and method. For both ratios, the robust estimator had the lowest overall error rate, whereas Mclust exhibited the highest degree of error. For each method except Mclust, error was slightly higher for the unequal sample size ratio, with the biggest such difference occurring for the trimmed clustering approach.

Table 2 contains the overall error rate by the N/I ratio, proportion of cases that were outliers, and clustering method. When there were no noise variables (N/I ratio = 0/1), all of the methods had somewhat lower error rates than was the case when the number of noise and informative variables were equal. In addition, across clustering methods and N/I ratio, the error rate increased concomitantly with increases in the percent of cases that were outliers. When there were no noise variables and no outliers, the Kmedoids method yielded the lowest error rate, followed by Kmeans and SPARCL. As the percent of outliers increased in value in the no noise condition, the increase in overall error rate was slower for the robust and DBSCAN clustering methods than for the other approaches included here. When the ratio of noise to informative variables was 1/1, the robust approach yielded the lowest error rate across percent of outlier conditions. The trimmed, SPARCL, Kmeans, and Kmedoids techniques all performed similarly to one another in the N/I ratio 1/1 condition, with Mclust having the highest error rate, and DBSCAN the second highest.

Table 3 includes the overall error rate by clustering method, N/I ratio, and cluster separation. Across methods and N/I ratio, the overall error rate was lower when the clusters had greater separation. In the no noise variable condition, the robust approach had a slightly lower error rate than did the other approaches when clusters were separated by 1 standard deviation. This advantage over the other methods was amplified with greater cluster separation, as can be seen in the top panel of Table 3. When cluster separation was 2 or 3 standard deviations, DBSCAN had the second lowest error rate, and Mclust exhibited the highest. When the noise to informative variable ratio was 1/1, the robust approach once again displayed the lowest error rate across group separation conditions, with Mclust having the highest. When group separation was 1 standard deviation and noise variables were present, DBSCAN had the second highest overall error rate, but with higher levels of separation its error rate was second lowest behind the robust approach. Mclust consistently displayed the highest error rate across all conditions.

Sensitivity Rates for Smaller Clusters

As noted previously, unequal cluster sizes has been associated with lower sensitivity for correctly classifying members of that group (Holden, Finch, & Kelley, 2011). Thus, an important outcome in the current study was the sensitivity rate for the smallest clusters. Clusters 2 and 3 (the small clusters when sizes differed) were compared to one another using a *t*-test, and the results were found not to be significantly different. Thus, they were averaged together for the purposes of the following analyses. The ANOVA identified the sample size ratio by group separation by method ($F_{7,282}$ = 15.284, p <0.001, η^2 = 0.275), and method by percent of cases that were outliers ($F_{7,284}$ = 15.645, p <0.001, η^2 = 0.278) as significantly contributing to the sensitivity rate. Table 4 includes the sensitivity rates for the clustering method by the percent of cases that were outliers. Each of the methods exhibited lower sensitivity rates under the condition of a higher percentage of outlying observations. This effect was most marked for Kmeans, which had an 18 point decline in sensitivity from the 0 outliers to the 40% outlier condition. The robust, SPARCL, and Kmedoids techniques each exhibited declines in sensitivity of approximately 10 points. In contrast, DBSCAN exhibited a decline in sensitivity of approximately 7 points, and had the highest values across all

Table 1. Overall Error Rate by Sample Size Ratio and Clustering Method.

Sample size ratio	Robust	Trimmed	SPARCL	Kmeans	Kmedoids	Mclust	DBSCAN
1/1/1	0.156	0.211	0.215	0.212	0.210	0.365	0.225
1/1/2	0.168	0.241	0.229	0.226	0.225	0.360	0.231

Table 2. Overall Error Rate by N/I Ratio, Proportion of Cases that were Outliers, and Clustering Method.

N/I Ratio	Proportion Outliers	Robust	Trimmed	SPARCL	Kmeans	Kmedoids	Mclust	DBSCAN
0/1	0.00	0.099	0.118	0.084	0.082	0.070	0.105	0.107
	0.10	0.102	0.130	0.131	0.127	0.121	0.284	0.125
	0.20	0.140	0.196	0.191	0.186	0.184	0.327	0.168
	0.30	0.144	0.256	0.251	0.246	0.244	0.362	0.210
	0.40	0.185	0.310	0.306	0.303	0.301	0.391	0.252
1/1	0.00	0.097	0.155	0.143	0.140	0.148	0.335	0.205
	0.10	0.129	0.197	0.202	0.200	0.205	0.407	0.247
	0.20	0.132	0.250	0.252	0.250	0.253	0.440	0.290
	0.30	0.213	0.301	0.305	0.303	0.303	0.478	0.321
	0.40	0.252	0.349	0.356	0.355	0.352	0.497	0.353

Table 3. Overall Error Rate by N/I Ratio, Cluster Separation, and Clustering Method.

N/I Ratio	Cluster Separation	Robust	Trimmed	SPARCL	Kmeans	Kmedoids	Mclust	DBSCAN
0/1	1.00	0.250	0.281	0.283	0.274	0.258	0.377	0.268
	2.00	0.113	0.170	0.156	0.154	0.154	0.288	0.132
	3.00	0.082	0.154	0.139	0.139	0.139	0.218	0.118
1/1	1.00	0.273	0.325	0.345	0.339	0.331	0.514	0.485
	2.00	0.143	0.219	0.211	0.211	0.223	0.398	0.192
	3.00	0.113	0.206	0.199	0.199	0.203	0.382	0.174

Table 4. Sensitivity for Clusters 2 and 3 by Proportion of Cases that were Outliers and Clustering Method

Proportion Outliers	Robust	Trimmed	SPARCL	Kmeans	Kmedoids	Mclust	DBSCAN
0.00	70.956	80.656	72.041	74.534	76.132	57.658	83.661
0.10	69.922	74.783	72.005	59.505	75.958	54.646	81.741
0.20	67.659	74.427	71.746	57.389	75.749	54.697	81.504
0.30	65.654	73.533	70.976	56.957	74.462	54.522	81.070
0.40	64.484	73.595	62.473	56.251	65.564	51.751	76.656

Table 5. Sensitivity for Clusters 2 & 3 by Sample Size Ratio, Cluster Separation, and Clustering Method

Sample Size Ratio	Cluster Separation	Robust	Trimmed	SPARCL	Kmeans	Kmedoids	Mclust	DBSCAN
1/1/1	1.00	65.395	70.510	65.160	59.805	71.948	47.736	76.541
	2.00	78.361	80.249	68.874	68.427	78.759	54.401	81.394
	3.00	80.211	82.262	75.352	76.499	81.137	66.174	88.098
1/1/2	1.00	53.746	68.454	59.127	49.802	66.393	43.440	71.774
	2.00	64.502	74.513	67.980	55.230	74.937	50.399	80.534
	3.00	68.817	76.413	70.658	66.122	80.434	60.010	87.453

outlier proportion conditions. Mclust performed the worst with respect to the sensitivity for clusters 2 and 3, regardless of the proportion of outlying observations.

Table 5 displays the sensitivity rates by sample size ratio, degree of cluster separation, and clustering method. As was evident in Table 4, Mclust consistently had the lowest sensitivity rates across the methods studied here, whereas DBSCAN had the largest values. For all methods, sensitivity rose concomitantly with increases in cluster separation, and was higher when the clusters were the same size in the population. The positive impact of increased cluster separation on sensitivity was apparent for all of the methods, but particularly for Mclust, Kmeans, and robust clustering, each of which exhibited increases of at least 15 points in sensitivity from cluster separation of 1 to 3 for both sample size ratio conditions. DBSCAN also exhibited a similarly large increase in sensitivity (almost 16 points) when the sample sizes were unequal, though it had a more modest increase when samples were equally sized.

Empirical example

In order to demonstrate the use of these clustering methods, we will apply them to a well-known dataset, the Wisconsin breast cancer diagnosis data
: (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)).
The data include measurements of 30 cell nuclei taken from 569 patients on whom a breast biopsy was conducted. Such variables as radius, texture, smoothness, compactness, and the like were included in the analysis. For each patient, a true diagnosis of cancer (yes or no) was made, and will serve as the criterion against which the cluster solutions will be compared. All of the clustering algorithms included in the simulation study were applied to the data, and their performance was assessed through overall accuracy, sensitivity (proportion of cases identified as cancerous that actually were), and specificity (proportion of cases identified as cancer free that actually were).

Of the 569 individuals in the sample, 212 (37.3%) were diagnosed with breast cancer. Rates of overall accuracy, sensitivity, and specificity appear in Table 6. With respect to overall accuracy, the trimmed approach had the highest value, followed closely by the robust method and SPARCL. The lowest accuracy rate was associated with Mclust. It is important to note, however, that even this lowest rate was still relatively high, with a proportion of 0.86 correctly classified. In terms of sensitivity for correctly identifying individuals with breast cancer, the robust clustering performed best by far, with a sensitivity rate of 0.99. The next highest value belonged to Mclust at 0.89. The lowest sensitivity was for Kmedoids. The highest specificity rates (correctly identifying individuals without cancer) were associated with SPARCL and the trimmed method, with the lowest specificity occurring for Mclust. It should also be noted that the robust clustering approach identified 52 observations as outliers, 37 of whom were cancer free, and 15 of whom were diagnosed with cancer. The R code used to conduct these analyses appears in the appendix to this manuscript. The data can be obtained from the following link http://www.rpubs.com/kstahl/wdbc_ann, or through the `kdevine` library in R.

Discussion

The goal of this simulation study was to examine the performance of several methods for cluster analysis in cases when the groups were of very unequal size, outliers were present, and when some of the variables used in the analysis were not associated with cluster differentiation in the population (i.e., noise variables). In addition, an empirical example was conducted in order to demonstrate the use of these approaches in practice, so that researchers interested in using them will have some applied context to work with. Results of the simulation study suggest that if the goal of the researcher is to obtain the most accurate classification of individuals overall, the robust clustering method may be the optimal method. When the data had no outliers nor any noise variables, this technique performed nearly as well as the best performers, Kmeans and Kmedoids. On the other hand, as the proportion of outliers increased the error rate of the robust method remained lower than that of other approaches, across all other conditions. In addition, the presence of noise variables was relatively less problematic for the robust approach when compared to the other methods studied here. Its error rate was the lowest across proportion of outliers, and the increase in error with the presence of more outliers was less notable for this approach than for the others.

The second outcome of interest in this study was the sensitivity rate, which measures how well individuals in a particular group were correctly classified. In this case, primary interest was in the smaller

groups in the sample. The simulation results revealed that DBSCAN yielded the most accurate sensitivity values across the methods studied here. In contrast, Mclust consistently yielded the lowest sensitivity rates, followed by Kmeans and the robust clustering algorithm. Thus, although the robust method had high accuracy rates overall, it was not particularly sensitive for the smaller groups in the sample when the groups were of unequal size, particularly for the lowest group separation condition.

Implications for Practice

The results described above present some implications for practice that researchers can take into their own work. First, the presence of outliers is problematic for clustering methods, and this issue is compounded when the groups are of unequal sizes. However, outliers have less of an impact on the robust clustering algorithm, trimmed clustering, and DBSCAN than on other methods studied here. Second, if the researcher's primary focus is on identifying key subgroups in the population, and it is anticipated that these groups are relatively small, then DBSCAN may be the optimal approach to use. Other good methods for this purpose are trimmed and Kmedoids clustering. On the other hand, if the goal is simply to maximize the overall accuracy rate, then the robust clustering algorithm will likely be the optimal method to use. Finally, the current results show that Mclust may have problems working in conditions similar to that used in this simulation study, namely 120 individuals with 3 groups. This sample size and number of clusters may simply be too small for the model based approach to obtain accurate parameter estimates.

Directions forFuture Research

The results of this study suggest several future avenues of work. First, a wider array of sample sizes and cluster structure should be examined. The goal of this study was to focus on the impact of outliers, noise variables, and unequal group sizes when the overall sample was small. For this reason sample sizes were not manipulated in order to keep the size of the study manageable. However, future work should consider other sample size conditions. In addition, different group separation values should also be examined. Values lower than 1, for example, could provide insights into the performance of these methods when clusters are similar to one another on some or all of the variables used in the analysis. Finally, future work should also investigate a wider variety of sample size ratios, some even smaller than those used here.

Conclusions

In practice, researchers are frequently faced with the problem of unequal cluster sizes in the population. Prior work has demonstrated clearly that such differences can lead to problems in correctly classifying individuals into the smaller groups (Holden, Finch, & Kelley, 2011). This problem can become particularly acute when outliers and noise variables are present in the data, and the overall sample size is small. This study was designed to help identify methods that might perform optimally with this challenging set of conditions. Given the results presented above, it appears that researchers should consider using an alternative to the standard Kmeans and Mclust approaches when they believe that the underlying population subgroups are unequal, outliers might be present, and some of the variables might not contribute to cluster separation. Instead, approaches such as robust clustering, DBSCAN, trimmed clustering, or Kmedoids may be preferable. Indeed, even when the data did not contain outliers or noise variables, these alternatives proved to be nearly as accurate at cluster identification as the more widely used Kmeans approach. It is hoped that the simulation study and empirical example provide researchers with insights into an additional set of tools that they can use in practice.

References

Aggarwal, C.C. (2016) *Outlier analysis, 2nd edn*. Springer, Berlin.

Breunig, M.M., Kriegel, H.P., Ng, R.T., & Sander, J. (2000). LOF: identifying density-based local outliers. *Association for Computing Machinery Sigmod Rec 29,* 93–104.

Brodinova, S., Filzmoser, P., Ortner, T., Breiteneder, C., & Rohm, M. (2019). Robust and sparse k-means clustering for high-dimensional data. *Advances in Data Analytics and Classification, 13(4)*, 905-932.

Candel, M.J. & van Breukelen, G.J. (2010). Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Statistics in Medicine, 29(14),* 1488-1501.

Ester, M., Kriegel, H-P, Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).* AAAI Press. 226– 231.

Evans, K., Love, T., & Thurston, S.W. (2015). Outlier identification in model-based cluster analysis. *Journal of Classification, 32(1),* 63-84.

Fraley, C. & Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97,* 611–631.

Galimberti, G., Manisi, A., & Soffritti, G. (2018). Modelling the role of variables in model- based cluster analysis. *Statistical Computing, 18(1),* 145–169

Garcia-Escudero, L.A., & Gordaliza, A. (1999). Robustness properties of k-means and trimmed k-means. *Journal of the American Statistical Association, 94(447),* 956–969.

Hahs-Vaughn, D. (2017). *Applied Multivariate Statistical Concepts.* New York: Routledge, Taylor & Francis group.

Hajnal, I. & Loosveldt, G. (1998). The evaluation of the sensitivity of some clustering techniques to irrelevant variables. *Advances in Methodology, Data analysis, and Statistics, 14,* 61-75.

Henry, D.B., Tolan, P.H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology, 19(1)*, 121-132.

Holden, J. E., Finch, W. H., & Kelley, K. (2011). A Comparison of Two-Group Classification Methods. *Educational and Psychological Measurement, 71,* 870-901.

Kaufman, L. & Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kaufman, L. & Rousseeuw, P.J. (1987). Clustering by means of medoids, in Y. Dodge (Ed.),
Statistical Data Analysis based on the L1 Norm, 405–416.

Lewandowski, K.E., Sperry, S.H., Cohen, B.M., & Ongur, D. (2014). Cognitive variability in psychotic disorders: A cross-diagnostic cluster analysis. Psychological Medicine, 44(15), 3239-3248.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 281–297.

Neykov, N., Filzmoser, P., Dimova, R., & Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational. Statistics and Data Analysis. 52(1),* 299–308.

Prokasky, A., Rudasill, K., Molfese, V.J., Putnam, S., Gartstein, M., & Rothbart, M. (2017). Identifying child temperament types using cluster analysis in three samples. *Journal of Research in Personality, 67*, 190-201.

R Core Team. (2018). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raftery, A.E., & Dean, N. (2006)Variable selection for model-based clustering. *Journal of the American Statistical Association, 101(473)*:168– 178

Rehm, F., Klawonn, F., & Kruse, R. (2007). A novel approach to noise clustering for outlier detection. *Soft Computing, 11(4),* 489-494.

Rocke, D.M. (1996). Robustness properties of S-estimators of multivariate location and shape in high dimension. *Annals of Statistics, 24(3),* 1327–1345.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B, 58(1),* 267-288.

Viswanath, P. & Babu, S. (2009). Rough-DBSCAN: A fast hybrid density based clustering method for large data sets. *Pattern Recognition Letters, 30(16),* 1477-1488.

Send correspondence to:
W Holmes Finch
Ball State University
Email: whfinch@bsu.edu