Journal of Classification 32 (2015)
DOI: 10.1007/s00357-015-9171-5

Outlier Identification in Model-Based Cluster Analysis

Katie Evans

Dupont, DuET Applied Statistics, Delaware USA

Tanzy Love

Dept. of Biostatistics & Computational Biology, University of Rochester, Rochester USA

Sally W. Thurston

Dept. of Biostatistics & Computational Biology, University of Rochester, Rochester USA

Abstract: In model-based clustering based on normal-mixture models, a few outlying observations can influence the cluster structure and number. This paper develops a method to identify these, however it does not attempt to identify clusters amidst a large field of noisy observations. We identify outliers as those observations in a cluster with minimal membership proportion or for which the cluster-specific variance with and without the observation is very different. Results from a simulation study demonstrate the ability of our method to detect true outliers without falsely identifying many non-outliers and improved performance over other approaches, under most scenarios. We use the contributed R package MCLUST for model-based clustering, but propose a modified prior for the cluster-specific variance which avoids degeneracies in estimation procedures. We also compare results from our outlier method to published results on National Hockey League data.

Keywords: Normal-mixture models; Influential points; MCLUST; Prior; National Hockey League.

The work on this paper was completed while K. Evans was finishing her PhD at the University of Rochester, Rochester USA.

Corresponding Author's Address: K. Evans, Dupont, DuET Applied Statistics, Chestnut Run Plaza 723/1144, 947 Centre Road, PO Box 2915, Willmington, DE 19805, Tel: (302) 999-3099, email: keighto@gmail.com.

Published online

1. Introduction

In model-based cluster analysis, a distinction can be made between identifying a few outlying observations in the data versus identifying clusters in the midst of a large field of noisy observations. This paper focuses on the former, where there is no gold standard method for defining, detecting, or handling outliers. In this paper, we present a new method for outlier detection in model-based cluster analysis.

1.1 Clustering Methods

Cluster analysis is a classification problem in which the number and properties of groups within the data are unknown. The goal of clustering methods is to discover an underlying structure in the data which (typically) consists of distinct groups of observations, such that observations within a group are similar to one another in some regard (Gnanadesikan 1989).

Distance-based clustering methods require selection of a distance metric (e.g. Manhattan, Euclidean) and linkage method, some of which may bias the results towards favoring small or large clusters (Gnanadesikan 1989). Instead, we focus on model-based clustering, based on normal mixture modeling. Results from this approach are also subject to the selection and transformation of variables, but not the scale. This method also provides likelihood-based criteria to evaluate and compare candidate models.

Because little is known about the number or variance structure of the groups in the data *a priori*, model-based clustering methods are heavily data-driven and results can be influenced by outliers. Some consequences of clustering with outliers include, but are not limited to, spurious 1-member clusters or variance structures which are unrealistically complex (Rehm, Klawonn, and Kruse 2007). Because these results will complicate the interpretation of clusters, it is important to identify outliers, and to consider their influence on cluster parameter estimates.

1.2 Current Outlier Methods

Dealing with outliers in model-based clustering includes both outlier identification and robust parameter estimation; however, an "outlier" in cluster analysis is not well-defined. In some cases "outlier" and "noise" are used interchangeably to describe numerous incongruous points which cover the entire range of the data (in all dimensions); that is, these points arise due to a random process which is independent of clustering. Many of the existing procedures were developed to identify true clusters amidst a field of noisy observations, that is, where there is a large proportion of outliers. However, since the estimated proportion of outliers can be specified, these methods

should be flexible enough to accommodate a small proportion of outliers, which is the focus of this paper.

We briefly review current approaches, starting with the most unique (which we were unable to compare against competing methods), then review noise approaches, followed by the distance approaches (which use a variety of robust estimation procedures.)

He, Xu, and Deng (2003) do not focus on a small or large proportion of unusual observations, but rather define entire "outlier clusters" if the number of observations in one or more clusters are not dominant in the data set. Shotwell and Slate (2011) propose outlier detection when using the Dirichlet Process Mixture (DPM) for clustering. From the estimated partitions from the DPM method, the outlier procedure is run on a per-cluster basis, where each cluster is then sub-divided into several non-outlier clusters and several outlier (or singleton) clusters. We did not attempt to evaluate these methods.

Breunig, Kriegel, Ng, and Sander (2000) propose a distance-based approach which considers how isolated a point is with respect to it's k nearest-neighbors (k selected by the analyst.) This results in a continuous "local-outlier factor" (LOF) where values below one indicate a clear inlier, but there is no criterion for determining a clear outlier. The extensions include: robustness to choice of k (Kriegel, Kröger, Schubert, and Zimek 2009) or scaling the LOF to be between $[0, 1]$ (Kriegel, Kröger, Schubert, and Zimek 2011). No extension proposes a cutoff criterion to provide a clear determination of outlier status.

Noise approaches: Banfield and Raftery (1993), Hennig (2004), and Coretto and Hennig (2011) propose methods to identify a noise component, while simultaneously clustering the non-noise observations. Banfield and Raftery model the noise component with a Poisson process and focus on robust estimation of cluster parameters. Hennig proposes a robust improper maximum likelihood estimator (RIMLE) method which selects a fixed constant for the noise instead of modeling the noise through a distribution. Coretto and Hennig focus on providing reliable estimators with the Gaussian-uniform mixture. Defining "breakdown" in parameter estimation as the number of outlying observations that can be added to a cluster before the cluster parameters become unrealistic (while keeping the number of clusters fixed), Hennig and Coretto (2008) show that Banfield and Raftery's method has a higher breakdown point than the *t*-mixture proposed by Peel and McLachlan (2000) (which will be discussed shortly).

Distance approaches: Rousseeuw and Van Zomeren (1990), Hardin and Rocke (2002), and Peel and McLachlan (2000) focus on outlier identification by finding robust shape and location parameter estimates, a robust-distance metric for multivariate outlier identification (based on the Mahalanobis Squared Distance (MSD)), and a proposed cut-off to determine bi-

nary outlier status. Rousseeuw and Van Zomeren use parameter estimates based on the Minimum Volume Ellipsoid (MVE), which covers half of the points in a cluster. The robust distance is calculated as the square-root of the MSD and the cut-off is based on the square-root of the $\chi^2_{D,\alpha}$ distribution (where D is the number of variables and increasing values of α increase the strictness of the criterion.) In contrast, Hardin and Rocke use parameters based on the Minimum Covariance Determinant (MCD), the subset of points which minimize the determinant of the covariance matrix. The robust distance is calculated as the MSD and the cut-off is based on the F distribution. Lastly, Peel and McLachlan use a *t*-mixture approach for robust estimation of cluster parameters. Their outlier metric is based on the Mahalanobis distance and the cut-off is based on the χ^2 distribution.

Svensén and Bishop (2005) propose an alternative *t*-mixture approach within a Bayesian framework which focuses on robust cluster assignment and estimation of cluster parameters, and does not provide a metric or criterion for identification of outliers.

Byers and Raftery (1998) and Wang and Raftery (2002) perform robust estimation through nearest-neighbor cleaning procedures. Wang and Raftery describe the bias present in Byers and Raftery and provide a modified procedure to address this behavior. While these methods provide a novel approach to identification of signal versus noise points in data, we have observed poor performance when applied in a setting with a small proportion of atypical observations, as discussed later.

While all of these methods offer important contributions to the topic of robust estimation and/or identification of outlying or noisy observations, the scope of our work focuses on the identification of a few outlying observations which deviate markedly from the rest of the sample or their cluster. These unique points should be further investigated before making any decisions about their inclusion in or exclusion from the analysis. In this setting, we find many of the current methods tend to identify too many observations as outliers.

We propose a post-hoc outlier identification method based on the degree to which a given observation influences the variance parameter estimates, which is cluster-specific and requires no parameter choice to define an outlier. Our method was developed to be applied in situations when at most a small percentage of observations are outliers. We also provide a cut-off to determine whether or not an observation should be flagged as an outlier.

1.3 Data

We illustrate our method using National Hockey League (NHL) data from the 1995-1996 season (as used in Breunig et al. 2000). There are 855

unique players in the database, with 125 of these players traded mid-season and contributing statistics from time spent with more than one team. We will follow the two analyses of Breunig et al., one using Points, Plus/Minus, and Penalty Minutes, and a second using Games Played, Goals Scored, and Goal Percentage. To make the tails less extreme for model-based outlier methods (Banfield and Raftery's and our proposed method), we use a $log(x+1)$ transformation on all variables, except Plus/Minus and Goal Percentage; however, some variables still exhibit non-normality due to an excess of zeros.

The rest of the paper is as follows: In Section 2 we describe existing model-based clustering procedures, our modified prior parametrization, and our new method for outlier identification. In Section 3, we present the results from our simulation study, our application to the NHL data, and a brief discussion. We present our conclusion in Section 4.

2. Methods

2.1 Normal-Mixture Models

We assume a Gaussian mixture model for the data X, with N observations and D variables. For G clusters, the likelihood is

$$\prod_{i=1}^{N} \sum_{k=1}^{G} \tau_k \phi_k(\mathbf{x}_i \mid \mu_k, \mathbf{\Sigma}_k),$$

where τ_k is the probability that an observation belongs to cluster k, ϕ_k is the Normal pdf centered at μ_k with variance-covariance matrix $\mathbf{\Sigma}_k$ (Fraley and Raftery 1999). One implementation of model-based clustering based on normal mixture modeling is the contributed R package MCLUST (Fraley and Raftery 2006). MCLUST provides 10 different constraints for the structure of $\mathbf{\Sigma}_k$ which range from a constant, spherical model for all groups (`EII`) to the cluster-specific unrestrained model (`VVV`).

In MCLUST, the initial clusters are determined via an agglomerative hierarchical clustering based on maximum likelihood criteria for normal-mixture models. From the initial clusters, Maximum Likelihood Estimation is carried out via the Expectation-Maximization (EM) algorithm for parameter estimation. Until convergence occurs, the EM algorithm iterates between the 'E'-step, which computes z_{ik}, the conditional probability that observation i belongs to group k given the current parameter estimates and the 'M'-step, which computes parameter estimates given $\mathbf{z}$. Once estimation is completed for each specified number of clusters and cluster variance structure, a final model (defined by the number of clusters and the cluster variance structure) is selected, based on the largest BIC value.

2.2 Regularization with a Prior

We have observed some instabilities when using MCLUST in practice. The number and structure of clusters are highly dependent on the results of the EM algorithm, which is sensitive to initial values. The initial values used in the EM algorithm can be affected by both the initial hierarchical clustering and specification of a prior. According to MCLUST documentation (Fraley and Raftery 1999), degeneracies and singularities which can interfere with estimation procedures may be avoided by incorporating prior information, without changing the results of models which were previously fit with no prior information. Fraley and Raftery (2007) discuss in detail how the prior is incorporated into MCLUST. When using a prior, the Maximum Likelihood Estimate (MLE) can no longer be used to estimate cluster parameters; instead the posterior mode will serve as the Maximum a Posterior (MAP), which is used in BIC calculations.

In normal-mixture modeling, disperse conjugate priors can be used for the mean and variance. The prior on the multivariate cluster mean (conditional on the cluster variance $\mathbf{\Sigma}_k$) is $\mu_k \mid \mathbf{\Sigma}_k \sim N(\mu_p, \mathbf{\Sigma}_k/\kappa_p)$. The prior for the univariate variance, used for spherical and diagonal variance structures, is $\sigma^2_{ki} \sim Inverse - Gamma(shape = \nu_p/2, scale = \lambda^2_{pi}/2)$ for $i = 1, ..., D$ variables. The prior for ellipsoidal variance structures is $\mathbf{\Sigma}_k \sim Inverse - Wishart(df = \nu_p, scale = \mathbf{\Lambda}_p)$. The hyper-parameters μ_p, κ_p, ν_p, and λ^2_{pi} (or $\mathbf{\Lambda}_p$) represent the mean, shrinkage, degrees of freedom, and scale for the prior distributions. Parameterizations of variance estimators with a prior are detailed in Fraley and Raftery (2007); Celeux and Govaert (1995).

In particular, for the diagonal variance structures (`EEI`, `EVI`, `VEI`, `VVI`), the Inverse Gamma scale parameter used for the variance of the G groups for all D variables, is determined by $\lambda^2_{pi} = trace(\tilde{V})/(D * G^{2/D})$, for all $i = 1, ..., D$ and is a function of the mean empirical variance of the entire data, $\tilde{V}$, calculated without regard to cluster. If the variance of any variable is very different from the variance of each of the other variables, this prior scale parameter serves as a penalty by reduction in the model BIC. Instead, we suggest that each variable has a unique prior scale parameter, such that, for $i = 1, ..., D$,

$$\lambda^2_{pi} = \frac{\tilde{V}_{ii}}{G^{2/D}} \ . \tag{1}$$

When tested in simulation settings, this modified prior preserves the results from models which could be fit without a prior, so that diagonal models are no longer penalized in the situation described above.

2.3 Outlier Identification

Based on our experience that the variance is likely to be the parameter most impacted by the presence of an outlier, we have developed a method for outlier detection based on leave-one-out re-estimation of cluster parameters. If an observation is near the center of the cluster, the cluster variance estimates will be very similar when calculated with or without this observation. However, if an observation lies much further from the center of the cluster to which it was assigned, the variance estimate of that cluster will be inflated when the observation is included.

For implementation, all cluster memberships are held constant and, for each cluster, the selected variance structure is estimated without each cluster member. Because it would be difficult to quantify and interpret changes to both the estimated variance structure and values, if the variance structure changes when an observation is removed, we fix the cluster variance structure at the best estimate without the point in question and re-estimate the value with all points.

For observation i, assigned to cluster g, consider the simple case of computing $R = \hat{V}_{g,-i}\hat{V}_g^{-1}$ from diagonal variance structure $\hat{V}_{g,-i}$, estimated without observation i, and $\hat{V}_g$, estimated with all observations:

$$\hat{V}_{g,-i} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \hat{V}_g = \begin{bmatrix} k_1 & 0 \\ 0 & k_2 \end{bmatrix}. \quad \text{Then,}$$

$$R = \hat{V}_{g,-i}\hat{V}_g^{-1} = I\hat{V}_g^{-1} = \begin{bmatrix} 1/k_2 & 0 \\ 0 & 1/k_1 \end{bmatrix}$$

and the eigenvalues of R will be $1/k_1, 1/k_2$. If no change occurs in the variance estimate when observation i is included, R is the identity matrix with all D eigenvalues equal to 1. Thus, we would expect observations near cluster centers to have eigenvalues near 1. If the inclusion of a point inflates the cluster variance estimate, we expect at least one of the eigenvalues of R to be less than 1. The minimum eigenvalue of R serves as our metric to capture inflation of the variance structure when an observation is included; this value can also be used in determining the degree to which an observation is an outlier. This procedure is outlined in Algorithm 1.

The outlier detection cut-off criterion C is computed separately for each cluster and is calibrated to return one cut-point per cluster. We compute C using the cluster-specific eigenvalues E^g, where $E^g_{(i)}$ is the i^{th} order statistic of the N_g-length vector E^g. Because of skewness in E^g, which has many values near 1 while the few values well below 1 have high variability, we compute thresholds based on trimmed sets of E^g (trimming only the ex-

Algorithm 1: Outline of algorithm for determining outliers via Eigenvalue method

Initialize: Determine initial cluster membership for all observations;
Return: Number of clusters: G and
Cluster variance estimates: $\hat{V}_g$ for $g = 1, ..., G$;
for *Observations* $i = 1, ..., N$ **do**
For observation i assigned to cluster g,
hold other $N-1$ cluster memberships constant;
(A) Determine variance structure when removing observation i;
With N-1 observations, use the M-step (from the EM algorithm) to calculate parameter estimates for all G groups under all specified variance structures Select best variance structure through maximization of the BIC ;
Return: Cluster variance estimate: $\hat{V}_{g,-i}$;
if Selected variance structure differs from original variance structure,
then re-estimate cluster variance $\hat{V}_g$ for $g = 1, .., G$ with all N observations under the best variance structure selected for $N-1$ observations ;
(B) Calculate the minimum eigenvalue E_i^g of R_{-i}, where
$R_{-i} = \hat{V}_{g,-i}\hat{V}_g^{-1}$
Define: the # of observations in cluster g as N_g, trimming value $T = \lceil\sqrt{N/G}\rceil$, and t_j for $j = 1, ..., 5$ to create ~ 4 equally spaced intervals from $1, ..., T$;
for $g = 1, ..., G$ **do**
if $N_g > T$ **then**
Compute thresholds with trimmed cluster-specific eigenvalues E^{g,t_j}
for $j = 1, ..., 5$, **do** $M_{t,g} = Mean(E^{g,t_j}) - 5 * SD(E^{g,t_j})$
for $j = 2, ..., 5$, **do** $M^*_{j,g} = M_{t_j,g}/M_{t_{j-1},g}$
If $\exists(M^*_{\cdot,g}) > (1 + 1/N_g)$
Then $k = \underset{j=2,...,5}{\text{argmax}} j : M^*_{t_j,g} > (1 + 1/N_g)$ and $C_g = M_{t_k,g}$
Else $C_g = M_{1,g}$
else Cluster g denoted as "outlier cluster", $C_g = E^g_{(N_g)}$;
Return: Observations with $\mathbf{E}^g \leq C_g$, for $g = 1, ..., G$

treme minimum values, defined later.) First, we define a maximum trimming value as $T = \lceil\sqrt{N/G}\rceil$. (For reference, $G = 2$, $N = 100, 250$, or 500 yields $T = 8, 12$, or 16, while $G = 4$, $N = 100, 250$, or 500 yields $T = 5, 8$, or 12.) We specify $t_1, ..., t_5$ to be the 5 integers which correspond as closely as possible to 4 equally spaced intervals from $1, ..., T$. Then, for each cluster g, thresholds are based on the cluster-specific trimmed means and standard deviations. That is, for cluster g and cluster-specific eigenvalues E^g, we define a trimmed set of eigenvalues as $E^{g,t} = (E^g_{(t)}, ..., E^g_{(N_g)})$ and then compute 5 cluster-specific thresholds (for $j = 1, ..., 5$) $M_{t,g} = Mean(E^{g,t_j}) - 5 * SD(E^{g,t_j})$. Here, $M_{1,g}$ represents the threshold based on the minimum trimming amount (t_1) and $M_{5,g}$ is based on the maxi-

mum trimming amount. Next, we compute ratios of the thresholds $M_{t,g}$ for $j = 2, ..., 5$ as $M^*_{j,g} = M_{t_j,g}/M_{t_{j-1},g}$. The ratios of equally spaced thresholds (based on trimmed means and standard deviations) define our boundaries for skewness in E^g to accommodate the potential presence of multiple outliers.

If the number of observations in cluster g is less than the trimming value T, then all points in the cluster g are considered outliers and the cluster-specific cut-off is $C_g = E^g_{(N_g)}$. If there are at least T points in cluster g, then, if some j satisfies the criterion $M^*_{j,g} > (1 + 1/N_g)$, $k = \underset{j:j=2,...,5}{\operatorname{argmax}} j : M^*_{t_j,g} > (1 + 1/N_g)$ and $C_g = M_{t_k,g}$; otherwise, if no $M^*_{j,g} > (1 + 1/N_g)$, $C_g = M_{1,g}$.

While it is easy to visually identify a separation in outlier vs non-outlier points when plotting E^g, the criterion C_g is calibrated to provide automatic detection of outliers. The use of thresholds $M_{\cdot,g}$ based on trimmed measurements addresses the skewness in E^g, while the threshold ratios $M^*_{\cdot,g}$ based on (approximately) equally spaced intervals determined by t_j allows for discovery of multiple outliers. We explored using unit intervals $t_j = 1, ..., T$ for computing M^*, but these unit increases only serve to minimize the effect of any one outlier and can lead to lower rates of success for identifying multiple outliers. Lastly, because the distribution of E^g is affected by the number of points in a cluster N_g, we selected $(1 + 1/N_g)$ to be a cluster-specific decision criterion for determining when the ratio of two thresholds M^* is large enough to suggest the presence of one or more outliers.

3. Results

We now present the results of implementing our outlier identification method, first for our simulation study, then for the NHL application.

3.1 Simulation Study

We first describe the design of our simulation study, then discuss consideration of competing methods, and finally present the results for the single and multiple outlier cases.

Simulation Design: The simulation study to examine the empirical properties of our outlier method and competing methods used two real clusters determined by two variables in the data. We considered the following settings: Number of non-outlier observations: (1A) $N = 99$, (1B) $N = 499$; Cluster membership proportion: (2A) Equal (50%, 50%), (2B) Unequal (25%, 75%); Cluster separation: (3A) "Close", Centered at (2,2) and (-2,-2), (3B) "Far", Centered at (3,3), (-3,-3); Cluster shape: (4A) Two spherical

clusters (0 correlation between variables), (4B) Two ellipsoidal clusters with parallel orientation ($-.8$ within-cluster correlation between the variables); Cluster spread: (5A) Equal variance for both clusters (V, V), (5B) Unequal variance for the two clusters (V, $2V$). For outliers, we systemically generated outliers at 4 or 5 SD from cluster means, on or off an axis of cluster variation. We then evaluated our metric by (A) considering 1 outlier at a time, (B) considering 5 or 8 outliers (for N=99 and 499 respectively), and (C) considering 7 or 15 outliers (for N=99 and 499 respectively).

In constructing our simulated data, we confirmed that there were two clusters determined by the non-outlier observations, and by applying our outlier identification method ensured that no observations in the null (non-outlier) data met our outlier criterion. This goal was achieved by simulating more observations than were needed, checking for outliers, replacing any identified outliers with the extra simulated observations, then repeating the process until no outliers were identified. We did *not* restrict MCLUST to only two clusters when determining the best clustering model on the null data plus outlier(s), and sometimes more than two clusters were chosen, especially when considering multiple outliers at one time.

The effectiveness of an outlier identification method is defined by a high rate of success for identifying a true outlier denoted "True+", and a low rate of mis-identifying non-outlier observations, denoted "False+." We learned that total sample size, proportion of cluster membership, shape and variance of clusters, and location of the outlier (in regards to axes of variation) all affect identification of outliers. Additionally, we observed that as the number of outliers in the dataset increased, MCLUST was more likely to select $g + 1$ clusters, where the additional cluster contained all atypical observations.

Eigenvalue Method: For our outlier identification criterion, for each cluster g, there is only one cut-point C_g to be used for outlier identification.

Competing Methods: To compare Rousseeuw and Van Zomeran (1990)'s MVE method and Hardin and Rocke (2002)'s MCD method to our Eigenvalue method, in our simulation study we implemented the MVE and MCD methods on a per-cluster basis to identify cluster-specific outliers (after determining cluster membership via model-based clustering methods). These methods require a significance level (for the χ^2 or F distribution, respectively), so we evaluated these methods at levels of $\alpha = .90, .95, .975, .99$. To compare Banfield and Raftery's Noise method we used a random initialization of noise points at $p_{noise} = .01, .03, .05$ where $p_{noise} = .01$ is the most stringent criteria. (We note that the Noise method is typically used in settings where noisy observations far outnumber non-noisy observations such that $p_{noise} > .50$.) Further discussion of the results from these methods will be given shortly.

We also evaluated Wang and Raftery (2002)'s NNVE procedure (using $p_{noise} = (0, .01, .03, .05)$) and Hennig and Coretto (2008)'s improper noise procedure (using the tuning value $\texttt{nnk}=(0, .01, .03, .05) * N$). For the NNVE procedure, in cases with unequal membership proportion between two clusters, all points in the less-populated cluster were flagged as outliers. For the improper noise procedure, we observed an unacceptably high False Positive rate when applied to our non-outlier data, even for the most stringent criterion. For the focus of this paper, these behaviors are unacceptable and we present no further discussion of these approaches.

We could not directly compare the success and error rates for Bruenig et al. (2000)'s LOF method, so we compared the behavior of our eigenvalues to the LOF values (not shown). In some settings, the eigenvalues and LOF values follow a similar trend, but there are distinct differences in their behavior when the cluster membership proportions are different or the clusters have an ellipsoidal shape. When cluster membership proportions are unequal, there is a clear separation of LOF values between the two clusters, which may lead to an entire cluster being flagged as an outlier cluster. In general, the eigenvalues from our method show a clear visual distinction between non-outlier and outlier points, but the separation is less apparent in LOF values, which is especially problematic as there is no binary decision criterion to determine outlier status from the LOF values.

A direct comparison to Shotwell and Slate (2011) could not be implemented, but we attempted our best approximation by varying the number of groups considered to be between $2 - 30$ and planning to identify any resulting clusters with < 3 observations as "outlier clusters." However, by using BIC maximization to select the number of groups, MCLUST always selected between 2 and 6 groups and no "outlier clusters" could be identified, except for the few instances of an outlier being the sole member of a cluster (which occured in $< 2\%$ of simulations.) In our simulation study, this method would have a very low error rate of identifying False Positives, but would also have a very low success rate of identifying a True Outlier, as these few instances of an outlier being assigned to its own cluster represent the best that Shotwell and Slate could do in this simulation.

Results from a comparison of our method to three other methods: In Figure 1, we present ROC curves illustrating the trade-off between "True+" and "False+", where "True+" is the proportion of data sets with 60% of the true outliers identified. In this multiple outlier setting, for $N = 99$, 2 outliers were placed at 4 SD from the cluster setting and 3 outliers were placed at 5 SD; for $N = 499$, 4 outliers were placed at 4 SD and 4 outliers were placed at 5 SD from the cluster center. This was done for two ellipsoidal or spherical clusters with unequal membership proportion $(75\%, 25\%)$, close separation of clusters, unequal between cluster variance $(2V, V)$, and multi-

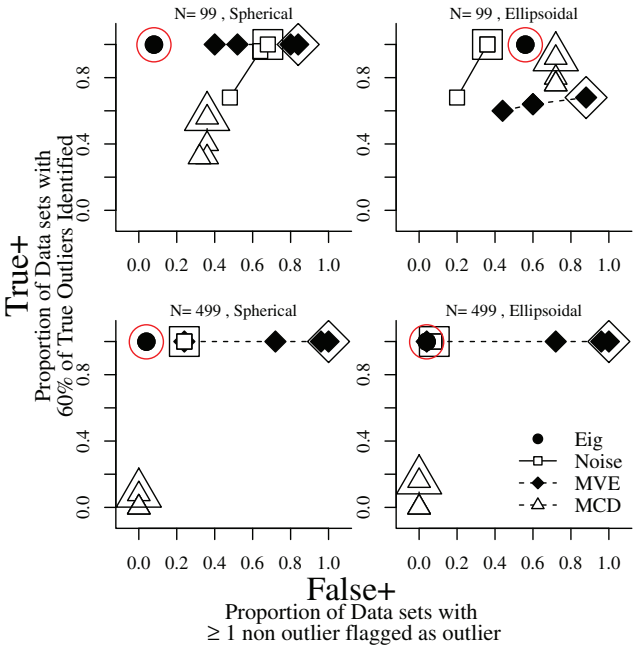

Figure 1. ROC curve of Success vs Error rates for 4 outlier identification methods. Results from N=99, 499 non-outliers, plus 5, 8 outliers (respectively). Outliers randomly placed 4/5SD from a cluster center, on/off axis of cluster variation.
Simulation settings: Unequal Membership Proportion $(75\%, 25\%)$,
Close: Centered at (2,2), (-2,-2), Unequal Between Cluster Variance $(2V, V)$.
Eig = Eigenvalue, Noise = Banfield and Raftery (Plotted for $p_{Noise} = .01, .03, .05$),
MVE = Rousseeuw and Van Zomeran (Plotted for $\alpha = .99, .975, .95, .90$),
MCD = Hardin and Rocke (Plotted for $\alpha = .99, .975, .95, .90$).
Outlined point character denotes selected criterion level.

ple outliers. Curves are plotted for our Eigenvalue method, Banfield and Raftery's Noise method, Rousseeuw and Van Zomeren's MVE method, and Hardin and Rocke's MCD method. The best method will be near the upper left-hand corner of the plot, representing high frequency of success with low frequency of errors.

In general, we found the MVE achieves high success when identifying the true outlier (especially with the larger sample size of $N = 500$), but the outlier cut-off criterion appeared too lenient as the proportion of data sets with at least one false positive was nearly 100% in every simulation setting and the number of false positives was also much larger than for any other method. We attempted to recalibrate the $\sqrt{\chi^2_{D,\alpha}}$ cut-off to $\sqrt{\chi^2_{2D,\alpha}}$, but this substantially decreased the success rate while decreasing the errors.

In contrast to the MVE approach, we found the MCD approach to be too conservative, especially at the smaller sample size of $N = 100$. (This method needs at least 25 observations per cluster to properly calculate the outlier cut-off criterion.) At the larger sample size for the single outlier case (results not shown), the MCD approach only achieved reasonable success in identifying true outliers at 5 SD, but did not outperform any other method.

Lastly, Banfield and Raftery's Noise typically had a high rate of identifying true outliers with a modest error rate, in comparison to the other competing methods. For the remainder of this section, results are only presented for the Eigenvalue method and Banfield and Raftery's Noise method ($p_{noise} = .03$), which was the alternative method most competitive against ours.

Simulation Results for a Single Outlier: Outlier detection results are discussed in detail for the effect of sample size, outlier placement, and cluster shape while fixing unequal membership proportion (75% vs 25%), close separation of clusters (centered at (2,2) and (-2,-2)), and equal between-cluster variability. These settings were selected because results were consistent when adding outliers to either cluster. For two spherical clusters, there was, as expected, essentially no difference in results for an outlier placed on versus off an axis of cluster variation, so those results are summarized over location.

When $N = 100$, our method can detect the true outlier at only 4 standard deviations away from the cluster center $89 - 100\%$ of the time, depending on cluster shape and outlier location. These rates change to $93 - 99\%$ when outliers are 5 standard deviations away. For $N = 500$ and an outlier at 4 SD, success rates of detecting one true outlier increase to $98 - 100\%$. For $N = 500$ and outliers at 5 SD, success rates are $99 - 100\%$. Error rates in this setting range from $10 - 34\%$ for $N = 100$ and $0 - 1\%$ for $N = 500$, with fewer than 3 false positives, on average, identified per occurrence.

In the same setting we used to report our results, the Banfield and Raftery's Noise approach identifies at least one false positive in $14 - 85\%$ of data sets, with, on average, $1 - 10$ false positives per occurrence at $N = 100$. (Results are not shown for other values of p_{noise}, but even at the most conservative $p_{noise} = .01$, false positives occur in $9 - 57\%$ with similar numbers of false positives per occurrence.) We note that the performance of Banfield and Raftery's Noise approach was quite variable in the single outlier case, sometimes experiencing only half the success rate of our method (compare the Ellipsoidal results for $N = 500$ in Table 1) and sometimes achieving a marginally higher success rate. Additionally, for two spherical clusters, the rate of false positives, as well as the number of false positives per occurrence, was nearly always higher than our method; for two ellipsoidal clusters, the two methods had comparable false positives rates.

Table 1. Success and Error rates for Single Outlier Identification for proposed Eigenvalue and Banfield & Raftery's Noise method. True+ reports proportion of datasets in which the 1 true outlier was correctly identified. False+ reports proportion of datasets in which at least 1 false positive error was made, with the conditional average number of false positives made shown in (). Simulation settings: Unequal Membership Proportion $(75\%, 25\%)$, Close: Centered at (2,2), (-2,-2), Equal Between Cluster Variance.

		Eig				Noise (p=.03)			
Cluster	Outlier	N=100		N=500		N=100		N=500	
Shape	Location	True+	False+	True+	False+	True+	False+	True+	False+
Spherical	4 SD	0.97	0.1(1.36)	0.98	0.01(2)	0.95	0.85(9.66)	1	0.46(3.08)
	5 SD	0.99	0.14(1.74)	0.99	0.01(2)	0.97	0.59(6.5)	1	0.15(1.69)
Ellipsoidal	4 SD On	0.89	0.2(1.77)	1	0(-)	0.71	0.14(1.54)	0.44	0(-)
	4 SD Off	1	0.26(1.38)	1	0.01(1)	0.95	0.21(1.99)	0.57	0(-)
	5 SD On	0.93	0.28(2.19)	1	0.01(3)	0.92	0.12(1.43)	1	0(-)
	5 SD Off	0.89	0.34(2.23)	1	0.01(2)	0.97	0.19(2.01)	1	0(-)

Simulation Results for Multiple Outliers: Next, we considered a modest number of multiple outliers in our simulated data (5 outliers for 99 non-outliers or 8 outliers for 499 non-outliers.) In summarizing this data, we were interested in (i) the proportion of data sets in which at least 60% of the true outliers were identified (3 or 5 for N=99, 499, respectively), (ii) the average number of true outliers identified, and (iii) the frequency and severity of false positives. In Table 2 we report the results for our method and the Noise method from the simulation setting for close clusters (centered at (2,2), (-2,-2)) with unequal cluster variance (2V, V), for each setting of varied membership proportion, spherical or ellipsoidal clusters, and different sample sizes. These results demonstrate that our method is often (though not always) more conservative than the Noise method. This is seen through reduced True and False positive rates and (typically) a smaller average number of True or False positive outliers identified.

Lastly, we considered adding more outliers to our simulated data, specifically 8 outliers for 99 non-outliers or 15 outliers for 499 non-outlier points. For outlier identification methods which rely on pre-determination of clusters (such as our method, MVE, and MCD), this setting was problematic, especially for the smaller sample size, which consisted of $\sim 8\%$ outliers. Here, Mclust struggled to find the true clusters, most often relegating all outliers to a third cluster ($> 50\%$ of the time), but often selecting > 3 clusters ($\sim 20\%$ of the time.) For the larger sample size, `Mclust` selected three clusters most often (87% of the time), but still struggled with > 3 clusters (12% of the time). In these instances of selecting 3 clusters in the data,

Table 2. Success and Error rates for Multiple Outlier Identification for proposed Eigenvalue and Banfield & Raftery's Noise method. True+ reports proportion of datasets in which $> 60\%$ of true outliers were correctly identified, with the average number of true outliers identified in (). For $N = 104$: 99 non-outliers, 5 outliers; $N = 507$: 499 non-outliers, 8 outliers. False+ reports proportion of datasets in which at least 1 false positive error was made, with the conditional average number of false positives made shown in (). Close clusters centered at (2,2), (-2,-2), Unequal Between Cluster Variance ($2V, V$.)

		Eig				Noise (p=.03)			
Membership	Cluster	N=104		N=507		N=104		N=507	
Proportion	Shape	True+	False+	True+	False+	True+	False+	True+	False+
(50%, 50%)	Spherical	0.96(4.48)	0.2(1)	0.92(5.64)	0.16(3)	0.92(4.6)	0.48(2)	1(6.04)	0.08(1)
	Ellipsoidal	1(4.76)	0.2(1.8)	1(8)	0.04(3)	1(5)	0.32(3)	1(8)	0(-)
(25%, 75%)	Spherical	1(4.84)	0.08(1)	1(6)	0(-)	0.96(4.76)	0.72(3.11)	1(6.52)	0.16(1.5)
	Ellipsoidal	1(4.56)	0.52(1.69)	1(8)	0.04(1)	0.96(4.8)	0.68(3.53)	1(8)	0.24(1.17)
(75%, 25%)	Spherical	0.84(3.92)	0.36(1.11)	0.92(5.76)	0.32(2)	0.88(4.4)	0.56(2.21)	1(6.04)	0(-)
	Ellipsoidal	1(4.88)	0.32(2.25)	1(7.88)	0.12(1.33)	0.96(4.8)	0.48(1.92)	1(8)	0(-)

it was most frequently the case that the two clusters of non-outlier observations were clearly identified, and the third cluster had an inflated variance and covered all outlier observations. The choice of additional clusters to model the outliers was also present in the first multiple outlier setting, but to a much lesser degree.

Use of an additional component to accommodate the outliers is reminscient of the the Noise approach, where the atypical observations are modeled by a normal component as compared to a Poisson component. However, in this instance, one additional normal component is often not enough to account for the degree of variability in the data, which is why > 3 normal components may be needed. In contrast, a Poisson component (in addition to the two normal clusters) can account for a greater degree of variability in the data and often this one additional component is all that is necessary.

This illustrates the shift from what we consider to be "outliers" versus "noise" in the data; demonstrating that our method is better suited to identify outliers, while a Noise approach is better suited to identify true clusters amidst a field of noisy observations.

3.2 Hockey Data

We followed two of the analyses of the National Hockey League data (with 855 observations) as performed by Breunig et al (2000). Our first analysis used Points, Plus-Minus, and Penalty Minutes as clustering variables, with a $log(x + 1)$ transformation of Points and Penalty Minutes to make the tails less extreme. We employed this transformation because the model-

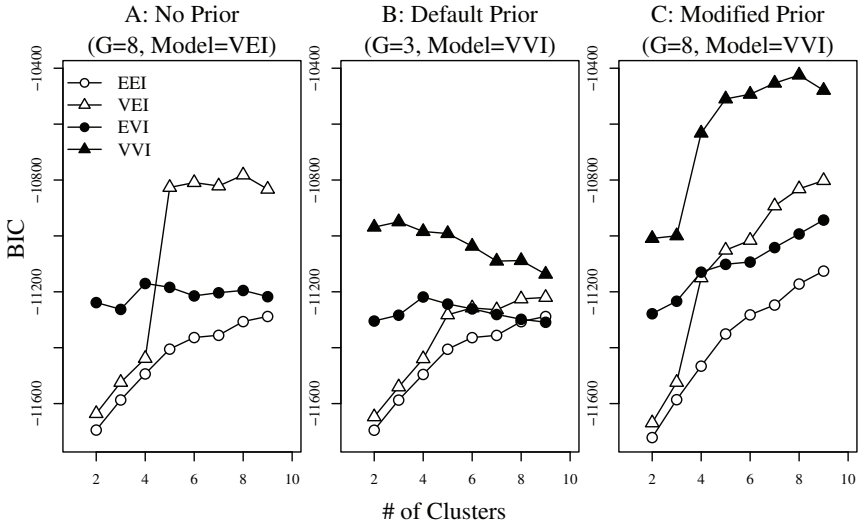

Figure 2. Effect of No Prior, Default Prior, Modified Prior for National Hockey League Data BIC of Model Fits for National Hockey League data. Only Diagonal Models shown to demonstrate effect of prior on BIC (BIC is used to select model-variance structure and number of clusters). Subtitles identify the # of clusters, cluster variance structure which maximize BIC. When a cluster-variance structure cannot be fit, it does not appear in the figure.

based framework is based on the assumption of normally distributed variables, but it is unclear which (if any) transformation Breunig et al. (2000) used.

In Figure 2, we present BIC plots for the diagonal model structures with no prior, the default prior, and our modified prior (given in Eq 1). The BIC was ultimately maximized at eight clusters, with marginal increase after five clusters for the most flexible ellipsoidal VVV variance structure (BIC for this structure not shown.) We report the back-transformed means of the selected model with five clusters (shown with percentage of players in each cluster) in Table 3A.

Outlier identification flagged Owen Nolan in Cluster 1, Peter White in Cluster 2, Doug Brown and Dave Reid in Cluster 4, and Sergei Federov and Vladimir Konstantinov in Cluster 5 (Figure 3A). Nolan has the lowest PlusMinus value (-33) in his cluster while also having the highest number of Points (~ 5), making his two-dimension performance atypical as compared to the other members in the cluster. White is clearly separated from the other members of his cluster by having the lowest Plus-Minus value (-14) and the fewest Penalty Minutes (0), but he also does not fit well with any of the other groups. Reid and Brown take the extreme minimum value

Table 3. National Hockey League Cluster Results: Membership proportion and back-transformed means for each cluster for the two analyses performed on the National Hockey League Data. Clustering variables for Analysis (A): Points, Plus/Minus, and Penalty Minutes. Clustering variables for Analysis (B): Goals, Games, and Goal. %

	Cluster	1	2	3	4	5
A:	Membership %	41	21	15	15	7
	Points	16.87	2.29	34.61	0	51.02
	Plus/Minus	-7.6	-0.82	7.66	-0.44	23.25
	Penalty Mins	54.66	6.02	47.99	2.03	72.63
	Cluster	1	2	3	4	
B:	Membership %	35	27	23	15	
	Goals	15.6	0	3.42	2.17	
	Games	73.31	9.66	57.52	17.3	
	Goal %	11.82	0	5.86	15.46	

in their cluster with the fewest Penalty Minutes. Federov and Konstantinov were assigned to Cluster 5 and have the largest Plus-Minus value ($49, 60$, respectively) in their cluster (and the entire data set) and Konstantinov's Points (34) was also below average within his cluster. The Plus-Minus statistic captures the number of goals scored for versus against a player's team while a player is on the ice, such that large, positive Plus-Minus values tend to occur when a player has a high number of Points. It is unusual for Konstantinov to have such a large, positive Plus-Minus value, with such a modest number of Points.

Of these flagged outliers, it is clear that White, Reid, Federov, and Konstantinov exhibit atypical behavior, either for their cluster or for the entire data set; while Nolan and Brown appear to be better explained by being on the perimeter of their assigned clusters.

Despite not knowing whether transformed or untransformed data was used by Breunig et al. (2000), they also report Konstantinov as a potential outlier. (We implemented the LOF method on the transformed data and observed that Konstantinov had the maximum LOF value.) Breunig et al. (2000) also noted that Matthew Barnaby may be an outlier, but in our analysis using the transformed data, neither our method nor our LOF implementation were able to identify him. Barnaby (assigned to Cluster 1) had the maximum number of Penalty Minutes in the data set (335), but the large variance of this cluster is able to account for Barnaby's large value.

For our second analysis, using Goals Scored, Games Played, and Goal Percentage (number of goals made out of number of shots taken) as clustering variables, we use a $log(x + 1)$ transformation on Goals Scored and Games Played. The back-transformed means of the four clusters (shown with percentage of players in each cluster) are in Table 3B.

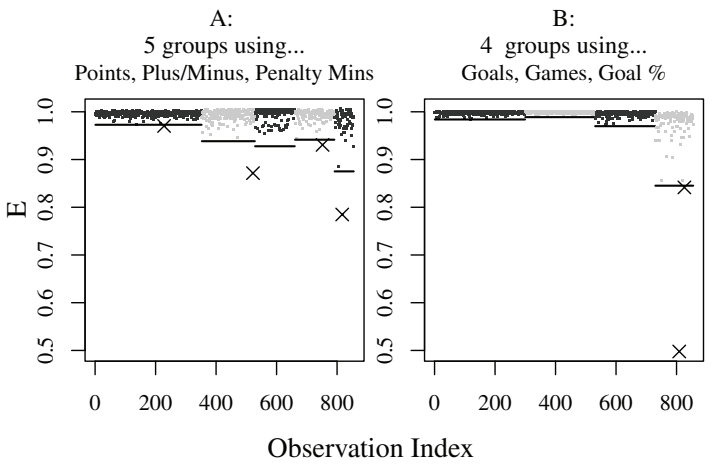

Figure 3. Outlier Identification Metric for National Hockey League Analyses. Greyscale distinguishes groups, outliers marked with $\times$

In this analysis, we identified Craig Janney in Cluster 1 and Larry Courville, Chris Osgood, Steve Poapst, and Gary Roberts in Cluster 4 as potential outliers (Figure 3B). Janney is barely below the outlier threshold and closer inspection of the data strongly suggests not flagging him as an outlier, but merely considering him on the perimeter of his cluster. Courville, Poapst, and Roberts all exhibit atypical behavior when jointly considering their Goals and Goal Percentage. Courville and Poapst have a small number of Goals with a high Goal Percentage, while Roberts has a high Goal Percentage and few Goals compared to other observations in the cluster. The outlier status of Osgood can be explained by the fact that in his 50 games played, he was the only player with a Goal Percentage of 100%. (In another analysis of this data, Knorr and Ng (1998) state that few players have a Goal Percentage higher than 20%.) However, because Osgood is a goalie and rarely has the opportunity to take shots, it is unique that he scored on the only shot he took. Breunig et al. (2000) also note that Mario Lemieux (assigned to Cluster 1) may be an outlier, but neither our method nor our implementation of Breunig et al. (2000)'s LOF method on the transformed data determined anything unusual about his statistics. Lemieux had the maximum number of Goals in his cluster (69), but, again, the large variance of this group was able to account for his large value.

We applied Banfield and Raftery's Noise method to the transformed versions of the data and did not identify any outliers in either analysis at the most lenient $p_{noise} = .05$. We also applied Rousseeuw and Van Zomeren's MVE approach and Hardin and Rocke's MCD approach to both versions

of the transformed data, but both of these methods encountered errors and failed to complete when there was a cluster in which the interquartile range was equal to zero for some variable. In the first analysis, this occurred in Cluster 3, with 132 players who all had 0 Points. In the second analysis, this occurred in Cluster 2, with 231 players all scoring 0 Goals.

3.3 Discussion

We examined the distributions of the Mahalanobis Distance of outlier and non-outlier observations to their cluster center in our simulation study to better understand mis-identified observations (i.e. true outliers which were not flagged, non-outliers which were flagged). Due to similarity of distances under various simulation settings, we only discuss distances from simulations with $N = 500$, 2 spherical clusters, unequal membership proportion, far separation of clusters, and equal between-cluster variability for the single-outlier setting. Our data points simulated to be non-outliers in the null setting have a mean MSD 1.96 (SD=1.84), while outliers placed at 4 SD have a mean MSD 16.50 (SD=1.71) and outliers placed at 5 SD a mean MSD 25.23 (SD=2.51).

Aggregating errors over all criterion levels, the distribution of the null distances for false positive errors has a mean MSD 8.22 (SD=2.28), which is markedly larger than other non-outliers under the same simulation settings. These points are not extreme enough to be identified as outliers in the null setting and can still be accommodated by the cluster variance. However, due to the nature of model-based clustering and model-selection criteria, the estimated variance structure of a cluster may change when one outlier is added to the data. Thus, a point that may have been on the cluster perimeter, but was still accommodated by the cluster variance, may not fit as well if a new variance structure is selected in the presence of the true outlier. For errors of missing the true outlier, which only happened 3 (out of 150) times in this setting, the distances of these points has a mean MSD of 16.23 (SD=2.19).

When adding a single outlier, for the smaller sample size (99 non-outliers), there was some minor variation in success and failure rates across cluster settings, but for the larger sample size (499 non-outliers) these rates were consistent across cluster settings. We also observed a higher rate of false positives for the spherical clusters than for ellipsoidal clusters. We attribute this effect to the feature of the spherical clusters being less separated (than the ellipsoidal clusters) so that when an outlier is added to one cluster, it may be too close to a point in the other cluster, which may be identified as a false positive. When adding multiple outliers, our results were largely impacted by the number of groups selected by MCLUST via the BIC.

We have noted that there are some settings in which Banfield and Raftery (1993)'s method out-performs our method in terms of success rate,

error rate, or both. In particular, when N=500, with two spherical clusters placed far apart with equal between-cluster variance, the Noise approach successfully identifies outliers placed at $4SD$ with higher frequency, while committing the same or lower frequency of errors for most criterion levels. Despite the better performance of Banfield and Raftery's method in this setting, we believe that our method provides additional information which can guide the user in a more thorough analysis. In particular, our method provides a cluster assignment for the observation in question, as well as a continuous metric to aid in determining how outlying the observation is.

In addition to providing a decision criterion for our method, we also recommend post-hoc visualizations to guide the user's decision on the inclusion or exclusion of an outlier. First, it is valuable to plot the eigenvalues (by cluster assignment) to visualize the distribution (shown in Figure 3 for the NHL analyses). Second, it is helpful to examine paired scatter plots of clustering variables to examine their joint distributions and to give explanation to the observation's outlier status.

4. Conclusion

This work has focused on outlier identification by quantifying the effect an observation has on cluster variance parameter estimates to define outlier status. By focusing on identifying a small proportion of outliers assigned to more densely populated (or well-defined) clusters, we address the behavior of model-based clustering and model-selection methods which may select a model which assigned an observation to a more densely populated cluster, even if it does not fit well within any of the existing clusters. This work focuses on identification of outlying values and leaves it to the discretion of the analyst to determine whether these outliers should be excluded from the analysis.

Other simulation studies for outlier detection in cluster analysis use sample sizes in the range of $N = 600 - 1700$ (Breunig et al. 2000; Hardin and Rocke 2002). Our method has been successful in detecting outliers even for a small sample size of $N = 100$ (99 non-outliers with 1 true outlier). Lastly, our approach is a unique combination of methods in that it does not use distances to define outlier status and it can identify outliers which were assigned to an existing cluster.

Many current approaches were developed to identify clusters amidst a large field of noisy observations, but these methods also offer parameter choices which allow for the detection of a small proportion of outliers. However, in our experience, the cut-offs for these methods do not behave as well for a small proportion of outliers as they do when a large proportion of outliers is present. When implementing methods proposed by Rousseeuw and

Van Zomeren (1990), Hardin and Rocke (2002), Byers and Raftery (1998), and Wang and Raftery (2002) we experienced a need for better calibration of the outlier decision criteria in this setting defined by a small proportion of outliers. We also occasionally encountered errors when cluster membership became too small or many members in a cluster had a constant value for some variable. Banfield and Raftery's Noise method was generally most competitive against ours, though it is difficult to recommend in all settings, as there was no apparent trend in which cluster features most affected Success and Error rates of this Noise method.

As compared to existing approaches, our Eigenvalue approach for outlier identification provides the most consistent ability to detect True Outliers across varying cluster features, with a smaller or comparable frequency of False Positive outlier identifications. In the NHL analyses, 11 outliers were identified; 8 of which substantially changed the parameter estimates for their respective clusters.

References

BANFIELD, J., and RAFTERY, A. (1993), "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics, 49(3)*, 803–821.

BREUNIG, M., KRIEGEL, H., NG, R., and SANDER, J. (2000), "LOF: Identifying Density-Based Local Outliers", *Sigmod Record, 29(2)*, 93–104.

BYERS, S., and RAFTERY, A. E. (1998), "Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes", *Journal of the American Statistical Association, 93(442),* 577–584.

CELEUX, G., and GOVAERT, G. (1995), "Gaussian Parsimonious Clustering Models", *Pattern Recognition, 28(5),* 781–793.

CORETTO, P., and HENNIG, C. (2011), "Maximum Likelihood Estimation of Heterogeneous Mixtures of Gaussian and Uniform Distributions," *Journal of Statistical Planning and Inference, 141(1),* 462–473.

FRALEY, C., and RAFTERY, A. (1999), "MCLUST: Software for Model-Based Cluster Analysis", *Journal of Classification, 16(2),* 297–306.

FRALEY, C., and RAFTERY, A. (2006), "MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering", Technical report, University of Washington.

FRALEY, C., and RAFTERY, A. (2007), "Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering", *Journal of Classification, 24(2),* 155–181.

GNANADESIKAN, R. (1989), "Discriminant Analysis and Clustering: Panel on Discriminant Analysis, Classification, and Clustering", *Statistical Science, 4(1)*, 34–69.

HARDIN, J., and ROCKE, D. (2002), "Outlier Detection in the Multiple Cluster Setting Using the Minimum Covariance Determinant Estimator", *Computational Statistics and Data Analysis, 44(4)*, 625–638.

HE, Z., XU, X., and DENG, S. (2003), "Discovering Cluster-Based Local Outliers", *Pattern Recognition Letters, 24(9-10)*, 1641–1650.

HENNIG, C. (2004), "Breakdown Points for Maximum Likelihood Estimators of Location-Scale Mixtures", *The Annals of Statistics, 32(4)*, 1313–1340.

HENNIG, C., and CORETTO, P. (2008), "The Noise Component in Model-Based Cluster Analysis", in *Data Analysis, Machine Learning and Applications*, eds. C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Berlin Heidelberg: Springer, pp. 127–138.

KNORR, E., and NG, R. (1998), "Algorithms for Mining Distance-Based Outliers in Large Datasets", in *VLDB '98 Proceedings of the 24th International Conference on Very Large Data Bases*, San Francisco: Morgan Kaufmann, pp. 392–403.

KRIEGEL, H., KRÖGER, P., SCHUBERT, E., and ZIMEK, A. (2009), "Loop: Local Outlier Probabilities", in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, New York: ACM, pp. 1649–1652.

KRIEGEL, H., KRÖGER, P., SCHUBERT, E., and ZIMEK, A. (2011), "Interpreting and Unifying Outlier Scores", in *Proceedings of the SIAM International Conference on Data Mining*, pp. 13–24.

PEEL, D., and MCLACHLAN, G.J. (2000), "Robust Mixture Modelling Using the t Distribution", *Statistics and Computing, 10(4)*, 339–348.

REHM, F., KLAWONN, F., and KRUSE, R. (2007), "A Novel Approach to Noise Clustering for Outlier Detection", *Soft Computing-A Fusion of Foundations, Methodologies and Applications, 11(5)*, 489–494.

ROUSSEEUW, P., and VAN ZOMEREN, B. (1990), "Unmasking Multivariate Outliers and Leverage Points", *Journal of the American Statistical Association, 85(411)*, 633–639.

SHOTWELL, M., and SLATE, E. (2011), "Bayesian Outlier Detection with Dirichlet Process Mixtures", *Bayesian Analysis, 6(4)*, 665–690.

SVENSÉN, M., and BISHOP, C. (2005), "Robust Bayesian Mixture Modelling", *Neurocomputing, 64*, 235–252.

WANG, N., and RAFTERY, A.E. (2002), "Nearest-Neighbor Variance Estimation (NNVE)", *Journal of the American Statistical Association, 97(460)*, 994–1019.