Clustering in the Presence of Scatter

Ranjan Maitra*
Department of Statistics, Iowa State University, IA 50011-1210, USA
**email:* maitra@iastate.edu

and

Ivan P. Ramler*
Department of Statistics, Iowa State University, IA 50011-1210, USA
**email:* ramleri@iastate.edu

SUMMARY: A new methodology is proposed for clustering datasets in the presence of scattered observations. Scattered observations are defined as unlike any other, so traditional approaches that force them into groups can lead to erroneous conclusions. Our suggested approach is a scheme which, under assumption of homogeneous spherical clusters, iteratively builds cores around their centers and groups points within each core while identifying points outside as scatter. In the absence of scatter, the algorithm reduces to k-means. We also provide methodology to initialize the algorithm and to estimate the number of clusters in the dataset. Results in experimental situations show excellent performance, especially when clusters are elliptically symmetric. The methodology is applied to the analysis of the United States Environmental Protection Agency (EPA)'s Toxic Release Inventory (TRI) reports on industrial releases of mercury for the year 2000.

KEY WORDS: Bayes Information Criterion, biweight estimator, exact-c-separation, k-clips, MCLUST, methylmercury, tight clustering

1. Introduction

Clustering or finding groups of similar observations in datasets is a well-studied issue in the statistics literature (Everitt *et al*, 2001; Fraley and Raftery, 2002; Hartigan, 1985; Kaufman and Rousseeuw, 1990; Kettenring, 2006; McLachlan and Bashford, 1988; Murtagh, 1985; Ramey, 1985). Some approaches are hierarchical (either agglomerative or *bottom-up*, or divisive or *top-down*) while others are non-hierarchical, optimally partitioning data using parametric assumptions and maximizing a loglikelihood function or minimizing some measure of distortion – such as the trace or determinant of the within-sum-of-squares-and-cross-products ($\boldsymbol{W}$) matrix (Freidman and Rubin, 1967; Scott and Symons, 1971) – of identified clusters. Implementation is usually via locally optimal algorithms such as k-means or its generalizations such as k-medoids (Chapter 2 of Kaufman and Rousseeuw, 1990), or by applying the expectation-maximization (EM) of a specified (typically Gaussian) mixture model (Celeux and Govaert, 1995; Fraley and Raftery, 1998; Fraley and Raftery, 2002; McLachlan and Peel, 2000).

Most common partitioning algorithms group all observations, an undesirable feature in datasets with *scatter points* (defined to be observations unlike any other in the dataset). Identifying such observations may be of scientific interest and including them in clusters with other observations may bias group summaries (such as means or variability measures) and other conclusions. Tseng and Wong (2005) give several examples of poor performance when clustering algorithms forcefully cluster every observation. The implications can be important, such as in the public health application which we discuss next.

1.1 *Categorizing Industrial Facilities that Release Mercury*

Public health officials have long been concerned about the adverse effects of prenatal mercury exposure on physiological and neurological development in children (Grandjean *et al*, 1997; Kjellstrom *et al*, 1986; National Academy of Sciences, 2000; Sörensen *et al*, 1999). The element is believed to originate in the environment through releases from industrial facilities and to travel long distances on atmospheric air currents, contaminating even distant areas (Fitzgerald *et al*, 1998). Bacteria convert some mercury to the even more lethal methylmercury which is transferred to humans through the consumption of contaminated fish and seafood. There is no safe limit for the latter chemical and it is more easily absorbed by the body. Developing effective policies that limit mercury releases is thus crucial, but requires a detailed understanding of what and how industries release mercury.

The TRI database for the year 2000 contains data on releases, in pounds, of mercury and mercury compounds to fugitive and stack air emissions, water, land, underground injection into wells and off-site disposals as reported by 1,409 eligible facilities from twenty-four different kinds of industries. A large proportion (33%) of these facilities combust fossil fuels to generate electricity while another 12% manufacture or process chemicals.

An unsupervised method of grouping similar facilities is an important tool for studying the characteristics of mercury releases. Facilities with similar patterns can be grouped together to better understand their effects on the environment and public health and lead to the framing of public policy targeted for maximum effect. However, a striking feature of the dataset is the number of reports that are unlike any other. Including them in other groups could skew results, obscure common characteristics and undermine the effectiveness of devised policies: there is thus a crucial need for an algorithm that accounts for scatter while clustering.

1.2 *Background and Related Work*

There are few existing methods that address the issue of clustering in the presence of scatter. A model-based approach using EM on a mixture-of-Gaussians augmented with a Poisson Process component for scatter (Banfield and Raftery, 1993; Dasgupta and Raftery, 1998; Fraley and Raftery, 1998) is implemented as part of the R package MCLUST (Fraley and Raftery, 2006). Unfortunately, the EM algorithm can be notoriously slow and difficult to apply to larger datasets. The density-based hierarchical clustering method (DHC) of Jiang *et al* (2003) extracts interesting patterns from noisy datasets without actually identifying scatter. The adaptive quality-based clustering (Adapt_Clust) of De Smet *et. al* (2002) iteratively finds cluster cores in zero-mean-unit-variance standardized data up to some terminating criterion, whereupon it classifies the remaining observations as scatter. Tseng and Wong (2005) developed Tight Clustering (hereafter, TCTW) by using resampling to sequentially determine cluster cores up to a targeted number (K_T) or until no further *tight and stable* cluster can be identified, at which point the rest are labeled as scatter. Along with K_T, the algorithm requires specifications for tightness (α in their paper, but we use α_T here), stability (β) and another tuning parameter (k_0) which severely influences performance. The last is set at one- to two-fold the true number of clusters, which is usually not known a priori. A data-driven approach to parameter estimation is not natural to implement because unlike model-based methods, the algorithm has no obvious objective function. Another drawback of this algorithm is that the sequential cluster identification does not account for already identified groupings, which can result in partitions that violate the distance assumption of the metric used (see Figure 1e for an illustration).

The optimal number of clusters (K) in a dataset often needs to be estimated in many practical settings. Several methods (Marriott, 1971; Milligan and Cooper, 1985; Tibshirani *et al*, 2001; Tibshirani and Walther, 2005) exist, but no method is viewed as clearly superior to the others. Additionally, many of them are inapplicable in the presence of scatter. For instance, the calculation of $\boldsymbol{W}$ – needed in the context of minimizing $K^2|\boldsymbol{W}|$ over different partitions for each K (Marriott, 1971), or in $tr(\boldsymbol{W})$ of the Gap statistic (Tibshirani *et al*, 2001) – is then not clear. For model-based approaches, the problem is sometimes reformulated in terms of model selection (McLachlan and Peel, 2000), with Schwarz's (1978) Bayes Information Criterion (BIC) as a popular choice. TCTW determines K algorithmically, terminating at $K = K_T$ or when no further tight and stable cluster is possible. In our experience however, the algorithm continues building clusters far beyond the true K. Hence, a good estimate of K is needed. In some cases, this may be provided by the researcher, however even experienced researchers often have no way of knowing what would be reasonable.

In this paper, we propose a k-means-type algorithm for "Clustering in the Presence of Scatter" (abbreviated as k-clips). Section 2 develops an iterative methodology together with an approach for its initialization. We estimate K using an adaptation of BIC which reduces to Marriott's criterion in the absence of scatter. While we use k-means with scatter for our algorithmic development, other algorithms using a central measure of tendency of a cluster (such as k-medoids with scatter) could also be developed similarly. Our algorithm is general enough to permit other strategies for initialization and estimation of K. The different elements of our proposed methodology are extensively evaluated and compared in Section 3 and Web Appendix Section 2 through simulation experiments with varying amounts of scatter and cluster separation. The mercury release dataset introduced in Section 1.1 is analyzed in Section 4. The paper concludes with some discussion.

2. Methodology

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be p-variate observations from the mixture distribution $g(\boldsymbol{x}) = \sum_{k=1}^{K} \zeta_k [\frac{1}{\sigma} f(\frac{\boldsymbol{x}-\boldsymbol{\mu}_k}{\sigma})] + \frac{\zeta_{K+1}}{V}$ where V is the volume of a bounded region $\mathcal{B}$ with uniform density, $\sigma > 0$ is a scale parameter, ζ_k is an identification parameter and $f(\boldsymbol{y}) = \psi(\boldsymbol{y}'\boldsymbol{y})$ with $\psi(\cdot)$ a real positive-valued function such that $f(\cdot)$ is a p-variate density. For *hard clustering* methods such as k-means, ζ_k is the observation's class indicator, while for *soft clustering* algorithms, ζ_k represents the probability of its inclusion in the kth cluster. The above formulation provides a convenient way to fix ideas in a statistical framework, stipulating that $\boldsymbol{X}_i$ comes from one of the K homogeneous populations or the $(K+1)$th group which comprises uniformly-distributed scatter and can not be summarized beyond this property. Our goal is to estimate K (if unknown), the means $\boldsymbol{\mu}_k$, the number of clustered observations (n^*), V and more importantly to classify observations into groups and scatter. We propose algorithms to achieve these objectives next.

2.1 **k-clips: A modified k-means algorithm for data with scatter**

Given K and initial cluster centers $\{\boldsymbol{\mu}_k; k = 1, \ldots, K\}$, our strategy is to build K p-dimensional uniform-volumed spheres (or cores) around each: these cores represent dense parts (*i.e.* clusters) of the dataset with their complement the domain of the scattered points. Points inside a core are assigned to the corresponding cluster, those outside all cores are nominated as scatter, and the cluster centers and the cores are updated. This procedure is iterated until convergence. The exact algorithm is as follows:

(1) *Building the cores.* Assign each observation $\boldsymbol{X}_i$ to cluster $k = \arg\min_{1 \leqslant \kappa \leqslant K} \parallel \boldsymbol{X}_i - \boldsymbol{\mu}_\kappa \parallel$ and denote this $\boldsymbol{X}_i$ as $\boldsymbol{X}_{i(k)}$. Obtain a robust estimate (s_{bw}) of the standard deviation (SD) of the observations, common for all dimensions, using the biweight estimator of Hoaglin *et al* (2000). Let $\tilde{\mu}_{k,j}$ be the median of the kth cluster and jth dimension ($k = 1, \ldots, K; j = 1, \ldots, p$), and $Y_{i(k),j} = X_{i(k),j} - \tilde{\mu}_{k,j}$ be the median-centered observations. Denote the common median absolute deviation (MAD) of these $Y_{i(k),j}$'s as $\tilde{s}$. For some constant w, let $u_{i(k),j} = \frac{Y_{i(k),j}}{w\tilde{s}}$. Then the robust biweight estimator of the SD is given by

$$s_{bw} = \frac{(np)^{\frac{1}{2}} \left[\sum_{|u_{i(k),j}|<1} Y^2_{i(k),j} \left(1 - u^2_{i(k),j}\right)^4\right]^{\frac{1}{2}}}{|\sum_{|u_{i(k),j}|<1} \left(1 - u^2_{i(k),j}\right)\left(1 - 5u^2_{i(k),j}\right)|}.$$

Next, create K p-dimensional spheres of common radius $r_K = s_{bw}\sqrt{\chi^2_{p,\alpha}}$, each centered at the current $\boldsymbol{\mu}_k$'s, where $\chi^2_{p,\alpha}$ is the $(1-\alpha)th$ quantile of the χ^2-distribution with p degrees of freedom, $0 < \alpha < 1$.

(2) *Assignments.* Observations outside all cores are labeled scatter; the rest are assigned to those clusters whose centers are closest to them.

(3) *Updates.* Recalculate $\{\boldsymbol{\mu}_k; k = 1, \ldots, K\}$ using the centers of the points currently in each cluster. Note that scatter points are not used to update the $\boldsymbol{\mu}$s.

(4) Repeat until convergence, upon which final cluster centers and classifications into groups or scatter are obtained.

Comments.

(1) Our algorithm locally minimizes, upon convergence, the approximate objective function given by $\frac{n^*p}{2}\log(\sigma^2 + 2\pi) + \frac{1}{2\sigma^2}\mathsf{tr}\,(\boldsymbol{W}^*) + (n - n^*)\log(V)$ where $\mathsf{tr}\,(\boldsymbol{W}^*)$ is the trace of the within-sums-of-squares-and-products matrix (SSP_W) based on the n^* clustered observations. A model-based interpretation for the above is provided in the hard-clustering context when each cluster density is assumed to be Gaussian with dispersion matrix $\sigma^2\boldsymbol{I}$ (see Web Appendix Section 1). Further, for known σ and in the absence of scatter, this objective function reduces to that for k-means. Note also that while our algorithm converged in all our experiments, convergence has not been rigorously proved, and needs to be established. Such development however, is perhaps beyond the scope of this paper.

(2) Hoaglin *et al.* (2000) show that s_{bw} is an efficient and robust scale estimator in several non-Gaussian situations. Our clustering scenario has scatter points and outliers; so s_{bw} is reasonable in determining the radius of the cores.

(3) The estimate s_{bw} uses a parameter w to modulate the influence of scatter points currently misclassified as clustered. Intuitively, higher choices of w correspond to higher (less robust) values of s_{bw} while lower choices result in estimates for σ that are biased downwards. Under Gaussian distributional assumptions, $w = 9$ means that observations more than 6 SD's from every cluster median in each dimension are ignored in the calculation of s_{bw}. In the absence of additional information, one may optimize the objective function over $w \in \{3, 4.5, 6, 7.5, 9\}$ to get a data-driven choice. We adopt this approach in this paper.

(4) The core volume is also controlled by α, with higher values identifying more observations as scatter. Under Gaussian density assumptions, $s_{bw}\sqrt{\chi^2_{p,\alpha}}$ approximates the radius of the densest $100(1-\alpha)\%$ sphere of concentration in each cluster. The distributional assumption helps motivate our algorithm since clustering makes most sense in the context of compact groups, for which the Gaussian density provides a reasonable frame of reference.

(5) As $w \uparrow \infty$ and $\alpha \downarrow 0$, the algorithm reduces to k-means.

The above differs from DHC in that it is a non-hierarchical algorithm classifying observations into both clusters and scatter. Adapt_Clust and TCTW identify clusters one-by-one, with the former designed to work only on mean-zero-unit-variance-standardized data. On the contrary, our algorithm modifies k-means to simultaneously partition any dataset into clusters and scatter, proceeding to convergence from initializing cluster centers, a strategy for choosing which we address next.

2.2 *Initialization of Cluster Centers*

Common initialization methods for k-means degrade with scatter, *cf.* Tseng and Wong (2005) who proposed to initialize using the means of the largest K groups obtained after cutting a hierarchically clustered tree into $K \times p$ groups Our own experience has shown this approach to not perform well with increased scatter. An alternative is to remove clutter using k-nearest neighbor cleaning (Byers and Raftery, 1998) followed by an algorithm such as the multi-stage deterministic initializer of Maitra (2007) which, in the context of k-means, finds a large number of local modes and then chooses K representatives from the most separated ones. Performance of this last approach in simulation experiments was uneven (see Web Appendix Section 2.2), so here we develop an adaptation of Maitra (2007) that accounts for the presence of scatter:

(1) Write $\boldsymbol{X}$ as the $n \times p$ matrix of observations, *i.e.* $\boldsymbol{X}' = \{\boldsymbol{X}_1 \vdots \boldsymbol{X}_2 \vdots \ldots \vdots \boldsymbol{X}_n\}$. For the jth column of $\boldsymbol{X}$ $(j = 1, 2, \ldots, p)$, obtain the $\tau = \lceil (np)^{1/(p+1)} \rceil$ equi-spaced quantiles, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Use one-dimensional k-means initialized with these quantiles to obtain the τ modes along each dimension. The product set $\mathcal{H}$ (cardinality K_0^*) of the resulting one-dimensional modes are our potential candidate starting points.

(2) For each point in $\boldsymbol{X}$, identify the closest member in $\mathcal{H}$. Eliminate those points from $\mathcal{H}$ that are not closest to at least $\mathcal{E}$ observations in $\boldsymbol{X}$. This eliminates scatter points from $\mathcal{H}$, yielding the reduced set $\mathcal{H}'$ with $\mid \mathcal{H}' \mid = K^*$. Assume for now that $K^* \geqslant K$.

(3) Initialize the algorithm of Section 2.1 with $\mathcal{H}'$ and apply to the dataset $\boldsymbol{X}$ to obtain K^* modes.

(4) Our next goal is to obtain K representatives from the most widely separated modes. We propose using hierarchical clustering with single linkage on the K^* local modes and cut the resulting tree into K groups. Classify all observations into one of the K groups. Use the group medoids as the initial cluster centers $\{\boldsymbol{\mu}_k; k = 1, 2, \ldots, K\}$ in the k-clips algorithm.

Comments.

(1) The choice of τ large relative to Maitra (2007) generates many univariate candidate starting points and avoids trapping one-dimensional k-means in local minima around scatter points. This strategy does not eliminate scatter, but increases the chance of also finding true modes at only a modest additional computational cost, since we use one-dimensional k-means.

(2) A liberal choice for τ means that $\mathcal{H}$ is large and includes some scatter. Large values of $\mathcal{E}$ (which may be thought of as a minimum cluster size) reduce the chance of retaining scatter in $\mathcal{H}'$, but $\mathcal{E}$ too large can result in $K^* < K$. We avoid this problem by successively running the initializer with decreasing $\mathcal{E}$ from an upper bound $\mathcal{E}_M \geqslant \mathcal{E}$ until $K^* \geqslant K$. Although no optimal choice for $\mathcal{E}_M$ can be prescribed *a priori*, we have found $\mathcal{E}_M = \frac{n}{Kp}$ to work well in our experiments.

(3) Data-driven choices of w and α allow for greater flexibility in determining the influence of scatter on the starting centers. Candidate values are all pairs of $w \in \{3, 9\}$ and $\alpha \in \{.01, .05, .1, .2\}$. We use the extreme candidate values for w to save compute time, since our objective here is merely to obtain starting values for the cluster means. The final initializer is chosen from among the candidate set by evaluating the objective function after one iteration of k-clips using moderate values of $w = 6$ and $\alpha = 0.05$. This last step is needed in order to obtain a summary assignment of scatter points in the calculation of the objective function.

(4) The hierarchical classification in Step 4 of the initialization algorithm does not exclude scatter, so we use medoids instead of means to help ensure that the starting centers are in the interior of the clusters.

Our initialization strategy is discussed and demonstrated in the context of k-clips, however, it is general enough to extend to other algorithms that need initialization in the presence of scatter. We now address the issue of estimating the number of clusters K.

2.3 *Estimating the Number of Clusters*

Our algorithms have so far assumed knowledge of K, which is rarely true in practice, so we turn to methods for optimally estimating K. Our proposal is to choose

$$\begin{aligned}\hat{K} = \operatorname*{argmin}_{k} \Bigg\{ \log\left(k+\frac{1}{p}\right) &+ \frac{n^*}{n}\left(\frac{1}{2}\log|\boldsymbol{W}^*|\right) \\ &+ \left(1-\frac{n^*}{n}\right)\log\hat{V}\Bigg\}\end{aligned} \tag{1}$$

where $|\boldsymbol{W}^*|$ is the determinant of SSP_W based on the n^* clustered observations, and $\hat{V}$ is the volume of the region containing the $n-n^*$ scattered points. We motivate (1) by modifying BIC (see Web Appendix Section 1) under the hard clustering model with Gaussian densities centered at $\{\boldsymbol{\mu}_k; k=1,2,\ldots,K\}$ and a common dispersion $\boldsymbol{\Sigma}$. In the context of a model without scatter, $n^*=n$ and $\boldsymbol{W}^*=\boldsymbol{W}$, yielding Marriott's criterion. Finally, the volume $\hat{V}$ is calculated from the p-dimensional cube with axes matching the ranges of $\boldsymbol{X}$ in each dimension, less the volume of the K p-variate spheres. In cases where any sphere intersects with the boundary of the cube, $\hat{V}$ is approximated using Monte Carlo simulation, specifically by generating uniform pseudo-random deviates within the region of the data and estimating V as the proportion of simulated points outside the cores.

3. Experimental Evaluations

The suggested methodology was extensively evaluated through a series of simulation experiments. The k-clips algorithm has three aspects: the main algorithm, initialization and estimation of the number of clusters. We only report evaluations on its overall performance here and refer to Web Appendix Section 2 for more detailed studies on each of these issues. Our assessment is presented graphically for bivariate examples, and numerically for all dimensions. Our numerical measure is a modification of the adjusted Rand measure ($\mathcal{R}_a$; Hubert and Arabie, 1985), where scatter points are considered to be sole members of clusters of size one. This modified measure is thus more severely influenced by misclassifications of scatter into groups with true cluster points and vice-versa than by erroneously creating additional clusters of scatter points only. $\mathcal{R}_a$ also more severely penalizes partitions with too few groups than with too many groups. Our experimental suite covered a wide range of dimensions and separation between clusters. Following Maitra's (2007) modification of Dasgupta's (1999) definition, we defined a set of p-variate Gaussian densities to be $exact-c-separated$ if for every pair $N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ and $N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$, $\|\boldsymbol{\mu}_i-\boldsymbol{\mu}_j\| \geqslant c\sqrt{p\max\left(\lambda_{\max}\{\boldsymbol{\Sigma}_i\},\lambda_{\max}(\boldsymbol{\Sigma}_j)\right\}}$, with equality holding for at least one pair and $\lambda_{\max}(\boldsymbol{\Sigma})$ denoting the largest eigenvalue of $\boldsymbol{\Sigma}$. In our experiments, c ranged from 0.8 (marginally separated) to 2.0 (well-separated) with $\boldsymbol{\Sigma}$'s enforced to be diagonal (but not spherical) for each cluster. Observations were generated with equal probability from each cluster, and realizations outside the 95% ellipsoids of concentration of all the clusters were eliminated. Finally, scatter points were uniformly generated via simple rejection from outside the union of these 95% confidence ellipsoids. The outer limits of this region were obtained by multiplying the observed ranges in each dimension by a $U(1,2)$ deviate. Such scatter formed between 15% and 50% of each dataset.

Experiments were done assuming known and unknown K and evaluating $\mathcal{R}_a$ for the derived groupings relative to the true. We used $\alpha=0.05$ and $5 \leqslant \mathcal{E} \leqslant n/pK$ for k-clips and compared its performance with groupings obtained using TCTW and the Mclust function of the MCLUST package in R. We used Mclust with the Poisson Process component for scatter initialized using a random estimate (see Page 15 of Fraley and Raftery, 2006) and a data-driven BIC-optimal choice of dispersion matrix. TCTW was implemented using Tseng and Wong's (2005) software. For the tuning parameters, we used a data-driven approach to choosing the (α_T,β) in $\{0,0.05,0.1\}\times\{0.6,0.7,0.8\}$ that maximized the likelihood model generating the data. The ranges for α_T and β corresponded to decreasing and increasing orders of tightness and stability respectively and were taken from different scenarios of possible values discussed in Tseng and Wong (2005). With unknown K, we set K_T in TCTW to be equal to the maximum number of groups considered for Mclust and k-clips. Further with TCTW, we set $K \leqslant k_0 \leqslant 2K$ for known K and $K_T \leqslant k_o \leqslant 2K_T$ with unknown K.

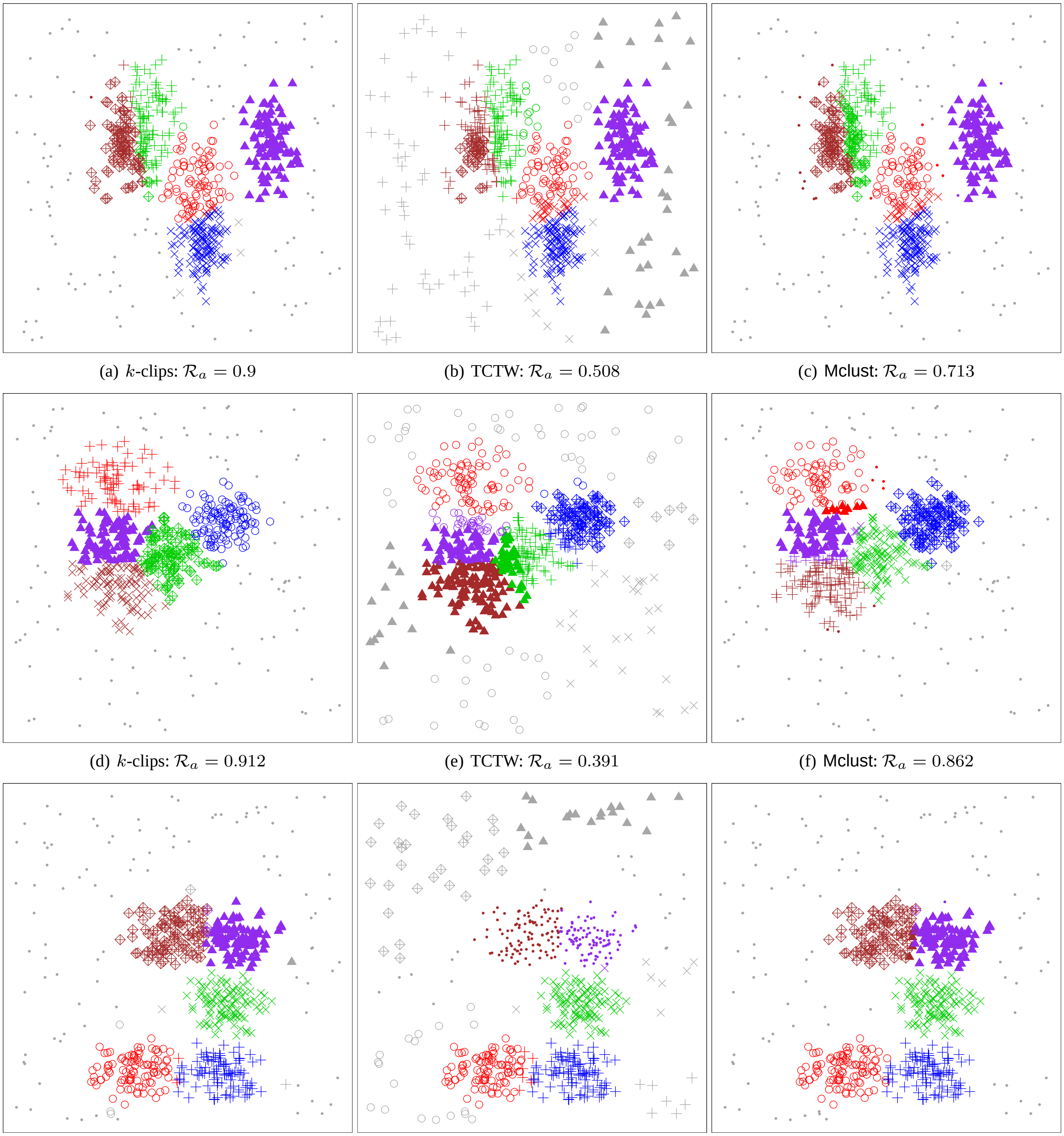
3.1 *Bivariate Experiments*

For each of the three experimental scenarios, we generated 100 scattered realizations along with 500 clustered observations drawn with equal probablity from one of five exact-c-separated Gaussians (Figure 1, colored symbols). For this suite, $K_T=15$. Clearly, k-clips always performed very well, even though Mclust was marginally better when $c=1.6$. However, the latter performed poorly with lower separation. The performance of TCTW was uniformly weak, with no scatter identified when $c=0.8$ and 1.2, and two true clusters identified as scatter when $c=1.6$. A disconcerting consequence of the sequential identification of clusters without accounting for already identified partitions is well-illustrated in Figure 1e: some of the true scatter in the lower left corner was identified ("∘") as belonging to a group (upper left) beyond its closest cluster ("▲").

When estimating K, k-clips was always correct while TCTW always found the maximum $K_T=15$ stable and tight clusters. These 15 partitions sub-divided several true clusters and misclassifed scatter into other true clusters, resulting in very poor $\mathcal{R}_a$'s. Mclust found the correct number of clusters for $c=0.8$ and 1.6, while $\hat{K}=3$ for $c=1.2$ leading to a considerably poor partitioning ($\mathcal{R}_a=0.476$, Figure 2f). The harsh penalization of too few clusters by $\mathcal{R}_a$ is demonstrated by the Mclust solution in Figure 2f, which performed better at identifying scatter and cluster than TCTW (Figure 2e) but had lower $\mathcal{R}_a$ even though TCTW identified 15 clusters and no scatter!

3.2 *Higher-dimensional Examples*

We performed higher-dimensional experiments with (p,K,n) = (5, 5, 500), (10, 7, 2,000) and (20, 15, 5,000) and varying the amounts of separation between clusters. The proportion (s) of scatter in the experiments ranged from 15% to 25% of the total observations. We also performed experiments with 50% scatter and $c=2.0$. In each case, we generated 25 sets of parameters (cluster means, dispersions, mixing proportions) thus obtaining 25 simulation datasets. Table 1 reports comprehensive measures on overall performance when K is known, with results for K unknown shown in Table 2. (For the latter, K_T was set to 15, 20 and 25 for $p=5,10$ and 20, respectively.) Note that TCTW did considerably worse than either k-clips or Mclust in



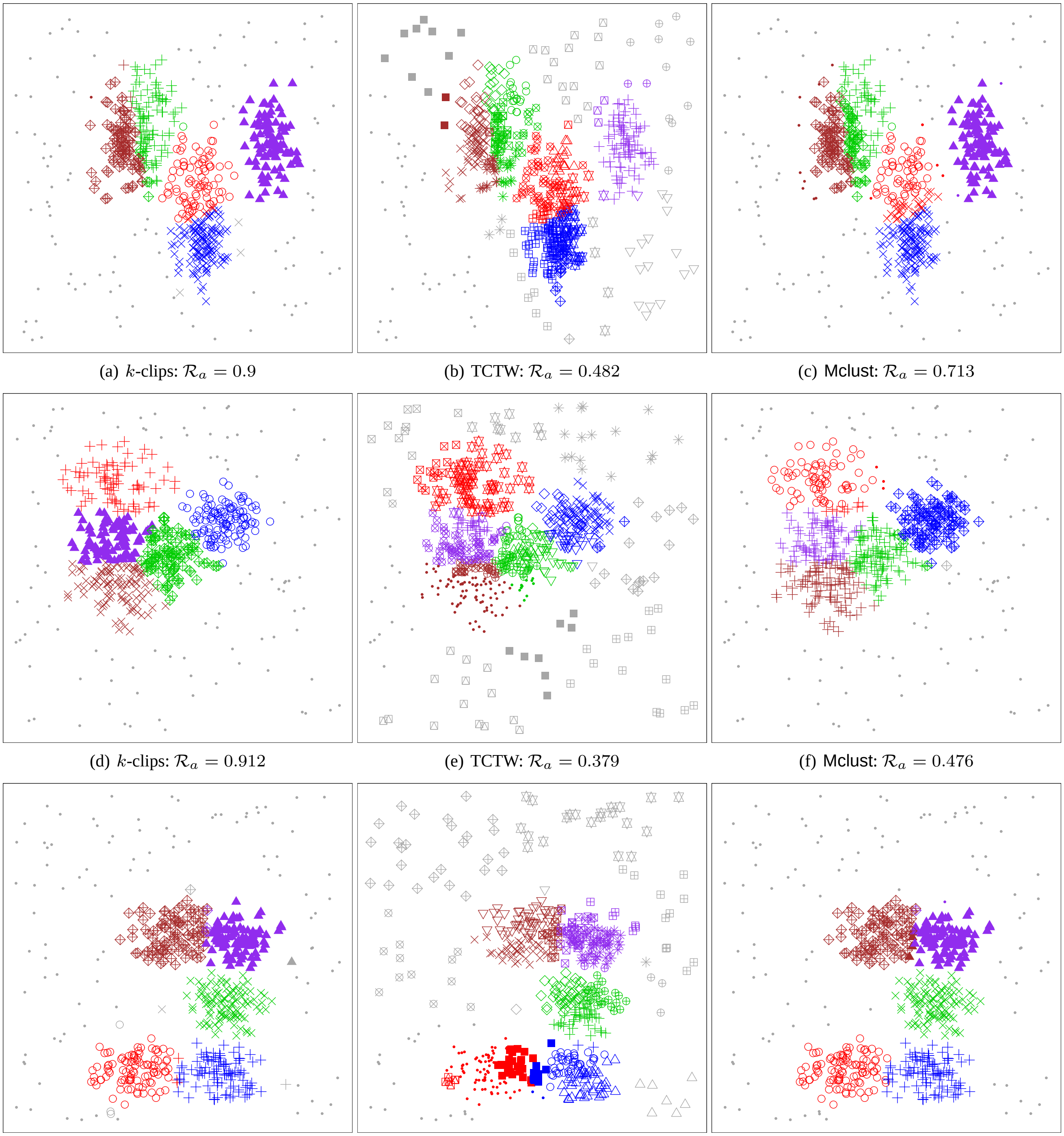
(a) k-clips: $\mathcal{R}_a = 0.9$ (b) TCTW: $\mathcal{R}_a = 0.508$ (c) Mclust: $\mathcal{R}_a = 0.713$

(d) k-clips: $\mathcal{R}_a = 0.912$ (e) TCTW: $\mathcal{R}_a = 0.391$ (f) Mclust: $\mathcal{R}_a = 0.862$

(g) k-clips: $\mathcal{R}_a = 0.952$ (h) TCTW: $\mathcal{R}_a = 0.615$ (i) Mclust: $\mathcal{R}_a = 0.975$

Figure 1. Results of k-clips (first column), TCTW (second column) and Mclust (third column) with $K = 5$ known clusters for $c = 0.8, 1.2,$and 1.6 for the first, second and third rows respectively. Small filled circles represent identified scatter, colors signify true clusters and characters as identified clusters.

almost all cases with known K, but occassionally performed best with unknown K. k-clips did substantially better than Mclust for cases with low separation, but Mclust performed marginally better than k-clips for higher c. In general, k-clips correctly estimated K when $p = 5, 10$, but tended to underestimate it for $p = 20$ and high values of c. On the other hand, for $p = 20$ Mclust tended to grossly underestimate K for low separation and slightly overestimate K for higher separation. Mclust $\mathcal{R}_a$ values were thus very poor for low separation but fairly high for well-separated clusters. TCTW often found the maximum K_T stable and tight clusters for $p = 5$


(a) k-clips: $\mathcal{R}_a = 0.9$

(b) TCTW: $\mathcal{R}_a = 0.482$

(c) Mclust: $\mathcal{R}_a = 0.713$

(d) k-clips: $\mathcal{R}_a = 0.912$

(e) TCTW: $\mathcal{R}_a = 0.379$

(f) Mclust: $\mathcal{R}_a = 0.476$

(g) k-clips: $\mathcal{R}_a = 0.952$

(h) TCTW: $\mathcal{R}_a = 0.390$

(i) Mclust: $\mathcal{R}_a = 0.975$

Figure 2. Results of k-clips (first column), TCTW (second column) and Mclust (third column) with K estimated for increasing separation $c = 0.8, 1.2,$and 1.6 for the first, second and third rows respectively. Small filled circles represent identified scatter, colors represent true clusters and characters denote identified clusters.

and usually substantially more than the true K clusters. Even then, performance was rather poor for cases with low separation, but like Mclust, improved substantially with higher separation. TCTW had very good $\mathcal{R}_a$ values for $p = 20$, but was again outclassed by k-clips and Mclust for higher proportions of scatter. The curse of dimensionality seemed to afflict k-clips slightly more than Mclust and TCTW for higher separation, as seen for $p = 20$, becoming more evident with K unknown. Note that in all our experiments, data were generated from the Gaussian-uniform mixture model explicitly assumed by Mclust – thus, it is expected that it would perform well.

Table 1

The median adjusted Rand $\left(\mathcal{R}_{\frac{1}{2}}\right)$ *values from 25 runs of* k*-clips and* Mclust *with known number of clusters and TCTW with* $K_T = K$ *target clusters.* $I_{\mathcal{R}}$ *is the interquartile of the 25* $\mathcal{R}_a$*'s and* η *is the number of runs where the method has the highest* $\mathcal{R}_a$*, with ties being split among the methods tying.*

Settings			k-clips			Mclust			TCTW		
	s	c	$\mathcal{R}_{\frac{1}{2}}$	$I_{\mathcal{R}}$	η	$\mathcal{R}_{\frac{1}{2}}$	$I_{\mathcal{R}}$	η	$\mathcal{R}_{\frac{1}{2}}$	$I_{\mathcal{R}}$	η
$p = 5, K = 5, n = 500$	.15	0.8	0.891	0.057	19	0.854	0.195	6	0.555	0.074	0
		1.2	0.973	0.032	17.5	0.951	0.086	7.5	0.676	0.094	0
		1.6	0.994	0.012	15	0.989	0.016	10	0.744	0.065	0
		2.0	0.997	0.006	14.5	0.995	0.009	10.5	0.752	0.046	0
	.25	0.8	0.871	0.084	20	0.797	0.131	5	0.490	0.150	0
		1.2	0.948	0.039	8.5	0.949	0.073	16.5	0.617	0.103	0
		1.6	0.970	0.018	7	0.981	0.029	18	0.673	0.073	0
		2.0	0.982	0.015	6	0.991	0.008	19	0.679	0.067	0
	.50	2.0	0.911	0.032	9	0.987	0.174	16	0.448	0.110	0
$p = 10, K = 7, n = 2000$	.15	0.8	0.963	0.018	22	0.521	0.400	3	0.666	0.048	0
		1.2	0.995	0.006	17	0.993	0.134	8	0.761	0.029	0
		1.6	0.999	0.002	13	0.999	0.147	12	0.802	0.028	0
		2.0	1.000	0.080	13	1.000	0.144	12	0.807	0.015	0
	.25	0.8	0.963	0.018	25	0.781	0.411	0	0.562	0.044	0
		1.2	0.995	0.006	18	0.988	0.156	7	0.650	0.073	0
		1.6	0.999	0.002	11.5	0.999	0.116	13.5	0.683	0.060	0
		2.0	1.000	0.080	13.5	1.000	0.134	11.5	0.675	0.083	0
	.50	2.0	1.000	0.002	11.5	1.000	0.001	13.5	0.422	0.060	0
$p = 20, K = 15, n = 5000$	.15	0.8	0.995	0.002	25	0.330	0.234	0	0.800	0.055	0
		1.2	0.929	0.064	15	0.929	0.019	10	0.858	0.054	0
		1.6	0.891	0.088	9	0.879	0.073	12	0.893	0.031	4
		2.0	0.887	0.085	7	0.923	0.066	15	0.866	0.052	3
	.25	0.8	0.994	0.002	25	0.394	0.338	0	0.678	0.063	0
		1.2	0.943	0.086	17	0.931	0.132	8	0.766	0.070	0
		1.6	0.862	0.081	7	0.921	0.060	18	0.765	0.086	0
		2.0	0.874	0.080	7	0.922	0.102	16	0.805	0.116	2
	.50	2.0	0.952	0.089	17	0.925	0.053	8	0.403	0.098	0

In summary, our experiments indicate that both k-clips and Mclust outperform TCTW with K known. For unknown K, k-clips is the best performer when groups are not well-separated, but is overtaken by TCTW or Mclust otherwise. A partial explanation is that k-clips sometimes estimated fewer than the true number of clusters, potentially affecting $\mathcal{R}_a$ severely. Even with well-separated clusters, however, large amounts of scatter uniformly reduced TCTW's performance, while honors were about even between k-clips and Mclust. Thus, k-clips complements Mclust by excelling for cases with poorly-separated clusters. Further, since k-clips modifies k-means, it is more practical to apply to large datasets than Mclust.

3.3 *Additional Experiments using Non-Gaussian Clusters*

The Web Appendix (Section 2.5) also reports detailed performance evaluations on two-dimensional simulation experiments where cluster distributions vary widely from the Gaussian. We summarize our findings here. Our clusters in these experiments were from distributions that were (a) symmetric but heavy-tailed, (b) very irregular, (c) constrained to lie on a sphere, or (d) highly-skewed. In (a), k-clips maintained the trend of good performance for both known and unknown K ($\mathcal{R}_a \geqslant 0.95$). For (b) and (c) with K known, k-clips performed at least moderately well ($\mathcal{R}_a \geqslant 0.70$). Performance however degraded ($\mathcal{R}_a \leqslant 0.62$) for unknown K as the algorithm tried to carve ellipsoidal clusters out of the irregular and spherically-constrained distributions by combining or splitting them. For both known and unknown K, performance on the highly-skewed datasets of (d) was very poor for moderately-separated clusters ($\mathcal{R}_a \leqslant 0.28$), but improved considerably for their better-separated counterparts ($\mathcal{R}_a \geqslant 0.72$). The methodology performed substantially better ($\mathcal{R}_a \geqslant 0.73$) when applied to data transformed to reduce skewness in each dimension.

We also tested Mclust and TCTW on these datasets. Mclust often performed worse than k-clips in the moderately-separated cases but better with higher separation. It had far more problems than k-clips in separating cluster from scatter observations with distributions in (a) and (b). For unknown K, it often grossly overestimated K for (b) and (c) achieving ellipsoidal shapes like k-clips by partitioning true clusters. For (c), Mclust did substantially better than k-clips for higher separation, perhaps because it has greater flexibility in accomodating lower-dimensional Gaussians (see Web Appendix Section 2.5). TCTW exhibited similar trends in many of these examples as it did for the Gaussian datasets, failing to identify scatter correctly, continually splitting apart clusters and occasionally failing to adhere to the distance metric. Like k-clips, Mclust and TCTW

Table 2

The median adjusted Rand $\left(\mathcal{R}_{\frac{1}{2}}\right)$ values from 25 runs of k-clips and Mclust *with unknown number of clusters and TCTW with K_T set to the maximum considered by the other methods. $I_{\mathcal{R}}$ is the interquartile of the 25 $\mathcal{R}_a$'s, η is the number of runs where the method has the highest $\mathcal{R}_a$, $\hat{K}_{\frac{1}{2}}$ is the median estimated number of clusters and $I_{\hat{K}}$ is the interquartile of the 25 $\hat{K}$'s over the 25 runs.*

Settings			k-clips					Mclust					TCTW				
	s	c	$\mathcal{R}_{\frac{1}{2}}$	$I_{\mathcal{R}}$	η	$\hat{K}_{\frac{1}{2}}$	$I_{\hat{K}}$	$\mathcal{R}_{\frac{1}{2}}$	$I_{\mathcal{R}}$	η	$\hat{K}_{\frac{1}{2}}$	$I_{\hat{K}}$	$\mathcal{R}_{\frac{1}{2}}$	$I_{\mathcal{R}}$	η	$\hat{K}_{\frac{1}{2}}$	$I_{\hat{K}}$
$p = 5, K = 5, n = 500$	.15	0.8	0.865	0.169	13	5	0	0.854	0.199	12	5	1	0.340	0.194	0	15	1
		1.2	0.971	0.041	14.5	5	0	0.951	0.032	10.5	5	1	0.555	0.156	0	15	1
		1.6	0.989	0.017	12.5	5	0	0.986	0.016	12.5	5	0	0.705	0.086	0	15	0
		2.0	0.924	0.138	7	5	1	0.995	0.016	18	5	0	0.760	0.064	0	15	0
	.25	0.8	0.807	0.124	11	5	1	0.797	0.147	14	5	1	0.500	0.219	0	15	0
		1.2	0.948	0.040	5.5	5	0	0.952	0.032	19.5	5	1	0.617	0.118	0	15	0
		1.6	0.970	0.019	3	5	0	0.984	0.027	22	5	0	0.742	0.148	0	15	0
		2.0	0.979	0.020	4	5	0	0.991	0.007	21	5	0	0.840	0.076	0	15	0
	.50	2.0	0.911	0.036	0	5	0	0.991	0.013	25	5	1	0.864	0.055	0	15	0
$p = 10, K = 7, n = 2000$	.15	0.8	0.963	0.018	20	7	0	0.648	0.466	5	7	4	0.761	0.136	0	14	5
		1.2	0.994	0.090	12	7	0	0.995	0.007	13	7	1	0.916	0.112	0	17	2
		1.6	0.911	0.126	5.5	6	1	0.999	0.036	17.5	8	1	0.969	0.055	2	17	3
		2.0	0.904	0.154	0.5	6	1	1.000	0.011	21.5	8	2	0.956	0.041	3	18	2
	.25	0.8	0.962	0.025	25	7	0	0.781	0.361	0	6	3	0.843	0.106	0	16	2
		1.2	0.994	0.005	20	7	0	0.991	0.022	5	8	1	0.956	0.036	0	16	2
		1.6	1.000	0.144	11	7	1	0.999	0.009	10	8	1	0.981	0.007	4	17	2
		2.0	0.918	0.137	5	6	1	0.999	0.003	15	7	1	0.980	0.009	5	18	3
	.50	2.0	1.000	0.002	10.5	7	0	1.000	0.008	14.5	8	1	0.912	0.150	0	11	3
$p = 20, K = 15, n = 5000$	.15	0.8	0.994	0.041	18	15	0	0.358	0.092	0	5	1	0.974	0.007	7	22	2
		1.2	0.858	0.161	0	13	2	0.985	0.022	15	17	2	0.982	0.004	10	21	2
		1.6	0.833	0.097	0	12	3	0.984	0.029	13	17	2	0.981	0.004	12	20	3
		2.0	0.789	0.077	0	11	1	0.986	0.029	15	17	3	0.981	0.003	10	21	3
	.25	0.8	0.994	0.002	25	15	0	0.407	0.143	0	5	2	0.940	0.020	0	17	1
		1.2	0.960	0.158	11.5	15	1	0.983	0.037	12.5	16	3	0.960	0.018	1	17	2
		1.6	0.880	0.092	1	15	3	0.979	0.028	19	18	3	0.962	0.010	5	16	2
		2.0	0.820	0.130	0	13	3	0.982	0.029	23	17	2	0.951	0.033	2	16	3
	.50	2.0	1.000	0.066	12	15	1	0.995	0.014	13	17	2	0.617	0.349	0	19	5

both performed poorly when applied directly to the highly-skewed datasets of (d), but did much better on data transformed to reduce skewness in each dimension, with similar results as in Section 3.1.

Our experiments illustrate some of the main properties of k-clips. The main algorithm performs well in identifying cluster cores, centers and scatter when the distribution is far removed from the Gaussian which is used to motivate the objective function in our algorithm. Robust methods of scale and hence core estimation allow the algorithm to take deviations from spherical and compact clusters in its stride. The algorithm does have a breaking-point, reached, for instance, when data are heavily-skewed and clusters only moderately-separated. In such cases, transforming the dataset can improve performance. Our BIC-type criterion for estimating the number of clusters is more sensitive to deviations from the Gaussian, performing well when the clusters are compact and regular but not quite spherical (Section 3.1 and 3.2) or spherical but not compact (Section 3.3), but rarely otherwise. We now turn our attention towards categorizing industrial releases of mercury.

4. Application to Mercury Release Data

The mercury release data were severely and differentially skewed in the variables of interest, with kurtosis for fugitive air emissions and total off-site releases of around 120 and 1,300 respectively. Given the results of the simulations, we followed Lin (1973) and used a shifted loglog transformation: $Y_{ij} = \log\left(1 + \log\left(1 + X_{ij}\right)\right)$, where X_{ij} denoted the jth type of release reported by the ith facility. Each transformed variable was scaled by its standard deviation to allow for the use of Euclidean distances. After transformation, there still remained a number of outliers, mostly facilities with massive mercury releases which needed to be accounted for while identifying similar facilities.

A total of fourteen clusters, ranging in size from 11 to 257 facilities, and 98 scatter facilities were optimally identified by k-clips. These clusters (henceforth k-clipsters) were estimated to be $exact - 0.89 - separated$, which indicated at least moderate separation of identified groups. The marginal distributions of the clusters, in terms of their deciles, with the intensities mapped onto the standardized shifted loglog scale described above are shown in Figure 3. Clearly, the mercury release characteristics for each group were quite distinct, providing confidence in the obtained groupings.

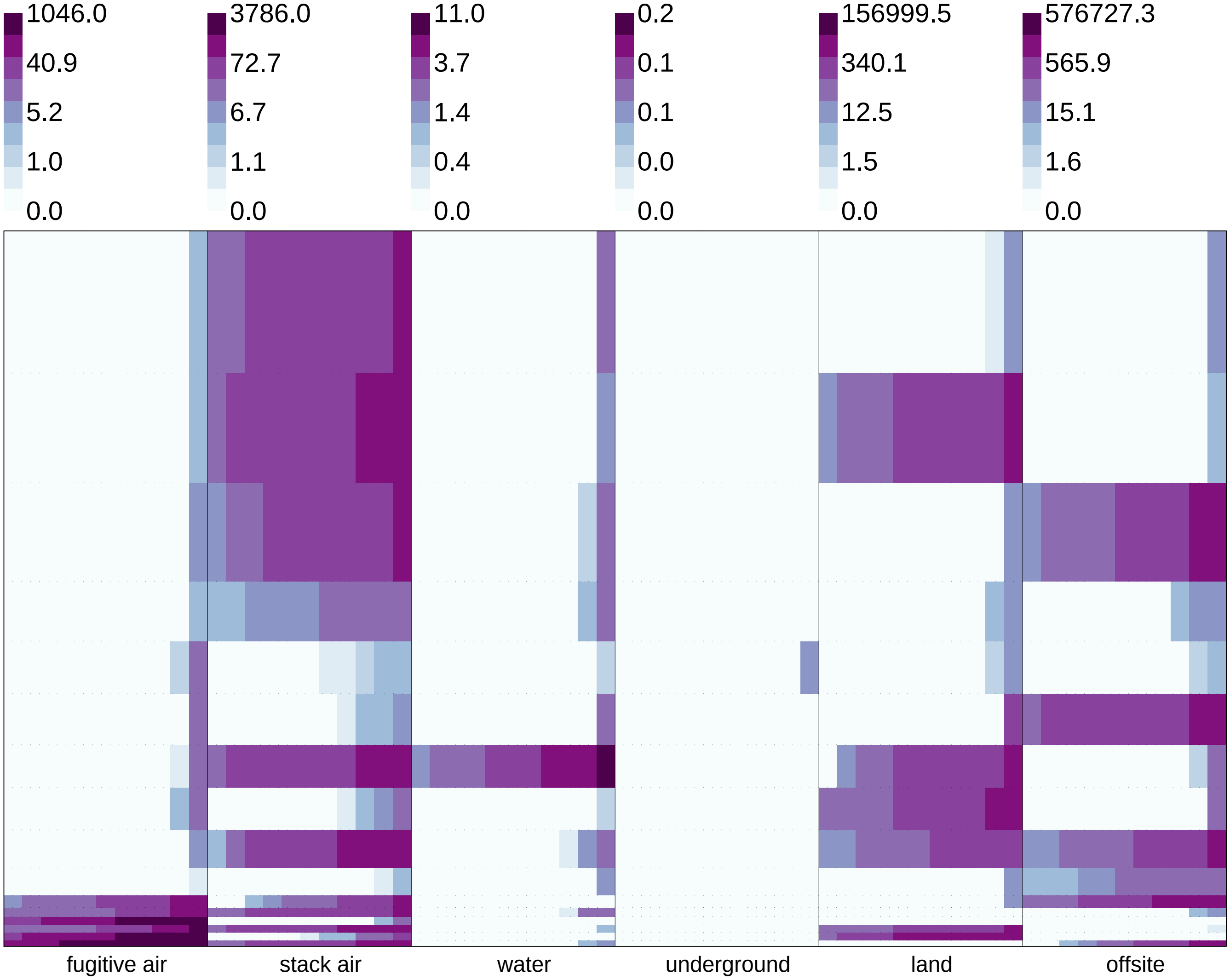


Figure 3. Distribution of members of each identified cluster with area of rectangles proportional to cluster size. The deciles of the marginal releases for each group are represented by the intensities on a standardized loglog scale separate for each release type.

The utility of groupings and scatter obtained via $TCTW$ was unclear, as the algorithm always built K_T groups, until we set $K_T = 200$ upon which it terminated at 199 clusters. Mclust estimated 56 optimal clusters (henceforth Mclusters), ranging in size from 5 through 94 facilities and 71 scattered points with $\mathcal{R}_a = 0.232$ when compared to k-clips. With so many Mclusters, interpreting the results was more challenging as not all clusters were noticeably distinct. Indeed, the four largest k-clipsters together comprised as many reports (60%) as the twenty-one largest Mclusters. Each k-clipster contained facilities that were in multiple Mclusters but very few (2 of 56) Mclusters contained a substantial number of observations from several k-clipsters. Further, the Mclusters were estimated to be $exact - 0.29 - separated$, so at least some of them were poorly-separated. Taken together, we conclude that Mclust subdivided many k-clipsters.

A review of the major k-clipsters showed that oil- and coal-combusting electric power facilities disproportionately populated the second (69.6% of 217 reports) and third largest clusters (58.2% of 212 reports) which were characterized by moderate to high stack air emissions, and high stack air and land releases, respectively. On the other hand, they formed a small proportion (16 or 12% of 134 reports) of the fourth largest ("clean") group, which also contained a substantially higher number of reports from California. The major group memberships thus appeared to be concordant with other intuitive facility characteristics and lent further confidence in the k-clips classification.

Maitra (2007) categorized industrial mercury releases via a mixture model with common Σ that made estimation with singleton groups possible, implicitly allowing for outliers. Similar to k-clips, his five largest groups accounted for almost 60% of facilities. His analysis showed that oil- and coal-combusting electric services dominated the groups characterized by high-volume mercury releases, also corroborated here by k-clips. However, his efforts to properly account for outliers appeared to be only partially successful as many

Table 3

Summary of different types of mercury releases (in pounds) of the major clusters as identified by k-clips and by Maitra (2007). For each cluster, the top row contains cluster sizes (n_c) and means of the different releases. The bottom row contains the number of common facilities in the k-clipster and its closest Maitra (2007) counterpart (first row) followed by the median for each cluster. In all cases, regular fonts (left side) indicate k-clips clusters and where applicable, italicized fonts (right side) represent Maitra (2007) clusters. Note that underground injections releases in each cluster are not displayed as all cluster means and medians were essentially zero. Also, the last k-clipster has no corresponding cluster in Maitra (2007).

n_c		Fugitive Air		Stack Air		Water		Land		Off-site	
257	*281*	0.008	*0.004*	99	*88*	0.021	*0.001*	0.114	*0.049*	0.067	*0.11*
[230]		0	*0*	34	*24.6*	0	*0*	0	*0*	0	*0*
217	*154*	0.009	*0.005*	225	*249*	0.021	*0.003*	153	*186*	0.415	*0.223*
[150]		0	*0*	120	*177*	0	*0*	53	*81.9*	0	*0*
212	*148*	0.009	*0.006*	93.6	*79.3*	0.075	*0.001*	3.59	*0.97*	3126	*4448*
[147]		0	*0*	37.5	*31*	0	*0*	0	*0*	20.2	*19.5*
134	*141*	0.043	*0.05*	0.275	*0.535*	0.025	*0.001*	0.126	*0.094*	0.261	*0.378*
[116]		0	*0*	0.05	*0.15*	0	*0*	0	*0*	0	*0*
124	*103*	0.019	*0.002*	0.273	*0.2*	0.041	*0.001*	1	*0.01*	1064	*1254*
[103]		0	*0*	0	*0*	0	*0*	0	*0*	27.9	*30*
97		0.011		4.36		0.073		0.289		0.434	
		0		3.7		0		0		0	

facilities with extremely high volumes of mercury releases were classified by k-clips as scatter and the k-clipsters were more similar. Table 4 compares the means and medians of the largest clusters of each method.

Maitra's (2007) largest cluster of 281 facilities shared 230 common reports with the largest k-clipster. These clusters shared similar characteristics, including moderately large stack air emissions. The remaining 51 reports had low stack air emissions (from 2–13 lbs) and total offsite releases (0–2 lbs) and were classified in the sixth largest k-clipster with other facilities also having moderately low stack air emissions and some off-site disposals. This cluster was the only large k-clipster not in loose correspondence with a Maitra (2007) cluster. Further, half of the primarily lime-manufacturing facilities were represented in this k-clipster. A quarter of the facilities from this industry were in the 51 discrepant reports above, thus k-clips identified a cluster more meaningful for targeted policy formulation. The fourth largest of Maitra's (2007) groups (141 reports) was characterized by low emissions for all categories. This "clean" group shared 116 of its cleanest facilities with the fourth largest k-clipster, thus making it a more truly "clean" group. The majority of the remaining reports from Maitra's (2007) "clean" group (24 of 25) fell in the sixth largest k-clipster characterized by moderately high stack air releases. Finally, 34 of the 98 scatter observations in k-clips were classified as singleton clusters in Maitra (2007) with another 28 in small clusters of two or three facilities each. The remaining 36 facilities were found in his bigger clusters along with facilities identified as clustered by k-clips. Of these, the largest number to appear in any one cluster was six. This group had 28 facilities and overlapped with a k-clipster of 18 facilities. The six observations had somewhat higher land and total offsite disposals resulting in skewed summaries for the Maitra (2007) group when compared to the above k-clipster.

In summary, k-clips found a new major cluster consisting of facilities with moderate levels of mercury releases that were grouped by Maitra (2007) as either "clean" or heavy polluters. While this new group was the most striking difference, a comprehensive assessment of the groupings indicated sharper boundaries between the k-clipsters. Well-demarcated clusters can lead to better understanding of factors governing mercury releases and more meaningful, effective public policies. In particular, determining what practices are used by industries with little to no levels of mercury releases without erroneously including the moderate polluters may better guide strategies to reduce the mercury effluents of the heavier polluters. Although Mclusters may help achieve the same goal, these less sharply distinguishable groups may not be significantly different, after accounting for measurement error, and could lead to confusing, contradictory and ineffective regulatory policies.

5. Discussion

The main contribution of this paper is the development of a modified k-means algorithm for clustering in the presence of scatter, *i.e.* observations that are unlike any other in the dataset. Several applications in the biological and other sciences need to cluster observations in the presence of scatter. As discussed in Tseng and Wong (2005), standard algorithms lead to erroneous conclusions when applied to such data. Our suggested methodology is an iterative scheme which requires initialization, for which we also provide a deterministic approach. We also developed a BIC-type criterion to estimate the number of clusters, which reduces to Marriott's (1971) criterion when scatter is *a priori* known to be absent. ISO/ANSI-compliant C software implementing k-clips and R code for all simulation datasets used in this paper and the Web Appendix are available upon request. Our algorithm is computer-intensive but can be implemented via modifications of efficient strategies for k-means. Further, while our methodology was developed in the context of adapting k-means, it can be readily retooled for other partitioning algorithms such as k-medoids.

Experimental evaluations of the algorithm in several scenarios were very promising: we almost uniformly outperformed TCTW even when the number of clusters was known and algorithm parameters for the latter were set to maximize the likelihood of the data under the true model. For experiments in which clusters were not well-separated, we typically outperformed Fraley and Raftery's (2006) Mclust even though the experimental datasets were generated using the model explicitly assumed by Mclust. Although Mclust and TCTW (for higher dimensions, with K unknown) performed marginally better for well-separated clusters, k-clips remained superior when clusters had some overlap. Our main algorithm also proved considerably robust to deviations from compact spherical clusters, despite the fact that its development was motivated using Gaussian distributional assumptions. We also estimated the number of clusters very satisfactorily in cases for compact or spherical clusters, but not so well when neither assumption was true. In summary k-clips complements existing clustering methods by excelling on possibly larger datasets whose clusters are not well-separated. Mclust is perhaps a better choice for smaller-sized datasets when clusters are well-separated or have lower-dimensional Gaussian representation. We caution only that the current method for estimating $\hat{K}$ may compromise k-clips's performance on non-spherical clusters, but there was no clear winner for such data.

Accounting for unusual data (*i.e.* scatter) can produce more meaningful classifications, enabling improved understanding of data and clearer distinctions between clusters. For example, our k-clips application to industrial release of mercury and mercury compounds in 2000 produced tighter, more interpretable clusters than a previous attempt (Maitra 2007) and could ultimately lead to improved policies for public health.

A few points remain to be addressed. As mentioned in Section 2.1, convergence of our algorithm needs to be rigorously established. There is some scope for optimism here, given that the algorithm converged in all experiments reported in this paper and the Web Appendix. Further, as seen in the Section 2.2 of the Web Appendix, our suggested initialization strategy did very well when clusters were not well-separated; performance was less emphatic when compared with an initialization strategy based upon using the nearest-neighbor cleaning of Byers and Raftery (1998) followed by the deterministic initialization strategy of Maitra (2007). One suggestion not implemented in our experiments is to obtain starting values using both strategies and to initialize k-clips with the one that optimizes the objective function. Any of these novel strategies could also potentially be modified for use in initializing MClust or TCTW. A second issue pertains to clustering in the presence of scatter using data that are constrained to lie in certain subspaces. Such applications arise, for instance, when the desired metric for clustering is correlation, which is equivalent to applying the Euclidean metric to data sphered after centering (note that Adapt_Clust is specifically designed for such data). Parts of the algorithm would translate readily but core building would need reconsideration. A third issue pertains to clustering massive datasets in the presence of scatter; in this case, it may be possible to adapt this approach within the framework of the multi-stage clustering approach of Maitra (2001). Finally, improved methods for estimating the number of clusters $\hat{K}$ could significantly improve the robustness of our method, since it was shown superior to existing methods for non-Gaussian clusters when K is known. Thus, while the methods suggested in this paper can be regarded as important statistical contributions for clustering datasets in the presence of scatter observations, some issues meriting further attention remain.

Supplementary Materials

The Web Appendix referenced in Sections 1, 2, 3 and 5 is available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgments

The authors very gratefully acknowledge the insightful comments and helpful suggestions of a Co-Editor, an associate editor and three referees which greatly improved this article. This material is based, in part, upon work supported by the National Science Foundation (NSF) under its CAREER Grant No. DMS-0437555 and NSF VIGRE Grant No. 0091953. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics* 49:803-21

Byers, S. and Raftery, A. E. (1998). Nearest-Neighbor Clutter Removal for Estimating Features in Spatial Point Processes. *Journal of the American Statistical Association* 93:577-584.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* 28:781-93.

Dasgupta, S. (1999) Learning mixtures of Gaussians. *Proceedings of IEEE Symposium on Foundations of Computer Science*, 633-44, New York.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., and Moreau, Y. (2002). Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* 18:735-46.

Everitt, B. S., Landau S. and Leesem, M. (2001). *Cluster Analysis* (4th ed.). Hodder Arnold. London.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which cluster method? Answers via model-based cluster analysis. *Computer Journal* 41:578-88.

Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* 97:611-31.

Fraley, C. and Raftery, A. E. (2006) MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering. Technical Report No. 504. University of Washington.

Fitzgerald, W. F., Engstrom, D. R., Mason, R. P. and Nater, E. A. (1998). The case for atmospheric mercury contamination in remote areas. *Environmental Science and Technology*. 32(1):1-7.

Freidman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association* 62:1159-78.

Grandjean, P., Weihe, P., White, R. F., Debes, F., Araki, S., Yokoyama, K., Murata, K., Sorensen, N., Dahl, R., and Jorgensen, P. J. (1997). Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicology and Teratology* 19(6):417-28.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics* 28, 126-130.

Hartigan, J. (1985). Statistical theory in clustering. *Journal of Classification* 2:63-76.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (2000). *Understanding*

Robust and Exploratory Data Analysis. John Wiley and Sons, Inc. New York, NY.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*. 2: 193-215.

Jiang, D., Pei, J, and Zhang, A. (2003). DHC: A Denisity-based Hierarchical Clustering Method for Time Series Gene Expression Data. *Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*. 393-400.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data*. John Wiley and Sons, Inc. New York, NY.

Kettenring, J. R. (2006). The Practice of Cluster Analysis. *Journal of Classification* 23:3-30.

Kjellstrom, T., Kennedy, P., Wallis, S., and Mantell, C. (1986). *Physical and mental development of children with prenatal exposure to mercury from fish. Stage 1: Preliminary tests at age 4.* Sewdish National Environmental Proctection Board, Sweden.

Lin, S. H. (1973). Statistical behavior of rain attenuation. *Bell System Techincal Journal* 52(4):557-81

Maitra, R. (2001). Clustering Massive Datasets With Applications in Software Metrics and Tomography. *Technometrics* 43:336-46.

Maitra, R. (2007). Initializing Partition-Optimization Algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics. http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70244*

Marriott, F. H. (1971). Practical problems in a method of cluster analysis. *Biometrics* 27:501-14.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, Inc. New York, NY.

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

Murtagh, F. (1985). *Multi-dimensional clustering algorithms*. Springer-Verlag (Berlin;New York).

National Academy of Sciences, (2000). *Toxicological Effects of Methylmercury*. National Academy Press, Washington DC.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL *http://www.R-project.org*.

Ramey, D. B. (1985). Nonparametric clustering techniques. In *Encyclopedia of Statistical Sciences.* 6:318-9. Wiley. New York, NY.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*. 27:387-97.

Sörensen, N., Murata, K., Budtz-Jrgensen, E., Weihe, P., and Grandjean, P. (1999). Prenatal exposure as a cardiovascular risk factor at seven years of age. *Epidemiology* 10(4):370-5.

Tibshirani, R., Walther, G. and Hastie, T. J. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society, Series B*, 63(2):411-423.

Tseng, G. C. and Wong, W. H. (2005). Tight Clustering: A resampling based approach for identifying stable and tight patterns in data. *Biometrics* 61:10-16.

Received December 2006. *Revised June* 2007.
Accepted March 2008.