RUTCOR
RESEARCH
REPORT

Scale-invariant Clustering with Minimum Volume Ellipsoids

Mahesh Kumar [a] James B. Orlin [b]

RRR 28-2005, September, 2005

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]Rutgers Business School & RUTCOR, Rutgers University, 180 University Avenue, Newark, NJ 07102, maheshk@rutgers.edu

[b]Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Bldg. E40-149, Cambridge , MA 02139, jorlin@mit.edu

Rutcor Research Report
RRR 28-2005, September, 2005

Scale-invariant Clustering with Minimum Volume Ellipsoids

Mahesh Kumar James B. Orlin

Abstract. This paper develops theory and algorithms concerning a new metric for clustering data. The metric minimizes the total volume of clusters, where the volume of a cluster is defined as the volume of the minimum volume ellipsoid (MVE) enclosing all data points in the cluster. This metric is scale-invariant, that is, the optimal clusters are invariant under an affine transformation of the data space. We introduce the concept of outliers in the new metric and show that the proposed method of treating outliers asymptotically recovers the data distribution when the data comes from a single multivariate Gaussian distribution. Two heuristic algorithms are presented that attempt to optimize the new metric. On a series of empirical studies on real and simulated data sets, we show that volume-based clustering outperforms k-means clustering.

Keywords: Minimum volume ellipsoid, Outliers, Scale-invariant clustering, Robust clustering, Iris data

1 Introduction

A drawback of most traditional clustering methods, such as k-means and hierarchical clustering [8], is that the clustering results are sensitive to the units of measurement, that is, changing the measurement units may lead to a different clustering structure, as shown in Figure 1. A commonly used approach to fix this problem is to normalize the data to unit variance on each variable before clustering. Normalization converts the original variables into unitless variables, but it can produce ad-hoc clusters as suggested in [10]. This motivates the development of a new scale-invariant clustering method that is invariant under an affine transformation of the data space.

Another drawback of a large number of clustering methods is that they do not account for outliers. Among methods that do account for outliers, most of them either cluster first and then identify the outliers, or identify the outliers first and then cluster the remaining data [2, 7, 15]. There arises a circularity in this approach that if we knew the correct clusters then we could identify the correct outliers, and if we knew the correct outliers then we could find the correct clusters.

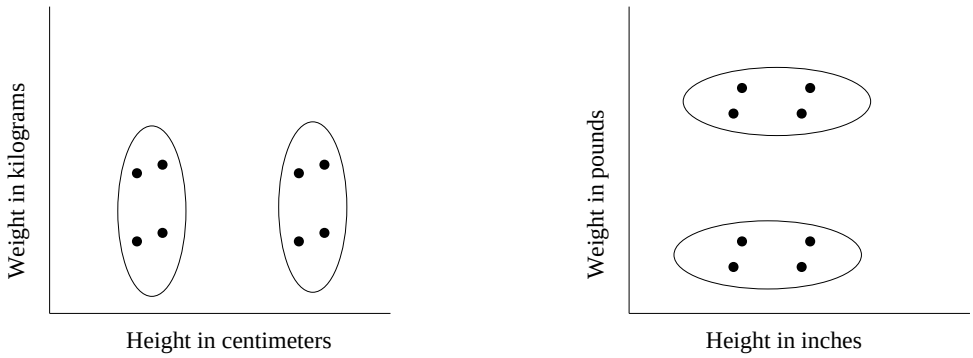

Figure 1: Clusters obtained using k-means on height and weight measurements of people

This paper focuses on two topics: (i) developing and analyzing a new criterion for clustering that has a scale-invariance property and (ii) incorporating methodology that simultaneously clusters and identifies outliers. We address these two issues via a new metric for clustering that is based on the volume computation of clusters. The *volume of a cluster* is defined as the volume of the minimum volume ellipsoid (MVE) enclosing all data points in the cluster. The goal is to minimize the sum of the volumes of these clusters. There are several ways to compute volume of a set of data points (for example, volume of a set of data points can be computed using a convex hull, sphere, rectangle, etc.). We chose the minimum volume ellipsoid in part because real-world data often exhibits a mixture of Gaussian distributions, which have equidensity contours in the shape of ellipsoids [18]. Further, representing clusters using ellipsoids provides a good visualization tool for the clusters [9]. We show that the new metric based on volume computation using MVE is scale invariant.

Given a data set to be clustered and a pre-specified value of α, we permit an α fraction of the data to be treated as outliers. The resulting problem is to find k clusters that include

at least $1 - \alpha$ fraction of the data and have the least total volume. In our empirical study, we show that this method of treating outliers performs well when the data comes from a mixture of Gaussian distributions. When there is only one Gaussian-distributed cluster, we show that the above method of treating outliers discovers the original data distribution asymptotically.

[14, 19] have shown that it is not easy to compute MVE enclosing a set of data points. The combinatorial nature of clustering problem adds further computational challenge. Using an approximate solution for the MVE developed by Sun and Freund [16], we present two heuristic algorithms, *kVolume* and *hVolume*, that are suitable for volume-based clustering. Using these algorithms, we are able to cluster one thousand data points in up to five dimensions in less than five minutes on a personal computer. We illustrate the effectiveness of volume-based clustering on the well known Iris data and simulated data sets, and contrast it to k-means clustering.

The rest of the paper is organized as follows. Section 2 summarizes previous work related to scale-invariant or volume-based clustering. Section 3 presents definition and theoretical properties of the volume-based metric. Section 4 presents two heuristic algorithms, *kVolume* and *hVolume*, that attempt to minimize the volume-based metric. In Section 5, we present empirical study results on Iris data and a series of simulated data sets. Finally, we provide a summary and future research directions in Section 6.

2 Literature Review

The first elaborate work on scale-invariant clustering was done by Friedman and Rubin [6]. They used the Mahalanobis distance function, which is scale-invariant. Marriott [11] generalized the work of Friedman and Rubin to propose several clustering criteria, some of which are scale-invariant. On similar lines, Fraley and Raftery [5] developed a model-based clustering approach for Gaussian clusters whose clustering criterion is scale-invariant. Knorr et al. [10] discussed the problem of scaling in data mining problems and proposed a robust space transformation approach using Donoho-Stahel estimator and showed its application in outlier detection.

The problem of clustering using ellipsoids was first considered by Rosen [13] in the context of separating patterns using convex programming. Barnes [1] provided a heuristic algorithm for this problem using the concept of eigenvalue decomposition. Jolion et al. [9] and Tamez-Pena and Perez [17] showed that using MVE as a basis for clustering produces robust clusters and showed its application in computer vision and image segmentation (also see [3] for a review of work on robust clustering). In all these papers, the authors faced serious computational challenge because it is not easy to compute the MVE for a set of data points.

Although, the MVE can be computed for a set of data points in polynomial time using semi-definite programming [19], in practice it take a lot of computer time even for medium size data sets. SAS offers a subroutine for MVE computation using an algorithm proposed by Rousseeuw and Leroy [14], but even this is not computationally fast enough to be used for clustering. Given the wide applications of computing the MVE, a number of approximation

algorithms were developed recently (see [16] and references therein for a review). An approximate MVE for a set of data points is an ellipsoid that encloses all the points and whose volume is only slightly higher than the volume of the MVE. In our empirical studies, we have used the approximation algorithm developed by Sun and Freund [16] using the notion of active sets.

3 Volume-based Clustering Metric

Given a set of n data points $x_1, x_2, \ldots x_n$ in $\Re^d$ to be grouped into K clusters, and an α between 0 and 1 that corresponds to the fraction of data to be treated as outliers, the objective of volume-based clustering is to minimize the sum of the volumes of the clusters where the volume of a cluster is defined as the volume of the MVE enclosing all points in the cluster, and the outliers are identified as defined below.

Definition 3.1. *Outliers: Given a set of data points S and an $\alpha > 0$ fraction of data to be treated as outliers, the outliers form a subset O of the data such that $|O| \leq \alpha|S|$ and the volume-based clustering obtains the least total cluster volume on the remaining data, $S \backslash O$.*

Let $vol(S)$ denote volume of the MVE enclosing all points in a set S, then volume-based clustering is formally defined as

$$\begin{aligned} Min \quad & \sum_{i=1}^{K} vol(S_i) \\ s.t. \quad & \sum_{i=1}^{K} |S_i| \geq (1-\alpha)n \\ & S_1, S_2, ..., S_K \subseteq \{x_1, \ldots, x_n\}. \end{aligned} \tag{1}$$

Here $S_1, S_2, ..., S_K$ are mutually disjoint subsets. An ellipsoid is mathematically defined using its center c and a symmetric positive definite covariance matrix Q as below.

$$E(c, Q) = \{x | (x-c)^t Q^{-1} (x-c) \leq 1\} \tag{2}$$

The volume of $E(c, Q)$ is equal to $det(Q)^{\frac{1}{2}}$ [19]. Thus the MVE of a set of points S is an ellipsoid $E(c, Q)$ for which $det(Q)^{\frac{1}{2}}$ is minimized and for which the inequality $(x-c)^t Q^{-1}(x-c) \leq 1$ holds for all $x \in S$. Using this definition of MVE, the formulation in Equation 1 can be rewritten as

$$\begin{aligned} Min \quad & \sum_{k=1}^{K} det(Q_k)^{\frac{1}{2}} \\ s.t. \quad & (x_i - c_k)^t Q_k^{-1} (x_i - c_k) \leq 1; \ \ \forall x_i \in S_k; \ \ \forall S_k \\ & \sum_{k=1}^{K} |S_k| \geq (1-\alpha)n \\ & S_1, S_2, ..., S_K \subseteq \{x_1, \ldots, x_n\} \\ & Q_k \in \Omega \quad k = 1, \ldots, K \end{aligned} \tag{3}$$

where Ω is the set of symmetric positive definite matrices.

Lemma 3.1. *Let $f : x \rightarrow Ax+u$ be an affine transformation of the data space, then for any dataset S*

$$vol(f(S)) = vol(S)|det(A)|. \tag{4}$$

Proof. Let $E(c^*, Q^*)$ be the MVE for set S in the untransformed space, that is, c^* and Q^* form an optimal solution of the following optimization problem

$$\begin{array}{ll} Min & det(Q)^{\frac{1}{2}} \\ s.t. & (x-c)^t Q^{-1}(x-c) \leq 1 \quad \forall x \in S \\ & Q \in \Omega \end{array} \tag{5}$$

and $vol(S) = det(Q^*)^{\frac{1}{2}}$. The MVE computation problem in the transformed space becomes

$$\begin{array}{ll} Min & det(Q)^{\frac{1}{2}} \\ s.t. & (Ax+u-c)^t Q^{-1}(Ax+u-c) \leq 1 \quad \forall x \in S \\ & Q \in \Omega \end{array} \tag{6}$$

It is easy to verify that $E(Ac^* + u, AQ^*A^t)$ is an optimal solution of the problem in Equation 6. Therefore, $vol(f(S)) = det(AQ^*A^t)^{\frac{1}{2}} = det(Q^*)^{\frac{1}{2}}|det(A)| = vol(S)|det(A)|$. □

Theorem 3.1. *The volume-based metric is invariant under an affine transformation of the data space.*

Proof. Let $f : x \rightarrow Ax + u$ be any affine transformation of the data space. The optimal clustering in the transformed space is given by

$$\begin{array}{ll} Min & \sum_{i=1}^{K} vol(f(S_i)) \\ s.t. & \sum_{i=1}^{K} |S_i| \geq (1-\alpha)n \\ & S_1, S_2, ..., S_K \subseteq \{x_1, \ldots, x_n\} \end{array} \tag{7}$$

From Lemma 3.1, $\sum_{i=1}^{K} vol(f(S_i)) = |\det(A)| \sum_{i=1}^{K} vol(S_i)$. Since $\det(A)$ is a constant, the optimal clusters in the formulation in Equation 7 are the same as the optimal clusters in the untransformed space according to formulation in Equation 1. □

3.1 Outliers Treatment

It is important to note that volume-based clustering incorporates the concept of outliers in its objective function; thus it facilitates algorithms that identify clusters and outliers simultaneously. We show that when the data consists of a single Gaussian distributed cluster,

the proposed method of identifying outliers is able to discover the correct data distribution asymptotically.

Let us assume that the data comes from a single multivariate Gaussian distribution $N(\mu, \Sigma)$, where μ is the mean of the distribution and Σ is its covariance matrix.

Theorem 3.2. *For a given $\alpha > 0$, as $n \to \infty$, the MVE enclosing at least $1 - \alpha$ fraction of data coming from $N(\mu, \Sigma)$ is given by ellipsoid $E(\mu, r\Sigma)$ for some constant r.*

Proof. As $n \to \infty$, the fraction of data enclosed in a region is equal to the cumulative density of the data distribution in that region. That means the theorem is equivalent to saying that the MVE that encloses at least $1 - \alpha$ cumulative density of $N(\mu, \Sigma)$ is given by an ellipsoid $E(\mu, r\Sigma)$ for some constant r.

Let us choose an r such that the cumulative density of the region inside $E = E(\mu, r\Sigma)$ is equal to $1 - \alpha$. To the contrary, let us assume that there exist an ellipsoid, $E' \neq E$ that has cumulative density of at least $1 - \alpha$ and whose volume is smaller than volume of E. This implies that

$$\int_{E' \setminus E} dx < \int_{E \setminus E'} dx. \tag{8}$$

The density function of $N(\mu, \Sigma)$ is given by $f(x) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$. From the definition of $E = E(\mu, r\Sigma)$ in Equation 2, it follows that for all points inside E, $f(x) \geq (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{r}{2}} = c$, where c is a constant, and for all points outside E, $f(x) < c$. This implies the following.

$$\begin{aligned}
\int_{E'} f(x)dx &= \int_{E \cap E'} f(x)dx + \int_{E' \setminus E} f(x)dx \\
&< \int_{E \cap E'} f(x)dx + \int_{E' \setminus E} cdx \\
&< \int_{E \cap E'} f(x)dx + \int_{E \setminus E'} cdx \\
&\leq \int_{E \cap E'} f(x)dx + \int_{E \setminus E'} f(x)dx \\
&= \int_{E} f(x)dx = 1 - \alpha.
\end{aligned} \tag{9}$$

which contradicts our assumption that E' has cumulative density of at least $1 - \alpha$, and hence the result follows. □

Doing a similar analysis for more than one cluster is not easy, and the natural extension of Theorem 1 to two clusters is not true. In Section 5, we present results from a series of empirical studies on two and three clusters that show that allowing a constant fraction of outliers improves the quality of clusters substantially over the case when outliers are not accounted for.

4 Volume-based Clustering Algorithms

In this section we propose two algorithms that attempt to minimize the total volume of clusters: (1) *kVolume*, an iterative algorithm that partitions the data into a specified number of clusters and (2) *hVolume*, a hierarchical clustering algorithm that produces a nested sequence of clusters. In both algorithms, we use the approximate MVE algorithm developed by Sun and Freund [16].

4.1 kVolume: An Iterative Algorithm

The kVolume algorithm starts with an initial clustering into K clusters and makes incremental improvements in total cluster volume until there is no improvement possible. It is a modification of the k-means algorithm.

4.1.1 Initial clustering

There are several algorithms available in the literature for initial partitioning of a data set [12]. A drawback in a majority of these algorithms is that the initial partition of data usually depends on the units of data measurement. In order to maintain the scale-invariance property of the final clusters, we must have an initial partition that has small total volume and at the same time is independent of the units of measurement. We propose *InitClust* algorithm for finding an initial clustering for the kVolume algorithm.

The InitClust algorithm starts with a single cluster consisting of all data points. The algorithm performs $K-1$ division steps, each of which increases the number of current clusters by one. At each step, the algorithm selects an existing cluster randomly and divides it into two clusters as described below. It randomly selects d data points from the selected cluster and computes a d-dimensional hyperplane that passes through the center of the selected d points and is perpendicular to the hyperplane passing through these d points. Then the algorithm partitions the selected cluster into two clusters, one on each side of this hyperplane. The InitClust algorithm is formally described in Algorithm 1

Algorithm 1 : $InitClust(x_1, \ldots, x_n; K)$

1: $NumClust = 1$
2: $C_1 = \{x_1, \ldots, x_n\}$
3: **while** $NumClust < K$ **do**
4: Select a cluster C_i randomly, $1 \leq i \leq NumClust$
5: Select d data points randomly from C_i
6: Calculate a hyperplane H passing through all d points
7: Divide C_i into two sub-clusters such that the sub-clusters lie on either side of H; Points that lie on the hyperplane can be randomly assigned to either sub-cluster
8: $NumClust = NumClust + 1$
9: **end while**

4.1.2 kVolume algorithm

The kVolume algorithm is an iterative algorithm that starts with an initial partition of data into K clusters and cycles through the following three steps.

- Step 1: For a given set of K clusters, compute the MVE $E(c_k, Q_k)$ for clusters $k = 1, \ldots, K$.
- Step 2: Compute d_i, the distance of x_i to its closest cluster according to the distance function in Equation 10. Sort x_i in the decreasing order of d_i and label the first α fraction of data as outliers.
- Step 3: Reassign each of the remaining data points to its closest cluster.

The distance of a data point x_i from cluster C_k is given by

$$d(x_i, C_k) = (x_i - c_k)^t Q_i^{-1} (x_i - c_k). \tag{10}$$

Note that $d(x, C) \leq 1$ if and only if x is inside the MVE of cluster C. This distance function is popularly known as the Mahalanobis distance, which has the scale-invariance property. We have chosen this distance function because it guarantees a decrease in the objective function in each iteration of kVolume, as shown in Lemma 4.1.

Lemma 4.1. *The total cluster volume decreases monotonically in successive iterations of the kVolume algorithm.*

Proof. Consider a cluster C_i that changed to C_i' after an iteration of kVolume. Let us consider a point x that moved from some other cluster (say C_j) to C_i. This implies that $d(x, C_i) < d(x, C_j)$. Further, we know that $d(x, C_j) \leq 1$ because x was initially inside C_j. The two inequalities together imply that $d(x, C_i) < 1$, which means that x was already inside the MVE of C_i before its movement. The above argument implies that MVE of C_i encloses all points in C_i'. Therefore, the volume of the MVE of C_i' is no more than the volume of the MVE of C_i, and hence the result follows. □

Proposition 4.1. *The kVolume algorithm converges in a finite number of iterations.*

Proof. The theorem follows from Lemma 4.1 and the fact that there are only finite number of different partitions of a set of data points. □

The kVolume algorithm is formally described in Algorithm 2.

A drawback of the kVolume algorithm is that the final clusters may depend on the initial partition. It can also produce empty clusters if all points in a cluster are reassigned to other clusters, thereby reducing the number of clusters. The k-means algorithm also has these shortcomings [12]. We propose the following approach, which is similar to the one that is often used in k-means. Run the kVolume algorithm a large number of times with different random initial partitions and pick the one that has the least total volume. We ignore those solutions that contain one or more empty clusters. Pena et al. [12] have shown that, if k-means is run a large number of times, the resulting clusters will be close to optimal and

Algorithm 2 : $kVolume(x_1, \ldots, x_n; K; \alpha)$

1: Get initial clustering using InitClust$(x_1, \ldots, x_n; K)$
2: Compute $E(c_k, Q_k)$ for all clusters $C_k, k = 1, \ldots, K$
3: Compute $d(x_i, C_k)$ for each data-cluster pair
4: $d_i = \min_{C_k} d(x_i, C_k)$
5: Sort d_i in decreasing order and label the first $\lfloor n\alpha \rfloor$ data points as outliers
6: Reassign each of the remaining $n - \lfloor n\alpha \rfloor$ data points to its closest cluster
7: **if** Clusters change **then**
8: **go to** Step 2;
9: **end if**

insensitive to the initial partition. In our empirical studies, we have found that this is also true for kVolume.

The time complexity of kVolume is $o(d^3nKI)$, where d is the dimension of data, n is the number of data points, K is the number of clusters, and I is the number of iterations the algorithm makes. In our empirical studies, we have found that the algorithm generally converges after a few (typically less than 20 iterations). Further, the algorithm computes volumes of clusters only once at the beginning of each iteration. The computation time is unaffected by α, the fraction of outliers in the data. On a standard PC, kVolume is able to cluster 1,000 data points in 5 dimensions in less than fives minutes.

4.2 hVolume: A Hierarchical Algorithm

Next we develop a hierarchical greedy heuristic that we call as the hVolume algorithm. The hVolume algorithm is a modification of Ward's agglomerative hierarchical clustering algorithm [20]. The key idea in most hierarchical clustering algorithms is to start with n singleton clusters (one cluster for each data point) and then at each step of the algorithm merge a pair of clusters that leads to the minimum increase in the objective function. One difficulty in using a standard hierarchical algorithm for volume-based clustering is that when we merge any two singleton clusters, the increase in total volume is zero because the volume of a cluster consisting of only two data points is zero. This may lead to a merger of two data points that are very far apart. We deal with this difficulty by modifying the standard hierarchical algorithm as follows.

Instead of starting with n singleton clusters, we start with $G > K$ clusters where each starting clustering has at least $d + 1$ points and positive volume. In our implementation of hVolume, we take $G = \lfloor \frac{n}{d+1} \rfloor$ and use kVolume(x, G, α) to obtain starting clusters.

Starting with $G > K$ initial clusters, each step of the hVolume algorithm combines a pair of clusters that leads to the minimum increase (or maximum decrease) in total volume of the resulting clusters. Thus each step of the algorithm reduces the number of clusters by one. The merging process continues until we have K clusters left. The algorithm is formally described in Algorithm 3.

The hVolume algorithm computes the MVE $O(n^2/d^2)$ times and is therefore slower than

Algorithm 3 : $hVolume(x_1, \ldots, x_n; K; \alpha)$

1: **initialization**: $G = \lfloor \frac{n}{d+1} \rfloor$; Find G clusters, $C_1, \ldots, C_G$, using $kVolume(x, G, \alpha)$
2: $\triangle(i,j) = vol(C_i \cup C_j, \alpha) - vol(C_i, \alpha) - vol(C_j, \alpha)$, for $1 \leq i \neq j \leq G$, where $vol(C, \alpha)$ is the volume of MVE enclosing at least $1 - \alpha$ fraction of points from cluster C.
3: **while** $G > K$ **do**
4: Merge the pair of clusters C_i and C_j for which $\triangle(i,j)$ is minimized;
5: $G = G - 1$;
6: Recalculate $\triangle(i,j)$ for $1 \leq i \neq j \leq G$
7: **end while**

the kVolume algorithm. An advantage of using hVolume is that it produces a set of nested clusters for each value of K. This approach is especially valuable when the number K of clusters is not specified as part of the input.

5 Empirical Study

In this section we present empirical study results on a series of simulated datasets and one real-world dataset. We compared clustering results using kVolume and hVolume algorithms against those using k-means algorithm. Since the results from k-means depend on the units of data measurement, we also compared against a version of k-means after normalizing the data to unit variance on each variable. We refer to the normalized version of k-means by k-means(norm). Note that while outliers for kVolume and hVolume were identified as proposed in this paper, the outliers for k-means were identified at the end of the algorithm based on their Euclidean distance from the closest cluster center. We ran kVolume and k-means 100 times with different starting partitions using InitiClust algorithm and picked the one that achieved the lowest value on the respective objective function.

We evaluate a clustering method by its misclassification error, i.e., the number of data points assigned to an incorrect cluster. We found that the misclassification error is significantly smaller for kVolume and hVolume than for k-means. We also found that kVolume generally performed slightly better than the hVolume algorithm. A likely reason for this is as follows. In our implementation, we ran kVolume with 100 different random initial partitions and picked the one that achieved the lowest total volume. This helped kVolume achieve better clusters than hVolume.

There are a large number of clustering methods in literature. We have chosen to compare our technique against k-means for two reasons: (i) k-means is the most popular clustering method in practice and (ii) we are comparing some of the first algorithms developed using the idea of volume calculation. Since k-means is also one of the first algorithms developed for Euclidean distance based clustering, it makes a fair comparison.

5.1 Simulated Data

We generated a series of simulated data sets for the case of two and three clusters in two and three dimensions using the multivariate Gaussian distribution function in MATLAB. One hundred data points were generated for each cluster. We considered several choices for μ and Σ, the parameters of Gaussian distribution. Each simulation experiment was run 100 times and the average misclassification error is reported for seven such experiments in Tables 1- 7.

Table 1: Misclassification error on two spherical clusters of equal size

$$\mu_1 = [0,0], \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \mu_2 = [3,0], \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	15.81	14.72	13.56
hVolume	18.36	15.12	14.91
k-means	13.35	12.76	11.98
k-means(norm)	14.51	13.75	12.98

Table 2: Misclassification error on two spherical clusters of different size

$$\mu_1 = [0,0], \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \mu_2 = [20,0], \Sigma_2 = \begin{bmatrix} 50 & 0 \\ 0 & 50 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	31.36	9.72	2.15
hVolume	35.12	9.17	2.31
k-means	10.06	8.96	7.67
k-means(norm)	12.59	10.87	9.36

We see that in all except one of these experiments, volume-based clustering performed significantly better than k-means. When there are two spherical clusters of equal size (Table 1), k-means performed slightly better than volume-based clustering. This is what we had expected because k-means is the best clustering method for spherical clusters of equal size. When there are two spherical clusters of different size (Table 2), volume-based clustering results improved significantly by allowing outliers, whereas k-means was not affected much by outliers.

Table 3: Misclassification error on one spherical and one elliptical cluster

$$\mu_1 = [0,0], \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad \mu_2 = [5,0], \Sigma_2 = \begin{bmatrix} 5 & 6 \\ 6 & 9 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	2.41	1.74	1.14
hVolume	3.96	1.91	1.63
k-means	33.20	27.54	21.56
k-means(norm)	43.81	37.11	30.85

Table 4: Misclassification error on two elliptical clusters of equal shape and size

$$\mu_1 = [0,0], \Sigma_1 = \begin{bmatrix} 5 & 6 \\ 6 & 9 \end{bmatrix}; \quad \mu_2 = [5,0], \Sigma_2 = \begin{bmatrix} 5 & 6 \\ 6 & 9 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	1.28	1.03	0.52
hVolume	2.11	1.87	1.12
k-means	56.97	54.39	50.07
k-means(norm)	63.72	60.88	56.49

5.2 Iris Data

In order to investigate the effectiveness of the new clustering algorithms in practice, we carried out a comparative analysis on a real data set, called Iris data, that was first published by Fisher [4]. Iris data consists of three species of Iris : Iris Setosa, Iris Versicolor, and Iris Virginica. There are fifty plants of each species with the following four measurements on each plant: petal length, petal width, sepal length, and sepal width. The goal is to separate the 150 plants into three clusters based on these four measurements, so that three clusters correspond to three species of Iris. Table 8 compares the misclassification error between volume-based clustering and k-means.

On further analysis of this data, we found that Setosa is a clearly separable cluster while the other two clusters, Versicolor and Virginica, have significant overlap with each other. All clustering methods were able to identify Setosa more or less correctly, but made mistakes on Versicolor and Virginica. While k-means is unaffected by outliers, volume-based clustering improved substantially by allowing outliers.

Table 5: Misclassification error on two elliptical clusters of different shape and size

$$\mu_1 = [0,0], \Sigma_1 = \begin{bmatrix} 6 & -7 \\ -7 & 9 \end{bmatrix}; \quad \mu_2 = [5,0], \Sigma_2 = \begin{bmatrix} 10 & 7 \\ 7 & 5 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	10.64	9.33	7.42
hVolume	15.17	12.06	7.13
k-means	38.22	36.16	33.78
k-means(norm)	43.54	42.09	40.28

Table 6: Misclassification error on three elliptical clusters of different shape and size

$$\mu_1 = [0,0], \Sigma_1 = \begin{bmatrix} 6 & -6 \\ -6 & 9 \end{bmatrix}; \ \mu_2 = [10,0], \Sigma_2 = \begin{bmatrix} 12 & 7 \\ 7 & 5 \end{bmatrix}; \ \mu_3 = [0,5], \Sigma_3 = \begin{bmatrix} 13 & -7 \\ -7 & 5 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	22.89	20.14	17.48
hVolume	29.18	23.59	19.73
k-means	77.01	75.01	68.94
k-means(norm)	67.59	66.09	61.43

6 Summary and Future Research

In this paper, we have proposed a new metric for clustering that is based on the minimum volume ellipsoid calculation of clusters. We showed that the new metric does not depend on the units of data measurement. We incorporated the concept of outliers in the new metric and showed that the proposed method of treating outliers recovers the data distribution asymptotically when the data comes from a single multivariate Gaussian distribution. We developed two clustering algorithms that simultaneously minimize the total cluster volume and identify outliers in the data. We demonstrated the effectiveness of the new clustering algorithms on both simulated as well as real data sets.

One drawback of the proposed clustering algorithms is that they take more computer time than most traditional clustering algorithms, primarily due to the computation time involved in finding MVE. This limits the use of volume-based clustering to small to medium size data sets. By developing better MVE algorithms, by designing better clustering algorithms, or perhaps by clustering a sample of points, one could make volume-based clustering usable for very large data sets. Another direction for future research would be to extend the theory and empirical studies presented in this paper to elliptical distributed data. Finally, it would

Table 7: Misclassification error on two elliptical clusters of different shape and size in three dimensions

$$\mu_1 = [0,0,0], \Sigma_1 = \begin{bmatrix} 6 & 3 & 4 \\ 3 & 8 & 2 \\ 4 & 2 & 5 \end{bmatrix}; \quad \mu_2 = [0,0,5], \Sigma_2 = \begin{bmatrix} 10 & 2 & 4 \\ 2 & 6 & 2 \\ 4 & 2 & 2 \end{bmatrix}$$

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	2.65	2.45	1.61
hVolume	2.91	2.14	1.58
k-means	55.06	51.38	47.42
k-means(norm)	66.29	62.34	58.50

Table 8: Misclassification error on Iris data

Clustering method	$\alpha = 0$	$\alpha = 0.05$	$\alpha = 0.10$
kVolume	11	2	2
hVolume	13	2	2
k-means	17	17	14
k-means(norm)	25	25	25

be useful to characterize real-world problems where volume-based clustering works well.

7 Acknowledgment

The authors are grateful to Nitin R. Patel, Center for E-business Research at MIT, and Research Resource Committee at Rutgers Business School for their valuable suggestions and financial support.

References

[1] E. R. Barnes. "An Algorithm for Separating Patterns by Ellipsoids," *IBM Journal of Research and Development*, 26(6):759-764, 1982.

[2] V. Barnett and T. Lewis. "Outliers in Statistical Data," *John Wiley*, 1994.

[3] R.N. Dave and R. Krishnapuram. "Robust Clustering Methods: A United View," *IEEE Transactions on Fuzzy Systems*, 5(2): 270-293, 1997.

[4] R. A. Fisher. "Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7(2):179-188, 1936.

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

[6] H. P. Friedman and J. Rubin. "On Some Invariant Criteria for Grouping Data", *American Statistical Association Journal*, 1159-1178, 1967.

[7] J. Hardin and D. Rocke. "Outlier Detection in Multiple Cluster Setting using the Minimum Covariance Determinant Estimator," *Computational Statistics and Data Analysis*, 44:625-638, 2004.

[8] A. K. Jain and R.C. Dubes. "Algorithms for Clustering Data," *Prentice Hall*, 1988.

[9] J. Jolion, P. Meer and S. Bataouche. "Robust Clustering with Applications in Computer Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):791-802, 1991.

[10] E. Knorr, R. Ng and R. Zamar. "Robust Space Transformations for Distance-based Operations," *In Proc. KDD*, 126-135, 2001.

[11] F.H.C. Marriott. "Optimization Methods of Cluster Analysis," *Biometrika*, 69(2):417-421, 1982.

[12] J. Pena, J. Lozano and P. Larranaga. "An Empirical Comparison of Four Initialization Methods for the k-means Algorithm," *Pattern Recognition Letters*, 50:1027-1040, 1999.

[13] J. B. Rosen. "Pattern Separation by Convex Programming," *Journal of Mathematical Analysis and Applications*, 10:123-134, 1965.

[14] P.J. Rousseeuw and A.M. Leroy. "Robust Regression and Outlier Detection," *John Wiley*, 1987.

[15] C.M. Santos-Pereira and A.M. Pires. "Detection of Outliers in Multivariate Data: A Method Based on Clustering and Robust Estimators," *Proc. 15th Symposium in Computational Statistics*, 291-296, 2002.

[16] P. Sun and R.M. Freund. "Computation of Minimum-Volume Covering Ellipsoids," *Operations Research*, 52(5):690-706, 2004.

[17] J.G. Tamez-Pena and A. Perez. "Robust Parallel Clustering Algorithm for Image Segmentation," *Proc. International Society for Optical Engineering*, 737-748, 1996.

[18] Y.L. Tong. "The Multivariate Normal Distribution," *Springer-Verlag*, 1990.

[19] L. Vandenberghe, S. Boyd and S.P. Wu. "Determinant Maximization with Linear Matrix Inequality Constraints," *SIAM Journal on Matrix Analysis and Applications*, 19(2):499-533, 1998.

[20] J.H. Ward. "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association*, 58: 236-244, 1963.