James-Stein Shrinkage to Improve K-means Cluster Analysis

Jinxin Gao
Eli Lilly and Company
Indianapolis, IN

David B. Hitchcock
University of South Carolina
Department of Statistics*

Abstract

We study a general algorithm to improve accuracy in cluster analysis that employs the James-Stein shrinkage effect in k-means clustering. We shrink the centroids of clusters toward the overall mean of all data using a James-Stein-type adjustment, and then the James-Stein shrinkage estimators act as the new centroids in the next clustering iteration until convergence. We compare the shrinkage results to the traditional k-means method. Monte Carlo simulation shows that the magnitude of the improvement depends on the within-cluster variance and especially on the effective dimension of the covariance matrix. Using the Rand index, we demonstrate that accuracy increases significantly in simulated data and in a real data example.

KEYWORDS: Centroids; Effective dimension; K-means clustering; Stein estimation.

*Corresponding author: David B. Hitchcock, Department of Statistics, University of South Carolina, Columbia, SC 29208 (email: hitchcock@stat.sc.edu) (Phone: 803-777-5346).

1 Introduction

Cluster analysis is a method of creating groups of objects, so that objects in the same group are similar and objects in different groups are distinct (Gan et al., 2007). Clustering and classification have a long history and have played an important role in many scientific disciplines. Most current statistical software has specific functions or procedures to perform cluster analysis.

There are several main classes of methods in cluster analysis, including hierarchical clustering, partitional clustering, and model-based clustering. In this paper, the primary emphasis is on the most popular partitioning method, k-means clustering. MacQueen (1967) introduced the k-means method as an alternative to hierarchical clustering methods (see also Hartigan and Wong, 1979). This method is more efficient than hierarchical clustering, especially for large data sets and high-dimensional data sets.

The basic algorithm for the k-means method is as follows:

1. Specify the number of clusters k and then randomly select k observations to initially represent the k cluster centers. Each observation is assigned to the cluster corresponding to the closest of these randomly selected objects to form k clusters.

2. The multivariate means (or "centroids") of the clusters are calculated, and each observation is reassigned (based on the new means) to the cluster whose mean is closest to it to form k new clusters.

3. Repeat step 2, until the algorithm stops when the means of the clusters are constant from one iteration to the next.

In the traditional k-means approach, "closeness" to the cluster centers is defined in terms of squared Euclidean distance, defined by:

$$d_E^2(\mathbf{x}, \bar{\boldsymbol{x}}_{\boldsymbol{c}}) = (\mathbf{x} - \bar{\boldsymbol{x}}_{\boldsymbol{c}})'(\mathbf{x} - \bar{\boldsymbol{x}}_{\boldsymbol{c}}) = \sum_k (x_{ik} - \bar{x}_{ck})^2,$$

where $\mathbf{x} = (x_1, \ldots, x_p)'$ is any particular observation and $\bar{\boldsymbol{x}}_{\boldsymbol{c}}$ is the centroid for, say, cluster c.

Compared to hierarchical clustering, the k-means method is more efficient. Tan (2005) showed that if the number of clusters k is much smaller than the number of observations n, the computation time will be linearly related to n, while the computation time of a hierarchical clustering will be related to n^2. This result makes the k-means method more useful for large data than hierarchical methods. In practice, k-means cluster analyses can be performed readily by many statistical software packages, the `kmeans` function in R (R Development Core Team, 2009) and the `FASTCLUS` procedure in SAS being two examples.

A number of alterations to the k-means algorithm have been developed in the statistical literature. Many previous approaches have sought to make the clustering more robust to outliers than the ordinary k-means algorithm, which relies on least-squares principles. For example, Kaufman and Rousseeuw (1987) developed the well-known k-medoids method, implemented in R by the function `pam`. Cuesta-Albertos et al. (1997) introduced trimmed k-means, in which a certain proportion of outlying objects were removed from the clustering. This trimming method was further developed by Cuesta-Albertos et al. (2008) and García-Escudero et al. (2009).

In addition, many authors have proposed implementations of K-means incorporating variable selection. For example, Krzanowski and Hand (2009) introduced a screening method to quickly select useful variables prior to clustering. Another variable selection method was proposed by Brusco and Cradit (2001). Variable weighting (e.g., DeSarbo et al., 1984; Makarenkov and Legendre, 2001) may achieve similar benefits to variable selection. Steinley and Brusco (2008a) introduced a variable selection method based on a "variance-to-range ratio" standardization. See Steinley and Brusco (2008b) for a comparison of the performances of eight variable selection or weighting methods. Steinley (2006) provides an excellent comprehensive review of the k-means method, including the properties and a multitude of variations thereof.

A distinctive aspect of our method is that it involves shrinkage, and therefore handles generally the situation of when high-noise multivariate data is to be clustered, as opposed to the specific situation of a few outlying objects. We would not specifically classify the proposed method as a robust method (characterized by a breakdown point in the presence of individual outliers). However, we do propose that this method significantly improves cluster recovery compared to traditional K-means in situations where the within-cluster variability is relatively high, when cluster recovery is especially difficult. An insightful related discussion is given in Steinley and Brusco (2008a), who debated whether the performance of a clustering algorithm should be judged based on how well the method recovered the underlying cluster structure or how closely the method came to optimizing an objective function of interest. These two approaches could produce different comparative conclusions. In this paper, however, we will focus on judging performance in cluster recovery.

We present in this paper a method to incorporate James-Stein shrinkage into the k-means algorithm. Section 2 reviews some fundamental facts about James-Stein estimation. In Section 3, we introduce the details of our method. Section 4 describes a basic simulation study that displays the improvement in clustering accuracy our method obtains relative to ordinary k-means, and describes how this improvement depends on characteristics of the data such as the within-cluster variance and the effective dimension of the covariance matrix. In Section 5 we include some supplementary simulations that consider specialized covariance structures. We illustrate the method on a real data set in Section 6, and summarize the results with a brief discussion in Section 7.

2 Background on James-Stein Estimation

In recent decades, the James-Stein approach has been widely used in the problem of statistical estimation. It originated in the context of point estimation of a multivariate normal mean.

2.1 Original James-Stein Estimator

For multivariate normal data, the sample mean maximizes the likelihood function and is the uniformly minimum variance unbiased estimator (UMVUE). However, James and Stein (1961) showed that the sample mean is inadmissible and their estimator, later named the James-Stein estimator, dominates the sample mean when the dimension of the data p is larger than 2.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ be a p-dimensional unknown mean parameter, and let $\mathbf{X} = (X_1, \ldots, X_p)'$ be a p-dimensional observation such that $\mathbf{X} \sim N(\boldsymbol{\theta}, \mathbf{I_p})$, and consider finding the best estimator $\hat{\boldsymbol{\theta}}$ based on the observation $\mathbf{X}$, where $\hat{\boldsymbol{\theta}} = \delta(\mathbf{X})$.

Under squared-error loss, the performance of an estimator $\hat{\boldsymbol{\theta}}$ may be judged by the risk function

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = MSE(\hat{\boldsymbol{\theta}}) = E[L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})].$$

Standard inference results show that the maximum likelihood estimator (MLE), the best linear unbiased estimator, and the least squares estimator all equal the sample mean, but Stein discovered a interesting and surprising phenomenon: If $p \leq 2$, then $\delta_0(\mathbf{X}) = \mathbf{X}$ is admissible; however, if $p > 2$,

$\delta_0(\mathbf{X}) = \mathbf{X}$ is inadmissible, and

$$\delta_{JS}(\mathbf{X}) = \left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right)\mathbf{X}$$

dominates the MLE. Since $\delta_{JS}(\mathbf{X})$ can be thought of as a weighted average of 0 and $\mathbf{X}$, the James-Stein estimator is also called a shrinkage estimator: $\delta_{JS}(\mathbf{X})$ shrinks $\mathbf{X}$ toward 0.

In certain contexts, it makes sense to shrink the usual estimator toward some meaningful nonzero quantity, and James-Stein-type estimators were developed in many such contexts (see, e.g., Lehmann and Casella, 1998, Section 5.6 for examples). Our method will be involve a shrinkage estimator of this nature.

2.2 General James-Stein Estimator

The original James-Stein estimator was obtained when $\mathbf{X} \sim N(\boldsymbol{\theta}, \mathbf{I_p})$. Bock (1975) derived a general James-Stein estimator when the elements of $\mathbf{X}$ may be correlated and have different variances.

Suppose an observation $\mathbf{X}$ is distributed according to the p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $\mathbf{Q}$, where $\mathbf{Q}$ is a symmetric positive definite covariance matrix. Bock showed that a general James-Stein estimator in this setting is

$$\delta_{JS}(\mathbf{X}) = \left(1 - \frac{\hat{p}-2}{\mathbf{X^T Q^{-1} X}}\right)\mathbf{X}$$

where $\hat{p}$ is the effective dimension of $\mathbf{Q}$, which equals the trace of $\mathbf{Q}$ divided by the maximum eigenvalue of $\mathbf{Q}$:

$$\hat{p} = \frac{tr(\mathbf{Q})}{\lambda_{max}(\mathbf{Q})}.$$

Bock showed that this general James-Stein estimator dominates the MLE $\mathbf{X}$ as long as $\hat{p} > 2$. Note that when $\mathbf{Q} = \mathbf{I_p}$, we have $\hat{p} = p$, the effective dimension becomes the actual dimension, and the general James-Stein estimator reduces to the original James-Stein estimator.

2.3 Positive-Part James-Stein Estimator

The fact that the shrinkage coefficient

$$1 - \frac{\hat{p} - 2}{\mathbf{X^T Q^{-1} X}}$$

may be negative is an inconvenient aspect of the original James-Stein estimator, and it can be shown that a restricted "positive-part" estimator (see Baranchik, 1964) is superior. For any scalar y, let y^+ be the nonnegative part of y:

$$y^+ = \begin{cases} y, & y \geq 0 \\ 0, & y < 0. \end{cases}$$

Then the positive-part James-Stein estimator is

$$\delta_{PJS}(\mathbf{X}) = \left(1 - \frac{\hat{p} - 2}{\mathbf{X^T Q^{-1} X}}\right)^+ \mathbf{X}.$$

As shown in, for example, Lehmann and Casella (1998) and Richards (1999), the positive-part James-Stein estimator dominates the original James-Stein estimator, and we will use a positive-part James-Stein estimator within our approach.

3 Methodology

The James-Stein estimator has been widely applied in engineering and economics. However, it has attracted little attention in cluster analysis. Here, we want to use James-Stein-type estimators as centroids in a k-means cluster analysis by shrinking the cluster sample means toward the overall sample mean.

In certain situations, the idea of shrinking is natural in cluster analysis. For instance, Hitchcock et al. (2007) and Hitchcock and Chen (2008) showed that shrinkage methods could aid in the clustering of functional data and binary data, respectively. Here, we construct a shrinkage method based on the James-Stein effect for the purpose of improving the clustering of continuous multivariate data.

Suppose observations $\mathbf{X_{i1}}, \mathbf{X_{i2}}, \ldots, \mathbf{X_{in_i}}$ are independently and identically distributed (iid) observations from k multivariate normal distributions with mean vectors $\boldsymbol{\mu_i}$ and covariance matrices $\mathbf{Q_i}$, where $i = 1, \ldots, k$. That is, we have observations from k subpopulations. Let the sample means of the k clusters produced by the k-means algorithm be $\mathbf{\bar{X}_1}, \mathbf{\bar{X}_2}, \ldots, \mathbf{\bar{X}_k}$. Let the overall sample mean be $\mathbf{\bar{X}}$. Define the James-Stein shrunken centroids as:

$$\mathbf{\bar{X}_i^{JS}} \quad = \quad \mathbf{\bar{X}} + \left[1 - \frac{\hat{p} - 2}{(\mathbf{\bar{X}_i} - \mathbf{\bar{X}})^{\mathbf{T}} \mathbf{Q_i^{-1}} (\mathbf{\bar{X}_i} - \mathbf{\bar{X}})}\right]^{+} (\mathbf{\bar{X}_i} - \mathbf{\bar{X}}), \tag{1}$$

Then we use the James-Stein shrinkage estimators $\mathbf{\bar{X}_i^{JS}}$ $(i = 1, \ldots, k)$ as the new centroids in the k-means method. Note that when the subpopulation covariance matrices are known, the true values $\mathbf{Q_i}$ may be used in this shrunken-centroid formula. When the $\mathbf{Q_i}$ are unknown (as is often the case in practice), the corresponding within-cluster sample covariance matrices $\mathbf{\hat{Q}_i}$ may be used in place of $\mathbf{Q_i}$ in the formula.

The specific algorithm we use is summarized as:

1. Classify the data into k clusters using the k-means method with k random starting points, and obtain the ordinary centroids $\mathbf{\bar{X}_i}$.

2. Shrink the resulting centroids $\mathbf{\bar{X}_i}$ towards the overall sample mean $\mathbf{\bar{X}}$ according to equation (3.1) and get the shrinkage centroids $\mathbf{\bar{X}_i^{JS}}$.

3. Classify data into k clusters using the k-means method with the shrinkage centroids $\mathbf{\bar{X}_i^{JS}}$. If two shrinkage centroids are not distinct, a tiny random jitter may be added to separate them (see details in Section 4).

4. Repeat steps 2 and 3 until convergence is achieved (see details in Section 4).

We emphasize that the within-cluster covariance matrices $\mathbf{Q_i}$ (or their sample estimates) are used only in the calculation of the shrunken cluster centroids. The distance between each observation and each centroid, on which the partitioning in k-means algorithm is based, is still defined as a Euclidean distance, as usual. Maronna and Jacovkis (1974) studied alternative distance measures in the k-means algorithm, and they found that none were as good as Euclidean distance. An alternative approach, which we have not used in this paper, would be to estimate the covariance component in $\mathbf{\bar{X}_i^{JS}}$ by centering the data across clusters (by subtracting the respective cluster centroid from each multivariate object) and then estimating a common covariance matrix using all the objects.

To judge the accuracy of the clustering results, we will calculate the Rand index of the clustering partitions resulting from both ordinary k-means and the shrinkage method. This index measures the similarity between the obtained partition and the true clustering structure underlying the data.

This index was originally defined by Rand (1971) and, following Tan (2005), may be written as follows: Let N_{00} be the number of pairs of objects coming from a different underlying subpopulation and being placed into a different cluster by the algorithm. Let N_{01} be the number of pairs of objects coming from a different underlying subpopulation and being placed into the same cluster by the algorithm. Let N_{10} be the number of pairs of objects coming from the same underlying subpopulation and being placed into a different cluster. Let N_{11} be the number of pairs of objects coming from the same underlying subpopulation and being placed into the same cluster. Then

$$Rand = \frac{N_{00} + N_{11}}{N_{00} + N_{01} + N_{10} + N_{11}}.$$

The index serves as a measure of concordance between the true underlying clustering structure and the result produced by a clustering algorithm.

4 Simulation Studies

In this section we cluster a variety of simulated data sets to compare the performance of our proposed method to that of ordinary k-means. As a basic template, we generated a simulated sample of $n = 50$ objects from two five-dimensional normally distributed subpopulations. The subpopulation means were $\mathbf{0} = (0, 0, 0, 0, 0)$ and $\boldsymbol{\delta} = (\delta, \delta, \delta, \delta, \delta)$, where δ was a fixed constant. The simplest covariance structure we studied was when the subpopulation covariance matrix was set to be $\mathbf{Q} = \sigma \mathbf{I_5}$, where $\mathbf{I_5}$ is a 5×5 identity matrix. The between-cluster dispersion was controlled by δ, and the within-cluster dispersion was controlled by σ. We also studied a variety of other covariance structures, as described below.

In the first step of our method, the initial 2-cluster partition of the objects was found using the k-means algorithm, implemented by the R function `kmeans`. Secondly, we shrunk the centroids towards the overall sample mean according to (1), resulting in the new centroids. Note that sometimes the new centroids might not be distinct and this would stop the k-means algorithm. In that case, we added a small amount of random noise $\mathbf{Y} \sim \mathbf{N}(\mathbf{0}, \mathbf{0.00001} \times \mathbf{I_5})$ to each centroid, implemented by the R function `jitter`, to separate them. We note that for the magnitudes of the data values simulated under these settings, this number 0.00001 represented a value small enough to separate the centroids numerically without creating any *practical* distance between them, relative to the overall data scale. On the other hand, if the data values (and thus the distance between clusters) is measured on a very small scale, then the `jitter` variance may need to be smaller. If the `jitter` variance is several orders of magnitude smaller than the corresponding variable's measurement scale, then this should be sufficient for our purpose.

Finally, we determined that the algorithm converged when the concordance between the partition produced at a given iteration and the partition *produced at the previous iteration* (as measured by the Rand index comparing those two partitions) was 1. In other words, the algorithm stops when the discrepancy between two consecutive partitions is zero (up to mere differences in cluster labels).

We use a variety of different settings for the simulations. In Section 4.1, we investigate the effect of the within-cluster variance σ and the effect of the covariance structure $\mathbf{Q}$ on the accuracy of the proposed method. We first study data with an uncorrelated covariance structure. In addition, we vary the number of underlying clusters, simulating data coming from three and five subpopulations. Then we examine data with a variety of correlated covariance structures, including autoregressive covariance structures. In Section 4.2, we investigate the effect of the effective dimension $\hat{p}$ of the covariance matrix. Further data structures, including both five-dimensional data and eight-dimensional data, as well as data having masking variables, are simulated and studied in Section 5.

4.1 Varying the Within-cluster Variance σ

4.1.1 Uncorrelated Data

In the following simulations, we fixed δ to be 2, and we allowed σ to take the values $0.1, 2, 4, 6, 8,$ or 10. We first examined data in which the variables were uncorrelated. The simulation results from the shrinkage method and the original k-means method are given in Table 1 and presented graphically in Figure 1.

[Figure 1 around here]

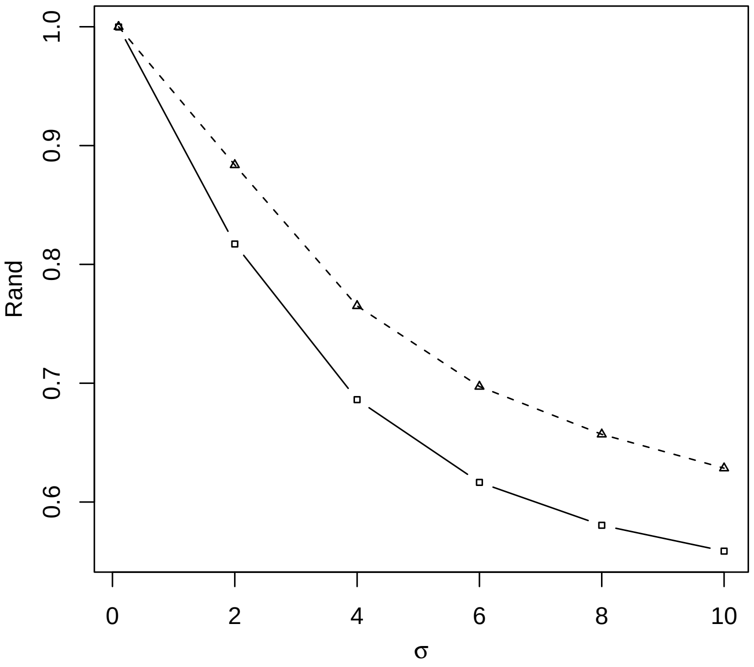


Figure 1: Average Rand index values for the k-means clusterings of 5000 simulated uncorrelated 5-dimensional data sets from 2 subpopulations, based on the ordinary method (squares) and the shrinkage method (triangles).

At each setting, we simulated 5000 data sets and the values in Table 1 represent the average Rand index for those 5000 sets. In examining the Rand indices, we see that the Rand indices based on the shrinkage approach are generally significantly higher than the corresponding ones based on the traditional approach. For $\sigma = 0.1$ (small dispersion within clusters), the Rand indices based on these two approaches are almost identical. For $\sigma = 2$ or 4 (medium dispersion within clusters), the Rand index based on the shrinkage approach is typically significantly better than that based on the traditional approach. For $\sigma = 6, 8$, or 10 (large dispersion within clusters),

Table 1: Average Rand index values for the k-means clustering of the simulated uncorrelated 5-dimensional data from 2 subpopulations with different σ values (Rand values averaged over 5000 simulated data sets and Monte Carlo standard errors given in parentheses).

Method	$\sigma = 0.1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 6$	$\sigma = 8$	$\sigma = 10$
Ordinary	0.9998	0.8171	0.6861	0.6165	0.5804	0.5587
	(0.00008)	(0.00153)	(0.00146)	(0.00125)	(0.00103)	(0.00089)
Shrinkage	1	0.8837	0.7651	0.6972	0.6570	0.6285
	(0)	(0.00086)	(0.00102)	(0.00100)	(0.00094)	(0.00090)

the Rand index based on the shrinkage approach is notably better than that based on the traditional approach. Additionally, we ran the simulation by increasing σ to 100 (extremely large dispersion within clusters), and the resulting Rand index based on the shrinkage approach was only slightly better for such large σ than that based on the traditional approach. These results indicate that with increasing σ, the magnitude of the improvement from the shrinkage approach first increases, then reaches its maximum, and finally decreases. We note that a similar pattern is observed in Hitchcock and Chen (2008).

4.1.2 Varying the Number of Clusters k

The simulations in Section 4.1.1 were done with $k = 2$ clusters. We also performed similar simulations in the case of the data coming from several clusters. In the following setting, we simulated $k = 3$ clusters, each containing 15 five-dimensional objects. The underlying subpopulation means were $\mathbf{0} = (0, 0, 0, 0, 0)$, $\mathbf{2} = (2, 2, 2, 2, 2)$, and $-\mathbf{2}$ here. The subpopulations were set to be normal with covariance matrix $\mathbf{Q} = \sigma\mathbf{I_5}$, and σ again varied as $0.1, 2, 4, 6, 8$, or 10. The Rand indices for the ordinary k-means and the k-means with shrinkage are shown in Table 2. Again, the shrinkage approach appears to lead to better clustering accuracy. The level of improvement from shrinkage (and the consistency of the improvement) was quite similar when $k = 3$ as it was when $k = 2$.

Table 2: Average Rand index values for the k-means clustering of the simulated uncorrelated 5-dimensional data from 3 subpopulations with different σ values (Rand values averaged over 5000 simulated data sets and Monte Carlo standard errors given in parentheses).

Method	$\sigma = 0.1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 6$	$\sigma = 8$	$\sigma = 10$
Ordinary	0.9118	0.8079	0.7292	0.6932	0.6687	0.6489
	(0.00177)	(0.00112)	(0.00062)	(0.00055)	(0.00057)	(0.00061)
Shrinkage	0.9171	0.8783	0.7661	0.7164	0.6843	0.6647
	(0.00178)	(0.00088)	(0.00068)	(0.00055)	(0.00054)	(0.00055)

We then simulated data coming from $k = 5$ clusters, each containing 10 five-dimensional objects. The underlying subpopulation means were set to be $\mathbf{0}$, $\mathbf{2}$, $-\mathbf{2}$ (defined as above), $(-2, 0, 2, 0, -2)$, and $(2, -2, 0, -2, 2)$. The subpopulations had the same covariance structures as in the 2-cluster and 3-cluster case. The top section of Table 3 shows the Rand indices for the ordinary k-means and shrinkage version. For most of the range of σ values examined, the shrinkage method produced better clustering accuracy, but at higher σ, the ordinary k-means had slightly higher Rand indices. A possible explanation for this is that for this data structure, the within-cluster variance was extremely high relative to the between-cluster separation. To investigate this, we simulated data from five clusters with the subpopulation mean vectors twice as large, creating somewhat greater between-cluster separation: $\mathbf{0}$, $\mathbf{4}$, $-\mathbf{4}$, $(-4, 0, 4, 0, -4)$, and $(4, -4, 0, -4, 4)$. The results are shown in the bottom part of Table 3. We see that, in this case, the shrinkage method outperformed ordinary k-means uniformly across every value of σ investigated.

Table 3: Average Rand index values for the k-means clustering of the simulated uncorrelated 5-dimensional data from 5 subpopulations with different σ values (Rand values averaged over 5000 simulated data sets and Monte Carlo standard errors given in parentheses).

Method	$\sigma = 0.1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 6$	$\sigma = 8$	$\sigma = 10$
Ordinary	0.9040	0.8399	0.7702	0.7405	0.7258	0.7161
($\delta = 2$)	(0.00117)	(0.00060)	(0.00050)	(0.00042)	(0.00039)	(0.00036)
Shrinkage	0.9094	0.8732	0.7930	0.7485	0.7178	0.6975
($\delta = 2$)	(0.00117)	(0.00053)	(0.00047)	(0.00052)	(0.00065)	(0.00070)
Ordinary	0.8963	0.9099	0.8926	0.8655	0.8389	0.8167
($\delta = 4$)	(0.00121)	(0.00093)	(0.00075)	(0.00066)	(0.00059)	(0.00057)
Shrinkage	0.8983	0.9446	0.9341	0.9027	0.8741	0.8471
($\delta = 4$)	(0.00120)	(0.00091)	(0.00067)	(0.00058)	(0.00053)	(0.00051)

4.1.3 Correlated Data

In addition to examining data in which the variables were uncorrelated, we also considered the case when the variables were correlated. To investigate this situation, we simulated data having covariance matrix $\mathbf{Q}$ with elements

$$\sigma_{ij} = \begin{cases} \sigma, & \text{if } i = j \\ r\sigma, & \text{if } i \neq j. \end{cases}$$

The simulation results for $r = 0.25$ are given in Table 4.

Table 4: Average Rand index values for the k-means clustering of the simulated correlated 5-dimensional data from 2 subpopulations with different σ values (Rand values averaged over 5000 simulated data sets and Monte Carlo standard errors given in parentheses).

Method	$\sigma = 0.1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 6$	$\sigma = 8$	$\sigma = 10$
Ordinary	0.9996	0.7295	0.6325	0.5911	0.5719	0.5565
	(0.00005)	(0.00116)	(0.00099)	(0.00085)	(0.00076)	(0.00069)
Shrinkage	0.9999	0.7582	0.6523	0.6072	0.5830	0.5657
	(0)	(0.00105)	(0.00095)	(0.00084)	(0.00076)	(0.00070)

Compared to Table 1, a similar pattern is observed. The Rand indices based on the shrinkage approach are generally somewhat higher than those based on the traditional approach. But as Tables 1 and 4 show, the improvement of the shrinkage approach with correlated data is less than the improvement with uncorrelated data. For instance, the biggest improvement for uncorrelated data is 0.0807, while the biggest improvement for correlated data is 0.0287.

4.2 Varying the Effective Dimension $\hat{p}$

Here, we investigate the relationship between $\hat{p}$ and the clustering accuracy. In our simulations in this section, we fixed the actual dimension p to be 5, and σ to be 4. We simulated data having the covariance matrix

$$\mathbf{Q} = \begin{bmatrix} \lambda & 0 & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 & 0 \\ 0 & 0 & \sigma & 0 & 0 \\ 0 & 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & 0 & \sigma \end{bmatrix}$$

We allowed λ to take the values 20.4σ, 6.67σ, 4.4σ, 2.86σ, 2.22σ, 1.82σ, 1.54σ, 1.33σ, 1.18σ and σ, and thus $\hat{p}$ was 1.2, 1.6, 2, 2.4, 2.8, 3.2, 3.6, 4, 4.4, and 5.

The simulated results are shown in Table 5, and presented graphically in Figure 2. Note that with increasing $\hat{p}$, the Rand index increases, because the variances of the components of $\mathbf{X}$ become more balanced and there are fewer outliers.

In examining the Rand indices, when $\hat{p}$ is less than or equal to 2, we do not observe any significant difference between the traditional approach and the shrinkage approach. When $\hat{p}$ is greater than 2, the Rand index values based on the shrinkage approach are higher than those based on the

[Figure 2 around here]

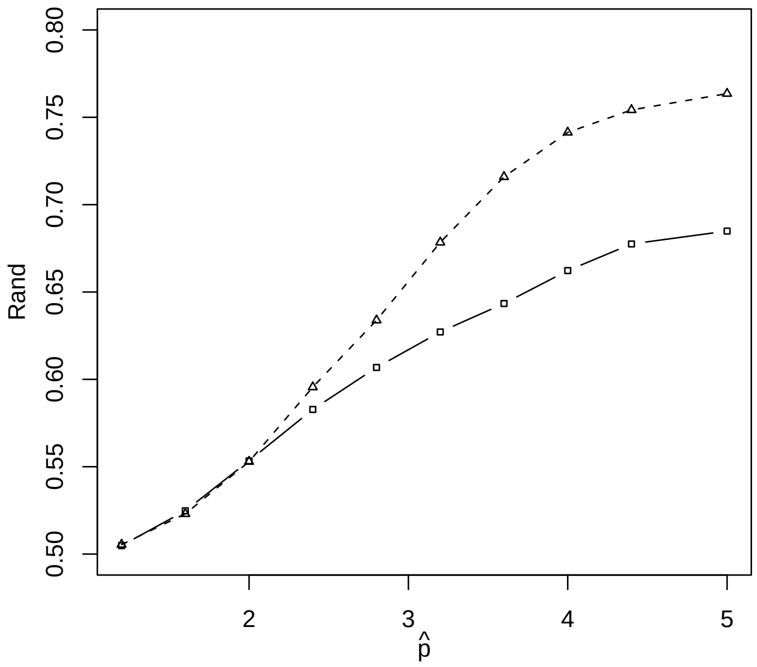


Figure 2: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets from 2 subpopulations, varying different $\hat{p}$, based on the ordinary method (square) and the shrinkage method (triangle).

traditional approach. In addition, the improvement from shrinkage (that is, the difference between the Rand indices of the two approaches) increases with $\hat{p}$, and reaches its maximum when $\hat{p}$ is 5, which is the actual dimension of the data p.

The simulated results for eight-dimensional data are presented graphically in Section 5. A similar pattern is observed: With increasing $\hat{p}$, the Rand index increases.

We see that the simulated results are approximately consistent with Bock's result in the estimation context: $\hat{p} = 2$ seems to be a turning point. When $\hat{p} \leq 2$, the MLE is dominant, so the centroids based on the MLE will give

Table 5: Average Rand index values for the k-means clustering of the simulated data from 2 subpopulations varying different $\hat{p}$ values (Rand values averaged over 5000 simulated data sets and Monte Carlo standard errors given in parentheses).

Method	$\hat{p} = 1.2$	$\hat{p} = 1.6$	$\hat{p} = 2$	$\hat{p} = 2.4$	$\hat{p} = 2.8$
Ordinary	0.5048	0.5248	0.5534	0.5828	0.6068
	(0.00028)	(0.00059)	(0.00094)	(0.00116)	(0.00127)
Shrinkage	0.5054	0.5230	0.5529	0.5956	0.6338
	(0.00029)	(0.00053)	(0.00089)	(0.00121)	(0.00132)
Method	$\hat{p} = 3.2$	$\hat{p} = 3.6$	$\hat{p} = 4$	$\hat{p} = 4.4$	$\hat{p} = 5$
Ordinary	0.6271	0.6434	0.6623	0.6775	0.6849
	(0.00135)	(0.00138)	(0.00141)	(0.00143)	(0.00144)
Shrinkage	0.6784	0.7159	0.7414	0.7542	0.7636
	(0.00128)	(0.00115)	(0.00107)	(0.00105)	(0.00103)

more accurate (or as accurate) clustering results. When $\hat{p} > 2$, the shrinkage estimator is dominant, and thus the centroids based on shrinkage will give more accurate clustering results.

4.3 Joint Effect of $\hat{p}$ and σ Simultaneously

We also investigated the joint effect of $\hat{p}$ and σ on clustering accuracy, and the simulated results are shown in Figure 3. The perspective plot (Figure 3) shows the relationship between the improvement (defined again as the Rand index based on the shrinkage method minus the Rand index based on the traditional method) and $\hat{p}$ and σ. In the perspective plot, the improvement seems to increase significantly as $\hat{p}$ increases, while the improvement increases only slightly as σ increases. Thus $\hat{p}$ seems to play a more important role than σ. In the perspective plot, the role of the contour line $\hat{p} = 2$ as a boundary is clear. When $\hat{p} \leq 2$, the improvement is negative or close to 0. When $\hat{p} > 2$, the improvement increases as $\hat{p}$ and σ increase.

[Figure 3 around here]

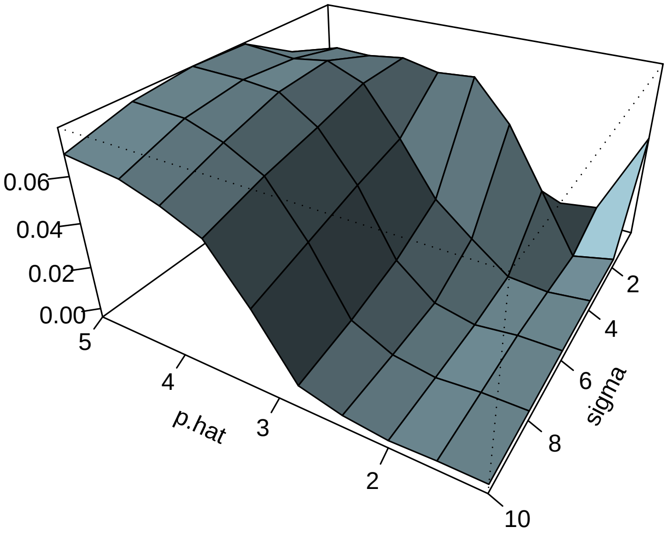


Figure 3: The perspective plot of the improvement of Rand index for the k-means clusterings of 5000 simulated 5-dimensional data sets from 2 subpopulations.

5 Supplementary Simulations

The purpose of this section is to display how the shrinkage method works (relative to the ordinary k-means approach) when the data come from an eclectic variety of covariance structures. We therefore explore simulation settings not considered in Section 4 and will indicate in which situations we can expect the shrinkage method to improve clustering accuracy substantially.

5.1 Additional Cluster Structures

For the first eight sets of additional simulations, we generated a sample of $n = 50$ objects from two five-dimensional normally distributed subpopulations. The subpopulation means were $\mathbf{0} = (0, 0, 0, 0, 0)$ and $\boldsymbol{\delta} = (2, 2, 2, 2, 2)$. The subpopulation covariance matrices were denoted $\mathbf{Q} = \mathbf{Q}_k, k = 1, \ldots, 8$, where each $\mathbf{Q}_k$ was a 5×5 matrix representing one of several various covariance structures. Each $\mathbf{Q}_k, k = 1, \ldots, 8$ is given as follows, and the values of the effective dimension $\hat{p}$ for $\mathbf{Q}_1, \ldots, \mathbf{Q}_8$ are listed in Table 6:

$$\mathbf{Q}_1 = \begin{bmatrix} \sigma & 0 & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 & 0 \\ 0 & 0 & \sigma & 0 & 0 \\ 0 & 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & 0 & \sigma \end{bmatrix} \quad \mathbf{Q}_2 = \begin{bmatrix} \sigma & 0.1\sigma & 0.1\sigma & 0.1\sigma & 0.1\sigma \\ 0.1\sigma & \sigma & 0.1\sigma & 0.1\sigma & 0.1\sigma \\ 0.1\sigma & 0.1\sigma & \sigma & 0.1\sigma & 0.1\sigma \\ 0.1\sigma & 0.1\sigma & 0.1\sigma & \sigma & 0.1\sigma \\ 0.1\sigma & 0.1\sigma & 0.1\sigma & 0.1\sigma & \sigma \end{bmatrix}$$

$$\mathbf{Q}_3 = \begin{bmatrix} \sigma & 0.2\sigma & 0.2\sigma & 0.2\sigma & 0.2\sigma \\ 0.2\sigma & \sigma & 0.2\sigma & 0.2\sigma & 0.2\sigma \\ 0.2\sigma & 0.2\sigma & \sigma & 0.2\sigma & 0.2\sigma \\ 0.2\sigma & 0.2\sigma & 0.2\sigma & \sigma & 0.2\sigma \\ 0.2\sigma & 0.2\sigma & 0.2\sigma & 0.2\sigma & \sigma \end{bmatrix} \mathbf{Q}_4 = \begin{bmatrix} \sigma & 0.3\sigma & 0.3\sigma & 0.3\sigma & 0.3\sigma \\ 0.3\sigma & \sigma & 0.3\sigma & 0.3\sigma & 0.3\sigma \\ 0.3\sigma & 0.3\sigma & \sigma & 0.3\sigma & 0.3\sigma \\ 0.3\sigma & 0.3\sigma & 0.3\sigma & \sigma & 0.3\sigma \\ 0.3\sigma & 0.3\sigma & 0.3\sigma & 0.3\sigma & \sigma \end{bmatrix}$$

$$\mathbf{Q}_5 = \begin{bmatrix} \sigma & -0.2\sigma & -0.2\sigma & -0.2\sigma & -0.2\sigma \\ -0.2\sigma & \sigma & -0.2\sigma & -0.2\sigma & -0.2\sigma \\ -0.2\sigma & -0.2\sigma & \sigma & -0.2\sigma & -0.2\sigma \\ -0.2\sigma & -0.2\sigma & -0.2\sigma & \sigma & -0.2\sigma \\ -0.2\sigma & -0.2\sigma & -0.2\sigma & -0.2\sigma & \sigma \end{bmatrix}$$

$$\mathbf{Q}_6 = \begin{bmatrix} \sigma & 0.4\sigma & 0.3\sigma & 0.2\sigma & 0.1\sigma \\ 0.4\sigma & \sigma & 0.4\sigma & 0.3\sigma & 0.2\sigma \\ 0.3\sigma & 0.4\sigma & \sigma & 0.4\sigma & 0.3\sigma \\ 0.2\sigma & 0.3\sigma & 0.4\sigma & \sigma & 0.4\sigma \\ 0.1\sigma & 0.2\sigma & 0.3\sigma & 0.4\sigma & \sigma \end{bmatrix}$$

$$\mathbf{Q}_7 = \begin{bmatrix} \sigma & -0.4\sigma & 0.3\sigma & 0.2\sigma & 0.1\sigma \\ -0.4\sigma & \sigma & -0.4\sigma & 0.3\sigma & 0.2\sigma \\ 0.3\sigma & -0.4\sigma & \sigma & -0.4\sigma & 0.3\sigma \\ 0.2\sigma & 0.3\sigma & -0.4\sigma & \sigma & -0.4\sigma \\ 0.1\sigma & 0.2\sigma & 0.3\sigma & -0.4\sigma & \sigma \end{bmatrix}$$

$$\mathbf{Q}_8 = \begin{bmatrix} \sigma & -0.4\sigma & 0.3\sigma & -0.2\sigma & 0.1\sigma \\ -0.4\sigma & \sigma & -0.4\sigma & 0.3\sigma & -0.2\sigma \\ 0.3\sigma & -0.4\sigma & \sigma & -0.4\sigma & 0.3\sigma \\ -0.2\sigma & 0.3\sigma & -0.4\sigma & \sigma & -0.4\sigma \\ 0.1\sigma & -0.2\sigma & 0.3\sigma & -0.4\sigma & \sigma \end{bmatrix}$$

Table 6: $\hat{p}$ values for different covariances.

Covariance	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
$\hat{p}$	5	3.57	2.78	2.27	4.17	2.25	2.57	2.25

These covariance matrices yield a variety of correlation structures for the simulated data. $\mathbf{Q}_1$ produces uncorrelated data (much like the simulations in Section 4.1.1), but in which we let the value of σ increase to 100, much larger than in Section 4.1.1. $\mathbf{Q}_2$ through $\mathbf{Q}_4$ include positive off-diagonal elements in an equicorrelation structure, which Dobson (2002) calls "exchangeable." The positive correlation values increase from $\mathbf{Q}_2$ to $\mathbf{Q}_4$ to show the effect on the shrinkage improvement of greater correlation among components. $\mathbf{Q}_5$ yields another equicorrelation (exchangeable) structure, but with negative correlations among components. $\mathbf{Q}_6$ yields an autoregressive-type structure. $\mathbf{Q}_7$ and $\mathbf{Q}_8$ include a variety of positive and negative correlations among components, with $\mathbf{Q}_8$ including more negative correlations.

Figure 4 through Figure 11, shown in the Appendix, display the comparative performances of the shrinkage method and the ordinary k-means method (as measured by average Rand index) for the various covariance structures.

From Figure 4, we see that the improvement from the shrinkage method gradually dissipates as the within-cluster variability gets very large, but the shrinkage method still does better than ordinary k-means for all values of σ in the plot. Figure 5, Figure 6, and Figure 7 show that with the equicorrelation structure, the improvement from shrinkage lessens somewhat as the correlations among components increase.

On the other hand, the improvement from shrinkage is quite sizable when there are negative correlations among components, as shown in Figure 8. This phenomenon is probably due to the low value of the largest eigenvalue of $\mathbf{Q}_5$ and the high value of the effective dimension $\hat{p}$.

However, Figure 9 indicates that the improvement from shrinkage is minimal when the autoregressive structure given by $\mathbf{Q}_6$ is specified. Any improvement from shrinkage is mixed under the less structured covariance patterns of $\mathbf{Q}_7$ and $\mathbf{Q}_8$.

Next, we considered generating 8-dimensional normal data having covariance matrix $\mathbf{Q}_9$. This matrix has $\sigma = 4$ and a variance component λ whose value determines the effective dimension of the data, much like the covariance structure of the simulated data in Section 4.2. This covariance matrix is:

$$\mathbf{Q}_9 = \begin{bmatrix} \lambda & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma \end{bmatrix}$$

We allowed λ to take the values 14σ, 7σ, 4.67σ, 3.5σ, 2.8σ, 2.33σ, 2σ, 1.75σ, 1.55σ, 1.40σ, 1.27σ, 1.16σ, 1.07σ and σ, and thus the corresponding $\hat{p}$ was 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, and 8.

The dependence on the value of $\hat{p}$ was about as marked for the 8-dimensional data as it was for the 5-dimensional data of Section 4.2. As shown in Figure 12, the shrinkage method only began leading to improvement when $\hat{p} \geq 3$, and the improvement became more sizable for the larger values of $\hat{p}$ in the plot.

Lastly, to investigate the performance of the shrinkage approach with data coming from a heavy-tailed multivariate distribution, we generated a sample of $n = 50$ objects from two five-dimensional t-distributed (with 5 degrees of freedom) subpopulations. The subpopulation means were $\mathbf{0} = (0, 0, 0, 0, 0)$ and $\boldsymbol{\delta} = (2, 2, 2, 2, 2)$. The subpopulation covariance matrix was $\mathbf{Q} = \mathbf{Q_1}$, where $\mathbf{Q}_1$ is the 5×5 matrix previously defined. Figure 13 shows the improvement from shrinkage is nearly identical to that under the corresponding situation with normal data that was shown in Figure 1.

5.2 Incorporating Masking Variables

Finally, we address the case of masking variables, or noise variables. Often not all the variables observed will play a role in the clustering structure. Here we simulated data following the same basic pattern as in Section 4.1.1: two clusters of 25 five-dimensional objects each. However, we let $p_M = 3$ of the $p = 5$ variables be masking variables, which do not play a role in the underlying separation of the objects into clusters. Hence only two of the variables were genuine clustering variables. This was achieved by setting the two underlying cluster means to $(2, 2, 0, 0, 0)$ and $(0, 0, 0, 0, 0)$. The results (as shown in Figure 14) indicate that the improvement due to shrinkage is consistent across all values of σ studied, similar to the corresponding situation displayed in Figure 1.

We also simulated data that included $p = 8$ variables, of which $p_M = 5$ were masking variables. The results are shown in Figure 15, and these similarly show consistent improvement due to the shrinkage method.

6 Example: Analysis of Expression Ratios of Yeast Genes

In this section, we apply the shrinkage approach to the yeast gene expression data collected by Alter, Brown and Botstein (2000). The data set contains 78 genes, and the variables are expression ratios measured 18 times (at 7-minute intervals). The details are described in Spellman et al. (1998). The data were log-transformed to make them approximately normally distributed.

Biologists believe that there are five groups in this gene data: genes 1 through 13 are believed to belong to the M/G1 group, genes 14-52 to the G1 group, genes 53-60 to the S group, genes 61-67 to the S/G2 group, and genes 68-78 to the G2/M group. (Here, the letter S denotes Synthesis; the letter M denotes Mitosis; the letter G denotes Gap.) While these classifications are by no means certain, for the purpose of this example we will treat them as the true underlying clusters.

To analyze these data, we treated them as 78 separate observations. Initially, we clustered the genes into 5 clusters using the ordinary approach, implemented by the R function `kmeans`. Then, we clustered the genes into 5 clusters using the shrinkage approach, implemented by our algorithm. For this real-data example, the within-cluster sample covariance matrices $\hat{\mathbf{Q}}_{\mathbf{i}}$ were used in place of $\mathbf{Q}_{\mathbf{i}}$ in formula (1). The results are listed in Table 7.

Table 7: The Rand index values for the yeast gene data using the ordinary k-means approach and the shrinkage approach.

Method	Rand
Ordinary	0.5818
Shrinkage	0.7343

According to the Rand index, the ordinary k-means method did not capture the supposed clustering structure extremely well: 58.18% of the possible pairs of curves were correctly "matched" by this approach. The shrinkage approach came closest to capturing the true grouping of the curves: 73.43% of the possible pairs of curves were correctly "matched" by this approach. The S group (genes 53-60) is the only cluster of the five that is particularly well-defined: these eight genes are classified into the same cluster. The S/G2 group (genes 61-67) is poorly defined. These seven genes are classified into four different clusters; the other three groups are moderately well-defined: these genes from the same group are classified into two different clusters.

In short, the shrinkage approach gave better accuracy than the traditional approach. Of course, conclusions based on the real data must be tempered by the uncertainty about the true number of clusters and the form of the true clustering structure.

7 Conclusion

We have developed an adjustment to the K-means clustering algorithm that relies on James-Stein shrinkage principles. It is particularly related to the multivariate normal mean point estimator of Bock (1975) and simulations show it has similar characteristics to Bock's estimator regarding when it improves on the usual estimator.

Based on various simulations, it appears the improvement from the shrinkage approach is greatest for data: (1) in which the variables are uncorrelated, (2) data with some negative off-diagonal covariance elements, (3) high-dimensional data, or (4) data with moderate to large within-cluster variance. The improvement from the shrinkage approach is less for data in which variables are strongly correlated, low-dimensional data, or data with either small or extremely large within-cluster variance. It is important to note, however, that in virtually all settings studied, the shrinkage approach was approximately as good as or better than the traditional approach. For a relatively wide variety of data, using shrinkage can significantly improve the accuracy of K-means cluster analysis.

Acknowledgments

The authors are grateful to two anonymous referees for comments which resulted in an improvement to this paper.

References

Aarts, E., and Korst, J., 1989. Simulated Annealing and Boltzmann Machines. John Wiley & Sons Inc., New York.

Alter, O., Brown, P. O., and Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA, 97, 10101-10106.

Baranchik, A. J., 1964. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Department of Statistics. Stanford University, Stanford, CA.

Bock, M. E., 1975. Minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. 3, 209-218.

Brusco, M. J. and Cradit, J. D., 2001. A variable-selection heuristic for K-means clustering. Psychometrika. 66, 249-270.

Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C., 1997. Trimmed K-means: An attempt to robustify quantizers. Ann. Statist. 25, 553-576.

Cuesta-Albertos, J. A., Matrán, C. and Mayo-Iscar, A., 2008. Robust estimation in the normal mixture model based on robust clustering. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 779-802.

DeSarbo, W. S., Carroll, J. D., Clark, L. A. and Green, P. E., 1984. Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. Psychometrika. 49, 187-215.

Dobson, A. J., 2002. An Introduction to Generalized Linear Models. Chapman & Hall/CRC Press, Boca Raton.

Gan, G., Ma, C., and Wu, J., 2007. Data Clustering: Theory, Algorithms, and Applications. SIAM-ASA, Alexandria, VA.

García-Escudero, L. A., Gordaliza, A., San Martn, R., Van Aelst, S. and Zamar, R., 2009. Robust linear clustering. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 301-318.

Hartigan, J. A. and Wong, M. A., 1979. A k-means clustering algorithm. J. R. Stat. Soc. Ser. C. Appl. Stat. 28, 100-108.

Hitchcock, D. B., Booth, J. G. and Casella, G., 2007. The effect of pre-smoothing functional data on cluster analysis. J. Stat. Comput. Simul. 77, 1043-1055.

Hitchcock, D. B. and Chen, Z., 2008. Smoothing dissimilarities to cluster binary data. Comput. Statist. Data Anal. 52, 4699-4711.

James, W., and Stein, C., 1961. Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Statist. Prob. 1, 361-379.

Kaufman, L. and Rousseeuw, P. J., 1987. Clustering by means of medoids. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1-Norm. North-Holland, Amsterdam, pp. 405-416.

Kaufman, L., and Rousseeuw, P. J., 1990. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons Inc., New York.

Krzanowski, W. J. and Hand, D. J., 2009. A simple method for screening variables before clustering of microarray data. Comput. Statist. Data Anal. 53, 2747-2753.

Liu, J. S., 2001. Monte Carlo Strategies in Scientific Computing. Springer, New York.

Lehmann, E. L., and Casella, G., 1998. Theory of Point Estimation. Springer-Verlag, New York.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, 1, pp. 281-297.

Makarenkov, V. and Legendre, P., 2001. Optimal variable weighting for ultrametric and additive trees and K-means partitioning: Methods and software. J. Classification. 18, 245-271.

Maronna, R. and Jacovkis, P. M., 1974. Multivariate clustering procedures with variable metrics. Biometrics. 30, 499-505.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 66, 846-850.

R Development Core Team, 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.r-project.org/).

Richards, J. A., 1999. An Introduction to James-Stein Estimation. Accessed at (http://ssg.mit.edu/group/alumni/johnrich/docs/jse.ps.gz) in April 2009.

SAS Institute Inc., 1999. SAS OnlineDoc, Version 8. SAS Institute Inc, Cary, NC.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell. 9, 3273-3297.

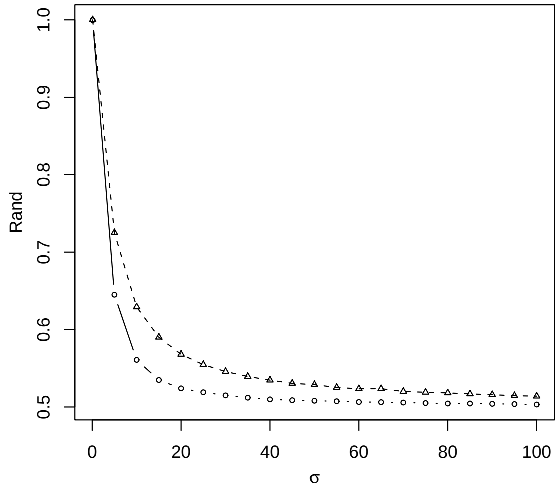
Steinley, D., 2006. K-means clustering: A half-century synthesis. British J. Math. Statist. Psych. 59, 1-34.

Steinley, D. and Brusco, M. J., 2008a. A new variable weighting and selection procedure for K-means cluster analysis. Multivariate Behav. Res. 43, 77-108.

Steinley, D. and Brusco, M. J., 2008b. Selection of variables in cluster analysis: An empirical comparison of eight procedures. Psychometrika. 73, 125-144.

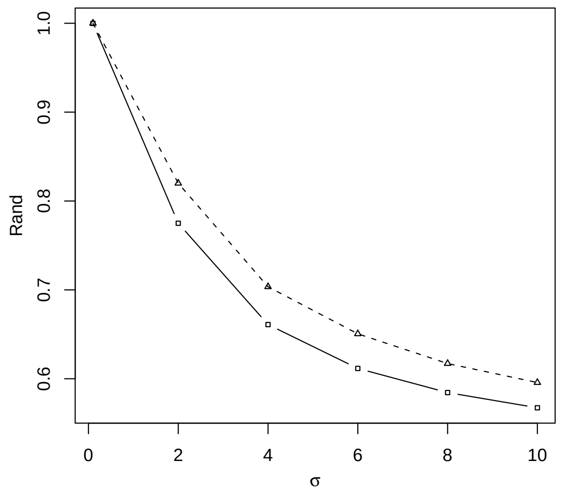
Tan, P., Steinbach, M., and Kumar, V., 2005. Introduction to Data Mining. Addison-Wesley, Reading, MA.

Appendix: Figures for Supplementary Simulations



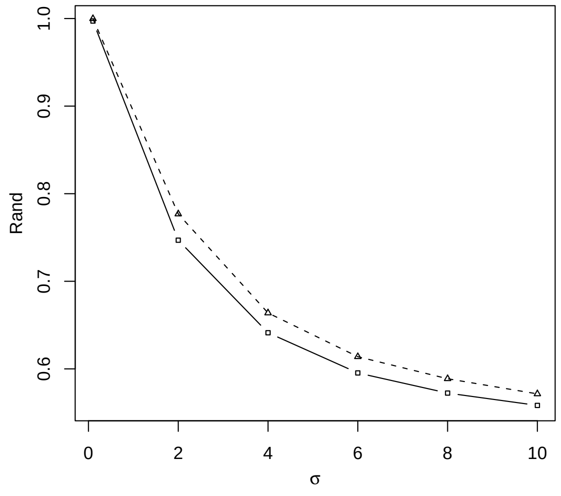
[Figure 4 around here]

Figure 4: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_1, based on the ordinary method (squares) and the shrinkage method (triangles).



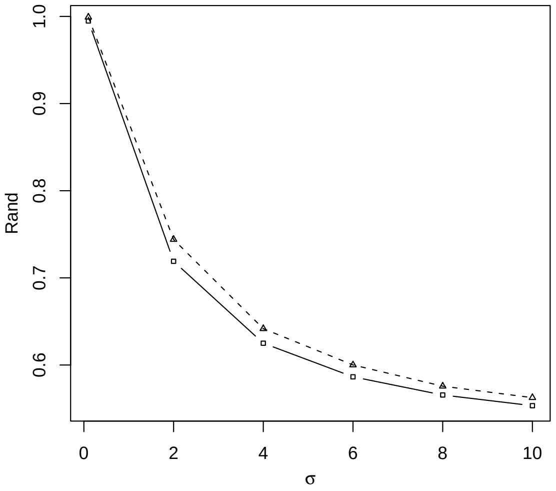
[Figure 5 around here]

Figure 5: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_2, based on the ordinary method (squares) and the shrinkage method (triangles).



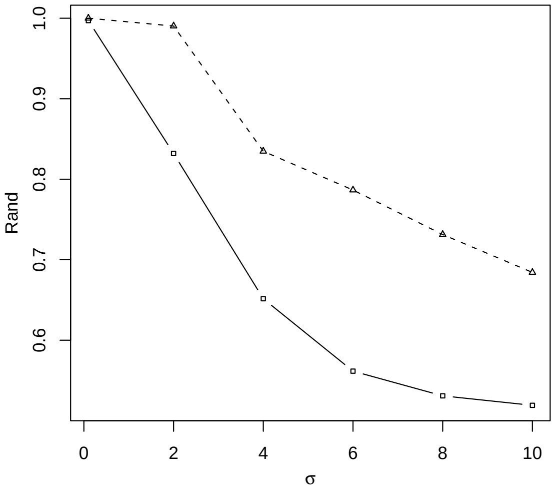
[Figure 6 around here]

Figure 6: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_3, based on the ordinary method (squares) and the shrinkage method (triangles).



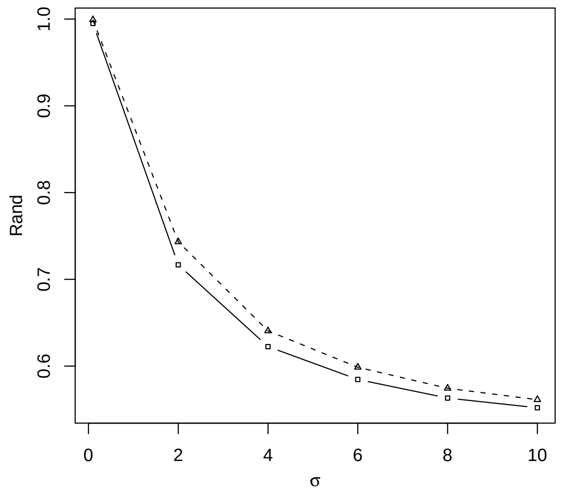
[Figure 7 around here]

Figure 7: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_4, based on the ordinary method (squares) and the shrinkage method (triangles).



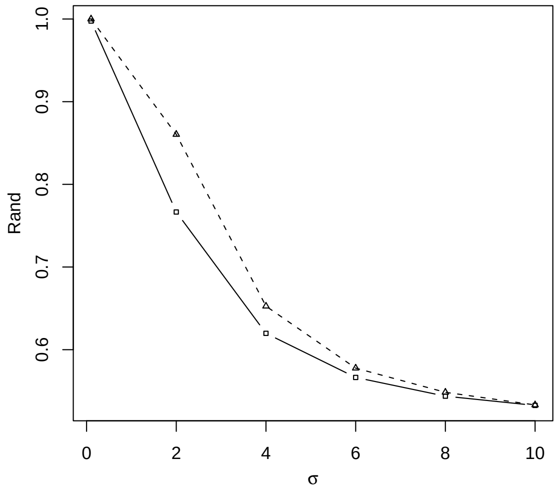
[Figure 8 around here]

Figure 8: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_5, based on the ordinary method (squares) and the shrinkage method (triangles).



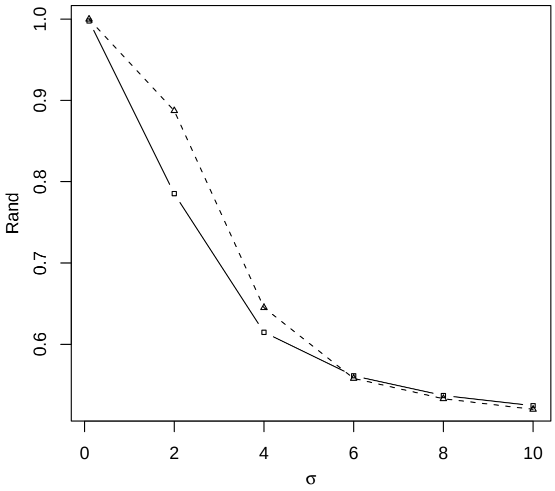
[Figure 9 around here]

Figure 9: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_6, based on the ordinary method (squares) and the shrinkage method (triangles).



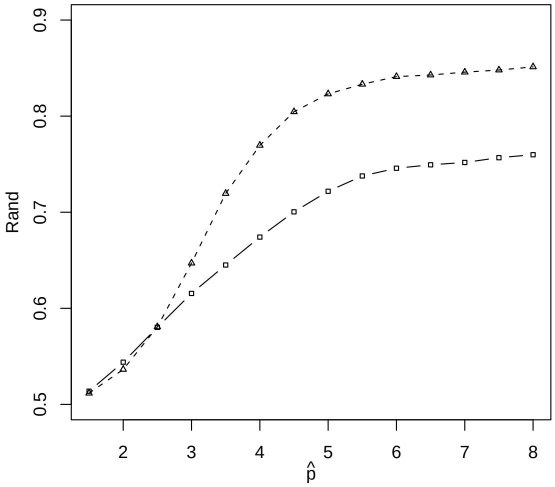
[Figure 10 around here]

Figure 10: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_7, based on the ordinary method (squares) and the shrinkage method (triangles).



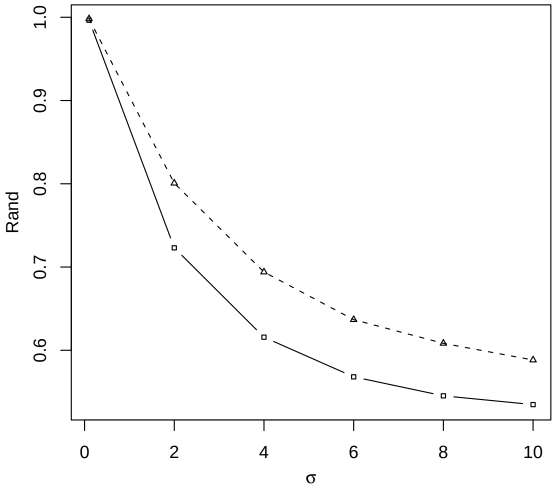
[Figure 11 around here]

Figure 11: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets with covariance Q_8, based on the ordinary method (squares) and the shrinkage method (triangles).



[Figure 12 around here]

Figure 12: Average Rand index values for the k-means clusterings of 5000 simulated 8-dimensional data sets with covariance Q_9, based on the ordinary method (squares) and the shrinkage method (triangles).



[Figure 13 around here]

Figure 13: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets from multivariate t-distribution, based on the ordinary method (squares) and the shrinkage method (triangles).

[Figure 14 around here]

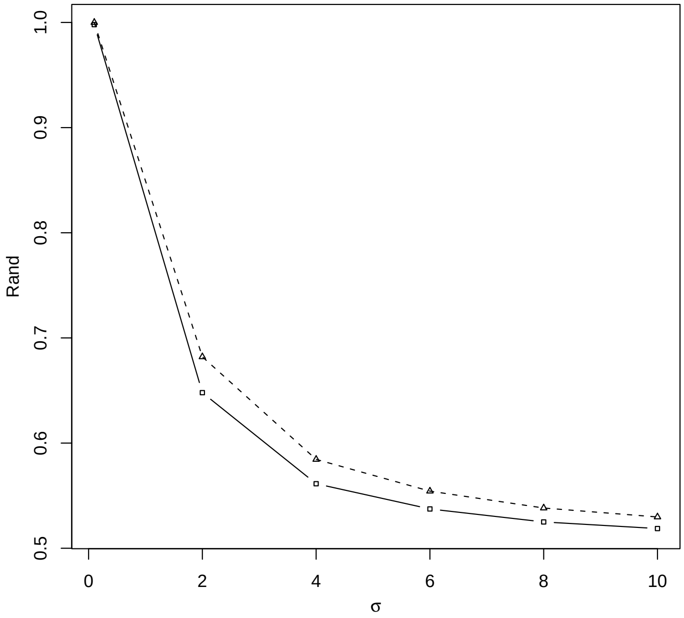


Figure 14: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets having 2 genuine clustering variables and 3 masking variables, based on the ordinary method (squares) and the shrinkage method (triangles).

[Figure 15 around here]

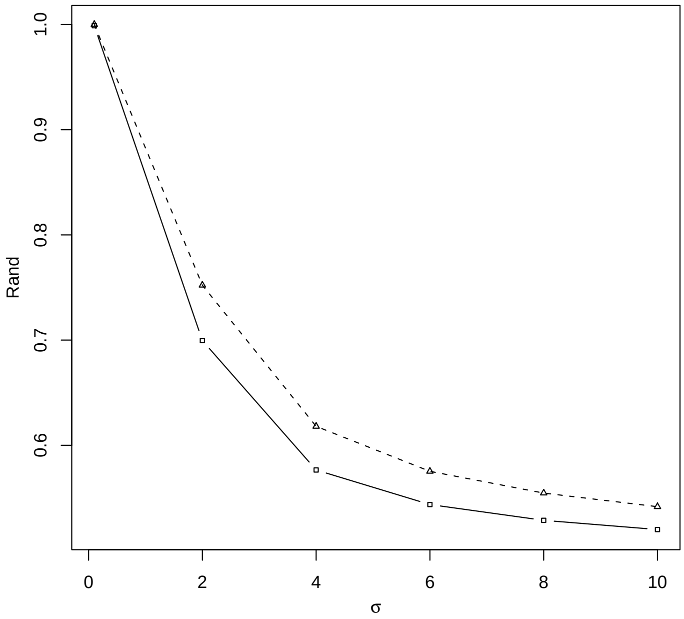


Figure 15: Average Rand index values for the k-means clusterings of 5000 simulated 5-dimensional data sets having 3 genuine clustering variables and 5 masking variables, based on the ordinary method (squares) and the shrinkage method (triangles).