Statistics and Its Interface Volume 6 (2013) 211–221

Penalized unsupervised learning with outliers

Daniela M. Witten

We consider the problem of performing unsupervised learning in the presence of outliers – that is, observations that do not come from the same distribution as the rest of the data. It is known that in this setting, standard approaches for unsupervised learning can yield unsatisfactory results. For instance, in the presence of severe outliers, K-means clustering will often assign each outlier to its own cluster, or alternatively may yield distorted clusters in order to accommodate the outliers. In this paper, we take a new approach to extending existing unsupervised learning techniques to accommodate outliers. Our approach is an extension of a recent proposal for outlier detection in the regression setting. We allow each observation to take on an "error" term, and we penalize the errors using a group lasso penalty in order to encourage most of the observations' errors to exactly equal zero. We show that this approach can be used in order to develop extensions of K-means clustering and principal components analysis that result in accurate outlier detection, as well as improved performance in the presence of outliers. These methods are illustrated in a simulation study and on two gene expression data sets, and connections with M-estimation are explored.

Keywords and phrases: Robust, Group lasso, Clustering, Principal components analysis, M-estimation.

1. INTRODUCTION

It has long been known that in the presence of outliers, classical statistical methods such as linear regression can fail. For this reason, many proposals for detecting outliers in regression have been made (for a survey, see Rousseeuw and Leroy, 1987; Rousseeuw and Hubert, 2011). Recently, She and Owen (2011) proposed a new approach for detecting outliers in regression problems. In regression, one generally assumes the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1}$$

where $\mathbf{y}$ is an outcome vector of length n, $\mathbf{X}$ is a $n \times p$ design matrix, and $\boldsymbol{\epsilon}$ a n-vector of error terms. If some outliers are present among the observations, then a more accurate model might be

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{2}$$

where $\boldsymbol{\gamma}$ is a sparse n-vector whose nonzero elements correspond to outlying observations. To fit the model (2), She and Owen (2011) propose solving the optimization problem

$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{y} - \boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{i=1}^{n} P(\gamma_i; \lambda) \right\}, \tag{3}$$

where $P(\gamma_i; \lambda)$ is a penalty on γ_i that encourages sparsity, and λ is a tuning parameter. It is shown in She and Owen (2011) that there is a close connection between M-estimation and (3). For instance, if $P(\gamma_i; \lambda) = \lambda \sum_{i=1}^{n} |\gamma_i|$, then solving (3) is equivalent to the M-estimate based on Huber's loss function (Huber, 1964).

In this paper, we consider the problem of performing unsupervised learning in the presence of outliers – that is, observations that do not come from the same distribution as the rest of the data. Specifically, we investigate K-means clustering and principal components analysis (PCA) in the presence of outliers. Both of these methods can perform very poorly when outliers are present. A number of proposals have been made for modifying these techniques to accommodate outliers (among others, Jolion and Rosenfeld, 1989; Dave, 1991; Garcia-Escudero and Gordaliza, 1999; Jiang, Tseng and Su, 2001; Fraley and Raftery, 2002; Tseng and Wong, 2005; Tseng, 2007). We instead propose an approach for unsupervised learning in the presence of outliers that is motivated by the work of She and Owen (2011). The flexible framework that we propose can be applied to K-means clustering, PCA, and other unsupervised learning techniques.

In recent years, much effort has focused upon the use of penalties to perform feature selection in regression problems, and to a lesser extent in the classification setting (for instance, see Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006; Witten and Tibshirani, 2011). Other work has involved transferring some of the ideas developed for performing feature selection in supervised contexts to the unsupervised setting (examples include Friedman, Hastie and Tibshirani, 2007; Pan and Shen, 2007; Xie, Pan and Shen, 2008; Witten, Tibshirani and Hastie, 2009; Witten and Tibshirani, 2010; Guo et al., 2011). In this paper, rather than using penalties to perform feature selection in the unsupervised setting, we develop an approach for performing *observation selection* in the unsupervised context – that is, an approach to obtain

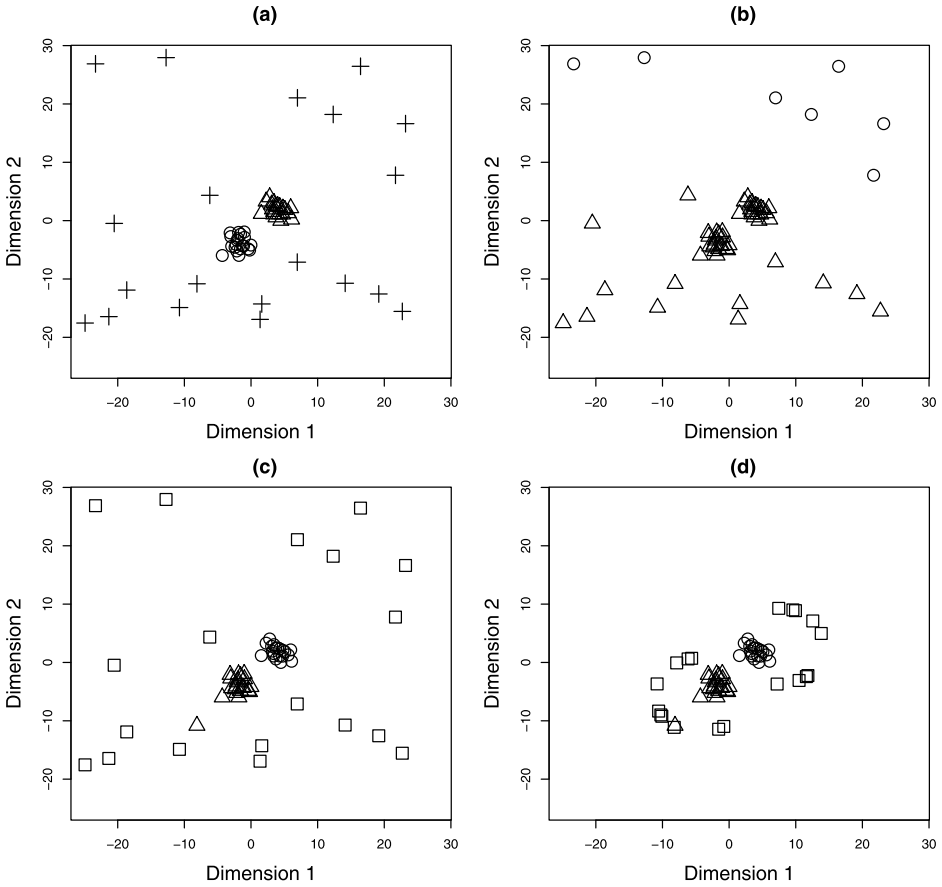

Figure 1. (a): Three sets of observations were generated in two dimensions: two clusters (shown as circles and triangles) and a set of outliers that belong to neither cluster (shown as crosses). (b): The cluster assignments from K-means are shown as triangles and circles. K-means clustering fails to correctly identify the two clusters, due to the presence of outliers. (c): The cluster assignments from our outlier clustering proposal (with a group lasso penalty) are shown as triangles and circles. The squares indicate observations that were assigned nonzero errors; i.e. that were identified as outliers. (d): $\mathbf{X}_i - \mathbf{E}_i$ is shown. The cluster assignments are shown as circles and triangles, and the observations with nonzero errors are shown as squares. The errors were assigned so that the resulting $\mathbf{X}_i - \mathbf{E}_i$ are quite consistent with the true clusters.

not sparsity in the features, but rather sparsity in the observations, where observations are excluded if they appear not to arise from the unsupervised model being fit to the majority of the observations.

Figure 1 illustrates the performances of K-means clustering and our outlier K-means clustering proposal on a small simulated example that consists of two clusters, along with a number of outliers. Our outlier K-means proposal is able to successfully identify outliers and cluster the non-outlying observations.

The rest of this paper is organized as follows. In Section 2, we propose our approach for unsupervised learning in the presence of outliers. In Section 3, we discuss K-means clustering in the presence of outliers as well as its connection to a generalized version of K-means clustering using Huber's loss function. A simulation study and an application to gene expression data are also presented. In Section 4 we discuss PCA in the presence of outliers, and present a simulation study. The discussion is in Section 5.

2. A PROPOSAL FOR UNSUPERVISED LEARNING WITH OUTLIERS

Suppose that we have a $n \times p$ data matrix $\mathbf{X}$, consisting of p feature measurements on a set of n observations. We wish to perform an unsupervised analysis of this data set, such as clustering or PCA. We assume that the procedure of interest solves an optimization problem of the form

$$\underset{\theta \in D}{\text{minimize}} \{f(\mathbf{X}, \theta)\}, \tag{4}$$

where θ represents a set of parameters for the unsupervised learning operation that is restricted to belong to a set D. For

instance, K-means clustering involves solving the problem

$$\underset{C_1,\ldots,C_K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k}{\text{minimize}} \left\{ \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{X}_i - \boldsymbol{\mu}_k\|^2 \right\} \tag{5}$$

where $\mathbf{X}_i \in \mathbb{R}^p$ denotes the ith observation (row) of the data matrix $\mathbf{X}$, $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^p$ denote the mean vectors for the K clusters, and $C_1, \ldots, C_K$ denote a partition of the n observations into K clusters, such that $C_k \cap C_{k'} = \emptyset$ for $k \neq k'$ and $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$. PCA can also be written as the solution to an optimization problem: the first K principal components of $\mathbf{X}$ are the columns of the matrix $\mathbf{V}$ that solves

$$\underset{\mathbf{D},\mathbf{U},\mathbf{V}}{\text{minimize}} \left\{ \frac{1}{2} \|\mathbf{X} - \mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 \right\}, \tag{6}$$

where $\mathbf{D}$ is a $K \times K$ diagonal matrix, and $\mathbf{U}$ and $\mathbf{V}$ are $n \times K$ and $p \times K$ orthogonal matrices, respectively.

Now, suppose that some of the observations in $\mathbf{X}$ are outliers. A simple model for this situation is as follows. We wish to learn the underlying signal in a $n \times p$ data matrix $\mathbf{W}$ via an unsupervised approach, but we do not observe $\mathbf{W}$; we instead observe $\mathbf{X} = \mathbf{W} + \mathbf{E}$, where $\mathbf{E}$ is a $n \times p$ matrix of *errors* for the observations. Most of the observations do not have errors, and so most rows of $\mathbf{E}$ contain only zeros. However, some subset of the observations are outliers, and hence contain errors. These correspond to nonzero rows of $\mathbf{E}$. These errors might potentially be very large.

If $\mathbf{E}$ were known then our task would be simple: we could just perform an unsupervised analysis on $\mathbf{W} = \mathbf{X} - \mathbf{E}$ instead of $\mathbf{X}$, leading to the optimization problem

$$\underset{\theta \in D}{\text{minimize}} \{ f(\mathbf{X} - \mathbf{E}, \theta) \} \tag{7}$$

instead of (4). Unfortunately, this is not possible because $\mathbf{E}$ is unknown. In particular, we do not know which observations are outliers, and much less the error terms associated with these outliers.

Therefore, rather than solving (7) as written, we propose to optimize it with respect to θ and $\mathbf{E}$ jointly. That is, we will estimate the errors for the observations, subject to a penalty intended to encourage just a few observations to be outliers. In particular, we propose to solve

$$\underset{\theta \in D, \mathbf{E}}{\text{minimize}} \left\{ f(\mathbf{X} - \mathbf{E}, \theta) + \sum_{i=1}^{n} P(\|\mathbf{E}_i\|_2; \lambda) \right\} \tag{8}$$

where λ is a nonnegative tuning parameter, and where $P(\cdot; \lambda)$ is a penalty function that encourages sparsity and that is applied to the ℓ_2 norm of the ith error term, i.e. it encourages the ith error term to be zero. There are a number of possible choices for the penalty function, which may be convex or non-convex; some possibilities are surveyed in She (2011) and Mazumder, Friedman and Hastie (2011). In the examples presented in this paper, we will take $P(\|\mathbf{E}_i\|_2; \lambda) = \lambda \|\mathbf{E}_i\|_2$. This is a *group lasso* penalty (Yuan and Lin, 2007). When $\lambda \to \infty$ then $E_{ij} = 0$ for all i and j, and so (8) will just reduce to (4). When $\lambda = 0$, then the errors can become arbitrarily large; each observation will be assigned an error, leading in general to useless results. However, for an intermediate value of λ, just a subset of the observations will be assigned nonzero errors. These observations with nonzero errors are thought to be outliers. Unsupervised learning is performed at the same time as the outliers are identified, so outliers are defined with respect to the unsupervised approach used.

Algorithm 1 A descent algorithm for unsupervised learning with outliers

1. Initialize $\mathbf{E}$, an $n \times p$ matrix of errors.
2. Iterate until convergence:
   (a) Holding $\mathbf{E}$ fixed, solve (8) with respect to θ. This amounts to performing unsupervised learning on the matrix $\mathbf{X} - \mathbf{E}$.
   (b) Holding θ fixed, solve (8) with respect to $\mathbf{E}$. This amounts to updating the estimates of which observations are outliers, given the current value of θ.
3. Perform unsupervised learning (solve (4)) on the observations that were assigned zero errors, i.e. on the observations in the set $\{i : \|\mathbf{E}_i\|_2 = 0\}$.

In general, we will take an iterative approach to solving the problem (8), as outlined in Algorithm 1. Each iteration of Step 2 of Algorithm 1 decreases the objective. In the next two sections, we will apply this formulation for unsupervised learning with outliers to two specific problems, K-means clustering and PCA.

3. K-MEANS CLUSTERING WITH OUTLIERS

3.1 The proposal

Suppose that we wish to cluster a set of observations, but we think that some of the observations are outliers that do not belong to any cluster. Rather than solving (5), we can instead solve the problem

$$\underset{C_1,\ldots,C_K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\mathbf{E}}{\text{minimize}} \left\{ \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{X}_i - \mathbf{E}_i - \boldsymbol{\mu}_k\|^2 + \sum_{i=1}^{n} P(\|\mathbf{E}_i\|_2; \lambda) \right\}. \tag{12}$$

As previously described, $\mathbf{E}$ is a matrix of elements that allows for "errors" in $\mathbf{X}$. That is, if the observation $\mathbf{X}_i$ does not seem to belong to any cluster, then $\mathbf{E}_i$ will take on a

Algorithm 2 A descent algorithm for outlier K-means clustering

1. Initialize the errors – e.g. take $\mathbf{E}_i = \mathbf{0}$ for the 90% of observations that are closest to the overall mean of the observations in terms of Euclidean distance, and $\mathbf{E}_i = \mathbf{X}_i$ for the others.
2. Iterate until the objective (12) converges:
   (a) Perform K-means clustering on the matrix $\mathbf{X} - \mathbf{E}$. That is, solve
   $$\underset{C_1,\ldots,C_K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{X}_i - \mathbf{E}_i - \boldsymbol{\mu}_k\|^2 \right\}. \quad (9)$$
   (b) For each observation, $i = 1, \ldots, n$, solve the problem
   $$\underset{\mathbf{E}_i}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{X}_i - \mathbf{E}_i - \boldsymbol{\mu}_{C(i)}\|^2 + P(\|\mathbf{E}_i\|_2; \lambda) \right\} \quad (10)$$
   where $C(i)$ indicates the current cluster assignment of the ith observation, i.e. $C(i) = k$ if $i \in C_k$. If $P(\|\mathbf{E}_i\|_2; \lambda) = \lambda\|\mathbf{E}_i\|_2$, the solution takes the form (Yuan and Lin, 2007)
   $$\mathbf{E}_i = (\mathbf{X}_i - \boldsymbol{\mu}_{C(i)}) \max\left(0, 1 - \frac{\lambda}{\|\mathbf{X}_i - \boldsymbol{\mu}_{C(i)}\|_2}\right). \quad (11)$$
3. Perform K-means clustering on the observations that were assigned zero errors, i.e. on the observations in the set $\{i : \|\mathbf{E}_i\|_2 = 0\}$.

nonzero value such that $\mathbf{X}_i - \mathbf{E}_i$ seems to belong to a cluster. $P(\|\mathbf{E}_i\|_2; \lambda)$ is a penalty function that encourages sparsity in $\|\mathbf{E}_i\|_2$; throughout this paper, we take $P(\|\mathbf{E}_i\|_2; \lambda) = \lambda\|\mathbf{E}_i\|_2$, a group lasso penalty. Then as $\lambda \to \infty$, (12) becomes equivalent to the K-means clustering criterion (5), since then the penalty for having a nonzero error becomes arbitrarily large and so all errors are zero. On the other hand, when $\lambda = 0$ we obtain a trivial result: there is no penalty for having nonzero errors, and so the errors will be such that the within-cluster sum of squares of $\mathbf{X} - \mathbf{E}$ is zero, i.e. $\sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{X}_i - \mathbf{E}_i - \boldsymbol{\mu}_k\|^2 = 0$. For an intermediate value of λ, some of the observations – those that do not seem to belong to any cluster – will be assigned nonzero errors. We refer to (12) as the *outlier K-means clustering optimization problem*, and the clustering procedure based on this criterion as *outlier K-means clustering*.

An example is shown in Figure 1, using a group lasso penalty. Outlier K-means clustering successfully identifies the outlying observations and assigns large errors to them, allowing for correct discovery of the two true clusters.

Algorithm 2 is a descent algorithm for performing outlier K-means clustering. That is, each step will decrease the objective of (12). Note that in Step 2(b), the problem (10) is convex if $P(\|\mathbf{E}_i\|_2; \lambda)$ is convex, in which case (11) solves it exactly. However, in Step 2(a), (9) is a non-convex problem, since K-means clustering is non-convex, and so only a local optimum will be obtained.

3.2 Tuning parameter selection for outlier K-means clustering

We now consider the problem of selecting the tuning parameter λ, assuming that K is known. We would like to be conservative in calling observations outliers – we do not want to call an observation an outlier unless we are quite confident that it does not belong to any cluster. We choose the largest value of λ (corresponding to the smallest number of observations called outliers) such that no observation that is not called an outlier appears to be one. That is, we choose λ at the largest value such that no observation assigned zero error has $\|\mathbf{X}_i - \boldsymbol{\mu}_{C(i)}\|_2$ larger than $m(\lambda) + 3s(\lambda)$, where for a given value of λ, $m(\lambda)$ is the mean of $\|\mathbf{X}_i - \boldsymbol{\mu}_{C(i)}\|_2$ over all observations with zero error, and $s(\lambda)$ is the standard deviation of $\|\mathbf{X}_i - \boldsymbol{\mu}_{C(i)}\|_2$ over all observations with zero error. To implement this approach, we perform outlier K-means clustering over a grid of λ values.

Throughout this paper, we assume that the number of clusters K is known, but in real applications this is often not the case. If K is unknown, then λ and K must be jointly selected in some way. Consider Figure 1. Given that $K = 2$, it is clear that a number of observations are outliers; however, if K were much larger, then each of the outliers shown in panel (a) could be assigned to its own cluster, and no observations would be assigned nonzero errors. More simply put, it is impossible to distinguish a data set comprised of K clusters and a single outlier from a data set comprised of $K + 1$ clusters and no outliers, unless one is willing to make an assumption about the number of clusters or the properties of those clusters. This is a complex issue, and in what follows we assume that K is known.

3.3 Connection with M-estimation

Our proposal for unsupervised learning in the presence of outliers is motivated by a recent proposal in the regression setting (She and Owen, 2011). In that paper, it was shown that there is a deep connection between performing regression in the presence of outliers, using the model (3), and M-estimation. We now show that there is a very close connection between our proposal for outlier K-means, and a generalized version of K-means given by

$$\underset{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \sum_{i \in C_k} \rho(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda) \right\} \quad (13)$$

where $\rho(t; \lambda)$ is some loss function (Garcia-Escudero and Gordaliza, 1999). Suppose that for a given penalty function $P(\cdot; \lambda)$, the problem

$$\underset{\mathbf{b}}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{b}\|^2 + P(\|\mathbf{b}\|_2; \lambda) \right\} \quad (14)$$

has the solution

$$\mathbf{b} = \Theta(\mathbf{y}; \lambda) \tag{15}$$

where $\Theta(\cdot; \lambda)$ is a thresholding function (discussed extensively in She, 2009; She, 2011). Consider the optimization problem (12) with $C_1, \ldots, C_K$ fixed. Then it is not hard to see that an iterative algorithm that successively holds $\mathbf{E}_1, \ldots, \mathbf{E}_n$ fixed and solves for $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ and then holds $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ fixed and solves for $\mathbf{E}_1, \ldots, \mathbf{E}_n$ will decrease the objective at each step. If this algorithm is iterated until the objective converges, then by inspection, the solution will satisfy, for $k = 1, \ldots, K$,

$$\sum_{i \in C_k} (\mathbf{X}_i - \boldsymbol{\mu}_k - \Theta(\mathbf{X}_i - \boldsymbol{\mu}_k; \lambda)) = 0. \tag{16}$$

Proposition 1. *Suppose that* $\Theta(\mathbf{X}_i - \boldsymbol{\mu}_k; \lambda) + \frac{\partial}{\partial \boldsymbol{\mu}_k} \rho(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda) = \mathbf{X}_i - \boldsymbol{\mu}_k$. *Then, (16) implies that*

$$\sum_{i \in C_k} \frac{\partial}{\partial \boldsymbol{\mu}_k} \rho(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda) = 0. \tag{17}$$

In other words, the solution to (12) with $C_1, \ldots, C_K$ held fixed satisfies the score equation associated with (13) with $C_1, \ldots, C_K$ held fixed.

Proposition 1 indicates that there is a connection between the outlier K-means clustering problem (12) and the generalized K-means problem (13) when $C_1, \ldots, C_K$ are held fixed. For example, consider the use of a group lasso penalty $P(\|\mathbf{E}_i\|_2; \lambda) = \lambda \|\mathbf{E}_i\|_2$, and let $\rho(t; \lambda)$ be Huber's loss function, given by (Huber, 1964)

$$\rho(t; \lambda) = \begin{cases} \lambda |t| - \lambda^2/2 & \text{if } |t| > \lambda \\ t^2/2 & \text{if } |t| \leq \lambda \end{cases}. \tag{18}$$

Then, it is easily shown that the condition of Proposition 1 is satisfied, since $\Theta(\mathbf{X}_i - \boldsymbol{\mu}_k; \lambda) = (\mathbf{X}_i - \boldsymbol{\mu}_k) \max(0, 1 - \frac{\lambda}{\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2})$, and $\frac{\partial}{\partial \boldsymbol{\mu}_k} \rho(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda) = \frac{\mathbf{X}_i - \boldsymbol{\mu}_k}{\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2} \rho'(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda)$ where

$$\rho'(t; \lambda) = \begin{cases} \lambda \, \text{sign}(t) & \text{if } |t| > \lambda \\ t & \text{if } |t| \leq \lambda \end{cases}. \tag{19}$$

In other words, with $C_1, \ldots, C_K$ held fixed, generalized K-means with Huber's loss function and outlier K-means with a group lasso penalty yield the same estimates $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$.

Now, holding $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ fixed, suppose we wish to solve (12) for $C_1, \ldots, C_K$ and $\mathbf{E}_1, \ldots, \mathbf{E}_n$.

Proposition 2. *Holding $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ fixed and minimizing (12), with a group lasso penalty, with respect to $C_1, \ldots, C_K$ and $\mathbf{E}_1, \ldots, \mathbf{E}_n$ amounts to assigning the ith observation to the class for which $\rho(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda)$ is smallest, where ρ is Huber's loss function.*

Proof. Minimizing (12) with respect to $\mathbf{E}_1, \ldots, \mathbf{E}_n$ and $C_1, \ldots, C_K$ amounts to assigning the ith observation to the class for which the quantity

$$\frac{1}{2} \|\mathbf{X}_i - \boldsymbol{\mu}_k - \mathbf{E}_i\|^2 + \lambda \|\mathbf{E}_i\|_2 \tag{20}$$

is minimized, where $\mathbf{E}_i = (\mathbf{X}_i - \boldsymbol{\mu}_k) \max(0, 1 - \frac{\lambda}{\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2})$. Note that (20) can be rewritten as

$$\begin{aligned} &\frac{1}{2} \min(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2, \lambda)^2 + \lambda \max(0, \|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2 - \lambda) \\ &\quad = \rho(\|\mathbf{X}_i - \boldsymbol{\mu}_k\|_2; \lambda). \end{aligned} \tag{21}$$

□

Together, Propositions 1 and 2 indicate that a generalized version of K-means with Huber's loss function (13) is essentially equivalent to our outlier K-means clustering proposal with a group lasso penalty.

3.4 A related proposal

Tseng (2007) proposed performing K-means clustering in the presence of outliers by solving

$$\underset{C_1, \ldots, C_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, S}{\text{minimize}} \left\{ \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{X}_i - \boldsymbol{\mu}_k\|^2 + \lambda |S| \right\}, \tag{22}$$

where λ is a nonnegative tuning parameter and S is a set of observations thought to be outliers. That is, if an observation is in S, then it does not belong to any of the clusters $C_1, \ldots, C_K$. In (22), $|S|$ indicates the number of observations in the set S. Solving (22) is not directly addressed, but a descent algorithm could be obtained by clustering all of the observations, assigning any observations that are more than a distance $\sqrt{\lambda}$ from the corresponding cluster mean to the set S, reclustering all of the observations not in S, and so on. It turns out that this algorithm is very closely related to our outlier K-means proposal (12) using a hard-thresholding penalty, given by $P(\|\mathbf{E}_i\|_2; \lambda) = 1_{\|\mathbf{E}_i\|_2 > 0} \lambda/2$ (see e.g. She, 2009). Using this penalty, Step 2(b) in Algorithm 2 yields $\mathbf{E}_i = 0$ if $\|\mathbf{X}_i - \boldsymbol{\mu}_{C(i)}\|_2 < \sqrt{\lambda}$ and $\mathbf{E}_i = \mathbf{X}_i - \boldsymbol{\mu}_{C(i)}$ otherwise. This is identical to the first iteration of Tseng (2007)'s procedure. However, further iterations of our procedure will differ, because in Step 2(a) clustering is performed on $\mathbf{X}_1 - \mathbf{E}_1, \ldots, \mathbf{X}_n - \mathbf{E}_n$ rather than simply on the observations currently assigned nonzero errors.

3.5 Simulation study

We generated data with 25 p-dimensional observations in each of K classes, along with q outliers. We used several values of K, p, and q: $K = 2$ with $p = 10$, $K = 5$ with $p = 50$, and $q = 0, 5, 10$. Each observation was generated independently; for an observation in class k,

$$\mathbf{X}_i \sim N(\boldsymbol{\mu}_k, \mathbf{I}), \qquad \boldsymbol{\mu}_k \sim N(0, \sigma^2 \mathbf{I}), \tag{23}$$

Table 1. In the simulation study described in the text, for various numbers of clusters (K) and outliers (q), K-means clustering (KM), outlier K-means clustering (OKM), and model-based clustering (MCLUST) were performed. Means and standard errors (over 50 repetitions) of the following quantities are reported: number of estimated outliers ($\hat{q}$), CER, and OER

K	q	Approach	$\hat{q}$	CER	OER
2	0	KM	–	0.043(0.008)	–
		OKM	0.52(0.077)	0.051(0.009)	0.01(0.002)
		MCLUST	9.78(1.508)	0.183(0.022)	0.196(0.03)
	5	KM	–	0.316(0.027)	–
		OKM	4.82(0.089)	0.103(0.022)	0.005(0.001)
		MCLUST	5.1(0.091)	0.104(0.021)	0.006(0.001)
	10	KM	–	0.372(0.024)	–
		OKM	3.84(0.573)	0.261(0.025)	0.103(0.01)
		MCLUST	8.48(0.493)	0.146(0.028)	0.026(0.008)
5	0	KM	–	0.036(0.003)	–
		OKM	2.28(0.128)	0.044(0.003)	0.018(0.001)
		MCLUST	30.38(1.58)	0.532(0.01)	0.243(0.013)
	5	KM	–	0.053(0.003)	–
		OKM	5.2(0.064)	0.033(0.003)	0.002(0)
		MCLUST	8.84(0.507)	0.663(0.009)	0.03(0.004)
	10	KM	–	0.072(0.004)	–
		OKM	10.22(0.066)	0.032(0.002)	0.002(0)
		MCLUST	10.44(0.115)	0.643(0.005)	0.004(0.001)

where $\sigma = 1$ if $K = 2$ and $\sigma = 0.5$ if $K = 5$. The outlying observations were also drawn according to (23), after random assignment to one of the K clusters, with an additional independent noise term for each feature. These noise terms were drawn according to a Unif$[(-6,-3) \cup (3,6)]$ distribution if $K = 2$, and according to a Unif$[(-2,-1) \cup (1,2)]$ distribution if $K = 5$. Three clustering approaches were compared:

1. *KM*: K-means clustering.
2. *OKM*: Outlier K-means clustering, with the tuning parameter selection approach described previously and using a group lasso penalty.
3. *MCLUST*: Model-based clustering, allowing for outliers, as described in Fraley and Raftery (2002). The `R` package `mclust` was used, under the assumption of spherical covariance matrices with common variance (the same assumption that is made by K-means clustering). Note that the software automatically determines the number of outliers.

To assess the accuracy of the clusters obtained by each method, the *clustering error rate* (CER) is used. CER measures the extent to which two partitions R and Q of a set of n observations are in agreement with each other. For instance, R might be a partition obtained by clustering, whereas Q could be the true class labels. Let $1_{R(i,i')}$ be an indicator for whether partition R places the ith and i'th observations in the same group, and define $1_{Q(i,i')}$ analogously. Then the CER is defined as $\sum_{i>i'} |1_{R(i,i')} - 1_{Q(i,i')}| / \binom{n}{2}$. It equals zero if the two partitions are in perfect agreement, and will take on a positive value if they disagree. Note that the CER is one minus the Rand Index (Rand, 1971). Table 1 reports the CER in each simulation setting, where the outliers are coded as a separate class. For purposes of computing the CER, the partition estimated by K-means uses only K classes since standard K-means clustering does not identify outliers; the other clustering methods use $K + 1$ classes. Table 1 also reports the outlier error rate (OER) – that is, the number of outliers erroneously thought to be non-outliers, plus the number of non-outliers erroneously thought to be outliers, divided by the total number of observations. Outlier K-means results in far lower CER than ordinary K-means, and tends to yield lower OER and CER than does model-based clustering (Table 1). However, a direct comparison between the performances of outlier K-means and MCLUST is challenging, since each approach identifies a different number of observations as outliers.

3.6 Application to gene expression data sets

We now study the performance of outlier K-means on two gene expression data sets.

The first data set consists of colon tissue samples for which 2,000 gene expression measurements are available, and can be obtained from http://genomics-pubs.princeton.edu/oncology/ (Alon et al., 1999). There are 40 tumor samples and 22 normal samples. The data were log-transformed, and observations were centered to have mean zero and standard deviation one. Applying our outlier K-means proposal with a group lasso penalty to this data set (with the automated tuning parameter selection procedure described earlier) identifies two outliers, namely observations 3 and 57; both are tumor samples. K-means clustering results in a CER of 0.508, whereas outlier K-means results in a CER of 0.183. Results are displayed in Figure 2.

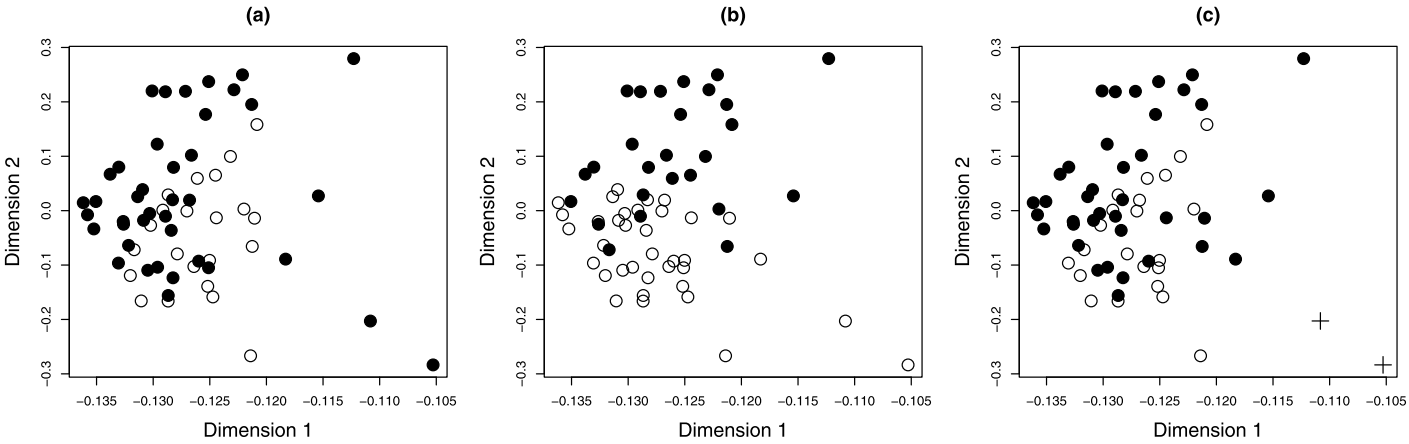

Figure 2. The colon gene expression data set of Alon et al. (1999). The 62 observations are projected onto the first two principal components. (a): Normal samples are shown as unfilled circles and tumor samples are shown as filled circles. (b): K-means clustering was performed, and the two clusters are indicated using filled and unfilled circles. (c): Outlier K-means was performed. The observations identified as outliers are shown as crosses, and the other observations are displayed as filled or unfilled circles according to the cluster labels.

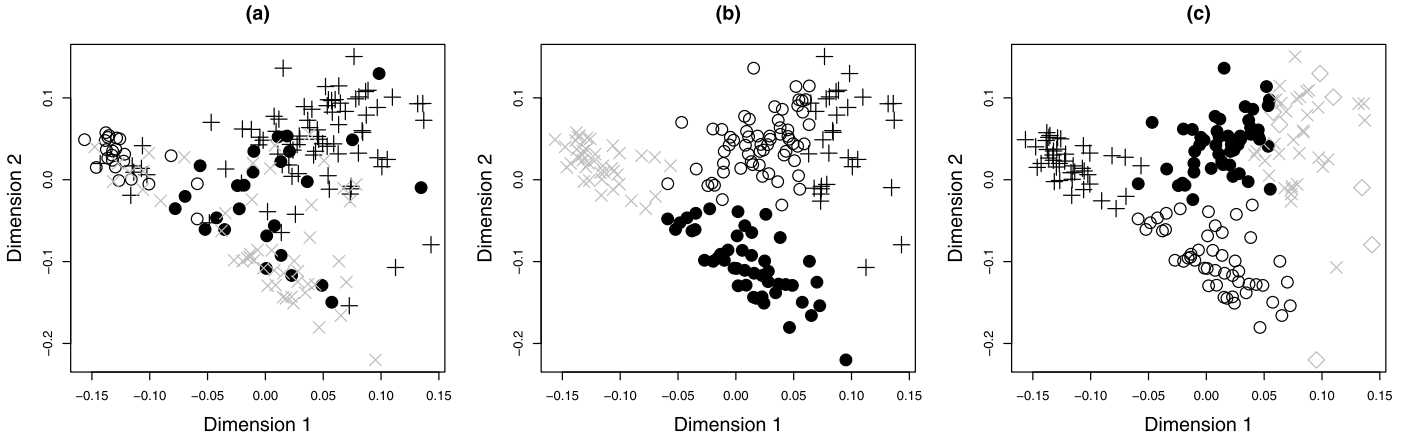

Figure 3. The glioma gene expression data set of Sun et al. (2006). The 180 observations are projected onto the first two principal components, after taking only the 1,093 genes with highest variance. (a): The four classes are indicated using distinct symbols. (b): K-means clustering was performed, and the resulting clusters are indicated using distinct symbols. (c): Outlier K-means was performed. Observations identified as outliers are indicated using grey diamonds, and the cluster labels of the other observations are indicated using distinct symbols.

We also performed K-means clustering on the glioma gene expression data set of Sun et al. (2006), which consists of 180 samples and 52,613 gene expression measurements. The samples fall into four classes, one non-tumor class and three types of glioma. The data are available from Gene Expression Omnibus (Barrett et al., 2005) with accession number GDS1962. Genes were standardized to have mean zero and standard deviation one before K-means clustering and outlier K-means clustering were performed on only the 2% of genes with highest variance before standardization. Outlier K-means identified seven outlying observations; these "outliers" were drawn from three of the four classes. The results are displayed in Figure 3.

4. PCA WITH OUTLIERS

4.1 The proposal

We now consider the problem of performing PCA when some of the observations are outliers. Rather than solving the problem (6), we instead solve the problem

$$\underset{\mathbf{D},\mathbf{U},\mathbf{V},\mathbf{E}}{\text{minimize}}\left\{\frac{1}{2}\|\mathbf{X}-\mathbf{E}-\mathbf{U}\mathbf{D}\mathbf{V}^T\|_F^2 + \sum_{i=1}^{n} P(\|\mathbf{E}_i\|_2;\lambda)\right\}, \tag{24}$$

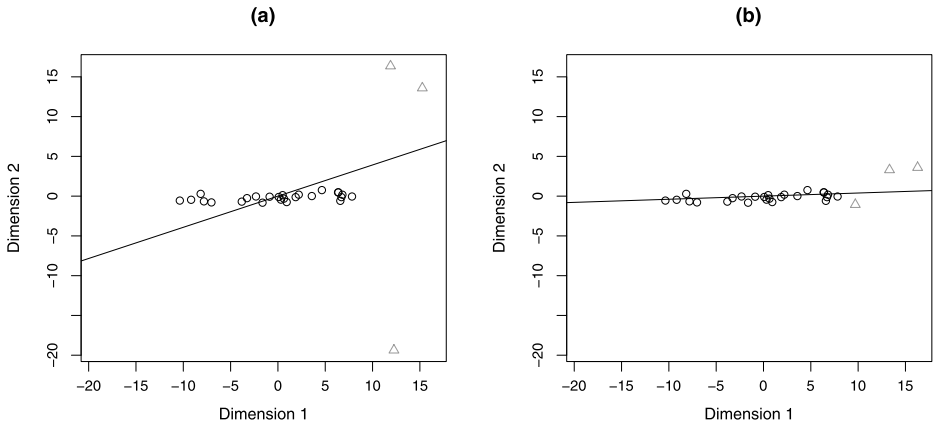

Figure 4. A two-dimensional example. In each figure, non-outliers are shown as black circles and outliers are shown as grey triangles. (a): The observations are plotted and the first estimated principal component is shown. (b): Outlier PCA was performed with a group lasso penalty, and $\mathbf{X}_i - \mathbf{E}_i$ is plotted for $i = 1, \ldots, n$. Only the three true outliers were assigned nonzero errors. The resulting estimated principal component is shown.

Algorithm 3 A descent algorithm for outlier PCA

1. Initialize the errors – e.g. take $\mathbf{E}_i = \mathbf{0}$ for the 90% of observations that are closest to the overall mean of the observations in terms of Euclidean distance, and $\mathbf{E}_i = \mathbf{X}_i$ for the others.
2. Iterate until the objective (24) converges:
   (a) Compute $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$, the components of the rank-K singular value decomposition of the matrix $\mathbf{X} - \mathbf{E}$.
   (b) For $i = 1, \ldots, n$, solve the problem
   (25)
   $$\underset{\mathbf{E}_i}{\text{minimize}} \left\{ \frac{1}{2} \|\mathbf{X}_i - \mathbf{E}_i - \mathbf{U}_i \mathbf{D} \mathbf{V}^T\|^2 + P(\|\mathbf{E}_i\|_2; \lambda) \right\},$$
   where $\mathbf{U}_i$ denotes the ith row of $\mathbf{U}$. If $P(\|\mathbf{E}_i\|_2; \lambda) = \lambda \|\mathbf{E}_i\|_2$, then the solution takes the form (Yuan and Lin, 2007)
   (26)
   $$\mathbf{E}_i = (\mathbf{X}_i - \mathbf{U}_i \mathbf{D} \mathbf{V}^T) \max\left(0, 1 - \frac{\lambda}{\|\mathbf{X}_i - \mathbf{U}_i \mathbf{D} \mathbf{V}^T\|_2}\right).$$
3. Compute the principal components of the observations that were assigned zero errors, i.e. perform PCA on the observations in the set $\{i : \|\mathbf{E}_i\|_2 = 0\}$.

where as in (6), $\mathbf{D}$ is a $K \times K$ diagonal matrix, and $\mathbf{U}$ and $\mathbf{V}$ are $n \times K$ and $p \times K$ orthogonal matrices, respectively. We call this the *outlier PCA optimization problem*, and columns of the matrix $\mathbf{V}$ obtained by solving this problem the *outlier principal components*. Algorithm 3 provides an iterative approach that will decrease the objective of (24) at each step, but in general will not attain the global optimum since (24) is non-convex (indeed, PCA itself as given in (6) is a non-convex problem). We illustrate outlier PCA on a simple toy example in Figure 4.

In the examples that follow, we assume that K is known, we take $P(\cdot; \lambda)$ to be a group lasso penalty, and we select λ to be the largest value (corresponding to the smallest number of observations declared outliers) such that no observation assigned zero error has $\|\mathbf{X}_i - \mathbf{U}_i \mathbf{D} \mathbf{V}^T\|_2$ greater than $m(\lambda) + 3s(\lambda)$. For a given value of λ, $m(\lambda)$ is defined to be the mean of $\|\mathbf{X}_i - \mathbf{U}_i \mathbf{D} \mathbf{V}^T\|_2$ over all observations assigned zero errors, and $s(\lambda)$ is defined to be the standard deviation of $\|\mathbf{X}_i - \mathbf{U}_i \mathbf{D} \mathbf{V}^T\|_2$ over all observations assigned zero errors. In other words, we choose the smallest possible number of outliers so that the low-rank model fits the observations assigned zero errors well.

4.2 Related work

In a series of recent papers, a number of authors have considered the problem of performing PCA in the case that an $n \times p$ data matrix $\mathbf{W}$ that is *exactly* low rank is observed with noise. That is, rather than observing $\mathbf{W}$, we instead observe $\mathbf{X} = \mathbf{W} + \mathbf{E}$, where $\mathbf{E}$ is a $n \times p$ sparse noise matrix that results from the corruption of certain elements scattered at random throughout the data matrix. For this problem, one can solve (Lin et al., 2009; Wright et al., 2009; Candes et al., 2011)

$$\underset{\mathbf{E}}{\text{minimize}} \left\{ \|\mathbf{X} - \mathbf{E}\|_* + \lambda \|\mathbf{E}\|_1 \right\}, \tag{27}$$

where λ is a nonnegative tuning parameter, and where $\|\cdot\|_*$ indicates the nuclear norm of a matrix, i.e. the sum of its singular values. It has been shown that under certain conditions one can *exactly* recover the low-rank matrix $\mathbf{W}$. We will refer to the solution to the problem (27) as *exact robust PCA*, to emphasize its assumption that the underlying matrix $\mathbf{W}$ is exactly low rank. If we modify (27) in order to allow the matrix $\mathbf{W}$ to be *approximately* low-rank rather than exactly low-rank, and to encourage sparsity in the rows of $\mathbf{E}$ rather than in the individual elements, then we obtain our outlier

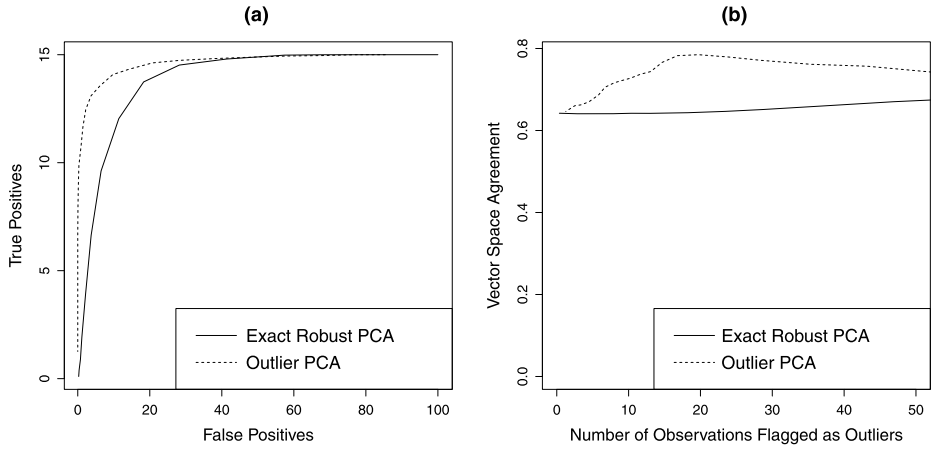

Figure 5. A comparison of outlier PCA with a group lasso penalty and exact robust PCA. Results are averaged over 50 simulated data sets. (a): Number of true positives versus number of false positives for outlier PCA and exact robust PCA. "True positive" refers to an outlying observation that was correctly assigned a nonzero error, whereas "false positive" refers to a non-outlying observation that was incorrectly assigned a nonzero error. (b): The vector space agreement, which in this case is simply the square of the inner product between the true and estimated principal components, is shown.

PCA proposal (24). We note that our proposal is closely related to a formulation in Xu, Caramanis and Sanghavi (2010).

To compare the performances of our outlier PCA proposal (24) to that of the exact robust PCA proposal (27), we implemented the algorithm described in Candes et al. (2011) to solve (27). Figure 5 shows a comparison of the results obtained for the two approaches on a simulated example where 100 observations are generated according to a rank-one model with Gaussian noise in 10 dimensions, and there are 15 outliers. Outlier PCA performs better because it can accommodate the fact that the data, even with the outliers removed, is only approximately rank-one, and because the group lasso penalty exploits the fact that we expect entire observations to be outliers.

4.3 Simulation study

To evaluate the performance of outlier PCA, we generated a $n \times p$ data matrix $\mathbf{X}$, with $n = 50$ or $n = 100$ observations in $p = 5$ dimensions, and $q = 0$, $q = 5$, or $q = 10$ outliers. The $n + q$ observations were generated according to $X_{ij} = 50u_{i1}v_{j1} + 10u_{i2}v_{j2} + \epsilon_{ij} + \delta_{ij}1_{i>n}$, where $\mathbf{u}_1, \mathbf{u}_2$ and $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal unit vectors of lengths $n+q$ and p, respectively, and $\epsilon_{ij}, \delta_{ij}$ are independent error terms: $\epsilon_{ij} \sim N(0,1)$ and $\delta_{ij} \sim \text{Unif}[(-5,-3) \cup (3,5)]$. The last q observations are outliers due to the additional noise terms δ_{ij}. Four approaches for estimating the first two principal components were compared:

1. *PCA*: PCA with $K = 2$.
2. *OPCA*: Outlier PCA with $K = 2$, using a group lasso penalty and the tuning parameter selection approach described previously. Let $\hat{q}$ denote the number of outliers identified.
3. *OSPCA*: "Outlier screening" followed by PCA. The $\hat{q}$ observations with the largest mean distance to all other observations were called "outliers". Then PCA with $K = 2$ was performed on the observations that were not identified as outliers.
4. *PCAOS*: PCA followed by "outlier screening". PCA with $K = 2$ was performed on the full set of $n+q$ observations. Then the $\hat{q}$ observations for which the rank-two approximation fit the worst (in terms of Euclidean distance) were called "outliers".

These latter two approaches were included in the comparisons since they constitute simple alternatives to outlier PCA.

To evaluate the accuracy of the principal component estimates, we computed the *vector space agreement* (VSA): namely, $\text{trace}(\mathbf{P}_{true}\mathbf{P}_{est})/2$, where $\mathbf{P}_{true}$ is the orthogonal projection matrix onto the space spanned by the two true principal components used to generate the data, and $\mathbf{P}_{est}$ is the orthogonal projection matrix onto the space spanned by the two estimated principal components. This quantity lies between 0 and 1. It will be 0 if the spaces spanned by the true and estimated principal components are orthogonal to each other, and will be 1 if the two spaces are identical. We also computed the OER, defined in the previous section. Results are shown in Table 2. In general, outlier PCA and OSPCA yield the highest VSA. Outlier PCA and PCAOS have comparable OERs; that of OSPCA is substantially worse.

5. DISCUSSION

In recent years, much effort has focused upon using shrinkage penalties to develop statistical methods that are sparse in the features. Such approaches were initially developed for the supervised setting (see e.g. Tibshirani, 1996;

Table 2. In the simulation study described in the text, for various numbers of observations (n) and outliers (q), we performed PCA, outlier PCA (OPCA), outlier screening followed by PCA (OSPCA), and PCA followed by outlier screening (PCAOS). The means and standard errors, over 50 simulated data sets, of three quantites are reported: the number of outliers selected by OPCA using the automated tuning parameter selection approach ($\hat{q}$), the vector space agreement (VSA), and OER. OER is not reported for PCA since PCA does not identify outliers

n	q	$\hat{q}$	Quantity	PCA	OPCA	OSPCA	PCAOS
50	0	0.24(0.084)	VSA	0.975(0.003)	0.974(0.003)	0.975(0.003)	0.975(0.003)
			OER	–	0.005(0.002)	0.005(0.002)	0.005(0.002)
	5	3.34(0.142)	VSA	0.662(0.019)	0.695(0.021)	0.714(0.019)	0.662(0.019)
			OER	–	0.038(0.002)	0.059(0.004)	0.033(0.002)
	10	6.44(0.368)	VSA	0.617(0.017)	0.646(0.02)	0.629(0.018)	0.617(0.017)
			OER	–	0.066(0.006)	0.083(0.005)	0.062(0.006)
100	0	0.48(0.104)	VSA	0.969(0.003)	0.969(0.003)	0.969(0.003)	0.969(0.003)
			OER	–	0.005(0.001)	0.005(0.001)	0.005(0.001)
	5	3.94(0.197)	VSA	0.683(0.021)	0.745(0.024)	0.73(0.024)	0.683(0.021)
			OER	–	0.019(0.001)	0.019(0.001)	0.017(0.002)
	10	8.86(0.631)	VSA	0.671(0.021)	0.728(0.023)	0.762(0.022)	0.671(0.021)
			OER	–	0.027(0.005)	0.028(0.005)	0.024(0.005)

Fan and Li, 2001; Zou and Hastie, 2005) but more recent work has focused upon feature selection in the unsupervised setting (see e.g. Pan and Shen, 2007; Xie et al., 2008; Witten et al., 2009; Witten and Tibshirani, 2010). Here instead we develop an approach for unsupervised learning that uses shrinkage penalties in order to obtain a result that is *sparse in the observations*, and is intended for use if outliers may be present among the observations. This general formulation can be used in order to combine outlier detection with any unsupervised learning method that can be written as the solution to an optimization problem.

Our proposal for unsupervised learning with outliers allows for any penalty to be applied to the ℓ_2 norm of the error associated with an individual observation. For simplicity, in the examples throughout the paper, we used a group lasso penalty in order to identify outliers. However, it is argued in She and Owen (2011) that convex penalties are inherently non-robust in the context of outlier detection in regression, and the use of non-convex penalties is espoused. Therefore, it may be preferable to use a non-convex penalty, such as hard-thresholding or SCAD, on the ℓ_2 norm of the error associated with each observation in (8). In the context of K-means clustering or PCA, this would require simply a different update in Step 2(b) of Algorithm 2 or 3 (She, 2009; She, 2011). Alternatively, if one has prior knowledge about the probability that a given observation is an outlier, then one might wish to modify the optimization problem (8) to allow a different penalty to be applied to each observation. Or if we believe that individual elements of the data matrix $\mathbf{X}$, rather than entire observations, contain errors, then we could apply a penalty to individual elements of $\mathbf{E}$, e.g. $\lambda \sum_{i=1}^{n} \sum_{j=1}^{p} |E_{ij}|$. Indeed, this approach was taken in Candes et al. (2011).

We demonstrated that there is a close connection between our proposal for K-means clustering with outliers and a generalized version of K-means discussed by Garcia-Escudero and Gordaliza (1999). Those authors showed that regardless of the loss function used, "generalized K-means do[es] not inherit the robustness properties of the M-estimator from which they came". This calls into question the extent to which our outlier K-means clustering proposal inherits those robustness properties; more investigation into this issue is needed.

Our proposal for unsupervised learning in the presence of outliers is simple and elegant, and can be applied to a range of problems. However, it also has some drawbacks. Though the problem (8) naturally lends itself to an iterative algorithm that decreases the objective at each step, this iterative approach will not, in general, yield the global optimum. Furthermore, the problem of selecting the tuning parameter that controls the number of outliers identified is quite challenging. Finally, in our simulation studies we have found that in many cases, outliers in the unsupervised context can be prohibitively difficult to identify, or else so easy to identify that even very simple methods (such as PCA followed by outlier screening) can do quite well.

ACKNOWLEDGMENTS

This work was supported by NIH grant DP5OD009145. The author thanks an anonymous reviewer for helpful comments that improved the quality of this manuscript.

Received 7 November 2011

REFERENCES

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sciences* **96** 6745–6750.

Barrett, T., Suzek, T., Troup, D., Wilhite, S., Ngau, W., Ledoux, P., Rudnev, D., Lash, A., Fujibuchi, W. and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Research* **33** D562–D566.

Candes, E., Li, X., Ma, Y. and Wright, J. (2011). Robust principal components analysis? *Journal of the ACM* **58**(3) 1–37. MR2811000

Dave, R. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters* **12** 657–664.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. MR1951635

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

Garcia-Escudero, L. and Gordaliza, A. (1999). Robustness properties of k means and trimmed k means. *Journal of the American Statistical Association* **94** 956–969. MR1723291

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**(1) 1–15. MR2804206

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Math. Stat.* **53** 73–101. MR0161415

Jiang, M., Tseng, S. and Su, C. (2001). Two-phase clustering process for outliers detection. *Pattern Recognition Leters* **22** 691–700.

Jolion, J. and Rosenfeld, A. (1989). Cluster detection in background noise. *Pattern Recognition* **22**(5) 603–607.

Lin, Z., Chen, M., Wu, L. and Ma, Y. (2009). The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices.

Mazumder, R., Friedman, J. and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106** 1125–1138.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8** 1145–1164.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66** 846–850.

Rousseeuw, P. and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1**(1) 73–79.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection*. Wiley, New York. MR0914792

She, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics* **3** 384–415. MR2501318

She, Y. (2011). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*.

She, Y. and Owen, A. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* **106**(494) 626–639. MR2847975

Sun, L., Hui, A., Su, Q., Vortmeyer, A., Kotliarov, Y., Pastorino, S., Passaniti, A., Menon, J., Walling, J., Bailey, R., Rosenblum, M., Mikkelsen, T. and Fine, H. (2006). Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell* **9** 287–300.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.* **58** 267–288. MR1379242

Tseng, G. (2007). Penalized and weighted k-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics* **23** 2247–2255.

Tseng, G. and Wong, W. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61** 10–16. MR2129196

Witten, D. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105**(490) 713–726. MR2724855

Witten, D. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, Series B* **73**(5) 753–772. MR2867457

Witten, D., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**(3) 515–534.

Wright, J., Ganesh, A., Rao, S., Peng, Y. and Ma, Y. (2009). Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. *Proc. of Neural Information Processing Systems*.

Xie, B., Pan, W. and Shen, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics* **2** 168–212. MR2386092

Xu, H., Caramanis, C. and Sanghavi, S. (2010). Robust PCA via outlier pursuit. *Advances in Neural Information Processing Systems* **23** 2496–2504.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68** 49–67. MR2212574

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Royal. Stat. Soc. B.* **67** 301–320. MR2137327

Daniela M. Witten
Department of Biostatistics
University of Washington
Seattle, WA 98195-7232
USA
E-mail address: dwitten@u.washington.edu