Statistics in the Twenty-First Century: Special Volume
In Honour of Distinguished Professor Dr. Mir Masoom Ali
On the Occasion of his 75th Birthday Anniversary
PJSOR, Vol. 8, No. 3, pages 359-379, July 2012

The BACON Approach for Rank-Deficient Data

Athanassios Kondylis[1]
Philip Morris International
ACR, Neuchatel, Switzerland
athanassios.kondylis@pmi.com

Ali S. Hadi
Department of Mathematics and Actuarial Science
The American University in Cairo, Egypt, and
Department of Statistical Sciences, Cornell University, USA.
ahadi@aucegypt.edu

Mark Werner
Department of Statistics
University of Georgia, USA.
mwerner@uga.edu

Abstract

Rank-deficient data are not uncommon in practice. They result from highly collinear variables and/or high-dimensional data. A special case of the latter occurs when the number of recorded variables exceeds the number of observations. The use of the BACON algorithm for outlier detection in multivariate data is extended here to include rank-deficient data. We present two approaches to identifying outliers in rank-deficient data based on the original BACON algorithm. The first algorithm projects the data onto a robust subspace of reduced dimension, while the second employs a ridge type regularization on the covariance matrix. Both algorithms are tested on real as well as simulated data sets with good results in terms of their effectiveness in outlier detection. They are also examined in terms of computational efficiency and found to be very fast, with particularly good scaling properties for increasing dimension.

Keywords: High-dimensional data, Mahalanobis distance, Outlier detection, Spatial median.

1. Introduction

Most statistical data analysis methods customarily assume that the available $n \times p$ data matrix $\mathbf{X}$ is of full-column rank, where n is the number of observations and p is the number of variables. For example, most multivariate analysis techniques assume a non-singular covariance matrix while in the regression setting, least squares regression assumes that $\mathbf{X}'\mathbf{X}$ is non-singular. This assumption automatically does not hold if the number of observations n is less than the number of variables p. This situation is often

[1] The opinions and conclusions of the authors are their own and do not necessarily reflect Philip Morris International's position.

encountered in high-dimensional data. There are practical situations when the available data sets have $p > n$ and sometimes even $p >> n$ (p is much larger than n). For example, recent development in communication and information technology, mostly in data instrumentation, have given rise to the availability of hyperdimensional data where there are a large number of often interrelated variables but a fewer number of observations. Examples of real-life data where $p > n$ include chemometrics and Near Infra-Red experiments in spectroscopy, image analysis and computer vision, gene expression experiments, and others. These types of data sets invalidate most multivariate and regression techniques. Note that the matrix $\mathbf{X}'\mathbf{X}$ may be singular even when $n \geq p$, as in the case where the columns of $\mathbf{X}$ are linearly dependent. Such data sets very often arise in environmental studies, econometrics and finance, so that the case $n \geq p$ with less than full rank is also of interest.

Much can be found in the statistical literature for the case where $n > p$. By comparison, fewer statistical techniques deal with the case of $p > n$, despite the constantly growing interest during the last years. Most of these techniques attempt to regularize the statistical problem and usually approximate the solution on lower dimensional subspaces. In regression, for instance, principal components (PC) and partial least squares (PLS) (see Tenenhaus, 1998) truncate the least squares solution to a few suitably extracted orthogonal directions. Another commonly used regularization method in Statistics is ridge regression (see Hoerl and Kennard, 1970), which regularizes the regression problem by adding a suitable constant in $\mathbf{X}'\mathbf{X}$. Similar to ridge regression, seen as an L_2 penalization method, other penalized estimation techniques have been used in order to solve statistical problems in high dimensions; for an excellent overview see Izenman (2008) and Hastie et al. (2009).

Another problem that arises in data analysis is the presence of outliers. Real-life data sets often contain outliers, whether or not $p > n$. The presence of outliers in the data can have dramatic effects on the results of the analysis. The aforementioned PC and PLS algorithms, for example, are not at all robust against outliers (see Kondylis and Hadi, 2006). It is therefore important to identify the outliers when they exist in the data. Outliers identification has a venerable history; see, for example, the books by Belsley et al. (1980), Cook and Weisberg (1982), Atkinson (1985), Rousseeuw and Leroy (1987), Chatterjee and Hadi (1988), and the articles by Gray (1986), Kianifard and Swallow (1989), Rousseeuw and van Zomeren (1990), Paul and Fung (1991), Hadi (1992a,b), Hadi and Simonoff (1993), Atkinson (1994), Hadi (1994), and Billor et al. (2000).

Yet, rank-deficient and high-dimensional data sets invalidate most outlier detection and robust estimation techniques (see, for example, Li and Chen (1985), Maronna and Zamar (2002) and the references therein). For example, despite the well-known robust properties (high breakdown point, affine equivariance, efficiency, etc.) of a wide range of robust variance-covariance estimates, most of them break down in high dimensions. Robust variance-covariance estimation in such data sets has been developed relatively recently, see for example Locantore et al. (1999) , Maronna and Zamar (2002), Croux and Ruiz-Gazen (2005), as well as Filzmoser et al. (2008) and the references therein.

A well established and effective method that deals with the identification of outliers and robust estimation in full-rank data is BACON (Blocked Adaptive Computationally-Efficient Outlier Nominators), proposed by Billor et al. (2000). This algorithm can be used in a variety of situations, such as the adaption for incomplete survey data by Beguin and Hulliger (2008). The BACON approach, however, assumes not only that the available data set is of full-column rank, but also that n/p is sufficiently large (e.g. $n/p > 4$ or 5). In this paper we adapt the BACON approach to obtain two procedures for outlier identification that can be used in cases where the data matrix is rank-deficient, which includes both $n >> p$ and $n \geq p$ cases. This will allow the users to identify outliers in these types of difficult but increasingly common situations, while still maintaining the computational efficiency of the BACON algorithm so that the proposed procedures are suitable for large, high-dimensional data sets. This latter situation is especially challenging for many classical algorithms since computational effort grows rapidly with increasing p.

The rest of the paper is organized as follows. Section 2 gives a brief description of the BACON approach for full-column rank data. Section 3 extends the BACON approach to the case where the data matrix is rank-deficient. We refer to it as the rank-deficient or RD-BACON approach. Two versions of the RD-BACON approach, RD1-BACON and RD2-BACON, are presented in this section. Section 4 gives illustrative examples of applications using two real data sets. In Section 5, a simulation study is used to assess the performance of the RD-BACON. In Section 6 the performance of the RD-BACON approach is compared to existing methods. Section 7 concerns the computational time for the proposed method, compared to the computational effort of its competitors. Finally, conclusions are given in Section 8.

2. The BACON Approach for Full-Rank Data

The BACON algorithm can be applied to multivariate data as well as to regression problems. We will focus on the former. The BACON algorithm assumes that the matrix $\mathbf{X}$ is of full-column rank and also comes from an elliptically symmetric distribution. The algorithm starts by selecting an initial basic subset, $\mathbf{X}_b$, of size $r > p$. There are two versions for selecting $\mathbf{X}_b$. In Version 1, the initial basic subset X_b consists of the r observations with the smallest values of the *Mahalanobis distances*,

$$d_i(\bar{\mathbf{x}},\mathbf{S}) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})}, \quad i = 1,\ldots,n, \qquad (1)$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$, and $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the respective mean and variance-covariance matrix of the variables in $\mathbf{X}$. In Version 2, $\mathbf{X}_b$ consists of the r observations with the smallest values of the Euclidean distances from the median,

$$d_i(\mathbf{x}_i,\mathbf{m}) = \left\|\mathbf{x}_i - \mathbf{m}\right\|, \quad i = 1,\ldots,n, \qquad (2)$$

where $\|\cdot\|$ denotes the Euclidean vector norm. In the original BACON algorithm, $\mathbf{m}$ is taken to be the vector containing the coordinatewise medians. Of course, other medians could also be used, such as the multivariate L_1 median or the spatial median (see Hössjer and Croux, 1995 for theoretical properties and computation). The latter is the L_1 location estimator $\mathbf{m}$ that solves the minimization problem

$$\min_{\mathbf{m}} \|\mathbf{x}_i - \mathbf{m}\|, \quad \text{for } i = 1, \ldots, n, \tag{3}$$

Using the L_1 median enables BACON to be equivariant regarding orthogonal transformations, but not with respect to changes in scale (i.e. it is not affine equivariant). The distances from the median provide a robust initial subset compared to the Mahalanobis distances, but the Mahalanobis distance is affine equivariant.

The initial basic subset $\mathbf{X}_b$ includes the m observations with the smallest distances d_i. The size of the initial basic subset is $r = cp$, where c is a multiplier, that is, at least c observations per parameter are used. Billor et al. (2000) suggest setting c equal to 3, 4, or 5. Let $\bar{\mathbf{x}}_b$ and $\mathbf{S}_b$ be the mean and covariance matrix of the observations in the current basic subset $\mathbf{X}_b$. Compute the robust distances

$$d_i(\bar{\mathbf{x}}_b, \mathbf{S}_b) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}}_b)' \mathbf{S}_b^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_b)}, \quad i = 1, \ldots, n, \tag{4}$$

The BACON algorithm lets the current basic subset $\mathbf{X}_b$ increase until it no longer changes. At each iteration, the basic subset includes the observations with

$$d_i(\bar{\mathbf{x}}_b, \mathbf{S}_b) < c_{npr} \chi_{p,\alpha/n} \tag{5}$$

where $\chi^2_{p,\alpha}$ is the $1-\alpha$ percentile of the chi-square distribution with p degrees of freedom, and

$$c_{npr} = c_1 + c_2 \tag{6}$$

is a correction factor, where

$$c_1 = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}, \qquad c_2 = \max\left\{0, \frac{n+p+1-2r}{n+p+1+2r}\right\}, \tag{7}$$

and r is the size of the current subset. The use of a chi-square cut-off value follows directly by the approximately normal distribution of the score vectors, assuming of course the original data are normally distributed.

The BACON algorithm for identifying outliers in full-rank data is given in Algorithm 1.

Algorithm 1: BACON for Full-Rank Data

Input: A full-rank matrix $\mathbf{X}_{n \times p}$ of multivariate data.

Step 1. Select the initial basic subset $\mathbf{X}_b$ of size $r > p$ using (1) or (2).

Step 2. Compute the distances $d_i(\bar{\mathbf{x}}_b, \mathbf{S}_b)$ in (4).

Step 3. Set the new basic subset to all points satisfying (5).

Step 4. Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5. Nominate the observations excluded from the basic subset as outliers.

Output: A set of observations nominated as outliers, if any.
Robust estimates of the location and scale $\bar{\mathbf{x}}_b$ and $\mathbf{S}_b$, respectively.

The observations excluded from the final subset are nominated as outliers. The distances $d_i(\bar{\mathbf{x}}_b, \mathbf{S}_b)$ at the final step can be used as robust distances. Furthermore, the mean and covariance matrix of the final basic subset, $\mathbf{x}_b$ and $\mathbf{S}_b$, can be viewed as robust estimators of location and scale, respectively.

As can be seen from (4), the BACON distances assume that the covariance matrix $\mathbf{S}_b$ is non-singular, or equivalently, that the subset $\mathbf{X}_b$ is of rank p, a condition which will not hold when $p > n$. Note that the matrix $\mathbf{S}_b$ could be singular even when $n > p$. Of course, if $p > n$, then the entire data matrix $\mathbf{X}$ and well as all subsets of $\mathbf{X}$ are rank-deficient and hence the corresponding $\mathbf{S}_b$ matrix is singular. Thus, when the matrix $\mathbf{S}_b$ is singular, the BACON algorithm can't be used. In Section 3 we extend the BACON approach to cases where $\mathbf{S}_b$ is singular.

3. Rank-Deficient Data

High-dimensional data where $p > n$ are automatically rank-deficient; the covariance matrix $\mathbf{S}$ as well as all subset matrices $\mathbf{S}_b$ are not invertible. Therefore, the initial distances in (1) as well as the distances in (4) cannot be computed. But rank-deficient data can also appear for $n \geq p$ when the dependence among the predictors is very high. For example, the Ionosphere data discussed in Section 6.3 has 351 observations in 31 variables, but is highly collinear.

The BACON algorithm for full-rank data can be extended to deal with rank-deficient data by regularizing the statistical problem in two main ways. The first, which we refer to as RD1-BACON, is based on applying the BACON algorithm to a subset of robust scores, denoted by $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k$ with $k << p$. This is done using the eigen decomposition of a robust scatter estimate. The second approach, which we refer to as RD2-BACON, is based on a ridge-like regularization of the matrices $\mathbf{S}$ and $\mathbf{S}_b$ by adding a small positive constant δ to its diagonal in order to solve singularities and recover the robust distances. Hence, the first technique is based on truncation on a small subset of robust orthogonal directions, while the second is based on a ridge-like regularization. The details of these two alternatives are given below.

3.1 The RD1-BACON Algorithm

The RD1-BACON algorithm is based on projecting the data along robust directions so that the outliers can be more easily detected. This set of robust directions is derived from the spatial sign covariance matrix, which is defined below. Of course, there are a variety of choices for a robust covariance matrix. We chose the spatial sign covariance matrix $\mathbf{C}$ because it is fast to compute, and is highly robust against outliers without totally discounting them. Its computational efficiency renders our algorithms suitable for large, high dimensional data sets. Other robust covariance matrices such as the MCD are clearly not suitable for high dimensional data, while computationally efficient covariance estimates that rely on special characteristics of the data such as sparsity or other features are not generally robust. More information about the spatial sign covariance estimate can be found in Visuri et al. (2000) and Locantore et al.(1999), among others.

Let $\mathbf{y} = \mathbf{x} - \mathbf{m}$, where $\mathbf{m}$ denotes the multivariate L_1 median or the spatial median (see Hössjer and Croux, 1995, for more information). Let $g(\mathbf{y})$ denote the spatial sign vector computed according to

$$g(\mathbf{y}) = \begin{cases} \dfrac{\mathbf{y}}{\|\mathbf{y}\|}, & \text{for } \mathbf{y} \neq \mathbf{0}, \\ \mathbf{0}, & \text{for } \mathbf{y} = \mathbf{0}, \end{cases} \tag{8}$$

with $\mathbf{0}$ the zero vector. Then the spatial sign covariance matrix is defined as

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i - \mathbf{m}) g(\mathbf{x}_i - \mathbf{m})^T. \tag{9}$$

As stated by Locantore et al.(1999), $g(\mathbf{x}_i - \mathbf{m})$ is the projection of the p-dimensional vector $\mathbf{x}_i$ onto the unit sphere centered at $\mathbf{m}$ to yield a robust projection. The core of this projection is given by the matrix $\mathbf{C}$, which is not affected by the outliers in $\mathbf{X}$.

Having obtained a robust covariance matrix, let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of $\mathbf{C}$, and let $\mathbf{V}$ be the corresponding matrix of normalized eigenvectors. Then, we have $\mathbf{C} = \mathbf{V}\ \mathbf{\Lambda}\mathbf{V}'$, where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues on its diagonal. For $k << p$, we compute the $n \times k$ robust scores matrix as follows

$$\widetilde{\mathbf{X}}_k = (\mathbf{X} - \mathbf{1}\mathbf{m}')\mathbf{V}_k, \tag{10}$$

where $\mathbf{m}$ is the coordinatewise median or the L_1 median from equation (3), $\mathbf{1}$ is the $n \times 1$ vector of 1's, and $\mathbf{V}_k$ is the matrix containing the first k vectors of $\mathbf{V}$.

The ordinary BACON presented in Algorithm 1 is then applied to $\widetilde{\mathbf{X}}_k$ in (10) to give the RD1-BACON algorithm. It worth noting that since BACON is computationally very efficient, it is expected that the RD1-BACON is computationally very efficient, as well. Indeed, in our experience, two or three iterations are sufficient for RD1-BACON to converge.

A parameter to be further defined in the RD1-BACON is the dimension reduction parameter k. A common way to determine k involves inspection of the eigenvalues $(\lambda_1, \ldots, \lambda_p)$ and choice of k so that for $\lambda_k > 0$ the truncated sequence $(\lambda_1, \ldots, \lambda_k)$ gives sufficient dimension reduction. Screeplots or barplots of the eigenvalues are quite often applied for such investigations. Yet, this is time consuming and for the purposes of reporting and comparison, could give different results due to subjectivity. Therefore, we propose to let k be the value such that the first k eigenvalues contribute to a predetermined high percentage of the total variation $\sigma*$, that is,

$$k = \arg\min_{k^*} \frac{\sum_{\kappa=1}^{k^*} \lambda_\kappa}{\sum_{\kappa=1}^{p} \lambda_\kappa} > \sigma^*. \tag{11}$$

Based mainly on personal experience and empirical investigation we chose k such that $\sigma^* = 97.5\%$ This has the effect of automatically eliminating components that contribute very little to the total variance, which is very important if $p > n$. We have obtained very good results using this automated selection rule, as will be demonstrated in the simulations in Section 5.

The RD1-BACON can thus be seen as an attempt to remove the directions with very small robust scales, as indicated by the eigenvalues of the positive semi-definite, robust covariance matrix $\mathbf{C}$.

In order to classify the observations of the basic subset as outlying or non-outlying, the regular BACON algorithm uses the cut-off value given by expression (5). In the rank-deficient case, we replace p by k, that is, the size of the initial basic subset $r = ck$. Note that the smaller the value of k, the larger the initial selection parameter c should be. In such cases, in order to make the algorithm more effective, one should compromise between k and c. However, sometimes k is still fairly large, so that even $3k > n$. We therefore select r as the minimum of ck and $h = \lfloor (n+k+1)/2 \rfloor$. We cap the initial subset at h since if it contains more than half the observations, there is increased likelihood that it might contain some outliers. Therefore, in order to make the algorithm more effective, we use

$$d_i < c_{nkr}\, \chi_{k,\alpha/max(p,n)}. \tag{12}$$

The RD1-BACON algorithm is given in Algorithm 2.

```
Algorithm 2  RD1-BACON for Rank-Deficient Data
Input: A rank-deficient matrix $\mathbf{X}_{n\times p}$ of multivariate data.
Step 1. Compute the robust scores matrix $\widetilde{\mathbf{X}}_k$ in (10).
Step 2. Apply the ordinary BACON Algorithm 1 to $\widetilde{\mathbf{X}}_k$ but using the
        critical value in (12) instead of (5).
Output: A set of observations nominated as outliers, if any.
        Robust estimates of the location and scale $\bar{\mathbf{x}}_b$ and $\mathbf{S}_b$, respectively.
```

3.2 The RD2-BACON Algorithm

The RD2-BACON algorithm is based on a ridge type regularization in the covariance matrix (see Hoerl and Kennard (1970) and Engl et al. (1996), Chapters 4 and 5). Using the same robust location estimate m as before, let the robustly centered data matrix be $\mathbf{Y} = \mathbf{X} - \mathbf{1m}'$. The spectral decomposition of this robustly centered covariance matrix $\mathbf{S} = \mathbf{Y}'\mathbf{Y}$ can be computed as

$$\mathbf{S} = \mathbf{V}\ \mathbf{\Lambda}\mathbf{V}', \tag{13}$$

where $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{S}$, and $\mathbf{V}$ is the corresponding orthonormal matrix of eigenvectors. Using expression (11), let k be such that $\sigma^* = 0.975$. In addition, set δ equal to the k^{th} largest eigenvalue and let $\mathbf{I}_p$ be the identity matrix of order p. Then the matrix in (13) is replaced by

$$\mathbf{S} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}', \tag{14}$$

where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda} + \delta\mathbf{I}_p$. Accordingly, the Mahalanobis distances in (1) can be replaced by

$$d_i\left(\mathbf{C},\mathbf{S}^*\right) = \sqrt{\left(\mathbf{x}_i - \mathbf{m}\right)'\left(\mathbf{S}^*\right)^{-1}\left(\mathbf{x}_i - \mathbf{m}\right)}, \quad i = 1,\ldots,n \tag{15}$$

The RD2-BACON algorithm starts by ordering the observations according to their distances in (15). The subset of $r << p$ with the smallest distances forms the initial basic subset. As before, we select r to be the minimum of ck and h. The spectral decomposition of $\mathbf{S}_b$ is computed and the distances in (4) are replaced by

$$d_i\left(\mathbf{x}_b,\mathbf{S}_b^*\right) = \sqrt{\left(\mathbf{x}_i - \overline{\mathbf{x}}_b\right)'\left(\mathbf{S}_b^*\right)^{-1}\left(\mathbf{x}_i - \overline{\mathbf{x}}_b\right)}, \quad i = 1,\ldots,n \tag{16}$$

wherze $\mathbf{S}_b^* = \mathbf{V}_b\boldsymbol{\Lambda}_b^*\mathbf{V}_b'$ and $\boldsymbol{\Lambda}_b^* = \boldsymbol{\Lambda}_b + \delta\mathbf{I}_p$. Note that since we now have an assumed outlier-free basic subset, we can use its mean $\overline{\mathbf{x}}_b$ rather than the coordinatewise median $\mathbf{m}$ (a robust mean is usually preferred to the median due to its higher statistical and computational efficiency).

Similar to the ordinary BACON algorithm, points are iteratively added to the basic subset if their distances are below the cutoff value, with convergence occurring when the basic subset no longer changes. However, since the distribution of the distances in equation (16) would be difficult to determine, we employ a nonparametric approach. The basic subset is thus chosen to include all observations satisfying

$$d_i \leq \text{med}(d) + c_\alpha \ \text{IQR}(d), \tag{17}$$

where c_α is a suitable constant chosen to control the desired null size α, and $\text{med}(d)$ and $\text{IQR}(d)$ are the median and the interquartile range, respectively, of the distances in $d = \{d_1, d_2, \ldots, d_n\}$. The subsets of RD2-BACON algorithm grow rapidly and according to our experience just a few iterations are sufficient to converge.

Note that when $p > n$, the computational burden of eigenvalue decomposition with large p reduces due to the n-dimensional observations space. That is, instead of computing the eigen decomposition of the matrix $\mathbf{S} = \mathbf{Y}'\mathbf{Y}$, which is $p \times p$, one can compute the eigen decomposition of the much smaller (in dimension) matrix

$$\mathbf{Y}\mathbf{Y}' = \mathbf{U}\boldsymbol{\Lambda}^*\mathbf{U}', \tag{18}$$

which is of order $n \times n$, where the matrix $\boldsymbol{\Lambda}$ in (13) and (18) are identical, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_p)$ is an $n \times p$ orthonormal matrix with columns corresponding to the eigenvectors of $\mathbf{Y}\mathbf{Y}'$. The eigenvectors $\mathbf{V}$ are then obtained according to

$$\mathbf{V} = \mathbf{Y}'\mathbf{U}\ \boldsymbol{\Lambda}^{-1/2}. \tag{19}$$

We shall see in Section 7 that both algorithms scale very well with increasing dimension, although RD2-BACON does not scale quite as well with increasing n, since it has to perform the eigen decomposition for each iteration. Fortunately though, it usually converges within a few iterations.

The RD2-BACON algorithm is given in Algorithm 3.

```
Algorithm 3 RD2-BACON for Rank-Deficient Data
Input: A rank-deficient matrix $\mathbf{X}_{n\times p}$ of multivariate data.
Step 1. Select the initial basic subset $X_b$ of size $r << p$[ using (2) or (15).
Step 2. Compute the distances $d_i\left(\bar{\mathbf{x}}_b, \mathbf{S}_b^*\right)$ in (16).
Step 3. Set the new basic subset to all points satisfying (17).
Step 4. Iterate Steps $2$ and $3$ until the size of the basic subset no longer changes.
Step 5. Nominate the observations excluded from the basic subset as outliers.
Output: A set of observations nominated as outliers, if any.
        Robust estimates of the location and scale $\bar{\mathbf{x}}_b$ and $\mathbf{S}_b$, respectively.
```

Finally, we note that the RD2-BACON algorithm is a natural generalization of the ordinary multivariate BACON algorithm. To see this, consider the case where $n \le p$ and no collinearity appears on the predictors. The ordinary multivariate BACON will not be able to give an inverse for matrix $\mathbf{S}$ and the final distances d_i. Therefore the RD2-BACON adds a constant δ in order to get $\mathbf{S}^*$ and d_i^*. When no collinearity is present, the d_i^* are proportional to d_i and almost any value of λ_j can play the role of δ.

4. Experience with Real Data

In this section, we apply the rank-deficient BACON algorithm for outlier detection on two real-life examples: the Climatological data (see Ramsay and Silverman, 2005) and the Octane data (see Tenenhaus,1998).

4.1 Climatological Data

The temperatures taken from 35 weather stations across Canada over a year (365 days) have been registered and are accessible through the FDA homepage at http://ego.psych.mcgill.ca/misc/fda/index.html, as well as in the *fda* package (Ramsay et al.,2009) of the R statistical software (R Core Team, 2012). The 35 weather stations represent Pacific, Continental, Atlantic, and Arctic climates. The temperatures for the latter are distinguished from the rest of the stations by their cold temperatures throughout the year.

This data set is a typical example of $p > n$ data which are are functioonal data (see Ramsay and Silverman, 2005). Data are essentially discretized curves rather than ordinary vectors. Our approach for outlier detection does not take into account the special functional form of the data but it identifies observations which are declared as outliers throughout the whole year.

The RD1-BACON and RD2-BACON algorithms are tested on these Canadian temperatures data based on daily records that constitute a 35×365 data matrix $\mathbf{X}$. It is not completely clear exactly which points should be outliers, but at least, the three Arctic stations Inuvik, Iqaluit, and Resolute are good candidates.

Figure 2 illustrates the final distances of the RD1-BACON algorithm in the left panel and the final distances of the RD2-BACON algorithm in the right panel. Together with these distances, the critical limits are depicted by a dashed line. The observations which lie above the latter are the detected outliers. The RD1-BACON detected 1 outlier, the most extreme (coldest) station of Resolute. The RD2-BACON detected the 3 Arctic stations (Inuvik, Iqaluit, and Resolute) as well as Churchill, which could also be reasonably viewed as an outlier, very cold winters but relatively warm summers.

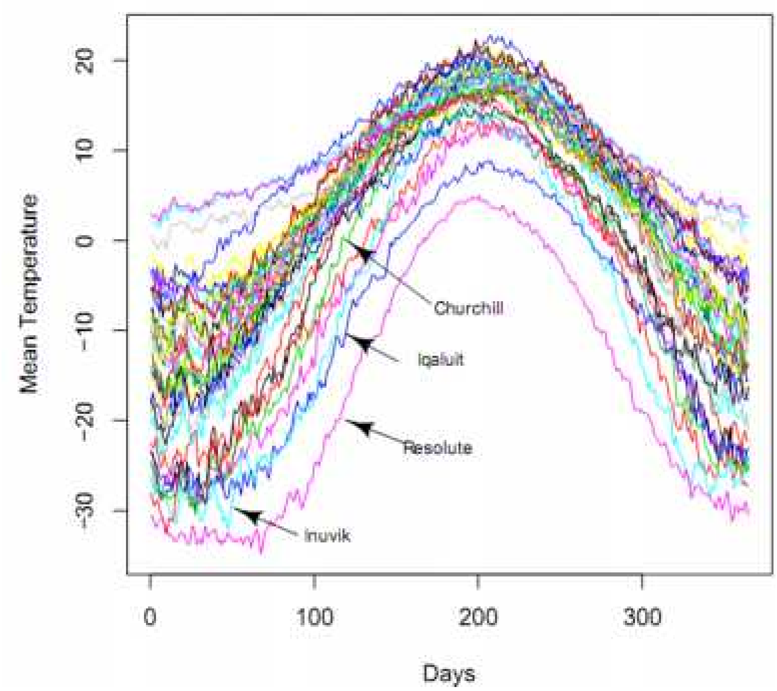


Figure 1: Temperature data. Temperatures over 365 days for 35 weather stations in Canada.

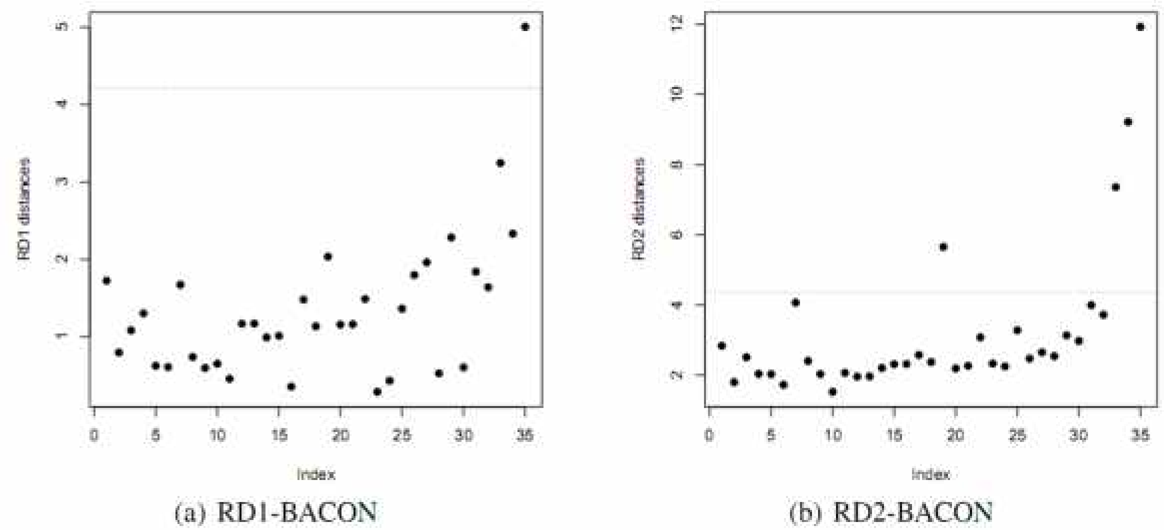
4.2 Octane Data

The octane data set consists of the Near Infra-Red spectra of 39 gasoline samples for which the octanes have been measured at 225 wavelengths in the spectral range 1102-1552 nm in steps of 2 nm. The resulting data are stored in a 39×225 data matrix. The octane data are commonly used in partial least squares regression (PLSR) (see Helland, 1988) and principal components regression (PCR) (see Jolliffe, 2002). Given a response vector, both methods construct predictive models relating the NIR spectra to chemical compounds. In this case these are the octanes. The spectra for the the 39 samples are illustrated in Figure 3, from which it can be clearly seen that there are 6 outliers --- the six curves with substantial deviations at the higher wavelength, to which alcohol has been added. From Figure 4 it is clear that the outliers are the observations with index numbers $25,26,36,37,38,$ and 39.

Both implementations of the BACON for rank-deficient data have detected the six outlying samples $(25,26,36,37,38,$ and $39)$ in the data. These are illustrated by the points above the critical value given by the dashed line in Figure 4.

5. Simulation Study

In this section we want to obtain a broader understanding of the proposed methods, based on a large scale simulation study. Our simulation experiments are described below, followed by a discussion of the simulation results.



(a) RD1-BACON (b) RD2-BACON

Figure 2: Temperatures data. (a) RD1-BACON robust distances indicated by points and the critical value given by the dashed line. (b) RD2-BACON robust distances indicated by points and the critical value given by the dashed line.

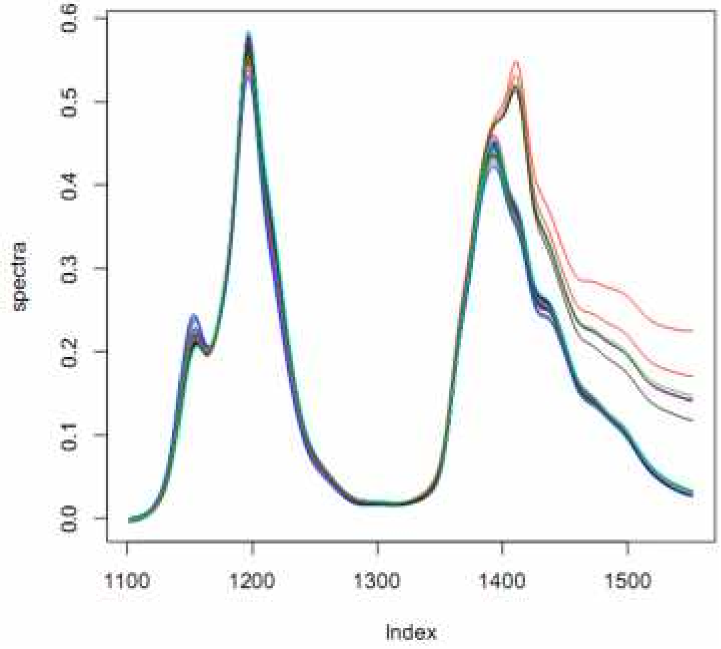
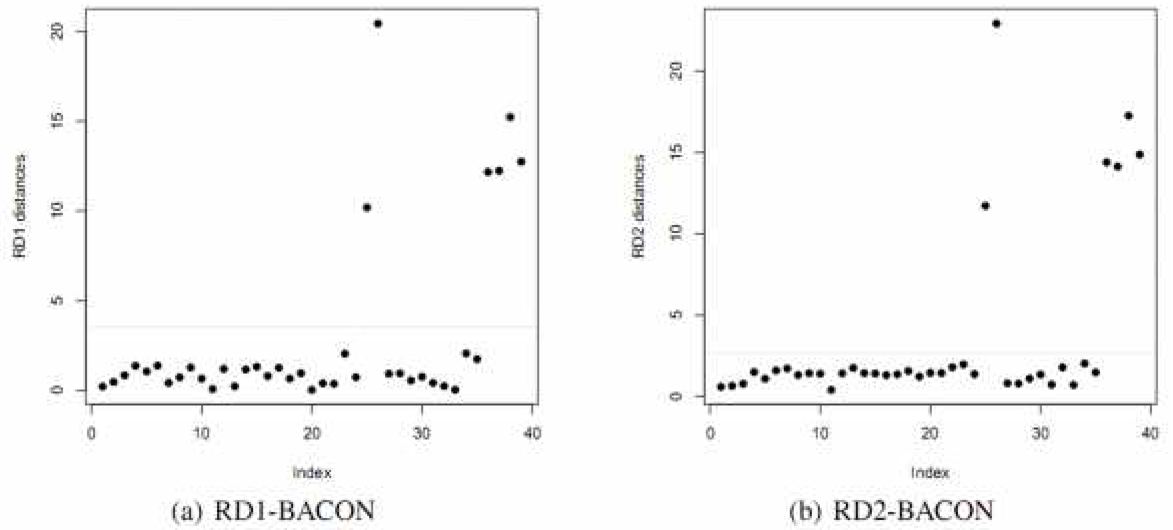


Figure 3: Octane data. Near Infra-Red spectra for 39 gasoline samples in the spectral range 1102-1552 nm in steps of 2 nm.

5.1. Design of the Experiments

The experimental design is described as follows: For the $n \times p$ matrix $\mathbf{X}$ we generate $p^* < p$ columns of random observations according to

$$x_{ij^*} \sim U(0,10), \, where \; i = 1,\ldots,n \; and \; j^* = 1,\ldots,p^*.$$



(a) RD1-BACON (b) RD2-BACON

Figure 4: Octane data. (a) RD1-BACON robust distances indicated by points and the critical value given by the dashed line. (b) RD2-BACON robust distances indicated by points and the critical value given by the dashed line.

For the remaining $(p - p^*)$ columns linear dependencies were introduced in the following manner. For $j = (p^* + 1),\ldots,p$ and $j^* = 1,\ldots,(p - p^*)$ we generate

$$\mathbf{x}_j = g\mathbf{x}_{j*} + \mathbf{e}, \text{ by setting } g \sim U(0,1) \text{ and } \mathbf{e} \sim N(0,1).$$

We then contaminate the data set by randomly replacing 10%, 15% and 20% of the observations. We denote by "% cont" the proportion of the contaminated units. These were generated according to $x_{ij^*} \sim U(12,20)$, with collinearity induced in exactly the same way with the non-contaminated data. We set the number of observations equal to 50 and we created moderately high-dimensional data ($n \le p$) by setting $p = 50$, and high-dimensional data ($n << p$), by setting $p = 100$. Finally, we fix a high level of induced linear dependencies among the predictors by setting $p^* = p/10$.

We run M = 1,000 simulations for all the combinations of these parameters. As performance measures, we report the average proportion of false negative, false positives, and $P_1 = \Pr$ (correctly classifying the entire data set). The latter is a very strict measure that gives the percentage of data sets (out of the 1,000 simulation runs) that were perfectly classified.

5.2 Simulation Results

We start with the null case of non-contaminated data, to see the performance of the algorithms without any outliers. Table 1 gives the average number of false positives for data sets without any outliers, measured over 1,000 simulations. It is clear that lack of outliers does not "force" the algorithm to classify some of the regular data as outliers.

Now we consider contaminated data. Table 2 shows that both RD1-BACON and RD2-BACON do a very good job of detecting outliers and regular data for the given simulation conditions. RD1-BACON generally has slightly higher average number of false negatives than RD2-BACON, but a slightly lower number of false positives. RD1-BACON generally has a higher probability P_1 of classifying the entire data set correctly than RD2-BACON. That is, there are more instances (of the 1,000 simulations) when RD1-BACON gets everything exactly correct, but for those data sets for which it doesn't get everything exactly correct, it makes relatively more mistakes than RD2-BACON. This implies that RD2-BACON consistently makes only a few mistakes whereas RD1-BACON is more variable. The difference in performance, however, is very small.

Table 1: Simulation Results (Null case): The probability of falsely declaring any point an outlier in data sets without any planted outliers.

p	RD1	RD2
50	0.001	0.007
100	0.004	0.008

Table 2: Simulation Results (Contaminated Cases): False negatives, false positives and $P_1 = \Pr$ (correctly classifying the entire data set). Note that 0.000 indicates an average less than 0.001, while 0 indicates no mistakes were made over any of the 1,000 simulations.

		Contamination Rate					
		10%		15%		20%	
	p	RD1	RD2	RD1	RD2	RD1	RD2
FN	50	0.027	0	0.014	0.000	0.005	0.001
	100	0.012	0	0.001	0	0	0
FP	50	0.000	0.002	0.000	0.001	0.000	0.000
	100	0.001	0.002	0	0.000	0.000	0.000
P_1	50	0.964	0.927	0.976	0.962	0.986	0.976
	100	0.989	0.929	0.996	0.966	0.994	0.997

6. Benchmarking to Other Robust Methods

The alternative approaches for outlier detection and robust location-scatter estimation that we consider are the robust location and dispersion given in Maronna and Zamar (2002), and the ROBPCA algorithm presented in Hubert et al.(2005). The latter is mainly used in Principal Components Analysis and it robustly estimates location and scatter by combining projection pursuit ideas with the MCD covariance estimator (see Rousseeuw,1984) ; it is indeed suitable for high-dimensional data. The OGK Maronna-

Zamar estimate, where OGK stands for Orthogonalized Gnanadesikan-Kettenring, is based on the approximately uncorrelated score vectors of a pseudo-correlation matrix. The latter is built by enforcing variances of all variables to be 1, while for the off-diagonal elements they use a robust measure of dispersion (denoted generally by σ) and determine the correlations in pairwise fashion X_i and X_j based on the identity

$$Cov(X_i, X_j) = \frac{1}{4}\left(\sigma(X_i + X_j)^2 - \sigma(X_i - X_j)^2\right),$$

popularized by Gnanadesikan and Kettenring (1972).

Various arguments can be used inside the ROBPCA and the OGK functions. We will try in what follows to simplify the calculation of both estimates by taking the appropriate arguments as suggested by the authors and our knowledge of the data to which we apply them. For example, we do not use the median and the MAD (where MAD denotes the median absolute deviation) as location and scale estimates in the OGK procedure since, as suggested by the authors, it worsens results, especially for high-dimensional data. We also use one reweighing step for OGK estimate and we detect outliers based on Mahalanobis distances (see Maronna and Zamar, 2002, page 308-309). For the ROBPCA algorithm we use prior knowledge concerning the value of the k components and we set the parameter α equal to its default value, that is 0.75. We take as outliers the observations which are finally flagged in the ROBPCA algorithm.

We revisit here the octane and the weather data in order to compare outlier detection results provided by the RD-BACON algorithms and its robust competitors. We display in Table 3 the detected outliers for each method together with the true outliers. The results are analyzed separately for the octane data and the weather data.

6.1 Octane Data

For the octane data, the ROBPCA method was run on three principal components with α set by default to 0.75. The OGK estimate was run based on a weighted mean for location and a truncated standard deviation for scale estimates as it is described in their paper (see Maronna and Zamar, 2002, page 310). We have, moreover, used one reweighing step. The detected outliers have had Mahalanobis distances larger than d_0, where

$$d_0 = \frac{\chi_p^2(\beta)\cdot \text{med}(d_1,\ldots,d_n)}{\chi_p^2(0.5)},$$

where $\chi_p^2(\beta)$ denotes the β^{th} quantile of a chi-square random variate with p degrees of freedom. Following Maronna and Zamar (2002) we set $\beta = 0.9$.

6.2 Weather Data

We followed the same steps concerning the OGK robust location and dispersion estimate. We investigated the ROBPCA method for 2 to 5 components. Generally, the fewer the number of components, the fewer outliers detected. Table 3 shows results from the ROBPCA method with 2 components, and $\alpha = 0.9$ to increase efficiency.

Table 3: Detected outliers in the real data sets for all robust methods. True outliers correspond to observations which have, to our knowledge, been referred to as outliers in previous analysis.

Octane data					Weather data			
TRUE	MZ	ROBPCA	RD1	RD2	MZ	ROBPCA	RD1	RD2
25	25	25	25	25	35	35	35	35
26	26	26	26	26	34	34	-	34
36	36	36	36	36	33	33	-	33
37	37	37	37	37	31	31	-	-
38	38	38	38	38	30	30	-	-
39	39	39	39	39	29	-	-	-
-	3	3	-	-	25	-	-	-
-	18	-	-	-	19	19	-	19
-	27	-	-	-	7	7	-	-
-	34	-	-	-	-	-	-	-

6.3 Ionospheric Data

The Ionospheric data set contains 351 observations on 35 variables. The first 34 continuous variables are used for the prediction, while the 35th variable classifies the observations into "good" or "bad" observations. The Ionospheric data set is available in package *dprep* (Acuna et al., 2009) in R (R Core Team, 2012), as well as from the UC-Irvine Machine Learning Repository(Asuncion and Newman, 2007). We did not include this data set in Section 4 because it is not rank-deficient, so any robust algorithm could be used.

This data set has already been analyzed in Maronna and Zamar (2002), among many others. It provides an excellent $n > p$ data set with extreme collinearity, where the use of the RD-BACON algorithm is necessary. In our analysis we followed Maronna and Zamar (2002), that is, we retained the 225 "good" observations while we removed variables 1, 2, and 27. These variables were removed because their $\text{MAD} = 0$. For the RD-BACON algorithms this poses no problem, yet variables 1,2, and 27 were removed in order to make results comparable to the ones given in Maronna and Zamar (2002).

Table 4 displays the observations with the largest Mahalanobis distances for various algorithms, and is essentially the same as the table given in Maronna and Zamar (2002). The first row shows the distances for the FMCD estimator (see Rousseeuw and van Driessen, 1999) with 500 subsamples for the MCD algorithm. The second and third rows display the distances from the OGK estimate with and without reweighing and for $\beta = 0.9$. For more details, see Maronna and Zamar (2002). The last 2 rows show the distances we obtained from the RD1-BACON and the RD2-BACON algorithms. We do not list the ROBPCA results here since they are essentially the same as the FMCD results.

We note that the Ionospheric data set is indeed quite complicated and lacking a gold standard, not possible for any method to claim superiority. We merely aim to show that our method produces results in line with other established methods. As noticed by Maronna and Zamar (2002), further research as well as subject matter knowledge are essential in order to understand this data set.

Table 4: Ionospheric data: Observations with the largest distances d_i from left to the right.

Estimate	Observations with largest distances								
FMCD	99	121	226	148	348	349	100	3	173
OGK (0.9)	264	121	324	99	257	163	173	157	110
OGK $_{(2)}$ (0.9)	121	99	148	349	242	264	173	4	358
RD1 – BACON	121	99	208	173	148	264	349	3	121
RD2 – BACON	121	99	148	349	4	208	302	348	3

6.4 Discussion

It is not our goal to compare the performance of the proposed method to other robust alternatives. If it was the case, we would have definitely prefered to do so in a simulation experiment where data generation and contamination would be under our control. For some of the given data sets, however, the notion of "true" outlier is not well defined. Nevertheless, our method should at least in principle agree with the other methods; these being well known and established robust methods in statistical practice. Therefore we would like to benchmark our proposal to these methods and briefly discuss the main results obtained from the three data sets under study.

For the octane data, the six outliers are known and easily identified by both versions of RD-BACON without any false positives, as can be seen in Table 3. The method of Maronna and Zamar (2002) identifies some false positives in addition to the correct ones, while the ROBPCA also correctly identifies the outliers as well as one false positive.

The true outliers for the Canadian weather stations have not been definitively declared as such in the literature, according to our knowledge. As can be seen in Table 3, RD1-BACON identifies the most extreme outlier, while RD2-BACON additionally identifies a few other unusual stations. ROBPCA identifies the same outliers as RD2-BACON as well as one additional station. The method of Maronna and Zamar (2002) identifies a few additional stations as outliers. Note that the general trend is in line with the Octane data example, where ROBPCA found only one additional outlier while the method of Maronna and Zamar (2002) found quite a number of additional outliers. Based on this, we think it is possible that some of the stations identified by Maronna and Zamar (2002) as outliers could be false positives.

Our results seem to be in line with other methods, particularly ROBPCA. This is also verified in the Ionospheric data where there is sufficient evidence that the robust methods agree, assigning large distances to the same suspected outliers.

7. Computation Times

It is helpful to compare the computation times of our algorithms with those of other established routines. Already examined in the preceding section, FMCD (Rousseeuw and van Driessen,1999) is one algorithm that has become very popular as a robust estimation technique. It is always helpful to compare a new algorithm against one that is well-known and for which standardized routines exist. We therefore compare the computation times of RD1-BACON and RD2-BACON with FMCD. We compared two implementations of FMCD in the R statistical software system (R Development Core Team, 2009): FMCD ^{a} is the *covMcd* routine in the *rrcov* package (Todorov and Filzmoser, 2009) , while FMCD ^{b} is the *cov.rob* routine in the *MASS* package (Venables and Ripley, 2002).

We examine the effect of p and n separately. Note that FMCD requires $p < n$ while the *cov.rob* implementation requires $p \leq 50$, so we split the investigation into two parts. In Table 5 we include FMCD in the comparison and restrict the dataset size accordingly. We thus maintain n = 1,000 and investigate p = 10, 20, 30, 40, and 50. Then hold $p = 50$ and investigate n = 1,000, 2,000, 5,000, and 10,000. We exclude FMCD from the comparisons in Table 6, for which we set n =100 and 500 and let p vary from the given value of n up to p = 1000. The results shown in these two tables are the average values from 10 simulations from a PC with 2 Ghz processing speed and 32-bit operating system.

It is clearly evident that not only are RD1-BACON and RD2-BACON very fast, but they also scale well with increasing dimension. RD1-BACON is always faster than RD2-BACON, which can be explained by the former's use of only the k most important eigenvalues/vectors regardless of the actual dataset dimension. That is, for high p and not so high n, RD1-BACON projects the data onto a subspace with dimension $k << p$, whereas RD2-BACON still utilizes all p dimensions during the computation of the covariance matrix and its inverse (as well as for other computations). As expected, FMCD is very slow for practical use in large datasets, and does not scale well with increasing dataset size (either n or p). Although a lot faster than FMCD, the OGK method is also rather slow and does not scale well with increasing p. This is likely due to its pairwise computation of the covariance matrix, which can become very burdensome in high dimensions. ROBPCA is generally pretty fast, although for certain configurations it can be unexpectedly slow, which we presume is due to it having difficulty finding the optimum parameters to investigate a given dataset. Nevertheless, it does scale very well with increasing dimension and is sometimes even faster than RD2-BACON, which is not as fast in high dimensions. ROBPCA is also faster than RD1-BACON for $n = 100$ and p = 500 and 1,000, but not when n is increased to $n = 500$.

8. Concluding Remarks and Further Research

The multivariate BACON algorithm for outlier detection has been extended here to include rank-deficient data. These arise from highly collinear variables as well as from high-dimensional data sets. Outlier detection in such cases commonly breaks down. Two algorithms have been presented and tested on real and simulated data sets. Both implementations regularize the statistical problem following rather different approaches.

Table 5: Computation time (seconds): $p < n$. FMCD^a is the *covMcd* routine in the *rrcov* package (Todorov and Filzmoser, 2009) and FMCD^b is the *cov.rob* routine in the *MASS* package (Venables and Ripley, 2002) are two implementations of FMCD in the R statistical software system.

n	p	RD1	RD2	FMCD^a	FMCD^b	OGK	ROBPCA
1,000	10	0.038	0.041	0.535	11.6	0.051	13.100
	20	0.050	0.060	1.394	43.2	0.230	2.674
	30	0.061	0.840	3.368	81.1	0.510	2.800
	40	0.059	0.114	6.121	123.6	0.889	2.985
	50	0.070	0.170	11.600	177.7	2.009	4.990
2,000	50	0.125	0.342	11.410	410.0	2.395	7.530
5,000	50	0.368	0.990	16.550	935.8	5.424	22.020
10,000	50	0.773	1.868	11.560	1,810.0	9.282	50.870

Table 6: Computation time (seconds): $p \geq n$

n	p	RD1	RD2	OGK	ROBPCA
100	100	0.032	0.124	3.074	0.599
	200	0.098	0.473	12.735	0.692
	500	0.885	4.728	80.032	0.738
	1,000	7.425	33.065	409.429	0.967
500	500	1.325	6.593	107.918	8.768
	1,000	9.082	53.747	469.593	11.906

The RD1-BACON implementation reduces the dimension of the data by projecting it onto a k-dimensional subspace, where $k < p$. The ordinary multivariate BACON can then be run in this subspace. The RD2-BACON algorithm is based on a ridge type operation in order to get a "pseudo"-inverse and calculate robust distances. Since the distribution of the latter is hard to determine, a nonparametric decision criterion has been utilized.

Both RD1-BACON and RD2-BACON demonstrate good performance at classifying both outliers and regular data for a range of parameters. Both algorithms are also very fast and are particularly good at handling high-dimensional data. RD1-BACON is somewhat faster than RD2-BACON, although this difference is small compared to other established algorithms. For datasets with dimension greater than a few hundred, though, RD1-BACON is noticeably faster than RD2-BACON. The simulation results justified the above remarks. Finally, in order to benchmark the proposed method to other well established robust techniques. To do so we used three real world examples. The resulting features demontrated that the BACON results have been in line with the other methods, particularly with ROBPCA.

Our goal has been rather limited, that is, extending the BACON algorithm for outlier detection in the rank-deficient setting. The proposed methods could be further used for robust estimation, but, this will be the goal of a following paper. In addition, the extension of the proposed methods to robust regression is undoubtedly very appealing, since the original BACON algorithm includes both robust multivariate estimation and robust regression.

References

1. Acuna, E., Members of the CASTLE Group at UPR-Mayaguez, and Rico, P. (2009). dprep: *Data preprocessing and visualization functions for classification*. R package version 2.1.
2. Asuncion, A. and Newman, D. (2007). UCI machine learning repository.
3. Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* Clarendon Press, Oxford.
4. Atkinson, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *J Am Stat Assoc*, 89:1329–1339.
5. Béguin, C. and Hulliger, B. (2008). The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data. *Surv Methodol*, 34(1): 91–103.
6. Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York.
7. Billor, N., Hadi, A. S., and Velleman, P. F. (2000). Bacon: Blocked adaptive computationally-efficient outlier nominators. *Comput Stat Data Anal*, 34: 279–298.
8. Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression*. John Wiley & Sons, New York.
9. Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.
10. Croux, C. and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the project-pursuit approach revisited. *J Multivar Anal*, 95: 206–226.
11. Engl, H. W., Hanke, M., and Neubauer, A. (1996). *Regularization of Inverse Problems.* Kluwer, Dordrecht.
12. Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Comput Stat Data Anal*, 52: 1694–1711.
13. Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biom*, 28: 81–124.
14. Gray, J. B. (1986). A simple graphic for assessing influence in regression. *J Stat Comput Simul*, 24: 121–134.
15. Hadi, A. S. (1992a). Identifying multiple outliers in multivariate data. *J Royal Stat Soc B*, 54:761–771.
16. Hadi, A. S. (1992b). A new measure of overall potential influence in linear regression. *Comput Stat Data Anal*, 14:1–27.

17. Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *J Royal Stat Soc B*, 56:393–396.
18. Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J Am Stat Assoc*, 88: 1264–1272.
19. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Elements of Statistical Learning.* Springer, New York.
20. Helland, I. (1988). On the structure of partial least squares regression. *Commun Stat Simul Comput*, 17: 581–607.
21. Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12: 55–67.
22. Hössjer, O. and Croux, C. (1995). Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *J Nonparametr Stat*, 4: 293– 308.
23. Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47:64–79.
24. Izenman, A. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* Springer Texts in Statistics, New York.
25. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-Verlag.
26. Kianifard, F. and Swallow, W.H. (1989). Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression. *Biom*, 45:571–585.
27. Kondylis, A. and Hadi, A. S. (2006). Derived components regression using the bacon algorithm. *Comput Stat Data Anal*, 51:556–569.
28. Li, G. and Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J Am Stat Assoc*, 80(391):759–766.
29. Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., and Cohen, K. (1999). Robust principal components analysis for functional data. *Test*, 8:1–73.
30. Maronna, R. A. and Zamar, R. (2002). Robust estimation of location and dispersion for high dimensional data. *Technometrics*, 44:307–317.
31. Paul, S. R. and Fung, K. Y. (1991). A generalized extreme studentized residual multiple-outlier-detection procedure in linear regression. *Technometrics*, 33: 339– 348.
32. R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
33. Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, Springer-Verlag, New York.
34. Ramsay, J. O., Wickham, H., Graves, S., and Hooker, G. (2009). fda: *Functional Data Analysis*. R package version 2.2.0.
35. Rousseeuw, P. (1984). Least median of squares regression. *J Am Stat Assoc*, 79:871– 880.
36. Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons, New York.
37. Rousseeuw, P. J. and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223.
38. Rousseeuw, P. J. and van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points (with discussion. *J Am Stat Assoc*, 85:633–651.

39. Tenenhaus, M. (1998). *La régression PLS. Théorie et pratique*. Technip, Paris.

40. Todorov, V. and Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *J Stat Softw*, 32(3):1–47.

41. Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

42. Visuri, S., Koivunen, V., and Oja, H. (2000). Sign and rank covariance matrices. *J Stat Plan Inference*, 91:557–575.