Functional outlier detection with robust functional principal component analysis

Article *in* Computational Statistics · March 2012

DOI: 10.1007/s00180-011-0239-3 · Source: RePEc

CITATIONS
32

READS
208

3 authors, including:

Pallavi Sawant
Kansas State University
1 PUBLICATION **32** CITATIONS
SEE PROFILE

Nedret Billor
Auburn University
46 PUBLICATIONS **835** CITATIONS
SEE PROFILE

Comput Stat (2012) 27:83–102
DOI 10.1007/s00180-011-0239-3

ORIGINAL PAPER

Functional outlier detection with robust functional principal component analysis

Pallavi Sawant · Nedret Billor · Hyejin Shin

Received: 12 May 2010 / Accepted: 2 February 2011 / Published online: 25 February 2011

Abstract Functional principal component analysis is the preliminary step to represent the data in a lower dimensional space and to capture the main modes of variability of the data by means of small number of components which are linear combinations of original variables. Sensitivity of the variance and the covariance functions to irregular observations make this method vulnerable to outliers and may not capture the variation of the regular observations. In this study, we propose a robust functional principal component analysis to find the linear combinations of the original variables that contain most of the information, even if there are outliers and to flag functional outliers. We demonstrate the performance of the proposed method on an extensive simulation study and two datasets from chemometrics and environment.

Keywords Functional data · Outliers · Principal component analysis · Robust methods

1 Introduction

In various areas such as chemometrics, biometrics, engineering, genetics, and e-commerce the data come from the observation of continuous phenomenons of time or space known as functional data. Due to advancement of new techniques it is now

P. Sawant · N. Billor (✉)
Department of Mathematics and Statistics, Auburn University, Auburn, AL, USA
e-mail: billone@auburn.edu

P. Sawant
e-mail: sawanpr@auburn.edu

H. Shin
Department of Statistics, Seoul National University, Seoul 151-747, Korea
e-mail: hjshin112@gmail.com

possible to record large number of variables simultaneously. The nature of this data in many applications is high dimensional where the number of variables (p) is greater than the number of observations (n). The focus of researchers is on analysis of such data due to the emergence of statistical problems while applying various statistical tools for data analysis.

In functional data, the first step is to represent the data in a lower dimensional space in order to have better interpretation. This is done by performing functional principal component analysis (FPCA) to capture the main modes of variability of the data by means of small number of components which are linear combinations of original variables that allow for better interpretation of various sources of variation. In the presence of outliers, dimension reduction via FPCA would yield untrustworthy results since FPCA is based on the covariance function which is known to be sensitive to outliers. As a result, the first few component functions are often attracted toward outliers, and therefore may not capture the main modes of variability of the homogeneous observations. This necessitates the need of the robust FPCA method. There has been proposed several robust FPCA methods in the statistical literature. The first study by Locantore et al. (1999) used the idea of spherical (SPHER) and elliptical (ELL) principal component analysis (PCA) to explore abnormalities in the curvature of the cornea in the human eye. However, these methods have two drawbacks. First drawback is that SPHER and ELL only estimate the principal components and not their eigenvalues. Second drawback is that SPHER and ELL PCA are influenced by outliers (Hubert et al. 2005). Febrero et al. (2007, 2008) also proposed two methods for outlier detection that are based on the idea of functional depths and distance measures which have computational difficulties. Sensitivity of the penalized functional PCA via empirical influence functions idea has been studied by Yamanishi and Tanaka (2005). Hyndman and Ullah (2007) used a robust PCA idea for forecasting the mortality and fertility rates. Recently, Gervini (2009) developed the robust FPCA for sparsely and irregularly observed functional data and used it for outlier detection. The main contribution of our work is to construct a robust PCA method to achieve dimension reduction of data and to develop tools for detection of functional outliers which could be in the form of shape or magnitude affecting functional statistics such as the mean and the covariance function.

The outline of this manuscript is as follows. In Sect. 2, a brief description of classical functional principal component analysis is given. The newly proposed robust PCA for functional data and functional outlier detection are described in Sect. 3. Section 4 consists of numerical examples containing two datasets and extensive simulation. Finally, we give our conclusion in Sect. 5.

2 Classical functional principal component analysis

Functional PCA is often the first step of the functional data analysis (FDA), followed by further FDA techniques. Thus, it is important to find those principal components functions that contain most of the information.

Consider a sample of functional data, denoted by $x_1(\cdot), \ldots, x_n(\cdot)$, observed at $t_1, \ldots, t_p$. For convenience and without loss of generality, we assume for this part that the $x_i(t)$ are subtracted by the sample mean function. The functional PCA problem is known as the following eigenequation

$$T_n f := \int_{\mathcal{T}} K_n(s,t) f(t) dt = \lambda f(s), \tag{1}$$

where K_n is the sample covariance function given by $K_n(s,t) = n^{-1} \sum_{i=1}^{n} x_i(s) x_i(t)$. The resulting eigenfunctions, denoted by f_j, are principal components and the principal component scores are given by

$$Z_{ij} = \int_{\mathcal{T}} x_i(t) f_j(t) dt. \tag{2}$$

In the classical FPCA approach, the first component function corresponds to the most important mode of variation that accounts for the highest percent of the total variation. The second component function, which must be orthogonal to the first one, corresponds to the second most important mode of variation. Continuing in this way produces all the principal components functions, which correspond to the eigenfunctions of the empirical covariance function.

3 Robust functional principal component analysis

Since the empirical covariance function is sensitive towards outlying points, the first component function and corresponding PC scores Z_{ij} in (2) would be often attracted towards outlying points, and may not capture the variation of the regular functional observations. Consequently, data reduction based on FPCA becomes unreliable in the presence of functional outliers. Therefore, the goal of this section is to develop functional principal component analysis that promises to be robust and efficient, and to employ this method for outlier detection in a functional data.

Let us first approximate each functional object by a linear combination of K known basis functions ϕ_k,

$$x_i(t) \doteq \sum_{k=1}^{K} c_{ik} \phi_k(t)$$

for sufficiently large K, but less than p. The simultaneous expansion of all n curves is expressed in matrix notation as

$$\mathbf{x}(t) = \mathbf{C}\boldsymbol{\phi},$$

where $\mathbf{x}$ is a vector-valued function with $x_i, 1 \leq i \leq n$, as its components, $\boldsymbol{\phi}$ is a vector-valued function having components $\phi_1, \ldots, \phi_K$, and $\mathbf{C} = \{c_{ij}\}_{i=1:n, j=1:K}$ is the coefficient matrix. Also, any eigenfunction of T_n can be represented as

$$f(s) = \sum_{k=1}^{K} b_k \phi_k(s)$$

and in matrix notation, $\mathbf{f}(s) = \boldsymbol{\phi}(s)^T\mathbf{b}$. Thus, the eigenequation (1) is simplified as

$$n^{-1}\mathbf{C}^T\mathbf{CWb} = \lambda\mathbf{b},$$

where $\mathbf{W} = \left\{\int_{\mathcal{T}} \phi_k(t)\phi_{k'}(t)dt\right\}_{k,k'=1:K}$. If the basis function is orthonormal then $\mathbf{W} = \mathbf{I}$. The functional PCA problem reduces to the standard multivariate PCA of the coefficient matrix $\mathbf{C}$ (Ramsay and Silverman 2005).

Instead of dealing with classical FPCA we consider to apply classical PCA in multivariate data analysis on $\mathbf{C}$ which would provide the equivalent information about the structure of the covariance function of $x(t)$. In addition, the classical PCA method for multivariate data can also be used to identify functional outliers in a functional dataset by implementing the classical PCA on the coefficient matrix, $\mathbf{C}$.

Because of the sensitivity of the classical PCA to outliers, the use of classical PCA would provide misleading results. Therefore, we consider the use of robust PCA approaches developed for multivariate data on the coefficient matrix, $\mathbf{C}$. The first study by Locantore et al. (1999) used the idea of spherical (SPHER) and elliptical (ELL) PCA to explore abnormalities in the curvature of the cornea in the human eye as a robust technique. However Hubert et al. (2005) showed that these methods are influenced by outliers when the data are high-dimensional or when there is a large percentage of contamination. Therefore, we propose to use some other robust PCA methods (ROBPCA by Hubert et al. 2005 and BACONPCA by Billor et al. 2005) on $\mathbf{C}$, which are known to have high breakdown points and computationally more efficient than the SPHER and ELL PCA methods.

The ROBPCA method (Hubert et al. 2005) combines ideas of both projection pursuit and robust covariance estimation based on Minimum Covariance Determinant (MCD) method (Rousseeuw and Van Driessen 1999) which is based on seeking h-subset whose classical covariance matrix has the smallest determinant. Assume that the original data matrix $\mathbf{C}_{n\times K}$, where n and K denote the number of objects and the original number of variables, respectively. This method has three important steps. In the first step, the data space is reduced to the affine subspace spanned by n observations so that we work in the space spanned by the k_0 columns which equals the rank of the mean-centered data matrix. The first step is especially useful when $K \geq n$, but even $K < n$ (Hubert et al. 2005). Second step focuses on finding the $h(< n)$ *least outlying* data points. By default, h is about $0.75\,n$ or chosen by the user where $n - h$ greater than the number of outliers in the dataset and h should be greater than $[(n + k_0 + 1)/2]$. If the data matrix is low dimensional, $n > K$, classical PCA method is applied to MCD based covariance matrix to obtain the $k < k_0$ robust principal components. If the data matrix is high dimensional, $n < K$, the MCD method can not be used directly due to singularity of the covariance matrix of any h-subset that results in *zero determinant*. Therefore Hubert et al. (2005) proposed to obtain the h *least outlying* data points by projecting the high-dimensional data points on many univariate directions v. On every direction, a robust center and scale of the projected data points, $c_i'v$, are computed, namely the univariate MCD of location, $\hat{\mu}_r$, and scale, $\hat{\sigma}_r$. Next, for every data point its standardized distance to that center is measured. Finally, for each data point its largest distance over all the directions is considered. This yields the outlyingness measure(Stahel 1981; Donoho 1982)

$$Out(c_i) = max_v \frac{| c_i' v - \hat{\mu}_r |}{\hat{\sigma}_r}, \quad i = 1, \ldots, n.$$

The h data points with smallest outlyingness are then retained and from the covariance matrix of this h-subset, the number of principal components to retain, k, is selected and then the data points are projected on the subspace spanned by the first k eigenvectors of the sample covariance matrix of h-subset. Third step consists of robustly estimating the covariance matrix of the mean-centered $\mathbf{C}^*_{n\times k}$ obtained in the second step using the MCD estimator and applying PCA on to this.

The BACONPCA method (Billor et al. 2005) also combines ideas of both projection pursuit and robust covariance estimation based on BACON (blocked adaptive computationally efficient outlier nominations) instead of MCD method. BACON method proposed by Billor et al. (2000) which was originally defined for low dimensional data requires very few steps regardless of sample size and comparable to previously published high-breakdown outlier detection methods, but obtained only in 4 or 5 iterations for datasets of sizes from 100 to 10,000 and dimensions from 5 to 20.

BACON algorithm consists of the following steps:

Step 1: Identify an initial basic subset of size $m > K$ of either the smallest Mahalanobis distances (affine equivariant, but not robust) or the smallest distances from the medians (robust but not affine equivariant) that can be safely be assumed free of outliers.

Step 2: Compute the discrepancies

$$d_i = \sqrt{(c_i - \bar{c}_b)' S_b^{-1} (c_i - \bar{c}_b)}, \quad i = 1, \ldots, n$$

where $\bar{c}_b$ and S_b are the mean and sample covariance matrix of the observations in the basic subset.

Step 3: Set the new basic subset to all points with discrepancy less than a correction factor based on chi square distribution with K degrees of freedom (Billor et al. 2000)

Step 4: The *stopping rule*: Iterate Steps 2 and 3 until the size of the basic subset no longer changes.

Step 5: Nominate the observations excluded by the final basic subset as outliers.

In BACONPCA, the data space is also reduced to $k_0 (< K)$ as in the first step in ROBPCA. Second step focuses on finding a *basic subset* of clean data points. If the data matrix is low dimensional, classical PCA method is applied to BACON based covariance matrix (S_b) to obtain the $k (< k_0)$ robust principal components. If the data matrix is high dimensional, original BACON method given above can not be used since the inverse of the covariance matrix based on the basic subset does not exist. This problem can be overcome similar to ROBPCA method by obtaining h *least outlying* data

points by projecting the high-dimensional data points on many univariate directions v. BACON based robust center, $\hat{\mu}_b$, and scale, $\hat{\sigma}_b$, of the projected data points $c_i'v$ are computed for every direction, v. For each data point, its largest standardized distance to its mean over all the directions is considered. This yields the outlyingness measure similar to the one given in ROBPCA. The h (chosen as default $= 0.75\,n$ or a value greater than $[(n + k_0 + 1)/2]$) data points with smallest outlyingness are then retained and from the covariance matrix of this final basic subset, the number of principal components to retain, k, is selected and then the data points are projected on the subspace spanned by the first k eigenvectors of the sample covariance matrix of the final basic subset obtained from the BACON algorithm. At the final stage, we robustly estimate the covariance matrix of the mean-centered $\mathbf{C}^*_{n\times k}$ obtained in the second step using the BACON method and applying PCA on to this. The first $k_1 (< k)$ eigenvectors of the BACON based covariance matrix, sorted in descending order of the eigenvalues, then yield robust loadings. Especially for large data sets (i.e. large n), MCD method requires more computational time than BACON does since MCD depends on resampling.

By using robust PCA methods described above, robust scores can be obtained from $\mathbf{Z} = \mathbf{CV}$ with $\mathbf{V}$ being a $K \times k_1$ matrix consisting of $k_1 (\leq K)$ robust eigenvectors. Then we can use robust PC scores, $z_i = V'(c_i - \hat{\mu}), i = 1, \ldots, n$ where $\hat{\mu}$ is a robust estimate (MCD or BACON) of the location parameter vector, for detection of outliers.

A diagnostic plot, called *orthogonal-score* plot (Hubert et al. 2005) is a very useful tool to summarize the results of the PCA analysis and to classify observations as *regular*, *good leverage*, *orthogonal outlier* and *bad leverage* in a multivariate dataset. In the context of PCA, a multivariate outlier is an observation which either lies far from the subspace spanned by the k_1 eigenvectors, and/or that the projected observation lies far from the bulk of the data within this subspace. To measure this degree of outlyingness, two distances are proposed: robust score distance and orthogonal distance (Hubert et al. 2005). The robust score distance is defined as:

$$Sd_i = \sqrt{\sum_{j=1}^{k} z_{ij}^2/\lambda_j}, \quad i = 1, \ldots, n,$$

where z_{ij} are the scores and λ_j are the eigenvalues. The orthogonal distance which measures the distance between an observation c_i and its projection in the k_1-dimensional PCA-subspace, Od_i, is given by:

$$Od_i = \|c_i - \hat{\mu} - \mathbf{V}z_i\|, \quad i = 1, \ldots, n,$$

where $\mathbf{V}$ is the $K \times k_1$ matrix of eigenvectors and z_i' is the ith row of the score matrix, $\mathbf{Z}$. So, the *orthogonal-score* plot is a scatter plot of the orthogonal distance Od_i versus the robust score distance Sd_i which classifies observations as *regular*, *good leverage*, *orthogonal outlier* and *bad leverage* for a multivariate data (Hubert et al. 2005). Bad leverage observation where Sd_i and Od_i are both large, means that this observation is far away from the regular observations and the PCA space. We use the cut-off values for Sd_i and Od_i on this plot suggested by Hubert et al. (2005) to flag outliers.

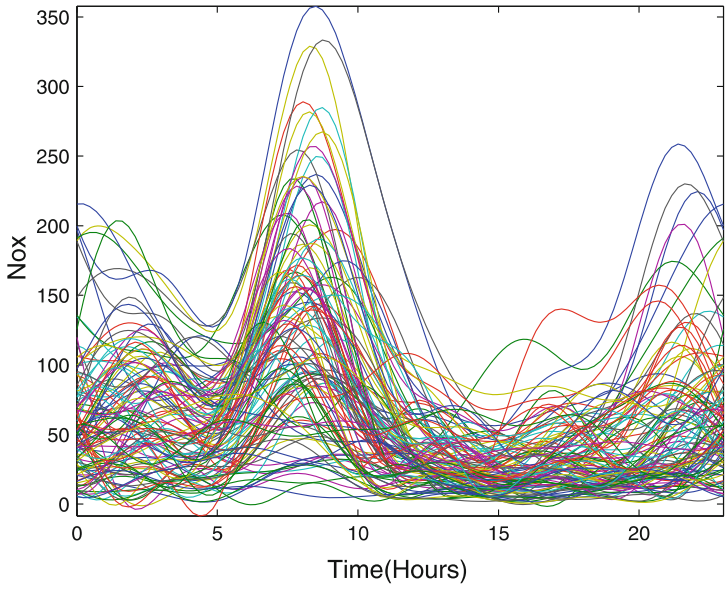

Fig. 1 Sample curves of NO_x data by using Fourier basis

In this study, the *orthogonal-score* plot based on the robust PC scores obtained from **C** coefficient matrix is also very useful tool to detect functional outliers (magnitude or shape) which correspond to the curves that deviate from the remaining curves. Numerical examples section illustrates how the multivariate outliers highlighted on this diagnostic plot can be associated with functional outliers.

4 Numerical examples

4.1 Low dimension: NO_x data

The aim of our analysis is to illustrate the performance of the robust FPCA on the NO_x data, which was used by Febrero et al. (2007, 2008). The dataset on NO_x emission levels, which comes from a control station near a power plant in Barcelona in year 2005, consists of 115 days of hourly measurements of NO_x levels ($\mu g/m^3$) for the period February 23, 2005 to June 26, 2005. Only 115 days NO_x levels are available due to missing observations problem for several consecutive hours of some days. To apply functional data analysis the discrete trajectories were approximated with the help of basis functions in order to get curves. Due to periodic nature of NO_x emission levels Fourier basis functions are used. The optimal number of bases based on generalized cross validation (GCV) method (Craven and Wahba 1979) is obtained as 10 (Fig. 2). The dataset of NO_x emission levels based on 10 Fourier bases is displayed in Fig. 1. Figure 1 depicts that NO_x levels increase in the morning and reach peak value around 8:00 a.m., then decrease until 2:00 p.m., and again increase in the evening. Moreover, we observe from Fig. 1 that the group of curves shows presence of a few trajectories that are in some way different from the rest.

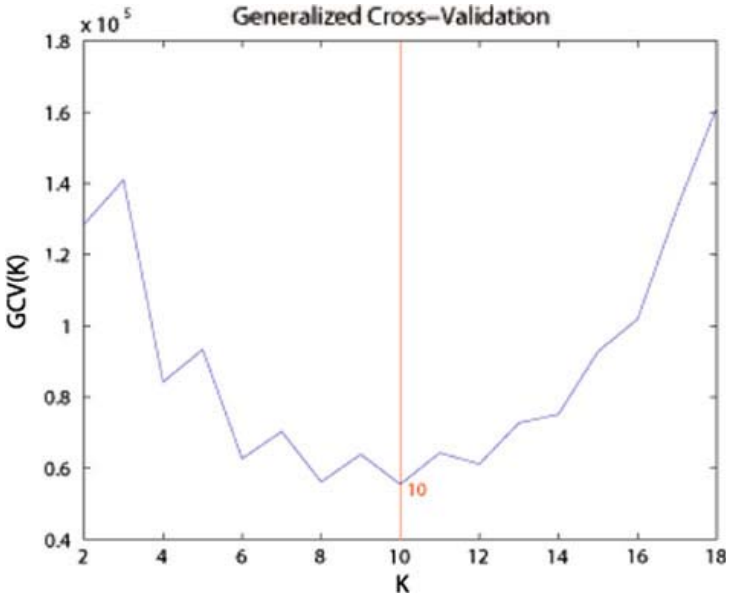

Fig. 2 Generalized cross validation by using Fourier basis for NO_x dataset

Table 1 Outliers detected by three PCA methods (.) denotes case number for NO_x dataset

Fourier basis		
CPCA	ROBPCA	BACONPCA
05/02(59)	03/18(22)	03/18(22)
	04/29(56)	04/29(56)
	03/11(16)	03/11(16)
	05/02(59)	05/02(59)
	03/09(15)	03/09(15)

We apply Classical Principal Component (CPCA), ROBPCA and BACONPCA on coefficient matrix, $\mathbf{C}$ of size 115×10. Five principal components, yielding explanation percentage more than 90%, were retained for the three PCA methods.

Figure 3 exhibits the results of the diagnostic plots for the three PCA methods by using Fourier bases. All bad leverage points, detected by these diagnostic plots (*orthogonal-score* plots) formed by using the three PCA methods are listed in Table 1. It is clear from Table 1 that both ROBPCA and BACONPCA detected similar outliers in NO_x dataset. These results also agree with the results obtained by Febrero et al. (2007).

However, we have detected one additional outlier (03/09) by using both ROBPCA and BACONPCA methods. The *orthogonal-score* plot based on CPCA detects only one bad leverage point, which clearly indicates that CPCA fails to detect the other outliers since the first components are affected badly by the existence of other outliers.

The highlighted bad-leverage outliers on the *orthogonal-score* plot formed based on robust PCA methods can be also observed in Fig. 4. It is clear that the bad-leverage points correspond to the functional observations that deviate from the remaining functional observations regardless of the type of functional outliers, that is, magnitude or shape.

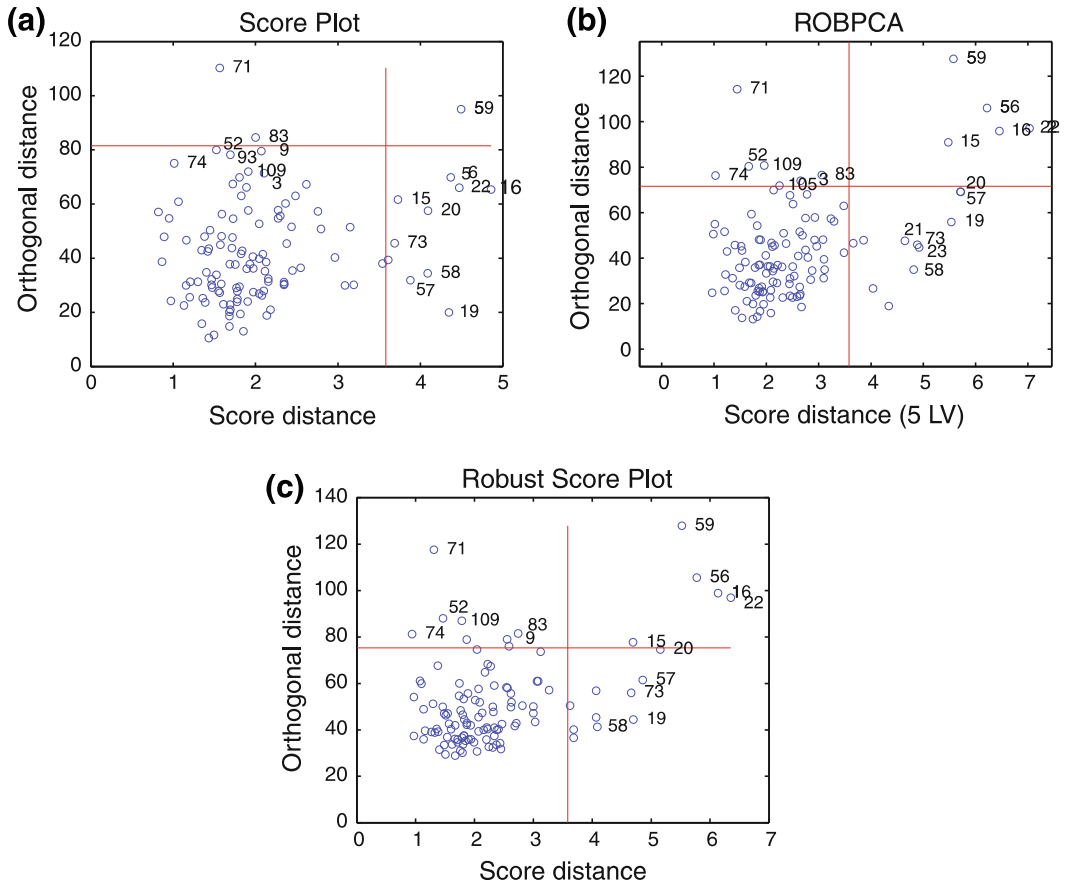

Fig. 3 Orthogonal-score plot for NO_x dataset by using Fourier basis computed with **a** CPCA, **b** ROBPCA, **c** BACONPCA

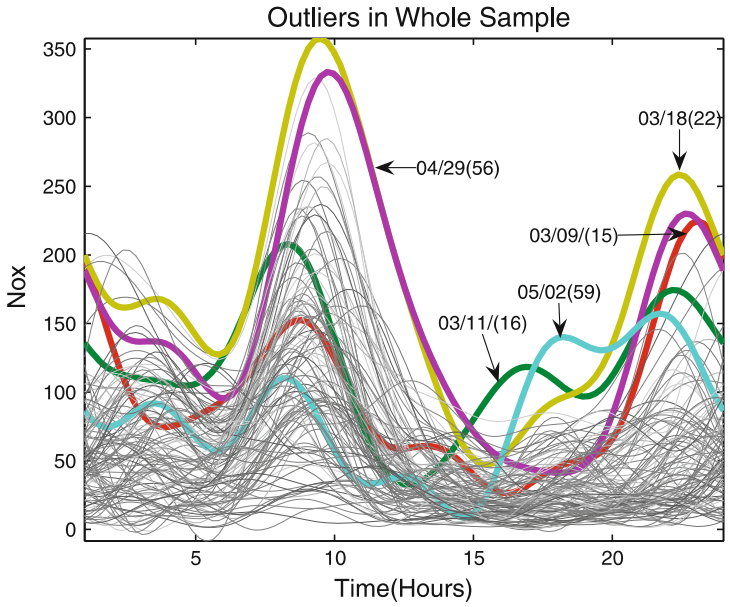

Fig. 4 Outliers detected by proposed method for NO_x dataset

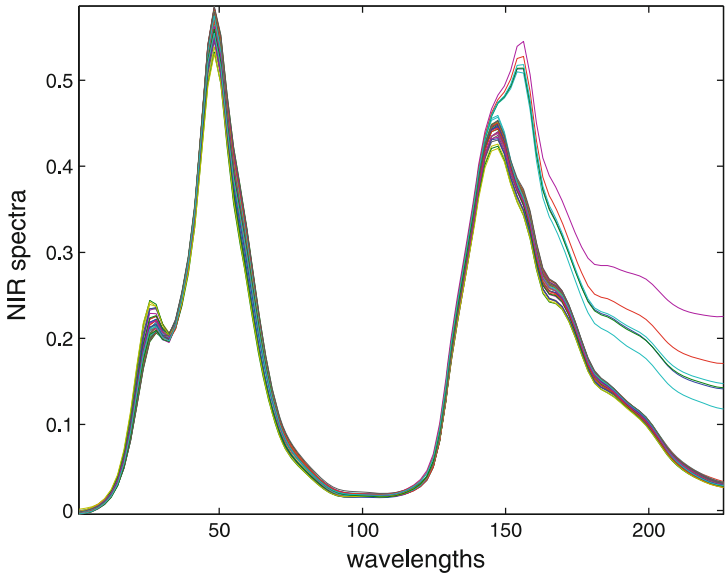

Fig. 5 Octane dataset by using B-spline basis

After detecting outliers, we checked for sources for abnormal values of these curves. A possible explanation of these abnormally large NO_x emissions on these particular days is provided. We found that Friday, March 11 is the beginning of a weekend. The Friday, March 18 and Saturday, March 19 are the beginning of the Eastern vacation in Spain in the year 2005. Also Friday, April 29, Saturday, April 30, Sunday, May 1, and Monday, May 2 correspond to long weekend. There is sudden increase in traffic on these small vacation periods. So we conclude that abnormal observations on specific days can be attributed to increase in traffic due to small vacation periods. We have also detected outlier on Wednesday, March 9. It is observed that high NO_x emissions are recorded on March 9 after 8:00 pm. Since the observation on March 10 is missing and thus not included in analysis, we could not pinpoint the reason behind this abnormal observation on March 9.

4.2 High dimension: octane data

In this section, robust and classical FPCA methods are applied on the octane dataset (Esbensen et al. 1994) that consists of near-infrared (NIR) absorbance spectra over 226 wavelengths of 39 gasoline samples with certain octane numbers.

Since data are non-periodic in nature, B-spline bases are used to convert discrete data to functional form. It is known that six of the samples (25, 26 and 36–39) contain added alcohol. The octane data are displayed in Fig. 5 which shows presence of a few functional outliers.

Initial dimension of the dataset is 39×226. Optimal number of bases based on GCV is obtained as $K = 136$. Data curves for octane dataset using 136 B-spline bases are depicted in Fig. 5. We apply CPCA, ROBPCA and BACONPCA on coefficient matrix, $\mathbf{C}$. Two principal components were retained, respectively each for CPCA, ROBPCA and BACONPCA, yielding a classical and robust explanation percentage more than 90%.

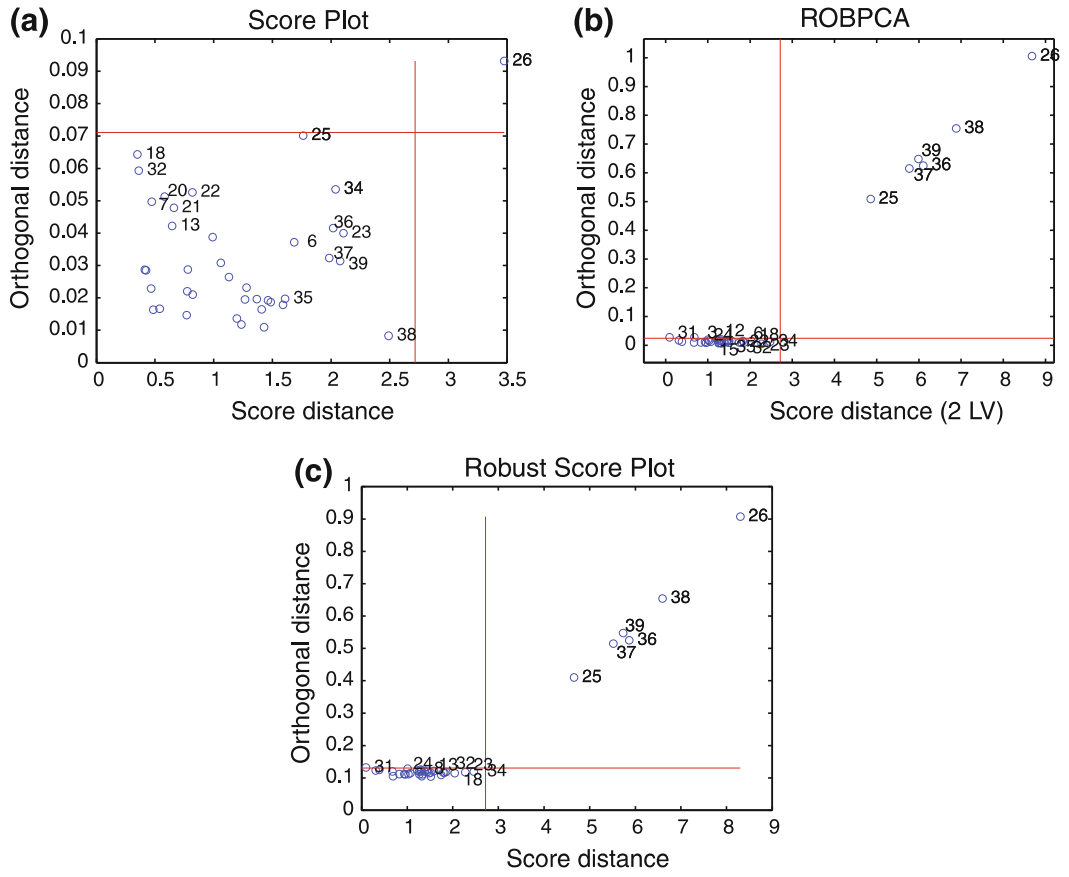

Fig. 6 Orthogonal-score plot for Octane dataset by using B-spline basis computed with **a** CPCA, **b** ROBPCA, **c** BACONPCA

The resulting diagnostic plots for the three PCA methods in Fig. 6 reveal that both ROBPCA and BACONPCA detected similar outliers (cases 25, 26 and 36-39). The orthogonal score plot based on CPCA detects only one bad leverage point (case 26). Figure 7 highlights the functional outliers identified by using the proposed methodology for the octane dataset. Moreover, we can clearly see that some part of the outliers is away from the rest of the data (Fig. 7) which is an example of partial contamination.

4.3 Simulation

The simulation study is conducted to compare the performance of ROBPCA and BACONPCA with CPCA on coefficient matrix. The simulation setting given by Fraiman and Muniz (2001), with few changes, is used here. For simulation we consider functional data $x_1, \ldots, x_n$ obtained as realizations from a stochastic process $X(\cdot)$. This functional dataset has continuous paths on [0,1]. Curves are generated from different models. Model 1 was generated without contamination and several other models were generated with different types of contaminations.

Model 1 (no contamination): $x_i(t) = g(t) + e_i(t)$, $1 \leq i \leq n$, where model error term $e_i(t)$ is a stochastic Gaussian process with zero mean and covariance function $\vartheta(s,t) = (1/2)(1/2)^{(0.9)|t-s|}$ and $g(t) = 4t$, with $t \in [0, 1]$.

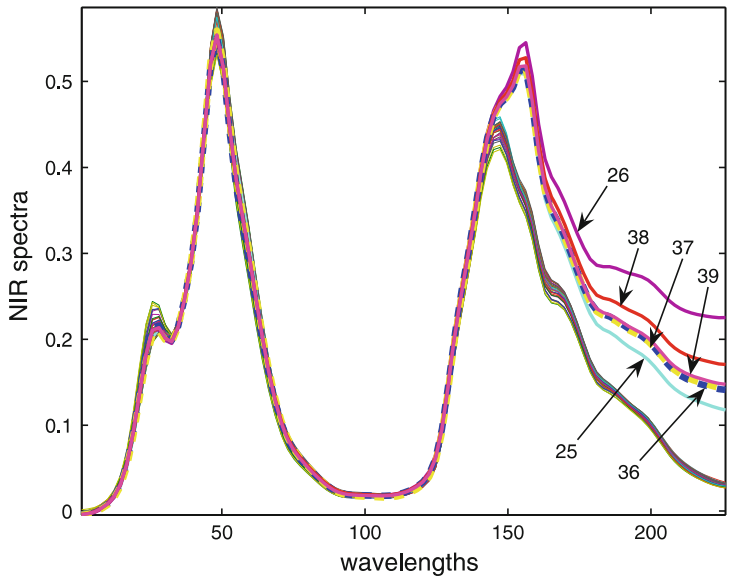

Fig. 7 Outliers detected by proposed method for Octane dataset

Model 2 (asymmetric contamination): $Y_i(t) = x_i(t) + c_i M, 1 \leq i \leq n$, where c_i is 1 with probability q and 0 with probability $1 - q$; M is the contamination size constant.
Model 3 (symmetric contamination): $Y_i(t) = x_i(t) + c_i \sigma_i M, 1 \leq i \leq n$, where c_i and M are defined as in model 2 and σ_i is a sequence of random variables independent of c_i taking values 1 and -1 with probability 1/2.
Model 4 (partially contaminated): $Y_i(t) = x_i(t) + c_i \sigma_i M$, if $t \geq T_i, 1 \leq i \leq n$, and $Y_i(t) = x_i(t)$, if $t < T_i$, where T_i is a random number generated from a uniform distribution on [0, 1].
Model 5 (Peak contamination): $Y_i(t) = x_i(t) + c_i \sigma_i M$, if $T_i \leq t \leq T_i + \ell, 1 \leq i \leq n$, and $Y_i(t) = x_i(t)$, if $t \notin [T_i, T_i + \ell]$, where $\ell = 2/30$ and T_i is a random number from a uniform distribution in $[0, 1 - \ell]$.

Figure 8 exhibits curves simulated from these five models. For each model, we generated 100 replications, with one setting each for low and high dimensional data. For low dimensional data we consider $n = 100$, $p = 12$ setting and for high dimensional data we consider setting with $n = 50$, $p = 200$. For the model 1 contamination percent is $q = 0$ and contamination constant is $M = 0$. For each contaminated model (2, 3, 4 and 5) we considered several levels of contamination: $q = 5, 10, 15$ percentage and contamination constants $M = 10$ and 25. The number of basis used for functional curves simulated from these five models are obtained from GCV method. Classical FPCA and robust FPCA based on ROBPCA and BACONPCA are used on the simulated functional data according to the five models.

GCV method finds the same number of basis for contaminated and uncontaminated data for symmetric and asymmetric models. For partial and peak contamination cases GCV method finds different number of basis for contaminated and uncontaminated models. The reason for this is the shape or pattern of the contaminated curves is different than the uncontaminated data.

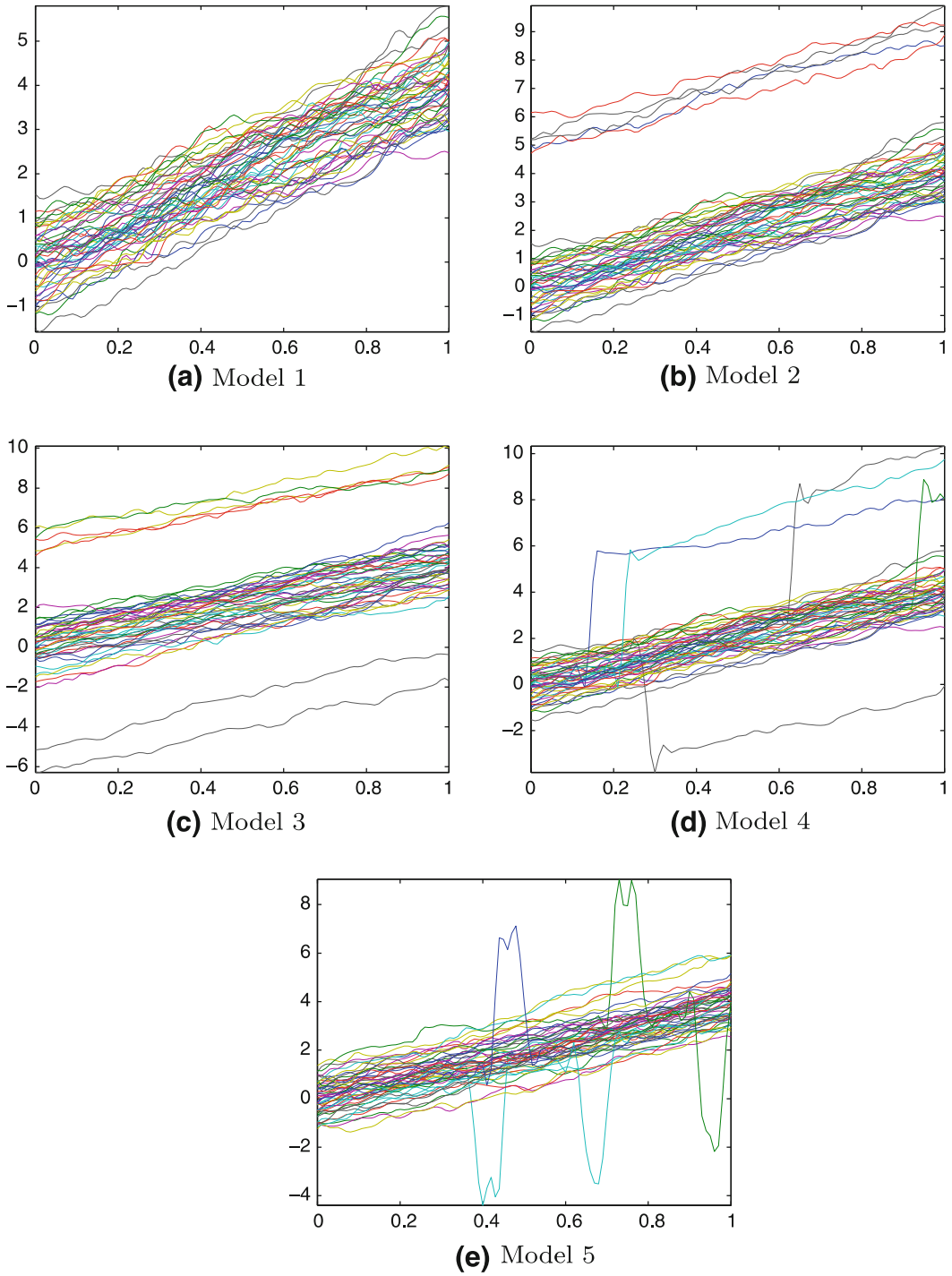

Fig. 8 Curves generated from model 1 (without contamination), model 2 (asymmetric contamination), model 3 (symmetric contamination), model 4 (partial contamination) and model 5 (peak contamination) with $n = 50$, $p = 100$, $M = 10$ and $q = 0.1$

Two quantitative measures of the goodness of the methods are considered. The first one is mean proportion of variability (MPV):

$$MPV = 1/N \sum_{m=1}^{N} \frac{\hat{\lambda}_1^m + \hat{\lambda}_2^m + \cdots + \hat{\lambda}_k^m}{\lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_p},$$

where N denotes the number of iterations, λ_j is the true jth eigenvalue of the covariance function, $\hat{\lambda}_j^m$ is the estimated value of λ_j at the mth replication. The $\hat{\lambda}_j^m$ is obtained by using classical or robust multivariate techniques on coefficient matrix of contaminated or uncontaminated model. For each setting, the optimal value for the mean proportion of explained variability is taken as 90% which corresponds to $k = 2$.

The second quantitative measure is the norm, that is, the square root of sum of squared error of $\hat{\lambda}_1$ given by $\sqrt{\sum_{m=1}^{N}(\hat{\lambda}_1^{(m)} - \lambda_1)^2}$, where λ_1 is the largest true eigenvalue of the covariance function and $\hat{\lambda}_1$ is the estimated largest eigenvalue of λ_1 at the mth replication. The $\hat{\lambda}_1^m$ is obtained by using classical or robust multivariate techniques on coefficient matrix of contaminated or uncontaminated model. The optimal value is zero or near zero.

Model 1 is compared with models 2, 3, 4 and 5. The simulation results of mean proportion of variability for four comparisons are given in Tables 2, 3, 4 and 5. It is clear that CPCA provides the best mean proportion of explained variability when there is no contamination in the data, which is expected. For the uncontaminated data robust methods also yield comparable results. However, when contamination is introduced to the data (models 2–5) the eigenvalues obtained with CPCA are overestimated. Since estimated percentages of MPV are larger than 100%. In ROBPCA and BACONPCA we obtain MPV of 90% for low dimensional data without and with contamination. For high dimensional data the mean percentage of explained variability is similarly 90% for without and with contamination. The main reason behind this is the optimal direction obtained by ROBPCA and BACONPCA are robust to outliers. CPCA clearly fails and provides the worst possible result because mean proportion of variability is above 100%. It is clear from Tables 2, 3, 4 and 5 that the MPV for ROBPCA at 15%

Table 2 Simulation results of the mean proportion of explained variability for no contamination (0%) and asymmetric contamination (5, 10 and 15%) for high and low dimensional cases

High dimension: $n = 50$, $p = 200$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.915	0.763	0.894	0.928	0.788	0.899
5	11.222	0.811	0.896	54.705	0.824	0.896
10	19.596	0.852	0.890	115.796	0.852	0.904
15	27.274	4.862	0.886	162.757	9.157	0.893
Low dimension: $n = 100$, $p = 12$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.970	0.875	0.934	0.945	0.856	0.908
5	9.783	0.891	0.932	57.942	0.875	0.913
10	19.316	0.919	0.936	110.982	0.893	0.908
15	26.071	1.310	0.937	159.057	0.912	0.909

Table 3 Simulation results of the mean proportion of explained variability for no contamination (0%) and symmetric contamination (5, 10 and 15%) for high and low dimensional cases

High dimension: $n = 50$, $p = 200$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.918	0.780	0.896	0.932	0.781	0.906
5	10.819	0.815	0.891	61.526	0.820	0.917
10	24.310	0.856	0.883	123.610	0.842	0.907
15	31.037	0.880	0.887	199.793	15.607	0.906
Low dimension: $n = 100$, $p = 12$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.957	0.867	0.919	0.946	0.844	0.908
5	10.740	0.877	0.921	62.115	0.867	0.909
10	21.128	0.892	0.953	126.060	0.887	0.909
15	30.824	0.928	0.928	186.967	0.911	0.921

Table 4 Simulation results of the mean proportion of explained variability for no contamination (0%) and partial contamination (5, 10 and 15%) for high and low dimensional cases

High dimension: $n = 50$, $p = 200$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.923	0.763	0.889	0.910	0.766	0.884
5	4.973	0.796	0.891	30.493	0.808	0.889
10	11.111	0.837	0.894	53.618	0.833	0.896
15	15.822	0.904	0.874	94.105	3.347	0.888
Low dimension: $n = 100$, $p = 12$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.941	0.857	0.908	0.958	0.865	0.918
5	6.493	0.892	0.924	32.535	0.858	0.912
10	10.754	0.891	0.909	66.224	0.895	0.916
15	15.693	0.904	0.903	98.966	0.918	0.918

contamination level is above 100% in few of the cases. BACONPCA gives better results than ROBPCA at 15% for both low and high dimensional case. From these results we can deduce that BACONPCA and ROBPCA outperform the CPCA.

Simulation results for the norm with $N = 100$ iterations and different contamination levels for comparison of model 1 versus model 2 are summarized in Fig. 9.

Table 5 Simulation results of the mean proportion of explained variability for no contamination (0%) and peak contamination (5, 10 and 15%) for high and low dimensional cases

High dimension: $n = 50$, $p = 200$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.914	0.771	0.889	0.892	0.738	0.865
5	2.280	0.831	0.883	10.300	0.769	0.853
10	3.048	0.897	0.890	16.264	0.819	0.867
15	3.686	0.934	0.873	20.717	0.972	0.867
Low dimension: $n = 100$, $p = 12$						
Contamination (%)	$M = 10$			$M = 25$		
	CPCA	ROBPCA	BACONPCA	CPCA	ROBPCA	BACONPCA
0	0.954	0.863	0.914	0.941	0.868	0.910
5	3.015	0.885	0.936	16.965	0.900	0.948
10	4.383	0.870	0.905	26.551	0.870	0.887
15	6.278	0.899	0.912	36.726	0.887	0.896

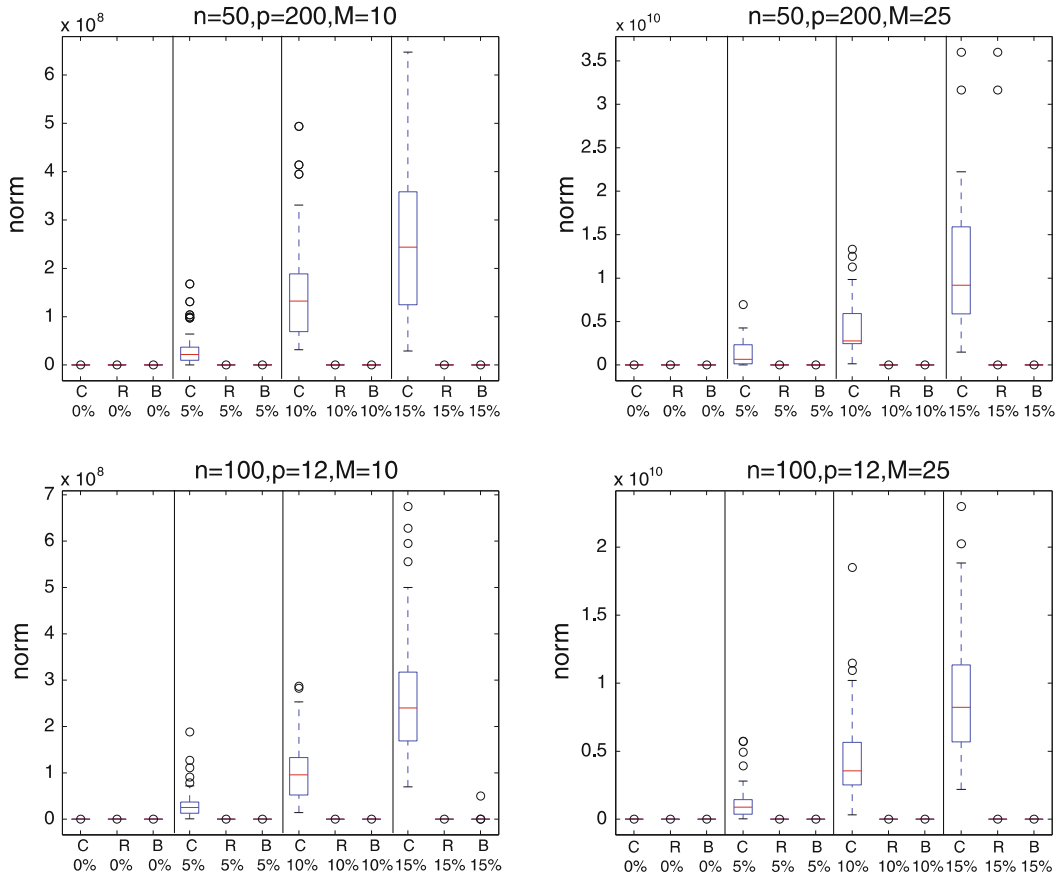

Fig. 9 *Boxplots* of norm for no contamination (0%) and symmetric contamination (5, 10 and 15) for high and low dimensional cases for CPCA (*C*) ROBPCA (*R*) and BACONPCA (*B*)

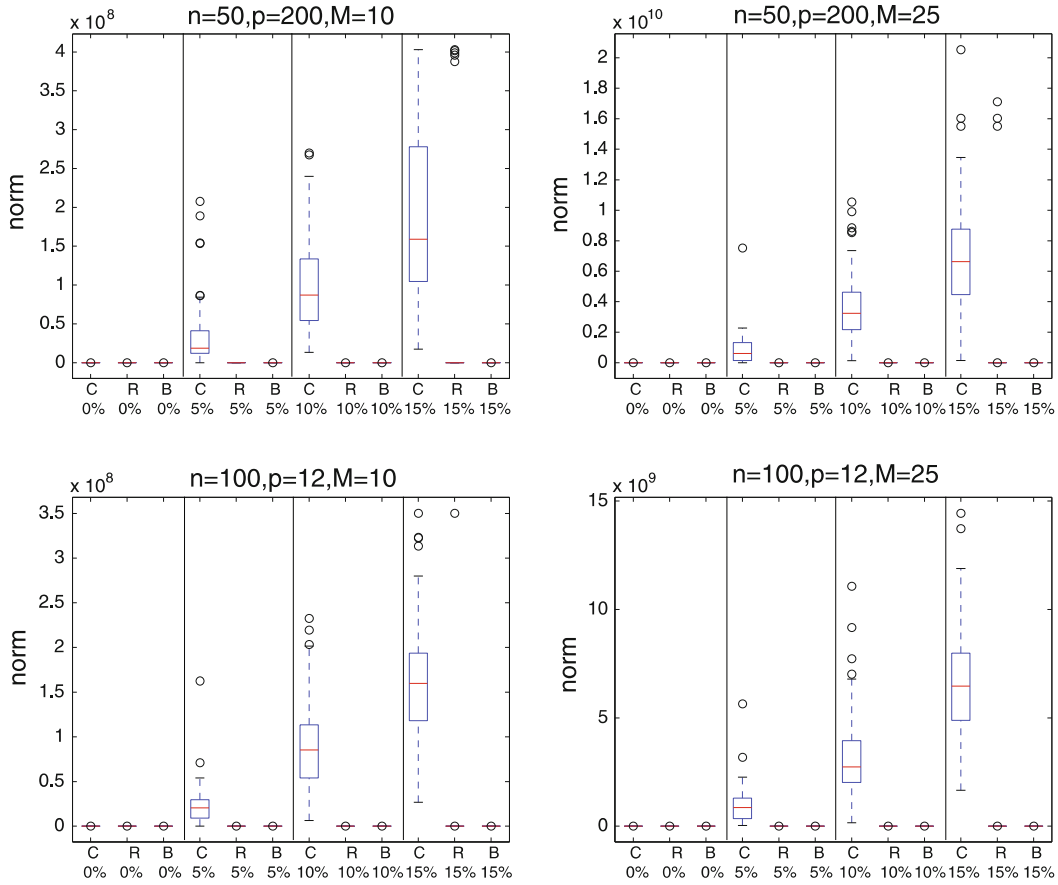

Fig. 10 *Boxplots* of norm for no contamination (0%) and asymmetric contamination (5, 10 and 15%) for high and low dimensional cases for CPCA (*C*) ROBPCA (*R*) and BACONPCA (*B*)

For this comparison, we used one high and one low dimensional settings with the two values of M (10 and 25). The ideal value of norm must be very small or near zero. We conclude that the norm is near zero when there is no contamination for all methods. This is an indication of ROBPCA and BACONPCA being also effective methods for uncontaminated data. The norm based on CPCA tends to increase as contamination level increases. For contaminated data, norms corresponding to ROBPCA and BACONPCA method yield minimum value which is near zero for high and low dimensional settings. The comparisons of model 1 versus models 3–5 for low and high dimensional settings yielded very similar results observed in Fig. 9, which are given in Figs. 10, 11 and 12. The only difference that we have in Fig. 11 is that the norm based on CPCA tends to remain small as contamination level increases for low dimensional data with partial contamination. We also observe the similar trend for the norm based on CPCA for both low and high dimensional cases with peak contamination (Fig. 12).

5 Conclusion and discussion

Robust FPCA methods based on multivariate robust PCA methods (ROBPCA and BACONPCA) have been developed for finding the components that contain the most

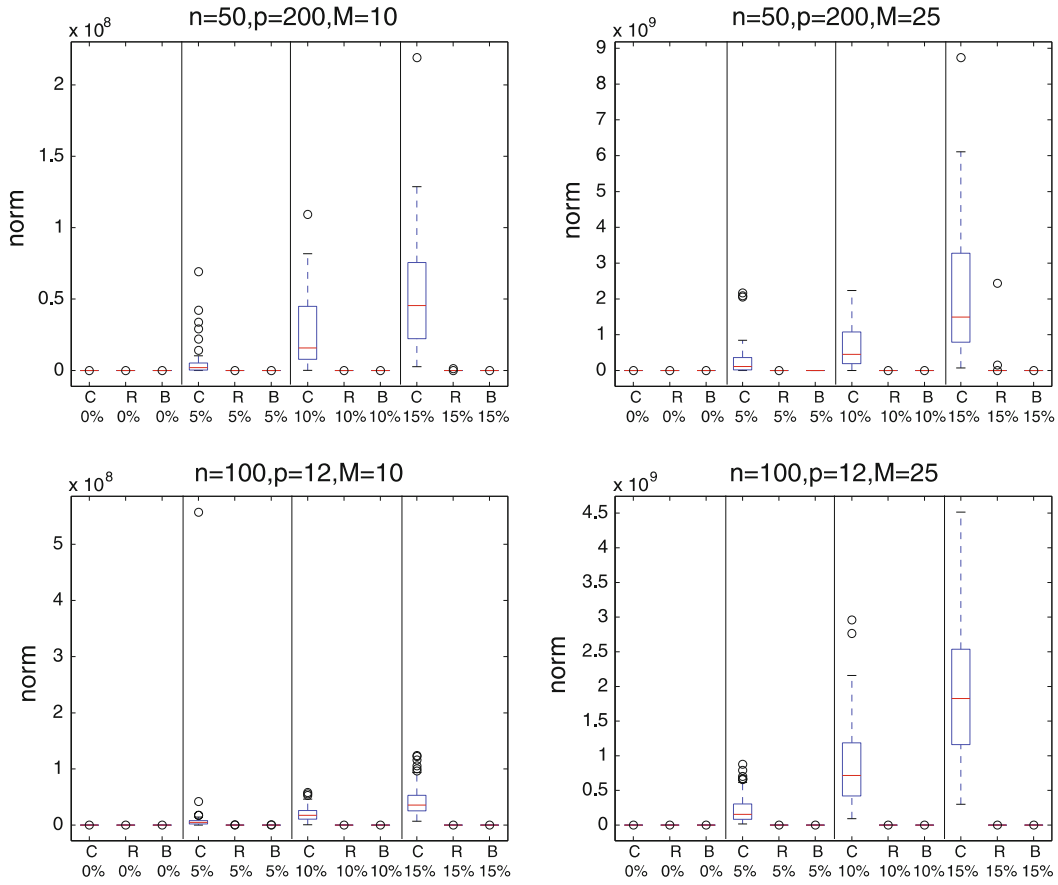

Fig. 11 *Boxplots* of norm for no contamination (0%) and partial contamination (5, 10 and 15%) for high and low dimensional cases for CPCA (*C*) ROBPCA (*R*) and BACONPCA (*B*)

of the information available in a functional data even if there are outliers and for detecting functional outliers.

An extensive simulation study was conducted and two real datasets from chemometrics and environment were employed to assess the performance of the robust FPCA. From the simulation study considering different contamination configurations (symmetric, asymmetric, partial and peak), we found that robust FPCA based on ROBPCA and BACONPCA successfully detects functional outliers and provides the components that provide the most of the variation in the existence of outliers.

We believe that this is the first paper focusing on detection of functional outliers via multivariate approaches. We also observe that the outliers that may arise in a functional dataset correspond to the bad leverages in a multivariate dataset.

The choices of basis function and the number of basis function, K, are important issues in functional data analysis. There are many possible choices for the functional basis, for example, splines, Fourier, wavelets. To choose the number of basis, K, several criteria for choosing the smoothing parameters are possibly adopted by summing those criteria across n curves. In this paper, we used the GCV criterion given by $\sum_{i=1}^{n} GCV_i(K)$. Note that GCV is not affected by magnitude outliers, but is affected by shape outliers because GCV approximates leave-one-out cross-validation of prediction error for each curve. Although we have experienced not much consequence of

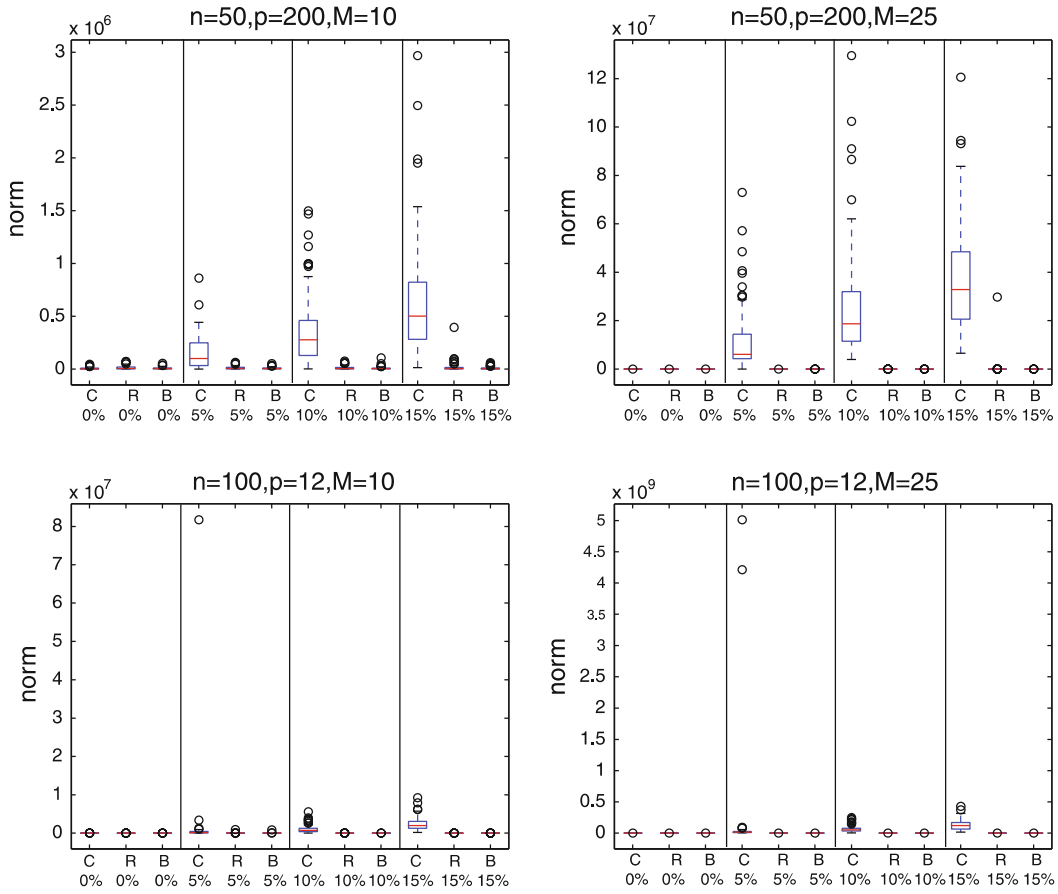

Fig. 12 *Boxplots* of norm for no contamination (0%) and peak contamination (5, 10 and 15%) for high and low dimensional cases for CPCA (*C*) ROBPCA (*R*) and BACONPCA (*B*)

this shortage of GCV on our simulation results for the shape contaminations (partial and peak contamination) in sense of the estimation and outlier detection we consider that this would require further investigation.

The MATLAB codes for *robpca* have been used for this study from LIBRA (The Matlab Library for Robust Analysis by Verboven and Hubert 2004). We have also used the Matlab codes for functional data analysis provided by Ramsay and Silverman (2001).

References

Billor N, Hadi AS, Velleman PF (2000) BACON: blocked adaptive computationally efficient outlier nominators. Comput Stat Data Anal 34:279–298

Billor N, Kiral G, Turkmen A (2005) Outlier detection using principal components. In: Twelfth international conference on statistics, combinatorics, mathematics and applications, Auburn (unpublished manuscript)

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31:377–403

Donoho DL (1982) Breakdown properties of multivariate location estimators. Ph.D. Qualifying paper, Harvard University

Esbensen KH, Schüonkopf S, Midtgaard T (1994) Multivariate analysis in practice. Camo, Trondheim

Febrero M, Galeano P, Gonzales-Mantegia W (2007) A functional analysis of NOx levels: location and scale estimation and outlier detection. Comput Stat 22:411–427
Febrero M, Galeano P, Gonzales-Mantegia W (2008) Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics 19:331–345
Fraiman R, Muniz G (2001) Trimmed means for functional data. Test 10:419–440
Gervini D (2009) Detecting and handling outlying trajectories in irregularly sampled functional datasets. Ann Appl Stat 3:1758–1775
Hubert M, Rousseeuw PJ, Branden KV (2005) ROBPCA: a new approach to robust principal component analysis. Technometrics 47(1):64–79
Hyndman RJ, Ullah MDS (2007) Robust forecasting of mortality and fertility rates: a functional data approach. Comput Stat Data Anal 51:4942–4956
Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, Cohen KL (1999) Robust principal component analysis for functional data. Test 8(1):1–73
Ramsay JO, Silverman BW (2001) Functional data analysis software. MATLAB edition. Online at http://www.psych.mcgill.ca/faculty/ramsay/software.html
Ramsay JO, Silverman BW (2005) Functional data analysis, 2nd edn. Springer, New York
Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41:212–223
Stahel WA (1981) Robust estimation: infinitesimal optimality and covariance matrix Estimators, Ph.D. thesis, ETH, Zürich
Verboven S, Hubert M (2004) A Matlab library for robust analysis. Chem Intell Lab Syst 75:127–136
Yamanishi Y, Tanaka Y (2005) Sensitivity analysis in functional principal component analysis. Comput Stat 20:311–326