Institut f. Stochastik und Wirtschaftsmathematik

1040 Wien, Wiedner Hauptstr. 8-10/105
AUSTRIA
http://www.isw.tuwien.ac.at

Evaluation of robust PCA for supervised audio outlier detection

S. Brodinova, T. Ortner, P. Filzmoser, M. Zaharieva, and Ch. Breiteneder

Forschungsbericht CS-2015-2

Juli 2015

Kontakt: P.Filzmoser@tuwien.ac.at

Evaluation of Robust PCA for Supervised Audio Outlier Detection

Sarka Brodinova[1], Thomas Ortner[1,2], Peter Filzmoser[2], Maia Zaharieva[1,3], and Christian Breiteneder[1]

[1] Interactive Media Systems Group, Vienna University of Technology
[2] Institute of Statistics and Mathematical Methods in Economics
[3] Multimedia Information Systems Group, University of Vienna

Abstract. Outliers often reveal crucial information about the underlying data such as the presence of unusual observations that require for in-depth analysis. The detection of outliers is especially challenging in real-world application scenarios dealing with high-dimensional and flat data bearing different subpopulations of potentially varying data distributions. In the context of high-dimensional data, PCA-based methods are commonly applied in order to reduce dimensionality and to reveal outliers. In this paper, we perform a thorough empirical evaluation of well-establish PCA-based methods for the detection of outliers in a challenging audio data set. In this evaluation we focus on various experimental data settings motivated by the requirements of real-world scenarios, such as varying number of outliers, available training data, and data characteristics in terms of potential subpopulations.

Keywords: outlier detection, pca, audio data, experiments

1 Introduction

The identification of outliers is an essential data mining task. Outliers do not only contaminate distributions and, thus, estimations based on the distributions, moreover, they often are the centre of attention. In many fields outliers carry significant, even crucial information for applications such as fraud detection, surveillance, and medical imaging [4]. In this paper, we utilize outlier detection in an automated highlight detection application for audio data. This is a first step towards the identification of key scenes in videos, where the audio is a fundamental component.

The detection of outliers gets considerably more difficult, when they are located in a high-dimensional space or there are less observations than variables available. In a high-dimensional space the data becomes sparse and the distances between observations differ very little [3]. The situation becomes even more complex when groups of outliers are present due to the emerging masking effect [1]. To justify the application of distance-based similarity measures in such a situation, the reduction of dimension is an inevitable course of action.

A well-established approach for dimensionality reduction is the use of principal component analysis (PCA), which transforms the variables to a smaller set of uncorrelated variables keeping as much of the total variance as possible [?]. This step removes the curse of high dimensionality [2] for this subspace. Nevertheless, it has been shown, that even though in theory distance functions loose their meaningfulness in high dimensionality [3], the orthogonal complement of the principal component (PC) space might still hold crucial differences in the distance and, thus, important information for outlier detection [31].

The main focus of this paper is the thorough empirical comparison of different PCA-based methods for high-dimensional and flat data, that are suitable for outlier detection in audio data. Most importantly, we will compare traditional PCA with robust versions of PCA since the estimation of the PC subspace might be affected by the outliers themselves. We compare the sensitivity of the selected PCA-based methods towards changes in the setup, namely the number of outliers, the size of the data sets, and the distribution of the data sets. Additionally, we evaluate the quality of the results as well as the necessary computational effort.

One of the most essential aspects of the outlier detection process, using any type of PCA, is the proper choice of number of components used for the construction of the PC space. Here we propose to manually classify a small number of observations and use those labels to estimate the best possible number of principal components without any prior knowledge regarding the data structure. This concept creates a reasonable situation for applications. Thus, an estimation for the optimal number of components is performed throughout all the sensitivity analyses including an analysis regarding the number of pre-classified observations itself. In addition, we propose an optimization of the critical values used for the outlier detection by employing the additional information from the classified observations, which greatly increases the robustness of the classification towards the number of chosen components. The requirement of a set of labelled observations, thus, results in a supervised outlier detection procedure.

This paper is organized as follows. Section 2 outlines related work. Section 3 describes in detail the evaluation setup for the performed experiments. We start with a data setup where we first compare classical and robust PCA using a set of training data to show the effects of a considerable portion of pre-classified and unclassified observations. Consequently, we change the setup to a situation where fewer observations are classified and show the sensitivity of the results towards the number of labelled observations as well as to other characteristics of the data set, such as the number of outlying observations and the group structure. Section 4 presents the results of the evaluations. Eventually, in Section 5 we provide conclusions for future applications.

2 Related Work

Several authors perform simulation studies to explore the performance of the classical and various robust PCA-based methods in different scenarios in the context of outlier detection, such as varying degree of data contamination, data dimensionality, and in the presence of missing data, e.g. [20][25][26][30]. For example, Pascoal et al. [20] compare the classical PCA approach [15] with five robust methods: spherical PCA [17], two projections pursuit techniques [5][6], and the ROBPCA approach [13] in several different contamination schemes. The experiments show, that overall the ROBPCA outperforms the compared methods in terms of estimated recall. Similarly, Sarpa [25] show that a robust PCA approach based on projection pursuit [10] outperforms the classical PCA even for data sets with more variables than observations. In a recent simulation study, Xu et al. [30] show that – for the generated data settings – the performance of ROBPCA and techniques based on projection pursuit degrades substantially in terms of expressed variance as the dimensionality of the data increases. However, the authors only consider the first few principal components and focus on a data setting where the observations and the variables are of the same magnitude. Usually, simulation studies are performed for very specific data settings (e.g. all observations/variables follow a predefined distribution). However, real data have a more complex data structure than synthetic data and, thus, outlier detection on real data is even more challenging.

Current evaluations on real data sets are often limited by the number of available data. As a result, a thorough investigation of different outlier detection methods for various data settings is barely feasible. For example, Sarpa [25] performs an evaluation on a small set of financial data with 120 observations. Hubert et al. [13] report evaluations on three low-sampled real data sets with varying dimensionality. While, in general, evaluations on multiple data sets provide an estimation of the robustness of the investigated approaches, no general conclusions about the sensitivity to specific data aspects can be made. Experiments with larger real data sets are commonly tailored to the sole evaluation of the performance of outlier detection methods for the particular data without any variation of the experiment settings (e.g. [9][27]). In contrast, in this paper we employ a large real data set in the simulation of different experiment settings and perform a thorough evaluation of the sensitivity of the explored approaches with respect to different data aspects such as varying percentage of outliers, available training data, and different data characteristics.

3 Evaluation Setup

In this section we present the evaluation setup of our experiments including the compared PCA-based approaches, the employed performance metrics, and the explored data set of high-dimensional audio data consisting of different subpopulations.

3.1 Compared Approaches

PCA is a well-studied and widely employed approach for dimensionality reduction. The data matrix $\mathbf{X}$ with n observations described by p variables is assumed to be column-centered. The original data is transformed into a subspace defining a new coordinate system $\mathbf{T} = \mathbf{XP} + \mathbf{E}$, with the $p \times k$ loading matrix $\mathbf{P}$, the $n \times k$ scores matrix $\mathbf{T} = [t_{ij}]$, and the error matrix $\mathbf{E}$. The j-th column of $\mathbf{P}$ is chosen such that the variance l_j of the corresponding j-th column of $\mathbf{T}$ is maximized, subject to orthogonality to the previous columns of $\mathbf{P}$. The newly constructed subspace of potentially lower dimension is determined by a pre-specified number k of PCs, i.e the first k score vectors.

In general, algorithms for estimating the PC space are based on either eigenvector decomposition of the empirical covariance matrix, singular value decomposition (SVD) of the (mean-centered) data matrix, or on projection-pursuit PCA [12]. We compare several approaches employing these techniques including both classically and robustly estimated PCs. Additionally, the compared methods are chosen in such a way that they are suitable for high-dimensional data with potentially more variables than observations.

Outlier detection based on PCA is commonly employed by means of two different distances for each observation derived from the PC space [13], namely score distance, SD, and orthogonal distance, OD:

$$SD_i^{(k)} = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}}, \quad OD_i^{(k)} = ||\mathbf{x}_i - \mathbf{P}\mathbf{t}_i||, \tag{1}$$

for $i = 1, \ldots, n$, where $\mathbf{t}_i = (t_{i1}, \ldots, t_{ik})^\top$ are the score vectors in the PC space, and $\mathbf{x}_i$ is the ith observation of $\mathbf{X}$. The index k refers to the number of PCs used to compute SD and OD. The score distance represents the (Mahalanobis) distance of an observation to the center of the data and, thus, it is a measure of outlyingness in the estimated subspace. The orthogonal distance measures the distance of the observations to the subspace. Two thresholds are used to detect outliers. First, the 97.5% quantile of the χ^2 distribution with k degrees of freedom where k is the number of PCs, i.e. $c_{SD}^{(k)} = \sqrt{\chi^2_{k,0.975}}$, is used as a critical value for SD. Second, $c_{OD}^{(k)} = (\hat{\mu} + \hat{\sigma} z_{0.975})^{3/2}$ represents the critical value for OD. The values $z_{0.975}$ is a quantile of the normal distribution and $\hat{\mu}$ (resp. $\hat{\sigma}$) can be obtained using the median (resp. Median Absolute Deviation (MAD)) of the values of $OD_i^{2/3}$ (see [13] for more details).

clPCA: classical (non-robust) PCA [15] employs the eigenvalue decomposition of the classical empirical covariance. In case of a data with a larger number of variables than observations, the SVD algorithm is a reasonable choice to obtain the solution. Therefore, the columns of the loading matrix $\mathbf{P}$ are the right singular vectors and the variance l_j corresponding to the j-th singular value. The direct relationship between SVD and PCA calculated on the classical covariance matrix is described in [29]. However, the classical covariance is sensitive to outliers [13] and the resulting PCs do not describe the true underlying data structure. Robustification of PCA in this context often achieves more reliable results.

OGK PCA: PCA based on robust covariance matrix estimation is the simplest way to obtain robust PCA. The orthogonalized Gnanadesikan-Kettenring (OGK) [18] estimator of covariance matrix achieves reliable results for high-dimensional data in terms of robustness and computational cost [18]. The method starts by robustly scaling the data, $\mathbf{Y} = \mathbf{X}\mathbf{D}^{-1}$, where $\mathbf{D} = diag\{\hat{\sigma}(X_1) \ldots \hat{\sigma}(X_p)\}$ is the robustly estimated univariate dispersion of each columns X_j of $\mathbf{X}$. Next, the Gnanadesikan-Kettenring estimator [11] is computed for all variable pairs of $\mathbf{Y}$, resulting in a robust correlation matrix, $\mathbf{U}$, where $U_{jk} = cov(Y_j, Y_k)$, $j, k = 1, \ldots, p$. The eigenvector decomposition of the correlation matrix $\mathbf{U} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^\top$ allows for the projection of the data onto the directions of the eigenvectors, $\mathbf{Z} = \mathbf{Y}\mathbf{E}$. Since the eigenvalues of $\mathbf{U}$ may be negative, the robust variances in the corresponding directions $\mathbf{L} = diag\{\hat{\sigma}(Z_1) \ldots \hat{\sigma}(Z_p)\}$ are estimated and replaced in order to obtain a positive-definite covariance matrix, $\mathbf{S_Y} = \mathbf{E}\mathbf{L}\mathbf{E}^\top$. Eventually, the covariance matrix is transformed back to the original space, $\mathbf{S_X} = \mathbf{D}\mathbf{E}\mathbf{L}\mathbf{E}^\top\mathbf{D}^\top$, where $\mathbf{DE}$ is the loading matrix of p orthogonal eigenvectors of dimension k and corresponds to the direction of the principal components.

GRID PCA: Robust PCA using the GRID search algorithm [5] employs the projection-pursuit (PP) principle to project the data on a direction which maximizes the robust variance of the projected data [16]. This method is suitable especially in case of data with more variables than observations. The basic idea of GRID is to iteratively search the optimal directions on a regular grid in the plane. The method starts by sorting the variables in decreasing order according to the robust dispersion. The first projection direction is found in the plane spanned by the first two sorted variables and it passes through the robust center and a grid point. Then, the remaining variables enter the search plane successively to obtain the first optimal direction. The algorithm searches the subsequent directions in a similar way by imposing orthogonality, until there is no improvement in maximizing the robust variance. The main advantage of the approach is the reduced computation time in high-dimensional data space when only the first k directions are of interest [5].

ROBPCA [13] is a method combining robust PP techniques [16] with robust covariance estimation. First, the data space is reduced to an affine subspace using a singular value decomposition [14]. If the number of variables p is larger than the number of observations n, this step allows to express the information with at most n dimensions. The next step aims at the identification of $h < n$ data points that are the least outlying observations. In order to identify such observations, a concept of outlyingness [7][28] is adapted by using the univariate Minimum Covariance Determinant (MCD) location and scale estimator [23]. The covariance matrix, $\mathbf{S}_0$, of the least outlying points is subsequently used to select a number of components k and to projected the data on the subspace determined by the first k eigenvectors of $\mathbf{S}_0$. The FAST-MCD algorithm [24] is employed to obtain a robust scatter matrix, $\mathbf{S} = \mathbf{P}\mathbf{L}\mathbf{P}^\top$, where $\mathbf{P}$ is the loading

Table 1. Confusion table denoting a potential result of outlier detection with respect to outliers.

		Predicted membership	
		Normal	Outlier
Actual membership	Normal	TN (True negative)	FP (False positive)
	Outlier	FN (False negative)	TP (True positive)

matrix of p orthogonal eigenvectors of dimension k and $\mathbf{L}$ the diagonal matrix of k eigenvalues.

PCOut [9] is a method already comprising an outlier detection algorithm, in contrast to the previously described approaches. Thus, there is no need to determine an optimal number of PCs or a cut-off value for the outliers. It is frequently used in problems related to high-dimensional, low-sample size data, such as in bioinformatics and genetics [19][22]. Although the method can compete with outlier detection methods in lower dimensions, it is advantageous for high-dimensional data [18][21] for its accuracy and computational speed [9]. PCOut performs high-dimensional outlier detection in two steps. The first step aims at the identification of *location* outliers, i.e. outliers that are far away from the center of the main body of the data. The second step focuses on the detection of *scatter* outliers, which are characterized by observations that are possibly generated from a model with the same location as the main data but with a different covariance structure. Since PCA is not scale-invariant, PCOut scales the data robustly using the median and MAD. Principal components that contribute to about 99% of the total variance are considered relevant. The remaining components are assumed to contain noise only. Eventually, outlier detection is performed in the principal component space of reduced dimension using again the robustly scaled principal components.

3.2 Performance Measures

The performance of outlier detection methods is commonly presented in a confusion matrix as shown in Table 1. We evaluate the performance of the compared approaches in terms of true positive rate, TPR, and false positive rate, FPR:

$$TPR = TP/(TP + FN), \qquad FPR = FP/(FP + TN), \tag{2}$$

where TP denotes *true positives* (i.e. correctly identified outliers), FN indicates *false negatives* (normal observations that were declared as outliers), and FP corresponds to *false positives.*

Additionally, we calculate the area under the Receiver Operating Characteristics (ROC) curve (AUC) [8] representing the trade-off between TPR and FPR

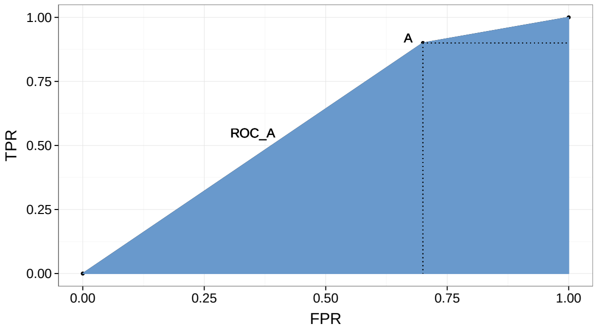


Fig. 1. Construction of ROC curve employed in our experiments and illustration of dividing the area into regular shapes to be summed up in order to obtain AUC.

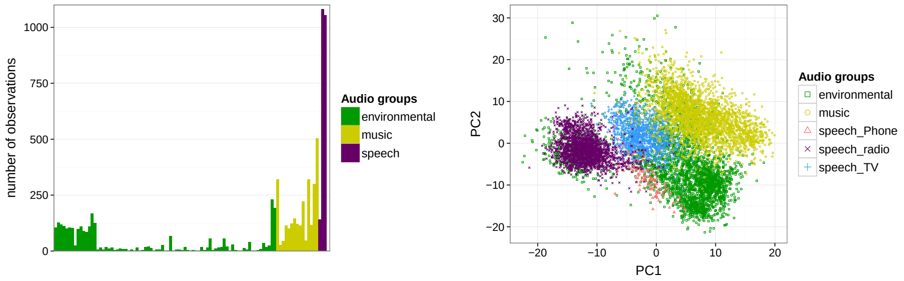
by a singe value. Fig. 1 illustrates the construction of a ROC curve employed in our experiments for an example evaluation A. The estimation of the corresponding AUC of A is obtained in such way that the area is divided into regular shapes and summed up, which results in the following estimation:

$$AUC = \frac{1}{2}(1 + TPR - FPR). \tag{3}$$

Note that when the algorithm does not detect any outliers (i.e. both TPR and FPR are zero) the AUC according to the above formula is equal to 0.5. Although the two extreme scenarios (no outliers detected and random prediction) are not identical, they are both not desired output in terms of effectiveness of outlier detection approaches.

3.3 Data set

In our experiments we employ a high-dimensional audio data set of approximately 8700 instances. Each observation is represented by a set of 50 high-dimensional features of 679 variables in total. The data set covers the three fundamental audio types: music, speech, and environmental sounds. The number of instances in the three groups is approximately equal. Data structure of each type is highly varying due to many subpopulations consisting of different numbers of observations, see Fig. 2(a). Fig 2(b) visualizes the projected data into the subspace spanned by the first two classical PCs and it reveals that all three audio types follow different distributions. We clearly see that, for instance, many of the environmental observations are mixed with the data points from the other two audio groups. The plot also shows different distributions of subpopulations within the speech data. Looking at the biggest speech subgroup corresponding to TV sounds, we observe that there are many observations from both environmental and music samples located in the same area. Hence, this is a challenging task to separate TV sounds from the remaining groups, especially in the context of outlier detection.



(a) Size of subgroups within the different audio groups

(b) Audio groups following the various distributions

Fig. 2. Visualization of data structure in terms of both data distribution and the size of subpopulations present in audio data.

The outlier detection approaches based on the two distance measures (SD and OD) employ three data sets: training, validation, and test set. The PC space spanned by k components is constructed with the observations coming from the training set. Additionally, we calculate two critical values for the orthogonal distance, c_{OD}^k, and for the score distance, c_{SD}^k. These measures are exclusively derived from loadings and scores of the training data. Next, the observations from the validation set are projected onto the constructed PC space spanned by k principal components. An observation having a distance larger than both corresponding critical values is declared as an outlier. This procedure is conducted with varying number of components k to select the optimal number of components, k_{opt}, in terms of maximizing AUC. Eventually, we perform an evaluation on the test data with respect to the optimal number of components k_{opt} from the validation set and the PC space spanned by k_{opt} determined by observations from the training set.

In contrast, the data setup for PCOut uses training and test sets only. Since the algorithm selects the number of PCs automatically, there is no need to employ the validation set. To ensure that the comparison of PCOut to the other methods is based on the same data, we consider that the training set in case of PCOut is identical to the combination of the training and the validation set in the previous data setup. In this way created training set is used to estimate all necessary parameters including median, MAD, the PC space holding about 99% of the total variance, and the boundaries indicating possible outliers being far away from the center of the main body of data as well as having different covariance structure. Eventually, we perform the evaluation on the test set with respect to these parameters estimated on the training set.

4 Experimental Results

In this section we present the results of the performed experiments. We explore the sensitivities of the investigated approaches with respect to the number of outliers, to the size of training and validation sets, and to the data characteristics. The assignment of the observations to the sets is done randomly and all evaluations are based on 100 replications. We report the evaluation in terms of average and standard error (SE) of both AUC and the number of PCs.

We rescale all data sets to make the variables comparable using the mean and standard deviation of the variables in the training set. The reason for applying a non-robust scaling is the presence of many variables which are almost constant but a small proportion of values has huge deviations. The robust MAD for such variables would be very small and this would artificially increase the whole data range during the scaling. As a consequence, many of the regular observations would be made indistinguishable from real outliers.

4.1 Sensitivity to the Number of Outliers

In our first experiment we investigate the sensitivity of the compared approaches to the number of outliers present in the selected data set. We consider the biggest speech subgroup as the main group of observations and we randomly select observations from both environmental and music samples as outliers. We split these data into training, validation and test sets: 0.33/0.33/0.33, corresponding to approximately 360 normal observations.

First, we calculate the PC space only on the main observations from the training set and vary the percentage of outliers ranging from 2% to 10% of the main observations in validation and test sets. Table 2 shows that clPCA performs similar or better than the robust PCA methods, while PCOut is capable of finding only approximately half of the outliers corresponding to low TPR. Although the performance of clPCA and its robust counterparts degrades slightly by decreasing the percentage of outliers, ROBPCA does not indicate such dependency. We also notice that SE remains at a very low level during the experiments for all methods.

Next, we consider that the training set is not free of outliers in order to explore their impact on the constructed PC space. From Table 3, we observe that the robust PCA methods clearly outperform clPCA which is able to identify only a small number of outliers. PCOut performs poor as before. While the number of outliers does not show any clear dependency on the resulting AUC, this is not the case for the number of PCs. GRID PCA reduces the number of selected PCs with decreasing contamination in contrast to the other methods. Additionally, ROBPCA tends to select a considerably lower number of PCs than its counterparts.

Considering the resulting AUC from both tables, the use of robust PCA methods is recommended when there is no guarantee that the training set is free of outliers. In a real-world scenario this can not always be satisfied and therefore we take this into account and the training set contains outliers in our further experiments.

Table 2. Evaluation of the sensitivity to % of outliers in terms of both average and standard error (SE) of AUC determined from the classification rates (TPR, FPR), and the number of PCs. PC space is calculated on the normal instances.

% outliers	**Method**	**AUC**	$\mathbf{SE_{AUC}}$	**k**	$\mathbf{SE_{k}}$	**TPR**	$\mathbf{SE_{TPR}}$	**FPR**	$\mathbf{SE_{FPR}}$
10	clPCA	0.948	0.002	122	1	0.953	0.005	0.058	0.003
	GRID PCA	0.943	0.002	144	2	0.943	0.004	0.056	0.002
	ROBPCA	0.907	0.002	57	2	0.918	0.007	0.103	0.004
	OGK PCA	0.936	0.002	191	4	0.946	0.005	0.074	0.003
	PCOut	0.602	0.006	-	-	0.516	0.060	0.311	0.002
5	clPCA	0.946	0.003	136	2	0.949	0.007	0.057	0.003
	GRID PCA	0.942	0.003	163	2	0.937	0.007	0.054	0.002
	ROBPCA	0.916	0.003	51	2	0.929	0.006	0.097	0.004
	OGK PCA	0.931	0.003	168	5	0.931	0.008	0.070	0.003
	PCOut	0.609	0.007	-	-	0.538	0.016	0.320	0.006
2	clPCA	0.912	0.007	164	4	0.865	0.016	0.040	0.003
	GRID PCA	0.937	0.007	190	4	0.860	0.015	0.039	0.003
	ROBPCA	0.913	0.005	39	3	0.911	0.010	0.085	0.005
	OGK PCA	0.905	0.007	129	6	0.862	0.015	0.052	0.004
	PCOut	0.604	0.009	-	-	0.531	0.021	0.323	0.006

4.2 Sensitivity to the Size of Sets

Since our date setup employs three data sets, we investigate whether varying the size of these sets considerably changes the performance of the compared approaches. Again, we consider the biggest speech subgroup as the main observations and we add 5% of instances from the other two audio groups (music and environmental sounds) as outliers. We divide the data into training, validation, and test sets according to different partitions ranging from 0.33/0.33/0.33 up to 0.05/0.05/0.90 corresponding to the size of sets from 378/378/380 to 57/57/1022 observations. Note that the percentage of outliers is the same in each data set.

Fig. 3 summarizes the results of the evaluation showing the performance in terms of AUC, Fig. 3(a), and number of PCs, Fig 3(b), necessary to distinguish outliers from normal observations. As expected from the previous experiment, clPCA fails since the training set contains outliers. In general, the performances of the robust PCA methods decrease with the reduction of the size of training and validation sets. GRID PCA achieves a high AUC and outperforms the remaining methods even if the size of the available training set is reduced to 170 instances. AUC falls rapidly when considering smaller data size. In contrast, ROBPCA yields still a reasonable AUC in the most extreme setting (57 observations). For a more detailed investigation, we visualize the distribution of the resulting AUC during the replications for each method. Fig. 4 illustrates that PCOut and clPCA perform similar in each situation since the distribution of observed intervals is almost identical. This does not hold for the other three methods. The proportion of AUC ranging from 0.9 to 1 representing the results of ROBPCA decreases with the size reduction of training and validation sets. Considering the performance

Table 3. Evaluation of the sensitivity to % of outliers in terms of both average and SE of AUC determined from the classification rates (TPR, FPR), and the number of PCs. PC space is calculated on the training set containing outliers.

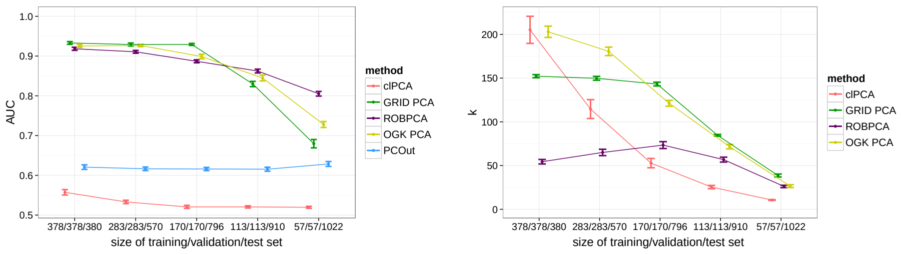
% outliers	**Method**	**AUC**	**SE$_{\mathbf{AUC}}$**	**k**	**SE$_{\mathbf{k}}$**	**TPR**	**SE$_{\mathbf{TPR}}$**	**FPR**	**SE$_{\mathbf{FPR}}$**
10	clPCA	0.531	0.005	152	16	0.067	0.011	0.004	0.001
	GRID PCA	0.921	0.003	140	1	0.896	0.006	0.053	0.001
	ROBPCA	0.891	0.004	75	3	0.887	0.007	0.106	0.005
	OGK PCA	0.929	0.003	281	4	0.926	0.006	0.068	0.002
	PCOut	0.624	0.004	-	-	0.351	0.007	0.103	0.002
5	clPCA	0.557	0.007	205	15	0.124	0.015	0.009	0.002
	GRID PCA	0.933	0.003	152	2	0.923	0.007	0.057	0.002
	ROBPCA	0.918	0.004	54	3	0.928	0.008	0.092	0.004
	OGK PCA	0.925	0.003	203	6	0.918	0.008	0.067	0.004
	PCOut	0.621	0.006	-	-	0.359	0.012	0.118	0.005
2	clPCA	0.565	0.009	173	15	0.141	0.019	0.011	0.002
	GRID PCA	0.914	0.006	174	4	0.872	0.013	0.044	0.002
	ROBPCA	0.918	0.005	39	3	0.916	0.011	0.081	0.005
	OGK PCA	0.908	0.007	139	7	0.865	0.016	0.050	0.004
	PCOut	0.615	0.009	-	-	0.420	0.022	0.191	0.011

of OGK PCA, we observe that the largest proportion of AUC between 0.9 and 1 is attained when the sample size of training set is 283, and subsequently reduced size to 57 instances causes that the majority of AUC achieves the values between 0.5 and 0.8. The results of GRID PCA reveal very large proportion of AUC from the interval $(0.9, 1]$ in the first three situations. However, when the size of training and validation sets is reduced from 113 to 57, almost half of the AUC values are in the interval $(0.7, 0.4]$.

Fig. 3(b) shows that the number of PCs selected by ROBPCA is independent from the size of the sets and it tends to chose less PCs, while the number of components in case of the other methods is affected by decreasing the number of observations in the training and validation sets. This is given by the method itself but also by the size of the employed training set. Moreover, the number of PCs selected by clPCA deviates considerably during the replications in the first three situations. In contrast, GRID PCA indicates small standard errors of the selected numbers of components.

4.3 Sensitivity to the Size of the Validation Set

The previous experiments employ a training set containing outliers to construct the PC space and calculate two critical values. That means, the available information about labels is required only for the validation set to select the optimal number of PCs. Additionally, the results from the experiment in Section 4.2 indicate that some of the compared approaches perform well even if the size of validation set is reduced to 170 or 57 instances. These two remarks motivate us to explore how many observations in the validation set need to be labeled to



(a) AUC

(b) Number of PCs

Fig. 3. Evaluation of the sensitivity to the size of training, validation, and test sets in terms of both average and SE of AUC and the selected number of PCs.

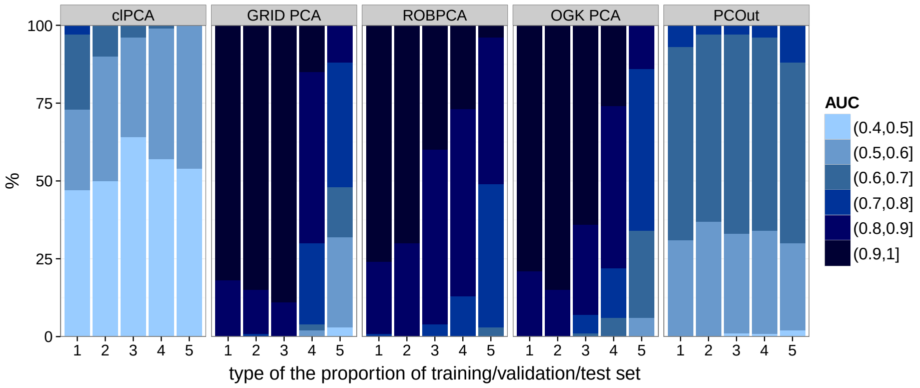
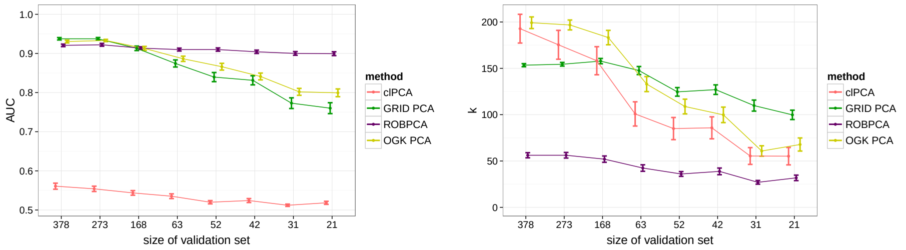


Fig. 4. Detailed investigation of the resulting AUC during the replications for different partitions of training, validation and test set (training/validation/test set) corresponding to the following size of sets: **1**: 378/378/380, **2**: 283/283/570, **3**: 170/170/796, **4**: 113/113/910, and **5**: 57/57/1022 observations.

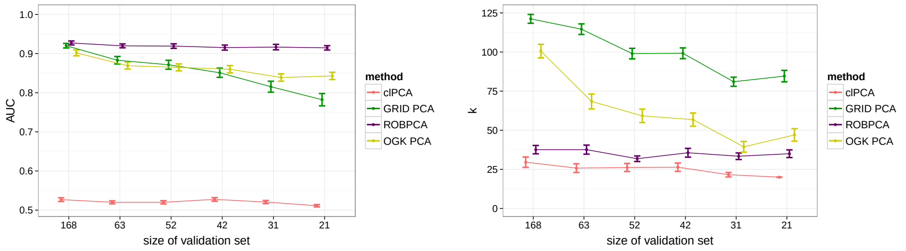
achieve satisfying results. Thus, we fix training and test sets to the same size, i.e. 378 observations, and vary the number of observation in the validation set from 21 up to 378 instances. As before, we simulate the biggest speech subgroup as the main observations and we add 5% from the other two audio groups as outliers. Note that PCOut is not included to this experiment because it does not use a validation set.

Fig. 5(a) shows that both GRID PCA and OGK PCA are sensitive to the size of the validation set. Additionally, the AUC deviates considerably with decreasing size of the validation set. In contrast, ROBPCA performs well independently from the number of instances in the validation set and achieves a high AUC even if the size of the validation set is small in comparison to the training and test set. Considering the larger size of validation set, i.e 378 and 273 observations, OGK outperforms its competitors. In general, the number of PCs (see Fig. 5(b)) decreases with reducing the size of validation set and deviates during the repli-



(a) AUC (b) Number of PCs

Fig. 5. Evaluation of the sensitivity to the size of validation sets in terms of both average and standard deviation of AUC and the selected number of PCs. The speech data are considered as main observations.



(a) AUC (b) Number of PCs

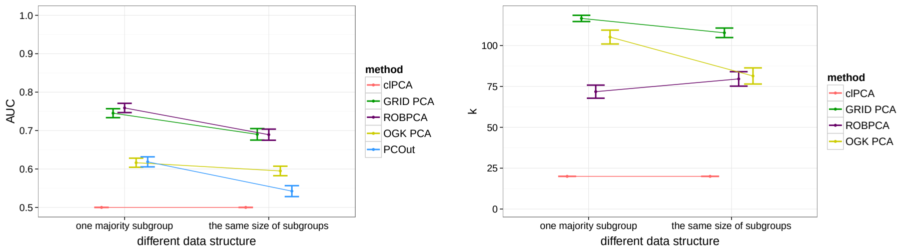
Fig. 6. Evaluation of the sensitivity to the size of validation sets in terms of both average and standard deviation of AUC and the selected number of PCs. Main observations are randomly selected from the music sample.

cations. However, ROBPCA indicates both small SE and slight decline in the selected number of PCs.

To stress our conclusion that available labeled validation data set can be small to achieve reasonable results, we change the main group to the music biggest subgroup and perform the same experiment considering the size of training and test set is fixed to 168 instances. Fig. 6 shows very similar performance and we clearly see that ROBPCA outperforms the other methods in all investigated situations.

4.4 Sensitivity to the Data Characteristics

In Section 3.3 we described the full data set consisting of many subpopulations. Therefore, we take this into account to explore the sensitivity of the compared approaches to the underlying data characteristics with respect to varying data structures. We simulate the main observations consisting of three randomly selected audio subgroups with different sample size and the percentage of outliers is fixed to 5% of main observations.



(a) AUC

(b) Number of PCs

Fig. 7. Evaluation of the sensitivity to different data characteristics in terms of both average and SE of AUC and the number of PCs.

First, we investigate the case of one majority subgroup present in the main observations and then we set up the same size of the subgroups. Fig. 7 visualizes both situations and shows that the performance is slightly better when one majority group is considered. Although ROBPCA and GRID PCA achieve a higher AUC than the other methods, the resulting AUC indicates that these two methods have difficulties of coping with the multi-group data structure since AUC is lower in comparison to the previous experiments. Additionally, we clearly see that clPca completely fails with AUC of 0.5 (i.e both TPR and FPR are zero). Looking at the SE of AUC, we observe that the values are considerably higher than in previous experiments. Overall, there is no clear dependency between the number of PCs and the different multi-group data structure.

4.5 Computation Time

Finally, we explore the average user CPU time of the compared approaches in seconds over 100 replication with respect to the number of PCs. For this experiments we follow the same setup as in Section 4.1 with 5% of outliers. Fig. 8 shows that computation time rises with increasing number of PCs for both ROBPCA and GRID PCA, while OGK PCA and clPCA indicate independence from the number of PCs since these methods need to compute all eigenvectors,and then select the number of PCs. Additionally, OGK PCA is a very time consuming algorithm in contrast to clPCA. PCOut is displayed by a single point since the algorithm selects the number of PCs automatically and it is the most computationally efficient. Although we clearly see that the price of using the robust PCA methods is an increasing computational effort, ROBPCA offers a reasonable compromise.

5 Discussion and Conclusion

In this paper we compared different PCA-based algorithms for outlier detection in the context of a high-dimensional audio data set. Since the classical PCA [**?**]

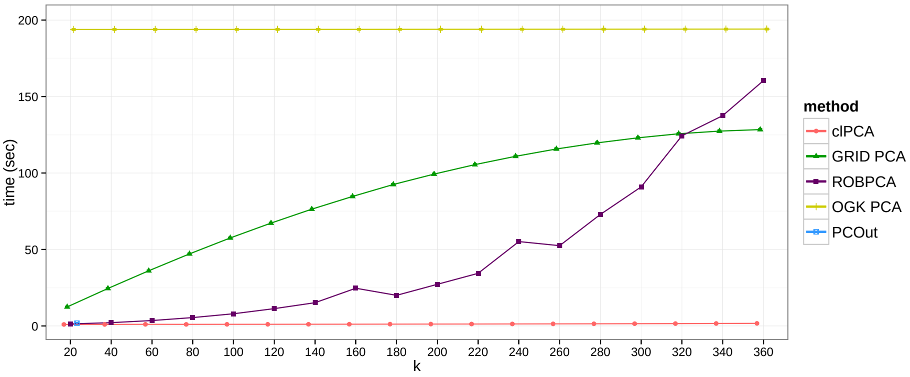


Fig. 8. Average computation time (over 100 replications) in seconds of the compared approaches and its dependency from the number of PCs.

is sensitive to the presence of outliers in the training data, we employed several, well-established robust PCA methods, such as GRID PCA [5], ROBPCA [13], OGK [18], and PCOUT [9], to better reveal outlying samples. We performed a thorough investigation of the sensitivity of employed approaches with respect to different data properties, number of outliers, and size of the available training data. In all of those settings, ROBPCA performed at the same level as the GRID and OGK algorithms. However, ROBPCA showed much lower sensitivity towards changes in the number of available training or validation observations. The reason for this property is the fewer necessary number of PCs to properly model the data structure. If the number of available observations is too low to create the necessary PCA space or to properly evaluate the used PCA space, the quality of the outcome decreases. We therefore recommend the usage of ROBPCA for outlier detection in similar setups. Especially Section 4.3 shows how much this procedure benefits from few pre-labelled observations by enabling the individual estimation of a proper number of PCs.

The observations from the validation set can be further utilized. Both, the critical value of the orthogonal distance and the critical value of the score distance are based on the assumption that the observations follow a multivariate normal distribution. Even though robust PCA allows for a certain violation of this assumption, the heterogeneous data structure suggests systematic deviations from normality. One could thus take advantage of the availability of validation data to adjust the critical values. For this purpose we use the AUC performance measure defined in Section 3.2 and maximise the AUC value for each fixed number of components, varying the critical values for score and orthogonal distances. Note that the only meaningful critical values are the distances given by the pre-classified outliers. All other possible values will increase the FPR, without affecting the TPR. Thus, the necessary computational effort is very eligible.

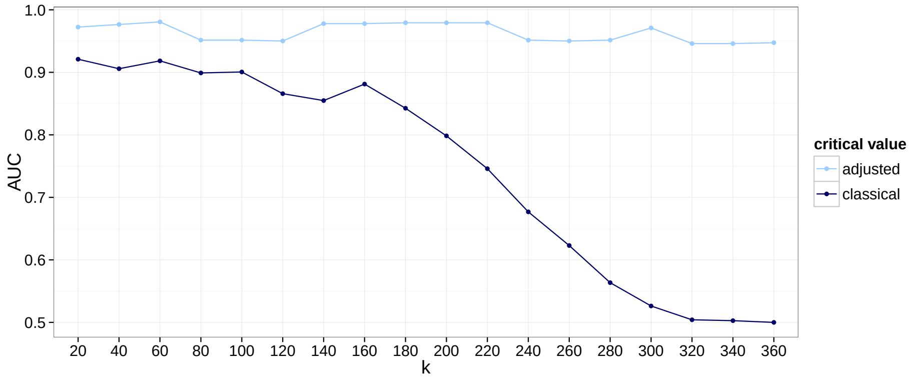


Fig. 9. Comparison of the AUC value, depending on the number of components. While the quality of the classification for the classical critical values is highly depending on the chosen number of components, the adjusted critical values remains at an almost constant level.

Even though the adjustment is not necessary for the comparison of PCA methods and might have masked some method specific behaviours, it provides advantages for the application itself. The main benefit is the resulting robustness towards the number of chosen principal components. Fig. 9 shows this effect for ROBPCA for one example of speech main observations with 5% outliers. It can clearly be seen that performing the outlier detection for a low number of principal components is sufficient. Thus, even though the adjustment needs computation time, the total computational effort decreases, for it is no longer necessary to calculate a whole range of different numbers of PCs. At the same time, the risk of choosing an inappropriate number of PCs vanishes. It can be shown that the adjustment will asymptotically always perform at least as good as the theoretical critical values with increasing numbers of validation observations and outperforms the theoretical critical values for skewed distributions.

Since we want to consider as few training and validation observations as possible in order to minimize the potential annotation effort, it highly depends on the situation whether or not this method should be applied. In general, if there are many training and validation observations available and the main observations are not normally distributed, one should consider adjusting the critical values as presented.

Acknowledgments This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010.

References

1. Becker, C., Gather, U.: The masking breakdown point of multivariate outliers. Journal of the American Statistical Association 94, 947–955 (1999)

2. Bellman, R.E.: Dynamic programming. Princeton University Press (1957)
3. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: International Conference on Database Theory. pp. 217–235. Springer Berlin Heidelberg (1999)
4. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys 41(3), 15:1–15:58 (2009)
5. Croux, C., Filzmoser, P., Oliveira, M.: Algorithms for projection-pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems 87, 218–225 (2007)
6. Croux, C., Ruiz-Gazen, A.: High breakdown estimators for principal components: the projection-pursuit approach revisited. Journal of Multivariate Analysis 95(1), 206–226 (2005)
7. Donoho, D.: Breakdown properties of multivariate location estimators. Ph.D. thesis, Harvard University (1982)
8. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006)
9. Filzmoser, P., Maronna, R., Werner, M.: Outlier identification in high dimensions. Computational Statistics and Data Analysis 52, 1694–1711 (2008)
10. Filzmoser, P., Serneels, S., Croux, C., Van Espen, P.: Robust multivariate methods: The projection pursuit approach. In: From Data and Information Analysis to Knowledge Engineering, pp. 270–277. Studies in Classification, Data Analysis, and Knowledge Organization, Springer (2006)
11. Gnanadesikan, R., Kattenring, J.R.: Robust estimates, residuals, and outliert detection with multiresponce data. Biometrics 28, 81–124 (1972)
12. Huber, P.: Projection pursuit. Annals of Statistics 13, 435–475 (1985)
13. Hubert, M., Rousseeuw, P., Vanden Branden, K.: ROBPCA: A new approach to robust principal component analysis. Technometrics 47, 64–79 (2005)
14. Hubert, M., Rousseeuw, P., Verboven, S.: A fast method for robust principal components with applications to chemometrics. Chemometrhics and Intelligent Laboratory Systems 60, 101–111 (2002)
15. Jolliffe, I.: Principal Component Analysis. Springer Series in Statistics, Springer (2002)
16. Li, G., Chen, Z.: Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and monte carlo. Journal of the American Statistical Association 80, 759–766 (1985)
17. Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., Boente, G., Fraiman, R., Brumback, B., Croux, C., et al.: Robust principal component analysis for functional data. Test 8(1), 1–73 (1999)
18. Maronna, R., Zamar, R.: Robust estimates of location and dispersion for high-dimensional data sets. Technometrics 43, 307–317 (2002)
19. Mittelstrass, K., Ried, J.S., Yu, Z., Krumsiek, J., Gieger, C., Prehn, C., Roemisch-Margl, W., Polonikov, A., Peters, A., Theis, F.J., et al.: Discovery of sexual dimorphisms in metabolic and genetic biomarkers. PLoS Genetics 7(8), e1002215 (2011)
20. Pascoal, C., Oliveira, M., Pacheco, A., Valadas, R.: Detection of outliers using robust principal component analysis: A simulation study. In: Combining Soft Computing and Statistical Methods in Data Analysis, pp. 499–507. Springer Berlin Heidelberg (2010)
21. Peña, D., Prieto, F.: Multivariate outlier detection and robust covariance matrix estimation. Technometrics 44, 286–310 (2001)

22. Rocha, C.M., Barros, A.S., Gil, A.M., Goodfellow, B.J., Humpfer, E., Spraul, M., Carreira, I.M., Melo, J.B., Bernardo, J., Gomes, A., et al.: Metabolic profiling of human lung cancer tissue by 1h high resolution magic angle spinning (hrmas) nmr spectroscopy. Journal of proteome research 9(1), 319–332 (2009)
23. Rousseeuw, P.J.: Least median of squares regression. Journal of the American Statistical Association 79, 871–880 (1984)
24. Rousseew, P., van Driessen, K.: A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223 (1999)
25. Sapra, S.K.: Robust vs. classical principal component analysis in the presence of outliers. Applied Economics Letters 17(6), 519–523 (2010)
26. Serneels, S., Verdonck, T.: Principal component analysis for data containing outliers and missing elements. Computational Statistics & Data Analysis 52(3), 1712–1727 (2008)
27. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. Tech. rep., DTIC Document (2003)
28. Stahel, W.: Breakdown of covariance estimators. Research report, Fachgruppe für Statistik, E.T.H. Zürich (1981)
29. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: A practical approach to microarray data analysis, pp. 91–109 (2003)
30. Xu, H., Caramanis, C., Mannor, S.: Outlier-robust pca: The high-dimensional case. IEEE Transactions on Information Theory 59(1), 546–572 (2013)
31. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining pp. 363–387 (2012)