Evaluation of robust PCA for supervised audio outlier detection

Sarka Brodinova, *Vienna University of Technology*, sarka.brodinova@tuwien.ac.at
Thomas Ortner, *Vienna University of Technology*, thomas.ortner@tuwien.ac.at
Peter Filzmoser, *Vienna University of Technology*, p.filzmoser@tuwien.ac.at
Maia Zaharieva, *Vienna University of Technology*, maia.zaharieva@tuwien.ac.at
Christian Breiteneder, *Vienna University of Technology*, christian.breiteneder@tuwien.ac.at

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 to reduce dimensionality and to reveal outliers. Thus, a thorough empirical evaluation of various PCA-based methods for the detection of outliers in a challenging audio data set is provided. The various experimental data settings are 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, Robust PCA, Audio data, Experiments

1 Introduction

Outlier identification is an essential data mining task. Outliers do not only contaminate distributions and, thus, estimations based on the distributions, moreover, they often are the prime focus of attention. In many fields outliers carry significant, even crucial information for applications such as fraud detection, surveillance, and medical imaging. In this paper, we employ 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.

Outlier detection gets considerably more difficult in a high-dimensional space or when there are less observations than variables available (flat data). In a high-dimensional space, data becomes sparse and distances between observations differ very little. To justify the application of distance-based similarity measures in such a situation, the reduction of dimensionality is an inevitable course of action. A well-established approach for this purpose is the use of principal component analysis (PCA), which transforms the original variables to a smaller set of uncorrelated variables keeping as much of the total variance as possible [8]. This step removes the curse of high dimensionality for this subspace. Nevertheless, it has been shown, that even though in theory distance functions loose their meaningfulness in high dimensionality,

the orthogonal complement of the principal component (PC) space might still hold crucial differences in the distance and, thus, important information for outlier detection [20].

The focus of this paper is the thorough empirical comparison of PCA-based methods for high-dimensional and flat data, that are suitable for outlier detection in audio data. We compare classical PCA with its robust versions in terms of sensitivity regarding changes in the setup such as the percentage of outliers and the size or the distribution of the data sets. A crucial aspect in this context is the proper choice of number of components used for the construction of the PC space. We propose to manually label a small number of observations and to use those labels to estimate the best possible number of PCs without any prior knowledge of the data structure. This concept creates a reasonable situation for real-world applications. Thus, an estimation for the optimal number of components is performed throughout all the experiments including an analysis regarding the number of pre-labeled observations itself. Furthermore, we outline an approach for the optimization of critical values used for outlier detection by employing the additional information from the labeled observations, which can greatly increase the robustness of the outlier detection towards the number of chosen components.

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 missing data, e.g. [12][15][16][19]. For example, Pascoal et al. [12] compare the classical PCA approach [8] with five robust methods: spherical PCA [10], two projections pursuit techniques [1][2], and the ROBPCA approach [6] in different contamination schemes. The results show that ROBPCA outperforms the compared methods in terms of estimated recall. Similarly, Sapra [15] shows that a robust PCA approach based on projection pursuit [4] outperforms the classical PCA even for data sets with more variables than observations. In a recent simulation study, Xu et al. [19] 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 more complex data structures 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, Sapra [15] performs an evaluation on a small set of financial data with 120 observations. Hubert et al. [6] report evaluations on three low-sampled real data sets with varying dimensionality. While 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 evaluation of the performance of outlier detection methods for a particular data without any variation of the experimental settings, e.g. [3][17]. In contrast, we employ a large real data set in the simulation of different experimental settings and perform a thorough evaluation of the sensitivity of the explored approaches with respect to varying data aspects.

3 Evaluation setup

Compared approaches

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 (PP) technique. We compare several approaches including both classically and

robustly estimated PCs which are suitable for high-dimensional flat data. PCA-based outlier detection can be employed using two different distances for each observation derived from the PC space [6]: score distance, SD, and orthogonal distance, OD:

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

where $\mathbf{t}_i = (t_{i1}, \ldots, t_{ik})^\top$ are the score vectors in the PC space, $\mathbf{x}_i$ is the ith observation of the data matrix $\mathbf{X}$, and the index k refers to the number of PCs. While SD represents the distance of observations in the estimated subspace to the center of data, OD measures the distance of the observations to the subspace. Two thresholds are used to detect outliers. For the SD, the 97.5% quantile of the χ^2 distribution with k degrees of freedom, i.e. $c_{SD}^{(k)} = (\chi^2_{k,0.975})^{1/2}$, and for the OD, 97.5% quantile of the standard normal distribution, $c_{OD}^{(k)} = (\hat{\mu} + \hat{\sigma} z_{0.975})^{3/2}$, can be taken as the critical values. The estimation of $\hat{\mu}$ (resp. $\hat{\sigma}$) can be obtained using the median (resp. MAD) of the values of $OD_i^{2/3}$ (see [6] for more details). If either threshold is exceeded, the respective observation is classified as an outlier.

clPCA: Classical (non-robust) PCA [8] for flat data is performed by means of SVD which is directly related to eigenvalue decomposition of the classical empirical covariance [18]. 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. However, the classical covariance is sensitive to outliers [6] and the resulting PCs do not describe the true data structure.

OGK PCA [11] is a PCA-based approach using robust covariance matrix estimation. 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 column X_j of the data matrix $\mathbf{X}$, and $\hat{\sigma}$ is computed by using τ-estimation of univariate dispersion. Next, the Gnanadesikan-Kettenring estimator [5] 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, $\mathbf{Z} = \mathbf{Y}\mathbf{E}$. Finally, 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 $\mathbf{L} = diag\{\hat{\sigma}(Z_1) \ldots \hat{\sigma}(Z_p)\}$ and $\mathbf{D}\mathbf{E}$ is the loading matrix of p orthogonal eigenvectors of dimension k and corresponds to the direction of the principal components.

GRID PCA [1] is a robust PCA approach using the GRID search algorithm. It employs the PP method to project the data on a direction which maximizes the robust variance of the projected data [9]. GRID first sorts 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. The remaining variables successively enter the search plane 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.

ROBPCA [6] combines robust PP techniques [9] with robust covariance estimation. First, the data space is reduced to an affine subspace using a SVD [7]. In the next step the least outlying observations are identified using the univariate Minimum Covariance Determinant (MCD) location and scale estimator [13]. The covariance matrix, $\mathbf{S}_0$, of the least outlying points is subsequently used to select a number of components k and to project the data on the subspace determined by the first k eigenvectors of $\mathbf{S}_0$. The FAST-MCD algorithm [14] is employed to obtain a robust scatter matrix, $\mathbf{S} = \mathbf{P}\mathbf{L}\mathbf{P}^\top$, where $\mathbf{P}$ is the loading matrix of p orthogonal eigenvectors of dimension k and $\mathbf{L}$ the diagonal matrix of k eigenvalues.

PCOut [3] is a method already comprising an outlier detection algorithm, in contrast to the previously described approaches. First, the observations being far away from the center of the main body of the data are identified, i.e. *location* outliers. Then, the detection of *scatter* outliers generated from a model with the same location as the main data but with a different covariance structure is conducted. Outlier detection is performed in the subspace using the robustly scaled PCs which contribute to about 99% of the total variance.

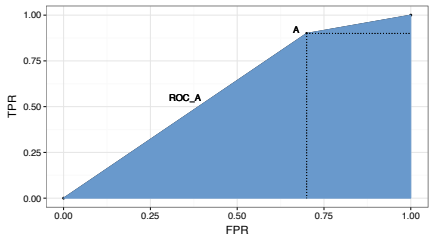

Figure 1. ROC curve construction.

Performance measures

We evaluate the performance of the compared approaches in terms of true positive rate, TPR, and false positive rate, FPR: $TPR = TP/(TP + FN), FPR = FP/(FP + TN)$, where TP refers to the *true positives* (correctly identified outliers), FN to the *false negatives* (outliers declared as normal observations), FP to the *false positives* (normal observations declared as outliers), and TN to the *true negatives* (correctly identified normal observations). Additionally, we calculate the area under the Receiver Operating Characteristics (ROC) curve (AUC) representing the trade-off between TPR and FPR by a single value. Figure 1 illustrates the construction of a ROC curve 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 $AUC = 1/2\ (1 + TPR - FPR)$. Note that when the algorithm does not detect any outlier (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 outlier detected and random prediction) are not identical, they are both not desired output in terms of effectiveness of outlier detection approaches. It should also be noted that the number of regular observations is much higher than the number of outliers. Thus, the defined AUC measure is much more sensitive towards changes in the total number of positively identified than towards negatively identified outliers. While this looks disproportional at first, the focus of the performed evaluations is the successful detection of outliers. Therefore, in this concept the high sensitivity towards single changes in TP is a welcome side effect.

Data set

We employ a high-dimensional, real-world audio data set of approximately $8,700$ observations to construct different challenging experimental settings, i.e. flat data, varying number of outliers and available training data, etc. The data set covers the three fundamental audio types: music, speech, and environmental sounds. Each observation is represented by a set of 50 (partially) multi-dimensional features, i.e. each feature consists of one or more variables, resulting in a feature vector of 679 dimensions in total. Features were selected in order to capture a wide range of audio properties and to represent the particular qualities of the three audio types equally well. The feature set comprises features that operate in the temporal and frequency domains, e.g. features for zero crossings, amplitude or brightness, features from the MPEG7 standard, perceptional features, and various cepstral coefficients.

The observations are approximately equally distributed across the three audio types. However, the underlying data structures are strongly varying due to present subpopulations of different sizes, e.g. different genres in the music samples and different voices in the speech samples. When constructing the data sets for the experiments and for the performance evaluation of the investigated approaches for outlier detection, we exploit the available labels, e.g. we define TV speech data, the largest subpopulation, as main group and select observations from environmental sounds as "outliers". This is a very challenging approach: While, usually, speech and music recordings can be easily separated by the employment of suitable features, this does not hold for environmental sounds. Environmental sounds cover a wide range

of noises that sometimes have great similarities with speech data, sometimes with music and often they are just different.

The outlier detection approaches based on the two distance measures (OD and SD) 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 the two critical values for the orthogonal distance, c^k_{OD}, and for the score distance, c^k_{SD}. 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 PCs. An observation having an orthogonal or score distance larger than the respective critical value 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. The use of validation set in this context prevents potential overfitting of the estimated parameter, k_{opt}, to the characteristics of the training data. Finally, 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. Finally, we perform the evaluation on the test set using the parameters estimated on the training set.

We rescale the data to make 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. The assignment of the observations to training, validation, and test sets is done randomly and all evaluations are based on 100 replications. Since we have a larger pool of available data, independent training and validation data were constructed repeatedly. We think this is preferable over cross-validation, which would typically be used in situations where independent validation data are not available.

4 Experimental results

In this section we present the performed experiments which focus on the sensitivity of the investigated approaches with respect to the percentage of outliers, size of training and validation sets, and data characteristics. We report results in terms of AUC, TPR, FPR, number of PCs, and the corresponding standard errors (SE) over the 100 randomly initialized replications for each experiment.

Sensitivity to the percentage of outliers

For this evaluation we consider TV recordings (the biggest speech subgroup) as regular observations and we randomly select observations from both environmental and music samples as outliers. We split the data equally into training, validation, and test sets, corresponding to approximately 360 regular observations per set.

In a first experiment, we calculate the PC space using only the regular observations from the training set and we consider different percentage of outliers for the validation and test sets: 2%, 5%, and 10% of the main observations (see Table 1). The results show that clPCA performs similar or better than the robust PCA methods, while PCOut is capable of finding only approximately half of the outliers (indicated by the 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. SE remains at a very low level during the experiments for all methods.

In a second experiment, we consider that the training set is not free of outliers in order to explore their impact on the constructed PC space. The results show that the robust PCA methods clearly outperform clPCA. PCOut performs as poorly as in the first experiment. 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 remaining methods.

%	Method	Pure training set				Training set with outliers			
		AUC (SE_{AUC})	**k (SE_k)**	**TPR (SE_{TPR})**	**FPR (SE_{FPR})**	**AUC (SE_{AUC})**	**k (SE_k)**	**TPR (SE_{TPR})**	**FPR (SE_{FPR})**
10	clPCA	0.948 (0.002)	122 (1)	0.953 (0.005)	0.058 (0.003)	0.531 (0.005)	152 (16)	0.067 (0.011)	0.004 (0.001)
	GRID PCA	0.943 (0.002)	144 (2)	0.943 (0.004)	0.056 (0.002)	0.921 (0.003)	140 (1)	0.896 (0.006)	0.053 (0.001)
	ROBPCA	0.907 (0.002)	57 (2)	0.918 (0.007)	0.103 (0.004)	0.891 (0.004)	75 (3)	0.887 (0.007)	0.106 (0.005)
	OGK PCA	0.936 (0.002)	191 (4)	0.946 (0.005)	0.074 (0.003)	0.929 (0.003)	281 (4)	0.926 (0.006)	0.068 (0.002)
	PCOut	0.602 (0.006)	- (-)	0.516 (0.060)	0.311 (0.002)	0.624 (0.004)	- (-)	0.351 (0.007)	0.103 (0.002)
5	clPCA	0.946 (0.003)	136 (2)	0.949 (0.007)	0.057 (0.003)	0.557 (0.007)	205 (15)	0.124 (0.015)	0.009 (0.002)
	GRID PCA	0.942 (0.003)	163 (2)	0.937 (0.007)	0.054 (0.002)	0.933 (0.003)	152 (2)	0.923 (0.007)	0.057 (0.002)
	ROBPCA	0.916 (0.003)	51 (2)	0.929 (0.006)	0.097 (0.004)	0.918 (0.004)	54 (3)	0.928 (0.008)	0.092 (0.004)
	OGK PCA	0.931 (0.003)	168 (5)	0.931 (0.008)	0.070 (0.003)	0.925 (0.003)	203 (6)	0.918 (0.008)	0.067 (0.004)
	PCOut	0.609 (0.007)	- (-)	0.538 (0.016)	0.320 (0.006)	0.621 (0.006)	- (-)	0.359 (0.012)	0.118 (0.005)
2	clPCA	0.912 (0.007)	164 (4)	0.865 (0.016)	0.040 (0.003)	0.565 (0.009)	173 (15)	0.141 (0.019)	0.011 (0.002)
	GRID PCA	0.937 (0.007)	190 (4)	0.860 (0.015)	0.039 (0.003)	0.914 (0.006)	174 (4)	0.872 (0.013)	0.044 (0.002)
	ROBPCA	0.913 (0.005)	39 (3)	0.911 (0.010)	0.085 (0.005)	0.918 (0.005)	39 (3)	0.916 (0.011)	0.081 (0.005)
	OGK PCA	0.905 (0.007)	129 (6)	0.862 (0.015)	0.052 (0.004)	0.908 (0.007)	139 (7)	0.865 (0.016)	0.050 (0.004)
	PCOut	0.604 (0.009)	- (-)	0.531 (0.021)	0.323 (0.006)	0.615 (0.009)	- (-)	0.420 (0.022)	0.191 (0.011)

Table 1. Evaluation results for different percentage of outliers (%).
px

ROBPCA tends to select a considerably lower number of PCs than its counterparts. The achieved results in terms of AUC suggest that 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. Therefore, we take this into account and all further experiments consider training set containing outliers.

Sensitivity to the size of training, validation, and test sets

In this experiment, we investigate whether varying the size of training, validation, and test sets considerably influences the performance of the compared approaches. Again, we consider the biggest speech subgroup as main observations and we add 5% from the instances from 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 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.

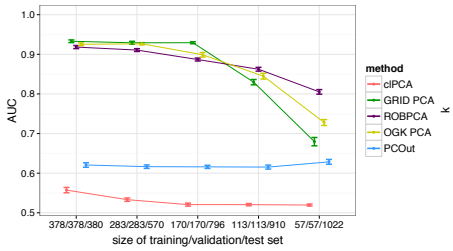

Figure 2. AUC

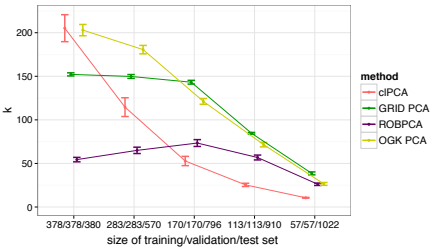

Figure 3. Number of PCs

Figure 4. Evaluation results for varying size of training, validation, and test sets.
px

Figure 4 shows the results of the evaluation in terms of AUC and number of PCs necessary to distinguish outliers from main observations. clPCA fails since the training set contains outliers. The performance of the robust PCA methods decreases 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

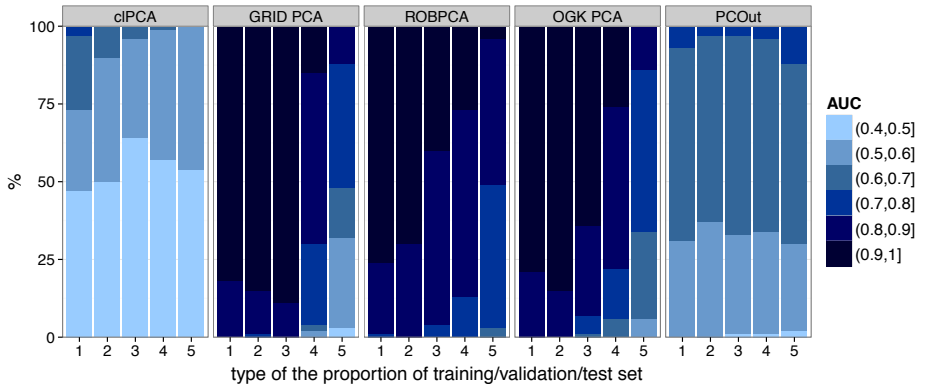

Figure 5. Detailed investigation of the resulting AUC during the replications for different partitions of training, validation, and test set (training/validation/test set) corresponding to the following size of sets: **1**: 378/378/380, **2**: 283/283/570, **3**: 170/170/796, **4**: 113/113/910, and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

px

more detailed investigation, we visualize the distribution of the resulting AUC during the 100 replications for each method. Figure 5 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 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 $(0.7, 0.4]$. Figure 3 shows that the number of PCs selected by ROBPCA is independent from the size of the sets and it tends to choose fewer 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 SE of the selected numbers of components.

Sensitivity to the size of the validation set

Our last experiment employed 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 indicated that some of the compared approaches perform well even if the size of validation set is reduced to 170 or 57 instances. These findings motivated us to explore how many observations in the validation set need to be labeled to achieve satisfying results. We fix training and test sets to the same size, 378 observations, and vary the number of observation in the validation set from 21 up to 378 instances. We simulate the biggest speech subgroup as the main observations and we add 5% from the other two audio groups as outliers. PCOut is not included to this experiment since it does not use a validation set.

Figure 6 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. In general, the number of PCs (see Figure 7) decreases with reducing the size of validation set and deviates during the replications. 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 music and perform the same experiment. The size of training and test set is fixed to 168 instances. Figure 8 and Figure 9 indicate very similar performance and ROBPCA outperforms the remaining methods in all investigated situations.

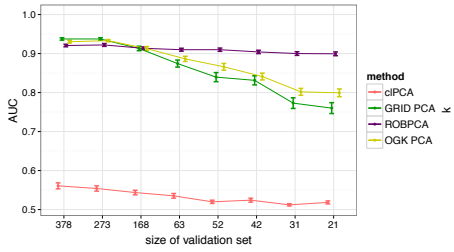

Figure 6. AUC

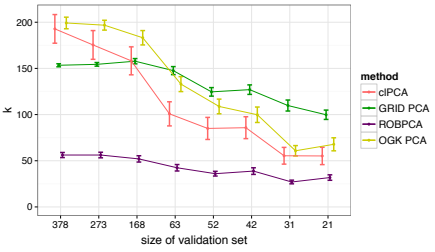

Figure 7. Number of PCs

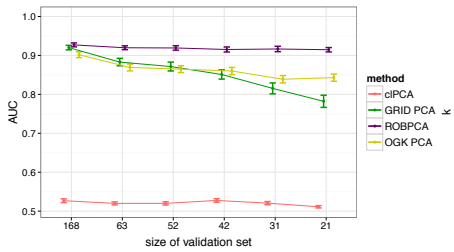

Figure 8. AUC

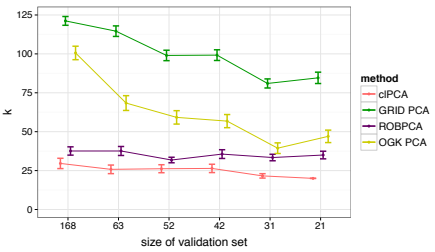

Figure 9. Number of PCs

Figure 10. Evaluation results for varying sizes of the validation set. Top row: main observations from speech data. Bottom row: music.

px

Sensitivity to the data characteristics

In this experiment we explore the sensitivity of the compared approaches to the underlying data characteristics with respect to varying data structures given by the different subpopulations in the audio dataset. 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 the corresponding main observations. We investigate the case of one majority subgroup present in the main observations and, in a next step, several subgroups. Figure 13 shows that the performance is slightly better when a single majority group is considered. Although ROBPCA and GRID PCA achieve a higher AUC, the results indicate that these two methods face difficulties in coping with multi-group data structures. clPca completely fails with AUC of 0.5. Additionally, the values for the the SE of AUC are considerably higher than in the previous experiments. Overall, there is no clear dependency between the number of PCs and the different multi-group data structure.

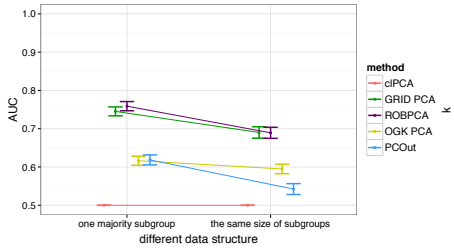

Figure 11. AUC

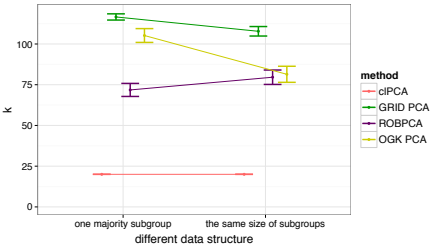

Figure 12. Number of PCs

Figure 13. Evaluation results for different data structures.
px

5 Discussion on critical values

The critical values are given by the quantiles of a χ^2 distribution for the SDs and the quantiles of the unknown distribution for the ODs which can be estimated by a robust Wilson-Hilferty approximation. Both critical values are based on the assumption of multivariate, normally distributed main observations. Those critical values are always an approximation since the distribution itself is estimated from the given observations. The central χ^2 distribution of the SDs and the non-central χ^2 distribution of ODs get distorted if the assumptions of normality are violated. In our experiments, we clearly observed data structures, which do not follow a normal distribution. We partly absorb this effect by using robust estimations. Therefore, the majority of observations can be properly modeled based on a normal distribution. To cope with the distorted distributions of the distances in addition to using robust estimations, we suggest to take advantage of the availability of validation data and to adjust the critical values. For this purpose, we can maximize the AUC performance for each fixed number of components, varying the critical values for SDs and ODs. Note, that the only meaningful critical values are the distances given by pre-labeled outliers. All other possible values will increase the FPR, without affecting the TPR. Thus, the necessary computational effort is very acceptable.

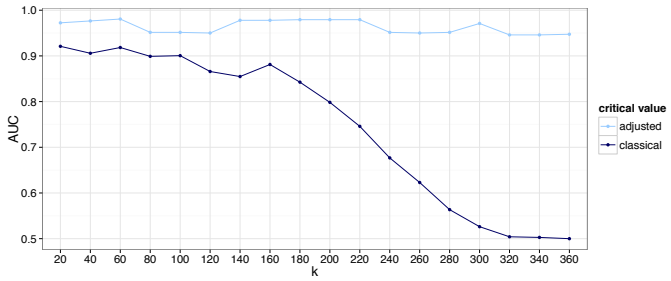

Figure 14. Comparison of AUC values depending on the number of components. While the quality of the classification for the classical critical values is highly depending on the chosen number of components, the adjusted critical values remains at an almost constant level.
px

The main benefit of this procedure is the resulting robustness towards the number of chosen PCs. Figure 14 shows this effect for ROBPCA for one example of speech main observations with 5% outliers. The experiment indicates that performing the outlier detection for a low number of PCs is sufficient. Thus, even though the adjustment needs computation time, the total computational effort decreases, since 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 with increasing number of observations. It can be easily shown that the adjustment will asymptotically always perform at least as good as the theoretical critical values with increasing numbers of validation observations. If the theoretical assumptions of multivariate normal distribution holds where observations with large Mahalanobis distance are classified as outliers, the adjustment converges to the provided theoretical critical value due to the law of large numbers. For any non-normal distribution it converges to the respective true critical value and, therefore, it outperforms the theoretical critical values, derived from false assumptions. However, for large number of observations, especially outlying observations, the adjustment converges. Thus, the method should only be used to analyze setups where enough outlying observations allow for a proper estimation of the ROC curve.

6 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 [8] is sensitive to the presence of outliers in the training data, we employed several, well-established robust PCA methods, such as GRID PCA [1], ROBPCA [6], OGK [11], and PCOUT [3], to better reveal outlying samples. We performed a thorough investigation of the sensitivity of the employed approaches with respect to different data properties, percentage 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 and 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 where few pre-labeled observations allow for the individual estimation of a proper number of PCs. Further utilization of pre-labeled observations is possible by adjusting the critical values for outlier detection if the observations do not follow a normal distribution. In such a situation, if the number of observations, especially outliers, is big enough, the adjustment can significantly improve the quality of the proposed procedures, providing a more robust set of critical values, which are able to cope with skewed distributions.

Acknowledgments

This work has been partly funded by the Vienna Science and Technology Fund (WWTF) through project ICT12-010 and by the K-project DEXHELPP through COMET - Competence Centers for Excellent Technologies, supported by BMVIT, BMWFW and the province Vienna. The COMET program is administrated by FFG.

Bibliography

[1] Croux, C., Filzmoser, P. and Oliveira, M. (2007) *Algorithms for Projection-Pursuit Robust Principal Component Analysis.* Chemometr. and Intell. Lab. Sys., **41**, 15:1–15:58.

[2] Croux, C. and Ruiz-Gazen, A. (2005) *High breakdown estimators for principal components: the projection-pursuit approach revisited.* Journal of Multivariate Analysis, **95(1)**, 206–226.

[3] Filzmoser, P., Maronna, R. and Werner, M. (2008) *Outlier Identification in High Dimensions.* Computational Statistics & Data Analysis, **52**, 1694–1711.

[4] Filzmoser, P., Serneels, S., Croux, C. and Van Espen, P. (2006) *Robust Multivariate Methods: The Projection Pursuit Approach* From Data and Information Analysis to Knowledge Engineering, 81–124.

[5] Gnanadesikan, R. and Kattenring, J. R. (1972) *Robust Estimates, Residuals, and Outlier Detection with Multiresponce Data.* Biometrics, **28**, 81–124.

[6] Hubert, M., Rousseeuw, P. and Vanden Branden, K. (2005) *ROBPCA: A New Approach to Robust Principal Component Analysis.* Technometrics, **47**, 64–79.

[7] Hubert, M., Rousseeuw, P. and Verboven, S. (2002) *A fast method for robust principal components with applications to chemometrics.* Chemometr. Intell. Lab., **60**, 101–111.

[8] Jolliffe, I.T. (2002) *Principal Component Analysis.* Springer.

[9] Li, G. and Chen, Z. (1985) *Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo.* Journal of the American Statistical Association, **80**, 759–766.

[10] Locantore, N. et al. (1999) *Robust principal component analysis for functional data.* Test, **8(1)**, 1–73.

[11] Maronna, R. and Zamar, R. (2002) *Robust estimates of location and dispersion for high-dimensional data sets.* Technometrics, **43**, 307–317.

[12] Pascoal, C., Oliveira, M., Pacheco, A. and Valadas, R. (2010) *Detection of outliers using robust principal component analysis: A simulation study.* Combining Soft Computing and Statistical Methods in Data Analysis, 499–507.

[13] Rousseeuw, P. J. (1984) *Least Median of Squares Regression.* Journal of the American Statistical Association, **79**, 871–880.

[14] Rousseew, P. and van Driessen, K. (1999) *A fast algorithm for the minimum covariance determinant estimator.* Journal of the American Statistical Association, **41**, 212–223.

[15] Sapra, K. S. (2010) *Robust vs. classical principal component analysis in the presence of outliers.* Applied Economics Letters, **17(6)**, 519–523.

[16] Serneels, S. and Verdonck, T. (2008) *Principal component analysis for data containing outliers and missing elements.* Computat. Statistics & Data Analysis, **52(3)**, 1712–1727.

[17] Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K. and Chang, L. (2003) *A novel anomaly detection scheme based on principal component classifier.* Tech. report, DTIC Document.

[18] Wall, M. E, Rechtsteiner, A. and Rocha, L. M. (2003) *Singular value decomposition and principal component analysis.* A practical approach to microarray data analysis.

[19] Xu, H., Caramanis, C. and Mannor, S. (2013) *Outlier-robust pca: The high-dimensional case.* Local journal of interesting topics research, **59(1)**, 546–572.

[20] Zimek, A., Schubert, E. and Kriegel, H.-P. (2012) *A survey on unsupervised outlier detection in high-dimensional numerical data.* Stat. Anal. and Data Mining, **5(5)**, 363–387.