Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 52 (2008) 1712 – 1727

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Principal component analysis for data containing outliers and missing elements

Sven Serneels[a,*], Tim Verdonck[b]

[a] ChemometriX Group, Department of Chemistry, University of Antwerp, Belgium
[b] Agoras Group, Department of Mathematics and Informatics, University of Antwerp, Belgium

Received 22 August 2006; received in revised form 21 May 2007; accepted 21 May 2007
Available online 24 May 2007

Abstract

Two approaches are presented to perform principal component analysis (PCA) on data which contain both outlying cases and missing elements. At first an eigendecomposition of a covariance matrix which can deal with such data is proposed, but this approach is not fit for data where the number of variables exceeds the number of cases. Alternatively, an expectation robust (ER) algorithm is proposed so as to adapt the existing methodology for robust PCA to data containing missing elements. According to an extensive simulation study, the ER approach performs well for all data sizes concerned. Using simulations and an example, it is shown that by virtue of the ER algorithm, the properties of the existing methods for robust PCA carry through to data with missing elements.

Keywords: Robustness; Principal component analysis; Robust PCA; Missing data; Incomplete data; Expectation maximization; Expectation robust

1. Introduction

Principal component analysis (PCA) is one of the key tools in multivariate statistical analysis. It aims at constructing components, each of which contain a maximal amount of variation from the data unexplained by the other components. The user thus hopes that the information in the data can be summarized into a few principal components, which is often the case in practice. Once the principal components have been determined, all further analysis can be carried out on them instead of on the original data, as they carry the relevant information in them. PCA is thus frequently considered a first step of a statistical data analysis which aims at compression of the data: decreasing their dimensionality without losing much information. Further analysis on the principal components can consist of various methods, such as clustering, discriminant analysis, regression, etc.

Principal components contain a maximal amount of variation from the data. In mathematics this means that principal components are defined according to a maximization criterion of variance. Let $\boldsymbol{X} \in \Re^{n \times p}$ be the data, consisting of n cases observed at p variables. Then the principal components $\boldsymbol{t}_i$ are defined as linear combinations of the

---

[*] Corresponding author at. Shell Global Solutions International B.V., Shell Research and Technology Centre, Amsterdam, P.O. Box 38000, 1030 BN Amsterdam, The Netherlands. Tel.: +31 20 6303856.

E-mail addresses: sven.serneels@shell.com (S. Serneels), tim.verdonck@ua.ac.be (T. Verdonck).

doi:10.1016/j.csda.2007.05.024

data $\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{p}_i$, where

$$\boldsymbol{p}_i = \arg \max_{\boldsymbol{a}} \{\mathrm{var}(\boldsymbol{X}\boldsymbol{a})\} \tag{1a}$$

under the constraints that

$$\|\boldsymbol{p}_i\| = 1 \quad \text{and} \quad \mathrm{cov}(\boldsymbol{X}\boldsymbol{p}_i, \boldsymbol{X}\boldsymbol{p}_j) = 0 \quad \text{for } j < i. \tag{1b}$$

Exact maximization of this criterion can be done by the Lagrange multiplier method and leads to the conclusion that the principal components are the eigenvectors of the variance–covariance matrix $\Sigma = n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ (here and elsewhere, we will assume the data to be centred).

Both the variance and the variance–covariance matrix are known to be sensitive to outliers. Hence, the same conclusion holds for PCA as a whole: it is a nonrobust method. A single bad outlier may cause that principal components are distorted so as to fit the outlier well, leading to bad interpretation of the results. Outliers can also cause the so-called *masking effect*: due to their presence, the model is distorted in such a way that based on the principal components, no outliers can be detected.

The sensitivity of principal components to outliers is well known and various robust alternatives to it have been proposed in literature. A topic which has, however, not yet been discussed in the context of robust principal components is how to deal with missing data. Especially in the biological and environmental sciences, missing data frequently occur. Data can be missing due to different reasons. In what follows, we will assume that the reason why a data point is missing is not related to its actual value, i.e. the data are *missing at random* (MAR) in the sense of Rubin (1976). Missing completely at random (MCAR) is a stronger hypothesis but we will assume that the data are at least MAR.

A good method to deal with data containing missing elements is the expectation maximization (EM) algorithm (Dempster et al., 1977). The EM algorithm basically consists of an iterative scheme where in each iteration two steps are carried out: (i) the missing elements are filled in by the values which they are expected to be (the *expectation step* or *E-step*) and (ii) the desired entity (e.g. the variance–covariance matrix) is estimated from the data in which missing elements have been filled in (called the *maximization step* or *M-step* if the estimates are obtained via maximum likelihood and the *robust estimation step* or *R-step* if the estimates are obtained by means of a robust estimation technique). Since the true values of the missing elements are unknown, the procedure is repeated until some convergence criterion is fulfilled. The EM algorithm has been applied to PCA (Walczak and Massart, 2001) on the one hand as well as to robust estimation of the variance–covariance matrix (Cheng and Victoria-Feser, 2002; Little, 1988) on the other hand (in the latter context called expectation-robust, ER).

In this article we investigate how the EM (or ER) approach to dealing with missing data can be extended to robust PCA. Two ways to solve this problem seem viable. On the one hand, it is possible to take the eigenvectors of a covariance matrix which has been estimated by the ER scheme. On the other hand, one can incorporate a robust PCA algorithm into the iterative EM scheme, thus obtaining a robust PCA method which can deal with missing elements. Either way, robust PCA for incomplete data always consists in some sense of inclusion of a robust estimator into an iterative scheme to estimate missing elements. Thus it is important to know the properties of the robust estimator used; in Section 2 we present a brief description of properties a robust estimator ideally should possess. In Section 3 the approach to robust PCA based on an ER covariance matrix is discussed, whereas in Section 4 we introduce the EM algorithm for robust PCA. As there is no agreement in the literature on how PCA should be robustified best, several robust PCA algorithms are being considered. In Section 4 we present an extensive simulation study which enables us to compare in an objective manner the different approaches. Finally, we give an example from the biological sciences.

2. Robustness properties of robust estimators

It is interesting to know which properties a robust method should have. These properties fall into three basic categories: properties related to the influence function, to the MaxBias curve and the statistical efficiency. The influence function (Hampel et al., 1986) is a tool which measures the effect an infinitesimally small amount of contaminated data has on an estimator as a function of the contaminated data's position in space. For an estimator to be robust, its influence function has to be bounded. However, note that in some special cases, this statement is not true: e.g. a location M-estimator with a bounded ψ function also has a bounded influence function, but a null breakdown point. A good property for a robust estimator is to have not only a bounded influence function, but also to have a smooth influence function. In practical terms this implies that the influence of contamination placed at z is approximately the same as the influence of a point of contamination placed at $z+\varepsilon$ with ε arbitrarily small. This property is also referred to as the *local shift sensitivity*: an estimator should have a small local shift sensitivity. Whereas the influence function measures infinitesimal robustness, the MaxBias curve (Rousseeuw, 1999) assesses global robustness. The MaxBias curve expresses how biased a robust estimator is with respect to the fraction of contaminated data, given that these are situated at the worst possible position in space. MaxBias curves all typically have an asymptote: there exists a fraction of contamination beyond which the estimator is totally unreliable and breaks down. This fraction is the *breakdown point* and cannot exceed 0.5 (except for estimators based on ranks or signs, see e.g. (Grize, 1978) or (Davies and Gather, 2005)). The breakdown point is the most cited property derived from the MaxBias curve, but the shape of the curve should also be considered: it is possible that an estimator breaks down at 0.5, but has at 0.2 a much higher bias than another estimator which breaks down at 0.3. In practice data containing 50% outliers seldomly occur so the latter estimator would be preferable for most applications. Finally, robust estimators invariably have a higher variance at the normal distribution than classical estimators. Depending on the design of the estimator, the increase in variance compared to the maximum likelihood (ML) estimator may or may not be drastic. A measure for the increase in variance is the *statistical efficiency* (or *efficacity*). The efficiency of an estimator is the sampling variance of the classical estimator, divided by the sampling variance of the robust estimator and lies between 0% and 100%.

3. Robust PCA for incomplete data based on an ER covariance matrix

As stated in the introduction, PCA corresponds to a spectral decomposition of the variance–covariance matrix as

$$\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{\mathrm{T}}, \tag{2}$$

where the matrix $\boldsymbol{P}$ contains as columns the eigenvectors $\boldsymbol{p}_i$ of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix where the diagonal elements λ_{ii} are the eigenvalues of $\boldsymbol{\Sigma}$ corresponding to $\boldsymbol{p}_i$. In order to construct a method for PCA which is robust and can deal with missing data, it suffices to obtain an estimate $\hat{\boldsymbol{\Sigma}}$ which fulfils both requirements, which can then be decomposed (2) in order to obtain the principal components.

The literature on robust covariance estimators for data containing missing elements is not abundant, but up to our knowledge, four approaches exist. The earliest proposal consists of inserting an M estimator into the EM algorithm (Little, 1988) (which is then called the ER algorithm). However, the M estimator for the covariance matrix used there is monotonic. A monotonic M estimator is an M estimator where the dispersion matrix is given by

$$n^{-1}\sum_{i=1}^{n} W[(\underline{x}_i-\hat{\boldsymbol{\mu}})\hat{\Sigma}^{-1}(\underline{x}_i-\hat{\boldsymbol{\mu}})^{\mathrm{T}}](\underline{x}_i-\hat{\boldsymbol{\mu}})^{\mathrm{T}}(\underline{x}_i-\hat{\boldsymbol{\mu}}) = \hat{\Sigma}, \tag{3}$$

where $\underline{x}_i$ are the rows of the data matrix $\boldsymbol{X}$, $\boldsymbol{\mu}$ is the location as a row vector and the function $W(t)t$ must be nondecreasing in t. Such M estimators are known to have a low breakdown point (i.e. the fraction of outliers the data may contain before the estimator yields unreliable results). The breakdown point of monotonic M estimators for the covariance matrix equals $1/(p+1)$ which implies that if the data are for instance, 100-variate, the estimator can only resist 1% of outliers in the data. To remedy this drawback, recently three other estimators have been proposed (Cheng and Victoria-Feser, 2002). The first proposal is inspired on results concerning robust covariance estimation, where it is noted that the robustness of an M estimator can be improved drastically if in the algorithm a more resistant estimator is used as a starting value for the iterative reweighting algorithm. Cheng and Victoria-Feser (2002) propose to use the minimum covariance determinant MCD estimator (Rousseeuw, 1985) with breakdown 0.5 for these purposes. They also note that alternatively, the MCD estimator could be extended to missing data as such, without being a part of the M algorithm. However, they prefer to use the MCD as a starting value for the M estimator as the MCD is known to have a rather poor statistical efficiency (Cheng and Victoria-Feser, 2002; Croux and Haesbroeck, 1999). Finally, Cheng and Victoria-Feser also propose to insert a translate-biweight S (TBS) (Rousseeuw and Yohai, 1984) estimator into the ER algorithm for missing data. Based on simulations and on an example from psychometrics they conclude that both estimators produce similar results. As the TBS estimator is slightly more complex from the mathematical point of view, we suggest to use the M estimator as a starting point for PCA. In our simulation study (Section 6) we include both the original M (Little, 1988) estimator and the M estimator with MCD ($k=0.5$) starting value (Cheng and Victoria-Feser, 2002). It is expected that the latter estimator outperforms the plain M estimator.

Finally we note that direct robust estimation of the covariance matrix requires many cases to be available with respect to the number of variables. In research fields such as chemometrics and bioinformatics, often data sets need to be treated where the number of variables vastly exceeds the number of cases, such that the methodology proposed in this section cannot be applied. For such undersampled problems we propose the strategy presented in Section 4: to include a robust PCA algorithm into the EM scheme.

4. Robust PCA for incomplete data based on insertion of a robust PCA algorithm into EM

As heeded before, the EM and ER algorithms basically consist of two parts which are executed until convergence: the missing values are filled in according to what they are expected to equal (the E-step), whereafter the (robust) estimation of the desired entity takes place (the M- or R-step). Assuming that the data structure is h dimensional (i.e. the data can be described well by h principal components), then the data are approximated well by $\hat{\boldsymbol{X}}_h = \boldsymbol{T}_h \boldsymbol{P}_h^{\mathrm{T}}$, where the matrices $\boldsymbol{T}_h$ and $\boldsymbol{P}_h$ contain the first h principal components and loadings, respectively. If the original data matrix $\boldsymbol{X}$ contains missing elements and is assumed to be well described by h principal components, then the corresponding elements of $\hat{\boldsymbol{X}}_h$ are viable estimates for the missing elements. On this principle the EM-PCA and ER-PCA algorithms can be based: initiate the missing elements by some means, do PCA, fill in the missing elements by the corresponding elements from $\hat{\boldsymbol{X}}_h$ and repeat this procedure till convergence.

In the classical case, Walczak and Massart (2001) use this strategy to perform PCA on data with missing elements. The first step is an initial estimation of each missing element by taking the sum of the means of the matching row and column and divide this by 2 (note that this strategy does only make sense for standardized data). Then the iterative part of the EM algorithm starts: alternatingly PCA is performed and the missing elements are filled in with the corresponding elements of $\hat{\boldsymbol{X}}_h$. As a convergence criterion, they take the squared relative difference between two vectors containing the successive estimates of the missing elements. If this difference is smaller than a certain tolerance limit to be chosen by the user, the algorithm stops.

Exactly the same strategy can be followed to construct an ER-PCA method which can cope with outlying cases and missing elements simultaneously. As in the classical algorithm, the missing elements first have to be filled in with some initial estimate. Taking the same starting value as in (Walczak and Massart, 2001) would lead to problems: if there are outliers in the data, most probably the column means will be affected by the outliers such that the initial estimate will be erroneous as well. In fact, using this strategy would lead to an increase in the percentage of outliers as some regular cases which contain missing elements, would become more outlying too. A first possible approach to obtain initial estimates of the missing values is by taking the means of the corresponding rows (again observe that standardization is necessary; the data have to be standardized on a robust scale such as the MAD). Secondly, one can also adapt the strategy from (Walczak and Massart, 2001) so as to be more robust: i.e. for each missing element a robust location estimator is applied to the corresponding column and row, whereafter the average of both values is filled in. As a robust location estimator one can opt for different choices: a median, a trimmed mean, an M estimator.

Once the initial estimates have been obtained, the iterative part of the ER algorithm is started, which is identical to the classical EM-PCA algorithm, except that now a robust PCA method is used to obtain the estimates for $\boldsymbol{T}_h$ and $\boldsymbol{P}_h$.

Finally, one also has to reconsider the convergence criterion. Missing elements determine convergence, but they can also belong to outlying cases, which do not have to be estimated well by the robust method. Robust PCA tries to fit the good data points well and the outliers badly, so that the updated values for these missing elements which are a part of an outlier, may differ significantly between two runs of the algorithm. Hence, only the missing elements belonging to normal cases should be taken into consideration for the convergence criterion. Estimates of missing elements which belong to cases whose distance in the score space exceeds the threshold ($\chi_{h,0.99}$) are omitted for the computation of the convergence criterion.

Our algorithm goes as follows:

- if $\boldsymbol{X}$ contains no missing elements
  (1) do robust PCA;
- else
  (1) obtain a first estimate $\acute{\boldsymbol{X}}^{(0)}$ where the missing elements are filled in with the initial estimates;
  (2) set $\ell = 1$;
  (3) until convergence, do:

(a) do robust PCA on $\acute{\boldsymbol{X}}_h^{(\ell-1)}$ (note that apart from the initial estimate $\acute{\boldsymbol{X}}^{(0)}$, $\acute{\boldsymbol{X}}_h$ is different for different complexities h), obtain scores $\boldsymbol{T}_h^{(\ell)}$, loadings $\boldsymbol{P}_h^{(\ell)}$ and eigenvalues $\boldsymbol{\Lambda}^{(\ell)}$;
(b) find missing elements which belong to outlying cases;
(c) fit $\boldsymbol{X}$ by $\hat{\boldsymbol{X}}_h^{(\ell)} = \boldsymbol{T}_h^{(\ell)} \boldsymbol{P}_h^{(\ell)\mathrm{T}}$;
(d) compare estimates of missing elements (which belong to 'good' cases) from $\hat{\boldsymbol{X}}_h^{(\ell)}$ with the corresponding (former) estimates from $\hat{\boldsymbol{X}}_h^{(\ell-1)}$;
(e) if convergence is not attained, construct a new estimate $\acute{\boldsymbol{X}}_h^{(\ell)}$ which contains the elements of $\boldsymbol{X}$ which were not missing and the elements of $\hat{\boldsymbol{X}}_h^{(\ell)}$ which correspond to the missing elements in $\boldsymbol{X}$;
(f) increase ℓ by 1 and return to (a).

In the algorithm described above, a robust PCA method is used. Several methods for robust PCA have been proposed in literature, all of which can be plugged into our algorithm. In order to evaluate which methods merit to be considered good candidates for the ER algorithm, we rely on theoretical and simulated results reported in the literature. Various comparisons, either based on theoretical properties, simulations or real data examples, have appeared. Not all of these comparisons favor the same method nor is there any comparative study in which all methods are treated, but nevertheless some methods can be ruled out due to results which have appeared. Hence, we will include those methods for robust PCA in our simulation study, which seem to be most likely to produce good results based on recent literature. We give a brief description of their known pros and cons in the following section.

5. Robust PCA estimators

The existing robust estimators for PCA are based on five different approaches: (i) taking the eigenvectors of a robust covariance matrix, (ii) projection pursuit (PP), (iii) a combination of both, (iv) robust subspace estimation and (v) spherical and elliptical PCA.

5.1. Robust PCA based on a robust covariance matrix

For robust PCA the same approach can be followed as in Section 3: construct a robust covariance matrix and then do a spectral decomposition on it. However, as these estimators can either not be computed or have bad properties for data where $p > n$, we will not take them in consideration here.

5.2. Projection pursuit

PP (Huber, 1985) is one of the most successful methods for robust estimation for high-dimensional data, as it basically reduces the multivariate robustness problem to many univariate ones via projection. For PCA, PP comes down to constructing all possible directions in the space spanned by the data variables (thus the directions are p vectors) and then numerically evaluating criterion (1a) for each of these vectors. The vector having the maximal value for the criterion is the first principal component. When the first principal component is found, the data are projected on the subspace which is orthogonal to it and the procedure is started over again to find the second PC. By this projection the second side condition in (1b) is respected. Successive PCs are found likewise.

In order to obtain robust principal components, it suffices to replace the variance in criterion (1a) by a robust counterpart. In practice a robust scale estimator is used. It is a good question which scale estimate should best be used. In the earliest proposal (Li and Chen, 1985), an M estimator for scale is chosen. However, the M estimators considered there do not have a high breakdown point. Simple high breakdown estimators of scale are the median absolute deviation and the Q_n estimator (Rousseeuw and Croux, 1994), which have a breakdown of 50%. Both these estimators have been examined for robust PCA and it was concluded that the Q_n estimator gave the best results (Croux and Ruiz-Gazen, 1996, 2005).

Of course, in practice not all directions can be constructed and PP algorithms always yield approximative solutions to the maximization problem. It is then a question how the algorithm should be designed such that it constructs a limited number of directions while yielding a good approximation to the true solution and without suffering from numerical imprecision. An early fast algorithm (Croux and Ruiz-Gazen, 1996) which only considered the n data points as possible directions was critiqued by Hubert et al. (2002) for numerical imprecision. These authors propose another PP algorithm based on a Householder transformation, called RAPCA. The algorithm is publicly available as a part of the LIBRA toolbox (Verboven and Hubert, 2005) for MATLAB, although we note that in this program, for $n > 40$ the A estimator of scale is used as projection index instead of Q_n (for a good overview of A estimators as well as some results concerning their efficiency, see Lax, 1985). As a counterpart to the RAPCA algorithm, Croux and Ruiz-Gazen have proposed a modification (Croux and Ruiz-Gazen, 2005) to their original algorithm (Croux and Ruiz-Gazen, 1996) (further referred to as the C-R algorithm) which no longer suffers from imprecision. The modified C-R algorithm is publicly available as a part of the TOMCAT toolbox (Daszykowski et al., 2007). Both algorithms yield approximations to the same estimator (except in these cases where RAPCA is based on the A estimator of scale). For several data sets from chemometrics, Stanimirova et al. (2004) found that the RAPCA method and the C-R algorithm yield virtually identical results.

The robustness properties of the PP estimator for PCA are well known. In fact, all properties depend on the scale estimator which is being used in the maximization criterion. It has been shown that the robust principal components and eigenvalues obtained by PP inherit the breakdown point (Li and Chen, 1985) and the efficiency (Croux and Ruiz-Gazen, 2005) of the scale estimator. Using the M estimator from the original proposal (Li and Chen, 1985) one obtains an efficient but not very robust method, whereas using the Q_n estimator one obtains a robust PCA method with a 50% breakdown point and en efficiency of 67% for the eigenvectors. To obtain better properties in both aspects, high-breakdown M estimators of scale (Croux, 1994) could be used as a projection index, but then at cost of high computation times and an unfavourable MaxBias function. The shape of the influence function also depends on the scale used; for the Q_n estimator the influence functions are bounded and smooth. For an underlying normal model, asymptotic normality of the estimators has been proven (Cui et al., 2003).

5.3. A hybrid method

Instead of using the PP approach to find the direction in space which maximizes the PCA criterion, it is also possible to use the PP method to find the points which are most outlying with respect to the bulk of the data. This approach is followed by Hubert et al. (2005). The outlyingness of a data point $\underline{x}_i$ (a row of $\boldsymbol{X}$) is defined as:

$$r(\underline{x}_i, \boldsymbol{X}) = \sup_{\boldsymbol{a}} \left\| \frac{\underline{x}_i \boldsymbol{a} - M(\boldsymbol{X}\boldsymbol{a})}{S(\boldsymbol{X}\boldsymbol{a})} \right\|, \tag{4}$$

where M and S denote location and scale estimators, respectively, and $\boldsymbol{a}$ is again a p variate direction. It is clear that the PP method can be used to determine r for each data point. Hubert et al. (2005) propose to use the univariate MCD estimators of location and scale for these purposes. Once r has been obtained for all data points, they omit the subset of $\eta n < n$ data points having the biggest values for r (η has to be chosen in advance). They then compute the classical principal components of this subset, project the data on these principal components and subsequently estimate the covariance matrix of these projected points by MCD (Rousseeuw, 1985). The eigenvectors of the last covariance matrix are then back-transformed to the original data dimensions, and are the robust principal components. This method is called ROBPCA by the authors, but as this general acronym for robust PCA can be found as a name for other robust PCA methods too, we will refer to this method as Hybrid PCA (HPCA).

The robustness properties of this method are also known; they derive from the robustness properties of the outlyingness (4) and from those of the MCD estimator. The influence function of HPCA is bounded but nonsmooth (Debruyne and Hubert, 2007). It even shows two distinct levels, which are due to the two steps in the estimation scheme (PP and MCD after projection). The breakdown point of the method equals η. Exact expressions for the efficiency are not available.

5.4. Robust subspace estimation

PCA is a bilinear least squares fit to the data. In fact the principal components satisfy the following minimization (Smilde et al., 2004):

$$(\boldsymbol{W}, \boldsymbol{P}) = \arg \min_{\boldsymbol{A},\boldsymbol{Q}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\boldsymbol{Q}^{\mathrm{T}}\|_{\mathrm{F}}, \tag{5}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm (side constraints are omitted). Eq. (5) can be seen as a minimization of a scale estimator (similar to regression), where in this case the standard deviation is taken as a scale estimator (similar to least squares). Note that finding lines of closest fit was indeed the objective when PCA was invented by Pearson (1901). By taking a robust scale instead of the standard deviation, the whole method is robustified. This idea was proposed recently by Maronna (2005), taking an M scale with robust starting value or an L (i.e. trimmed) scale as robust scale estimators. In simulations the method shows good performance; it is observed that the robust M estimator outperforms the L estimator. This robust PCA method is the only one which has also been extended to orthogonal regression. If an L estimator is used, the subspace found is found by similar computational means as the least trimmed squares (Rousseeuw and Leroy, 1987) estimator for regression. Hence this method is sometimes referred to as the LTS subspace PCA estimate. In this article we discuss both estimators proposed by Maronna (2005); hence we will refer to them as *MAR-L* and *MAR-M*. As the method has only recently been proposed, its robustness properties are not yet known, apart from some preliminary results suggesting that the method inherits the breakdown point of the scale estimator.

5.5. Spherical and elliptical PCA

Spherical and elliptical PCA (Locantore et al., 1998) consist of the following simple idea: project the data onto a unit sphere or an ellipse, and then carry out classical PCA on these projected points. Spherical PCA comes down to a spatial sign transformation of the data and yields consistent estimates; the elliptical PCA estimates are not consistent (see discussion to Locantore et al., 1998). Spherical PCA is very fast in the computational sense and very robust, but its efficiency and bias properties can be expected to be rather bad, like those of the spatial sign transform applied to partial least squares regression (Serneels et al., 2006).

5.6. Comparisons

As theoretical results allowing an objective comparison between the different robust PCA approaches are only partially available (especially the MaxBias curves have not been reported), one also has to rely on simulation results. However, also the simulation results reported do not cover the whole range of robust PCA methods. Hubert et al. (2005) compare their hybrid PCA method to RAPCA, spherical and elliptical PCA. They do not mention which algorithm for RAPCA was used but as in both of their settings n exceeded 40 we suppose the A estimator of scale was used as a projection index. Their results indicate a slightly better performance of HPCA over the three other methods. Maronna (2005) compares to spherical and elliptical PCA, to PP with the median absolute deviation and an M scale as a projection index (but not Q_n nor the A scale) as well as some estimates based on the robust covariance matrix. He finds that MAR-M outperforms MAR-L; moreover, both Maronna estimators seem to be least sensitive to the position in space where the outliers are situated. In terms of efficiency, the Maronna methods often (but not always) perform better than the PP-based methods. The Maronna methods are seen to outperform the PP methods for data where the eigenvalues decrease continuously, whereas the PP methods seem to perform slightly better if the eigenvalues significantly decrease after the first few (an exact threshold is not given). Hubert et al. (2005) also find this result: for their data PP (in the guise of RAPCA) is very precise for the first few eigenvalues but HPCA starts to outperform it later on. Finally, Engelen et al. (2005) provide a comparison of MAR-L to HPCA, where they find that according to the maxsub criterion (Hubert et al., 2005; Krzanowski, 1979) HPCA slightly outperforms MAR-L. However, we note that they fail to provide a comparison to the MAR-M estimator which is, according to Maronna, better than the MAR-L estimator. Even using the MAR-L estimator, in several of their settings the difference is already very small.

6. Simulation study

6.1. Design

Several simulation studies have appeared concerning robust PCA, however, not all of them are set up in the same way. In some simulation studies the data matrix $\boldsymbol{X}$ is generated from a multivariate normal distribution to which a certain percentage of outliers is added. A problem for such a design is that the correct number of principal components is not known. If it is estimated for a normally distributed $\boldsymbol{X}$, the optimal complexity may differ according to the method. The most straightforward way to set up a simulation for PCA is to construct data with a given complexity as $\boldsymbol{X}=\boldsymbol{T}_h\boldsymbol{P}_h^{\mathrm{T}}+\boldsymbol{E}=\boldsymbol{Y}+\boldsymbol{E}$, where the optimal complexity h is chosen in advance. This implies that $\boldsymbol{X}$ is chosen as a matrix of rank h plus unstructured noise.

The true principal components $\boldsymbol{T}_h$ are taken as n cases from the h-variate standard normal distribution

$$\boldsymbol{T}_h \sim N_h(\boldsymbol{0}_h, \boldsymbol{I}_h)$$

with $\boldsymbol{I}_h$ the $h \times h$ identity matrix and $\boldsymbol{0}_h$ the null vector with h components. To make sure that the components are uncorrelated, they are orthogonalized. Then the loadings $\boldsymbol{P}_h$ are defined as an orthogonal $p \times h$ matrix of uniformly distributed pseudorandom numbers. After having defined the principal components and the loadings, we can construct the classical data matrix $\boldsymbol{Y} = \boldsymbol{T}_h\boldsymbol{P}_h^{\mathrm{T}}$.

For adding outliers, we replace the last $100\alpha\%$ of observations of $\boldsymbol{Y}$ with data from another distribution. The contaminated data $\check{\boldsymbol{Y}}$ are constructed as $\check{\boldsymbol{Y}} = (\boldsymbol{Y}_{(1)}^{\mathrm{T}}\ \boldsymbol{Y}_{(2)}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{Y}_{(1)}$ contains the first $100(1-\alpha)\%$ of observations of $\boldsymbol{X}$ and $\boldsymbol{Y}_{(2)}$ is an $\alpha n \times p$ matrix taken from

$$\boldsymbol{Y}_{(2)} \sim N_h(\boldsymbol{15}_h, 8 * \boldsymbol{I}_h)$$

with $\boldsymbol{15}_h$ a vector containing h elements equal to 15. We thus used the multivariate normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{15}_h$ and covariance matrix $\boldsymbol{\Sigma} = 8 * \boldsymbol{I}_h$ for generating outlying values. Then standard normally distributed random noise (divided by 100) was added to our contaminated data matrix. The diagonal matrix $\boldsymbol{\Lambda}_h = (\boldsymbol{YP}_h)^{\mathrm{T}}(\boldsymbol{YP}_h)$ contains the first h "true" eigenvalues. These eigenvalues can now also be estimated from the (un)contaminated data after noise addition, by the different methods.

Missing values were added by randomly replacing elements of $\check{\mathbf{X}}$ by NaN's, so that in the simulation the MAR assumption on missingness holds. However, for highly contaminated situations (i.e. simulations for 30% outliers and 30% of missing data), this may cause problems as it is probable that the missing elements only occur in the part of the data generated from $\boldsymbol{Y}_{(1)}$ (regular cases), such that the missing elements are filled in with a median based on data which contains over 50% of outliers. That is because the 30% of missing values are added randomly to the whole data matrix, and may thus be more concentrated than 30% in individual columns. As the missing values are then filled in with a value affected by the outliers, this "first guess" data matrix (starting value for the algorithm) can contain for some columns even more than 50% outliers, such that the robust estimators will show breakdown. In order to avoid such situations of breakdown, which do not reflect the intrinsic performance of the robust estimators, we corrected the simulation setup as follows. After addition of the missing values, for each column the median of the nonmissing elements is estimated. If it exceeds 2, insertion of missing elements is repeated for the corresponding matrix until it does not exceed this threshold.

In a simulation study one wants to know how well a certain method operates and so a measure of performance is necessary. It is not clear which measure can be used best as a criterion of performance for robust PCA (the different articles cited above use different criteria). Of course a performance criterion either depends on the estimated principal components or on the eigenvalues. We propose here a combined measure of performance based on the eigenvalues and the percentage of explained variance. Because in our simulation setting the complexity is known, the first h components can replace the original p variables with little loss of information. The exact proportion of total variance explained by the principal components is equal to

$$\frac{\sum_{i=1}^{h} \boldsymbol{\lambda}_i}{\|\boldsymbol{Y}\|_{\mathrm{F}}}, \tag{6}$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm and $\boldsymbol{\lambda}_i$ are the diagonal elements of $\boldsymbol{\Lambda}$. The proportion explained by the fitted principal components can be calculated as

$$\frac{\sum_{i=1}^{h} \hat{\boldsymbol{\lambda}}_i}{\sum_{j=1}^{p} \hat{\boldsymbol{\lambda}}_j}, \tag{7}$$

where $\hat{\boldsymbol{\lambda}}_i$ are the estimated eigenvalues.

For our first performance criterion we compare (6) with 7

$$\left(\frac{\sum_{i=1}^{h} \hat{\boldsymbol{\lambda}}_i}{\sum_{j=1}^{p} \hat{\boldsymbol{\lambda}}_j}\right) \Big/ \left(\frac{\sum_{i=1}^{h} \boldsymbol{\lambda}_i}{\|\boldsymbol{Y}\|_{\mathrm{F}}}\right) \tag{8}$$

and this should be as close to 1 as possible.

This measure seems viable to compare the quality of the different techniques. However, it does not guarantee that the individual principal components are similar to those known from the simulation setup: it only guarantees that they *jointly* explain enough information from the data. In our setting, we have orthogonalized the scores $\boldsymbol{T}$ as well as the loadings $\boldsymbol{P}$, so that $\boldsymbol{T}^{\mathrm{T}}\boldsymbol{T}$ and $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}$ equal the identity matrix. Consequently the corresponding matrix $\boldsymbol{\Lambda}_h$, which contains the first h eigenvalues, also equals the identity matrix. In such a way always two principal components were created which were equally important. Assume for instance that the true eigenvalues are given by $\boldsymbol{\lambda}=(1, 1, 0, \ldots, 0)$. According to performance criterion (8) the vector (1.99, 0.01, 0.00001, 0.000001, ...) for example would be very acceptable, but we see immediately that the first component is much more important than the second, which does not at all correspond to the true situation. Hence this estimate is very bad although it would be deemed acceptable according to criterion (8). Therefore, apart from the previous criterion, in the current simulation study we also use the ratio of eigenvalues

$$\hat{\lambda}_1/\hat{\lambda}_2 \tag{9}$$

as an ancillary criterion. In our simulations, results for the second criterion are to be considered good if they are close to 1. In our opinion, an estimate which performs well for both criteria can be considered to yield an overall good performance.

We conducted simulations for the following methods for robust PCA suitable for the analysis of incomplete data:

(1) RAPCA: PP based on a reflection algorithm;
(2) HPCA: combine PP and MCD;
(3) CRPCA: direct PP algorithm;
(4) MAR-M: based on robust subspace estimation (M-estimator for scale);
(5) MAR-L: based on robust subspace estimation (L-estimator for scale);
(6) CV-M: an ER covariance matrix (Huber M-estimator as starting value);
(7) CV-MCD: an ER covariance matrix (MCD as starting value).

Also the classical PCA method was always executed (in the EM algorithm) to compare its results with our robust alternatives. All simulations were run in MATLAB 7.0 (The MathWorks, Natick, MA). The algorithms for RAPCA and HPCA were taken from the LIBRA M file library (Verboven and Hubert, 2005) whereas the algorithm for CRPCA was taken from the TOMCAT toolbox (Daszykowski et al., 2007). All remaining algorithms were implemented by the authors. The ER algorithm was initiated by filling in the missing elements with the average of the corresponding row and column medians.

All simulations were performed on data sets containing 5%, 10%, 20% and 30% of missing values. This process was repeated for levels of outlier fractions of 5%, 10% and 30%, respectively. The simulations were repeated 500 times; the means of these 500 runs are reported.

Some of the methods allow to change the fraction of outliers they can resist. The greater this value, the more robust, but also the less efficient the method is. Because not all considered methods are able to change such a parameter (for example for both PP techniques no such variable exist) for the methods which do take a modifiable parameter we have used the standard settings as reported by the authors or provided in the different toolboxes. One can argue that for low levels of contamination, the estimators which allow adaptation of a parameter could be tuned to be more efficient. However, we have chosen this value for the tuning parameter such that the same settings can be used throughout all simulations (if a value of 25% is used the tuneable estimators would break down for 30% of outliers). Note that in practice the number of outliers is usually not known in advance.

Data size has been reported to affect the results of robust estimators (see also previous section). Hence we have repeated the simulations for different data dimensions:

- configuration A: $n = 100$, $p = 5$, $h = 2$;
- configuration B: $n = 40$, $p = 10$, $h = 2$;
- configuration C: $n = 40$, $p = 200$, $h = 2$.

Data configurations have been chosen diligently. For the HPCA method, it is known that its robustness properties (e.g. that its breakdown point equals η) only hold if the data dimensions satisfy the following requirement (Hubert et al., 2005). Let $k = n\lceil 1-\eta\rceil$. Then for HPCA its robustness properties hold if $k > \lfloor (n+h+1)/2 \rfloor$. In order to compare the

Table 1
Simulation results for the different methods for uncontaminated data of size 100 × 5

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.97	1.2	0.96	1.2	0.93	1.5	0.90	1.9
HPCA	0.98	1.2	0.98	1.3	0.95	1.6	0.92	2.1
CRPCA	0.97	1.2	0.96	1.2	0.93	1.5	0.90	1.9
MAR-M	0.96	1.2	0.94	1.2	0.91	1.4	0.87	1.6
MAR-L	0.94	1.3	0.92	1.3	0.87	1.5	0.83	1.7
CV-M	0.97	1.1	0.94	1.2	0.86	1.2	0.80	1.3
CV-MCD	0.97	1.1	0.94	1.2	0.86	1.2	0.80	1.3
Classical	0.99	1.0	0.99	1.0	0.99	1.0	0.99	1.1

Table 2
Simulation results for the different methods for uncontaminated data of size 40 × 200

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.95	1.2	0.95	1.2	0.94	1.2	0.93	1.3
HPCA	0.98	1.4	0.97	1.5	0.97	1.6	0.96	1.7
CRPCA	0.94	1.2	0.93	1.2	0.92	1.2	0.91	1.3
MAR-M	0.94	1.3	0.91	1.3	0.85	1.4	0.79	1.5
MAR-L	0.92	1.6	0.89	1.6	0.84	1.6	0.78	1.7
Classical	0.98	1.0	0.98	1.0	0.98	1.0	0.98	1.0

methods fairly, the data configurations have been chosen so that the condition is satisfied. However, one has to keep in mind that the HPCA method may behave differently than what is observed here for data with a very limited number of cases.

6.2. Results

The results of each of these simulation settings are reported in a set of tables. Means over 500 simulations are reported. In each table, results for the different methods are presented with respect to varying percentages of missing data. Tables correspond to the three data sizes considered and to the percentages of outliers considered. In order to save space in the article, results for data configurations without outliers are only shown for configurations A and C, as in this case simulation results for all data configurations showed the same trend.

At first it can be observed, as expected, that the classical estimator yields the best results if the data are not contaminated (see Tables 1 and 2; recollect that the theoretically optimal value for all numbers reported in Tables 1 through 10 is 1). Applying any of the robust estimators to uncontaminated data with missing elements seems to be safe as all of these yield acceptable and quite comparable results. The three PP-based methods seem to be best according to criterion (8) whereas the Victoria-Feser ER estimates perform generally best according to criterion (9). The remaining estimators are in between and perform well according to both criteria. However, in our opinion the differences observed here are too small so as to draw a conclusion concerning which robust estimator is most efficient for uncontaminated data.

As soon as data are contaminated, classical PCA breaks down. This can be seen from the results for data contaminated with 5% of outliers, shown in Tables 3–5.It can be seen that all robust methods provide reliable results. When they can be computed, the CV methods seem to perform best according to criterion 2, whereas the PP-based methods perform best according to criterion 1. The MAR-M method provides a compromise. All methods perform worse as the percentage of missing values increases but even in the case of 30% of missing values, all robust methods are still reliable, an indication that the ER algorithm proposed here can deal with outliers and missing values simultaneously.

Table 3
Simulation results for the different methods for 5% contaminated data of size 100 × 5

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.97	1.2	0.96	1.2	0.93	1.4	0.90	1.8
HPCA	0.98	1.2	0.98	1.3	0.95	1.5	0.92	2.0
CRPCA	0.97	1.2	0.96	1.2	0.93	1.5	0.90	1.9
MAR-M	0.96	1.2	0.94	1.2	0.91	1.4	0.87	1.6
MAR-L	0.95	1.3	0.92	1.3	0.88	1.4	0.84	1.7
CV-M	0.97	1.2	0.94	1.2	0.85	1.2	0.78	1.3
CV-MCD	0.97	1.2	0.94	1.2	0.85	1.2	0.78	1.3
Classical	0.98	36.7	0.97	20.1	0.94	9.5	0.92	6.9

Table 4
Simulation results for the different methods for 5% contaminated data of size 40 × 10

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.96	1.3	0.94	1.3	0.89	1.5	0.84	1.8
HPCA	0.98	1.6	0.97	1.7	0.93	2.0	0.89	2.4
CRPCA	0.96	1.3	0.94	1.3	0.89	1.5	0.84	1.8
MAR-M	0.95	1.3	0.92	1.4	0.87	1.5	0.81	1.6
MAR-L	0.94	1.6	0.91	1.6	0.84	1.7	0.78	1.8
CV-M	0.95	1.3	0.88	1.3	0.77	1.4	0.78	1.4
CV-MCD	0.95	1.3	0.88	1.3	0.77	1.4	0.78	1.4
Classical	1.00	44.1	1.00	25.8	1.00	11.3	1.00	6.5

Table 5
Simulation results for the different methods for 5% contaminated data of size 40 × 200

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.95	1.3	0.95	1.3	0.94	1.3	0.93	1.4
HPCA	0.97	1.4	0.97	1.5	0.97	1.6	0.96	1.6
CRPCA	0.94	1.3	0.94	1.3	0.93	1.3	0.91	1.4
MAR-M	0.94	1.3	0.91	1.3	0.85	1.4	0.79	1.5
MAR-L	0.93	1.6	0.90	1.6	0.84	1.6	0.78	1.7
Classical	1.00	23.6	1.00	14.4	1.00	7.8	1.00	5.2

For 10% of outliers, results are shown for data configurations A and C (the simulations for configuration B confirmed the trend observed in configurations A). The results are shown in Tables 6 and 7.One could draw the same conclusions as for 5% contaminated data: the classical estimator breaks down and the robust ones are reliable. From this percentage of outliers on, it is also clear that also in the context of missing values, the Maronna M estimator (slightly) outperforms the Maronna L estimator, as was reported for complete data by Maronna (2005). The Maronna methods seem to depend more strongly on data dimensions as they yield inferior results for the high-dimensional data configuration C (see Tables 5 and 7). In this case it is clear that the PP and hybrid PP methods (RAPCA, HPCA, CRPCA) are preferable, with slight but not very significant differences among them. Again, the ER algorithm gives a good estimation of the principal components even for a high percentage of missing values.

Finally, for 30% contaminated data, results for all data configurations are shown (see Tables 8–10).At first, it can be observed that the CV methods cannot cope with 30% of outliers as (viz. Tables 8 and 9). Here we retrieve some recent

Table 6
Simulation results for the different methods for 10% contaminated data of size 100 × 5

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.97	1.2	0.96	1.2	0.93	1.5	0.90	1.9
HPCA	0.98	1.2	0.97	1.3	0.95	1.5	0.92	1.9
CRPCA	0.97	1.2	0.96	1.3	0.93	1.5	0.90	1.9
MAR-M	0.96	1.2	0.94	1.2	0.90	1.4	0.87	1.6
MAR-L	0.95	1.3	0.92	1.3	0.88	1.5	0.83	1.7
CV-M	0.96	1.3	0.93	1.4	0.84	1.5	0.77	1.5
CV-MCD	0.96	1.3	0.93	1.4	0.84	1.5	0.77	1.5
Classical	0.97	33.2	0.95	19.7	0.92	10.8	0.87	7.2

Table 7
Simulation results for the different methods for 10% contaminated data of size 40 × 200

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.95	1.3	0.95	1.3	0.94	1.3	0.93	1.4
HPCA	0.97	1.4	0.97	1.5	0.96	1.5	0.96	1.6
CRPCA	0.94	1.3	0.94	1.3	0.93	1.3	0.90	1.4
MAR-M	0.94	1.3	0.91	1.3	0.84	1.4	0.78	1.5
MAR-L	0.93	1.6	0.90	1.6	0.83	1.6	0.77	1.6
Classical	0.96	38.8	0.94	23.8	0.90	12.8	0.85	8.3

Table 8
Simulation results for the different methods for 30% contaminated data of size 100 × 5

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.97	1.3	0.95	1.3	0.92	1.5	0.89	1.9
HPCA	0.97	1.2	0.95	1.2	0.92	1.4	0.89	1.6
CRPCA	0.97	1.2	0.96	1.3	0.92	1.5	0.89	1.8
MAR-M	0.95	1.2	0.91	1.3	0.84	1.4	0.76	1.5
MAR-L	0.95	1.2	0.92	1.2	0.85	1.4	0.78	1.5
CV-M	0.98	21.1	0.97	24.1	0.90	13.7	0.86	10.6
CV-MCD	0.98	21.1	0.97	24.1	0.90	13.7	0.86	10.6
Classical	0.96	36.0	0.94	22.1	0.89	12.4	0.84	8.1

results concerning the CV estimates stating that they have a maximal breakdown point of 30% (Copt and Victoria-Feser, 2003). The starting value of the algorithm (M or RM) does not seem to improve results. Note that for complete data, it is known that for computing monotonic M estimators, the starting value only influences the number of iterations, not the estimate itself. These results seem to carry through to incomplete data.

Apart from the breakdown of the CV estimator, it can be observed that all robust estimators resist the high degree of contamination well (see Table 8). As is the case for 10% of contamination, the Maronna methods depend most on data dimensions and perform inferior to the methods based on PP as the ratio np^{-1} decreases (see Tables 9 and 10). Nonetheless their decrease in performance is still far from breakdown behaviour.

Table 9
Simulation results for the different methods for 30% contaminated data of size 40 × 10

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.95	1.5	0.93	1.5	0.87	1.6	0.82	1.9
HPCA	0.96	1.4	0.93	1.4	0.89	1.5	0.84	1.7
CRPCA	0.95	1.5	0.93	1.5	0.87	1.6	0.82	1.9
MAR-M	0.93	1.4	0.88	1.5	0.77	1.5	0.66	1.6
MAR-L	0.94	1.4	0.89	1.5	0.78	1.5	0.68	1.7
CV-M	0.96	18.3	0.94	25.7	0.95	50.0	0.98	99.4
CV-MCD	0.96	18.3	0.94	25.7	0.95	50.0	0.98	99.4
Classical	0.96	38.7	0.93	24.5	0.87	13.9	0.81	9.0

Table 10
Simulation results for the different methods for 30% contaminated data of size 40 × 200

Missings	0.05		0.10		0.20		0.30	
Method/criterion	(8)	(9)	(8)	(9)	(8)	(9)	(8)	(9)
RAPCA	0.95	1.4	0.95	1.4	0.94	1.4	0.92	1.5
HPCA	0.95	1.3	0.93	1.3	0.89	1.4	0.85	1.5
CRPCA	0.94	1.4	0.93	1.4	0.92	1.5	0.90	1.6
MAR-M	0.92	1.4	0.86	1.4	0.75	1.4	0.62	1.5
MAR-L	0.92	1.4	0.87	1.4	0.75	1.4	0.63	1.5
Classical	0.93	89.8	0.90	55.3	0.82	29.4	0.74	18.6

7. Example

In a recent study (Garamszegi et al., 2005) data were collected to assess the male sexual expression of barn swallows (*Hirundo rustica*). Two data sets were collected, one relating to physical properties and one relating to the birds' singing capacities. Although preliminary results are available as regards the quantitative relation which is supposed to exist between both blocks (Garamszegi et al., 2005), further statistical analysis may be expected to provide more insight into the relation between the song and physical characteristics. In the current example we will limit ourselves to the data set describing the birds' physical properties.

The data consist of 43 samples measured at 13 different variables which describe some properties which are in the biological sense supposed to be related to male sexual expression. The variables are of a heterogeneous character, e.g. two variables are included which relate to the birds' keel and one which expresses the tail length. But also a variable about the number of chewing lice (*Hirundoecus malleus*) present on the birds and a variable describing the hematocrit level are measured. Not all variables could be measured at all birds such that missing data are present.

Normal distribution along the variables can hardly be expected: at first the variable counting the chewing lice can be expected to be Poisson distributed rather than normal. Moreover, birds are not a preparate made in a laboratory and therefore a wide natural variation exists. Drawing a sample of over 40 birds likely leads to the presence of some outliers in the data. The goal of a biological analysis is to detect the general trend in the data rather than the properties of the outliers (although the latter are interesting in their own right).

We will show the effect of robust PCA in order to identify the outliers. At first the data were centred and scaled to unit variance. For the classical analysis, scaling was done using the classical standard deviation whereas in the robust case the median absolute deviation was used. Both classical PCA and Maronna PCA based on an M estimator of scale (MAR-M) with robust starting value were performed on the data. In order to capture the information in the data fully, MAR-M needs seven principal components whereas PCA requires eight components (based on a scree plot). However, for the sake of simplicity, we will not go into detail considering the individual (pseudo)eigenvalues and principal components, but show the results of a PC1 vs. PC2 bi-plot. The scores plots are shown in Fig. 1. Classical

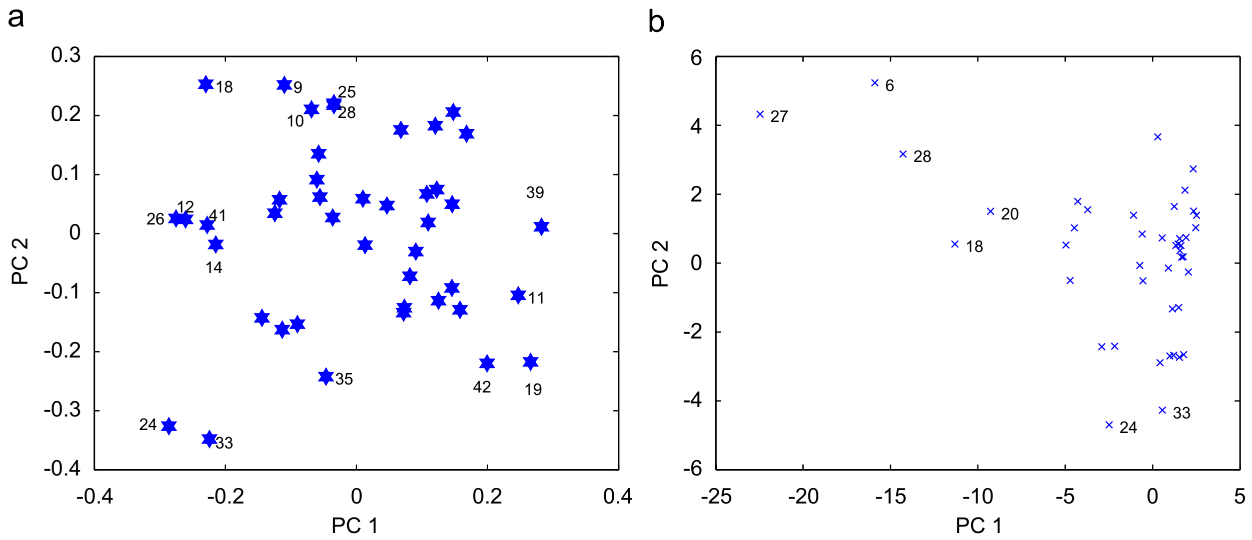

Fig. 1. Scores plots of PC1 vs. PC2 for the swallow data, constructed based on PCA (a) and MAR-M (b) included in the respective versions of the EM algorithm to deal with missings.

PCA shows a rather blurry image where points are almost spread out over the whole space. Two points are clearly identified as outliers: points 24 and 33. The remaining points do not appear as clear outliers, although some points may be considered slightly more excentrical than the bulk of the data. On the contrary, the scores plot based on MAR-M robust PCA clearly identifies several points as outliers, most strikingly points 27, 6, 28, 24 and 33. Moreover, points 24 and 33 are outlying in the direction of PC2 whereas points 27, 6 and 28 are outlying in the direction of both PCs. From the biological point of view, the outlyingness of each of these points, as well as the reason why two types of outliers can be observed, can be interpreted. Points 27, 6 and 28 are samples of birds which have an exceptionally high amount of feather lice. Points 24 and 33 are birds which have exceptionally small keels. Classical PCA is thus only able to identify the latter two outliers. For the other outliers, the masking effect occurs so that they can no longer be identified. To conclude, we note that for some of the points which are slightly excentrical in the classical scores plot, no biological interpretation for their possible outlyingness can be given.

The example shows that the advantages of robust PCA over classical PCA can readily be transposed to data with missing elements using the methodology described in this article. Some of the other robust PCA approaches for incomplete data could have been used here as well, but the goal of this example is to show that correct robust inference can be drawn for data with missing elements. By choosing a different approach of the several robust methods for incomplete data proposed in this article, one does not obtain identical results (as the methods are intrinsically different), but similar conclusions are drawn. We applied the CV-MCD, HPCA and CPCA methods and concluded that all were capable of detecting the outliers mentioned above.

8. Summary and conclusions

In this article, we have provided methodology to perform PCA on data which contain outlying cases and missing elements. We have provided two approaches to tackle the problem. At first, we suggested to take the eigendecomposition of a covariance matrix estimate which can deal with such data. However, this approach does not apply to data where the number of variables exceeds the number of cases. Therefore, we developed a new ER algorithm, based on the EM algorithm for classical PCA as proposed by Walczak and Massart (2001). The algorithm can treat missing elements and can accommodate different robust PCA approaches.

In a simulation study, we investigated the properties of the different approaches. At first, we observed that even for $n > p$, the approach based on an eigendecomposition of an ER covariance matrix did not significantly outperform our ER algorithm. The different robust PCA methods did not provide very distinct results when included in our algorithm. Only for highly (30%) contaminated data some of the approaches can be ruled out. Our recommendation is thus to use the ER algorithm in combination with either Croux PP-PCA (Croux and Ruiz-Gazen, 2005), RAPCA (Hubert et al., 2002), HPCA (Hubert et al., 2005) or Maronna PCA based on an M scale (Maronna, 2005). Slight differences between these methods have been highlighted in the simulation study.

The basic conclusion to draw is that all results for robust estimators carry through to incomplete data if these estimators are plugged into our ER algorithm. One of these well known properties is that robust estimators do not suffer from the masking effect, i.e. they yield a proper outlier detection regardless of the data configuration. In an example from the biological sciences, it is illustrated that the classical EM-PCA does not yield correct outlier detection whereas robust ER-PCA does. For these purposes it does not really matter a lot which robust PCA method to choose. The results shown are obtained with the Maronna PCA based on an M estimate of scale, but equally informative outlier detection could have been obtained by plugging in e.g. the HPCA or Croux PCA methods into the ER algorithm.

Acknowledgements

Research financed by a Ph.D. Grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). The authors thank Dieter Heylen for providing the data used in Section 6. We also thank the reviewers for their helpful and constructive comments.

References

Cheng, T.-S., Victoria-Feser, M.-P., 2002. High-breakdown estimation of multivariate mean and covariance with missing observations. British J. Math. Statist. Psych. 55, 317–335.

Copt, S., Victoria-Feser, M.-P., 2003. Fast algorithms for computing high breakdown covariance matrices with missing data. Cahiers du département d'économétrie août 2003. Faculté des sciences économiques et sociales. Université de Genève, Geneva, Switzerland.

Croux, C., 1994. Efficient high-breakdown M-estimators of scale. Statist. Prob. Lett. 19, 371–379.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J. Multivariate Anal. 71, 161–190.

Croux, C., Ruiz-Gazen, A., 1996. A fast algorithm for robust principal components based on projection pursuit. In: Prat, A. (Ed.), COMPSTAT: Proceedings in Computational Statistics. Physica, Heidelberg, pp. 211–216.

Croux, C., Ruiz-Gazen, A., 2005. High breakdown estimators for principal components: the projection-pursuit approach revisited. J. Multivariate Anal. 95, 206–226.

Cui, H., He, X., Ng, K.W., 2003. Asymptotic distributions of principal components based on robust dispersions. Biometrika 90, 953–966.

Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P.J., Croux, C., Walczak, B, 2007. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. Chemometr. Intell. Lab. Syst. 85, 269–277.

Davies, L.P., Gather, U., 2005. Breakdown and groups. Ann. Statist. 33, 977–1035.

Debruyne, M., Hubert, M., 2007. The influence function of Stahel–Donoho type methods for robust covariance estimation and PCA. Scand. J. Statist., submitted for publication.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood for incomplete data via the EM algorithm (with discussions). J. Roy. Statist. Soc. Ser. B 39, 1–38.

Engelen, S., Hubert, M., Vanden Branden, K., 2005. A comparison of three procedures for robust PCA in high dimensions. Austr. J. Statist. 34, 117–126.

Garamszegi, L.G., Heylen, D., Møller, A.P., Eens, M., de Lope, F., 2005. Age-dependent health status and song characteristics in the barn swallow. Behavioral Ecology 16, 580–591.

Grize, Y.L., 1978. Robustheitseigenschaften von Korrelationsschätzungen. Diplomarbeit, Eidgenössische Technische Hochschule (ETH). Zürich, Switzerland.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley, New York.

Huber, P., 1985. Projection pursuit. Ann. Statist. 13, 435–475.

Hubert, M., Rousseeuw, P.J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemometr. Intell. Lab. Syst. 60, 101–111.

Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: a new approach to robust principal components analysis. Technometrics 47, 64–79.

Krzanowski, W.J., 1979. Between-groups comparison of principal components. J. Amer. Statist. Assoc. 74, 703–707.

Lax, D.A., 1985. Robust estimators of scale: finite-sample performance in long-tailed symmetric distributions. J. Amer. Statist. Assoc. 80, 736–741.

Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo. J. Amer. Statist. Assoc. 80, 759–766.

Little, R.J.A., 1988. Robust estimation of the mean and covariance matrix from data with missing values. Appl. Statist. 37, 23–38.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L., 1998. Principal component analysis for functional data. Test 8, 1–73.

Maronna, R., 2005. Principal components and orthogonal regression based on robust scales. Technometrics 47, 264–273.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2, 559–572.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W., (Eds.), Mathematical Statistics and Applications, vol. B. Reidel, Dordrecht, pp. 283–297.

Rousseeuw, P.J., 1999. Maxbias Curve. In: Kotz, S., Read, C., Banks, D., (Eds.), Encyclopedia of Statistical Sciences, Update vol. 3. Wiley, New York, pp. 441–443.

Rousseeuw, P.J., Croux, C., 1994. Alternatives to the median absolute deviation. J. Amer. Statist. Assoc. 88, 1273–1283.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P.J., Yohai, V.J., 1984. Robust regression by means of S-estimators. In: Franke, J.W., Hardle, P.J., Martin, R.D., (Eds.), Robust and Nonlinear Time Series Analysis. Springer, New York, pp. 256–272.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Serneels, S., De Nolf, E., Van Espen, P.J., 2006. Spatial sign pre-processing: a simple way to impart moderate robustness to multivariate estimators. J. Chem. Info. Model. 46, 1402–1409.

Smilde, A.K., Geladi, P., Bro, R., 2004. Multi-way Analysis with Applications in the Chemical Sciences. Wiley, Chichester, UK.

Stanimirova, I., Walczak, B., Massart, D.L., Simeonov, V., 2004. A comparison between two robust PCA algorithms. Chemometr. Intell. Lab. Syst. 71, 83–95.

Verboven, S., Hubert, M., 2005. LIBRA: a MATLAB library for robust analysis. Chemometr. Intell. Lab. Syst. 75, 127–136.

Walczak, B., Massart, D.L., 2001. Dealing with missing data. Part I. Chemometr. Intell. Lab. Syst. 58, 15–27.