Available online at www.sciencedirect.com
ScienceDirect
Talanta 72 (2007) 172–178
Talanta
www.elsevier.com/locate/talanta

Dealing with missing values and outliers in principal component analysis

I. Stanimirova, M. Daszykowski, B. Walczak *

Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland

Received 19 June 2006; received in revised form 3 October 2006; accepted 5 October 2006
Available online 7 November 2006

Abstract

An efficient methodology for dealing with missing values and outlying observations simultaneously in principal component analysis (PCA) is proposed. The concept described in the paper consists of using a robust technique to obtain robust principal components combined with the expectation maximization approach to process data with missing elements. It is shown that the proposed strategy works well for highly contaminated data containing different amounts of missing elements. The authors come to this conclusion on the basis of the results obtained from a simulation study and from analysis of a real environmental data set.

Keywords: Expectation maximization approach; Robust PCA; Missing elements

1. Introduction

Principal component analysis (PCA) is one of the most used tools in chemometrics thanks to its very attractive properties. It allows relatively easy projecting of data from a higher to a lower dimensional space and then reconstructing them without any preliminary assumptions about the data distribution. As a least squares method, PCA is optimal in terms of mean squared error and its parameters are obtained from the data directly. However, despite these features, PCA is known to posses some shortcomings. One is that it is strongly affected by the presence of outliers, i.e. objects exhibiting far different values for some of the measured variables in comparison with the majority of objects. Therefore, the obtained principal components will not describe the majority of the data well and one cannot get a proper insight into the data structure. A way to deal with this problem is to remove the outlying objects observed on the score plots and to repeat the PCA analysis again. Another, more efficient way is to apply a robust, i.e. not sensitive to outliers, variant of PCA. Several robust PCA approaches have been proposed and their robust properties have been extensively tested over the last years. They are either based on projection pursuit approach [1–6] or on obtaining robust estimates of the covariance matrix [7–10] or on a combination of both [11].

Another shortcoming of the classical approach to PCA is that it fails to process missing elements, which frequently occur in the experimental data. What is often done in this situation is to delete the experimental variable (data column) or object (data row) containing the missing elements or to replace the missing element with the corresponding column's or row's mean. However, the deleting process leads to loss of supposedly important information and is undesired when the amount of missing elements is large. From statistical point of view, any replacement action provides a wrong estimation of the mean, standard deviation, sample covariance and moreover destroys the correlation structure of the data. From practical point of view, this can lead to improper conclusions and can set incorrect hypothesis still in the beginning of the analysis. For instance, in environmental data analysis a possible pollution can go unnoticed or a costly remediation procedure can take place when this is unnecessary. There are better ways for dealing with missing data. Different methods treating the problem of missing information in principal component analysis have been discussed in the literature [12–16]. Among them is the self-consistent iterative procedure called expectation maximization (EM). It is a flexible method, which allows computing model's parameters while filling in the missing information directly.

* Corresponding author. Tel.: +48 32 359 2115; fax: +48 32 259 9978.
E-mail address: beata@us.edu.pl (B. Walczak).

0039-9140/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2006.10.011

Dealing with missing data and outlying observations simultaneously is an ongoing problem. The presence of outlying observations in the data disturbs a correct replacement of missing elements as well as the presence of missing elements does not allow adequate outliers' identification.

In this paper, we propose a computationally efficient and relatively simple method for exploring data containing missing elements and outliers in terms of PCA. The proposed approach, denoted later in the text as EM-SPCA, consists of implementing a robust version of PCA, such as spherical PCA, in the expectation maximization algorithm. For convenience, the results obtained from EM-SPCA are compared to expectation maximization classical PCA (EM-PCA) by the use of simulated and real data sets.

2. Theory

2.1. Classical principal component analysis (PCA)

Principal component analysis, is mostly used for data compression and visualization [17]. The main goal of this method is to explain the information contained in the data by a set of the so-called principal components, PCs. The PCs are mutually orthogonal and are a linear combination of the original data variables. Since the orthogonal PCs are constructed to maximize the description of the data variance, the first PC describes the largest portion of the data variability, whereas the following ones, the information not explained by the previous PCs. Moreover, each PC carries different information about the data variability. In most applications, PCA allows compression of the data to a few PCs that can be later used to visualize the data structure and they contain almost the same information as the original data. Therefore, PCA is usually the first step in almost any data analysis, and PCs serve as an input to most of the chemometric techniques, e.g. clustering methods, neural networks, etc. Despite of the compression issue that PCA offers, it is vulnerable to outliers in the data due to its least squares nature. Outliers strongly affect the data variance and the true correlation structure of the data. This implies a crucial impact upon overall picture of the data structure since the outliers are responsible for rotating the PCs axes towards them. The robust versions of PCA are considered as a remedy to the problem with outliers.

2.2. Robust PCA by means of spherical PCA

The spherical PCA, SPCA, aims to construct a robust PCA model [18], i.e. a PCA model not influenced by outlying objects. In spherical PCA, this goal is fulfilled by projecting the data objects onto a hyper-sphere of unit radius with center in the robust center of data. In order to define a robust center of the data, the L1-median estimator is used. The L1-median center is a point in the multivariate space that minimizes the sum of all the Euclidian distances between this point and data objects [19].

The authors of SPCA proposed another approach to robust PCA called elliptical PCA. In this method, data are firstly scaled, i.e. each element of data variable is divided by its corresponding robust scale measure (for example the Q_n estimator [20]) and SPCA is then performed. In elliptical PCA, EPCA, the objects are projected onto a hyper-ellipse instead of on a hyper-sphere. The radii of a hyper-ellipse are proportional to the robust scales of the data variables. However, it is pointed out by Boente and Fraiman [21] that elliptical PCA has problems with consistency. Therefore, only spherical PCA is considered in this paper.

In the first step of the SPCA approach, the data are centered about L1-median. The projection of objects onto a hyper-sphere with a robust center at L1-median and with a unit radius, $\boldsymbol{x}_i^{\mathrm{p}}$, can be presented as:

$$\boldsymbol{x}_i^{\mathrm{p}} = \frac{\boldsymbol{x}_i - \mu_{L1}(\mathbf{X})}{||\boldsymbol{x}_i - \mu_{L1}(\mathbf{X})||} + \mu_{L1}(\mathbf{X}) \tag{1}$$

where $\boldsymbol{x}_i$ is the ith data object, $1/||\boldsymbol{x}_i - \mu_{\mathrm{L1}}(\mathbf{X})||^2$ is its weight, $\mu_{\mathrm{L1}}(\mathbf{X})$ is the L1-median center and $||\cdot||$ denotes the Euclidean norm.

At this point, it is important to notice that the denominator of the above formula informs about the Euclidean distance of each object from the robust data center. Therefore, objects being far from the robust data center are far from data majority as well, and thus, they receive small weights. By using such a weighing scheme, the influence of outliers upon the PCA model is diminished.

Later on, classical PCA is performed on the data projected onto a sphere what essentially corresponds to PCA applied to weighted data. After the PCA decomposition, robust scores and robust loadings are obtained.

Once the robust PCA model is constructed, it is possible to perform outlier identification on the basis of two types of distances, namely robust and orthogonal distances, describing every data object [22]. In fact, the robust distance is the Mahalanobis distance of an object, which informs about its location with respect to the robust data center within the space of the robust scores. Robust distance (RD_i) of the ith object is expressed as:

$$RD_i = \sqrt{\sum_{a=1}^{f} \frac{t_{ia}^2}{v_a}} \tag{2}$$

where $a = 1, 2, \ldots, f$ denotes the number of robust principal components, $\boldsymbol{t}_a$ the ath principal component and the v_a is its eigenvalue.

The orthogonal distance of an object describes how far the object from the space of the robust model is. Orthogonal distances for all objects are computed as follows:

$$\mathrm{OD} = ||\mathbf{X} - \mu_{L1}(\mathbf{X}) - \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}}|| \tag{3}$$

where the robust scores matrix, $\boldsymbol{T}$, and robust loadings matrix, $\boldsymbol{P}$, contain f factors.

Taking both distances into account, the data objects can be classified into four categories. The regular objects are characterized by small robust and orthogonal distances. Good leverage objects have large robust distances and small orthogonal distances. High residuals objects have small robust distances, but large orthogonal distances, whereas bad leverage observations have large both robust and orthogonal distances. In order to identify objects with large distances compared to the majority of

objects, one can use robust z-scores. The robust z-scores are obtained by centering each type of distance, $\boldsymbol{d}$, about its median, $\mu(\boldsymbol{d})$ and dividing the elements by their corresponding robust scale, $\sigma(\boldsymbol{d})$ (e.g. the Q_n robust scale) as follows:

$$z = \frac{|\boldsymbol{d} - \mu(\boldsymbol{d})|}{\sigma(\boldsymbol{d})}. \quad (4)$$

To identify objects with large distances, one can use a default cut-off value equal to three.

The Q_n estimator is a robust measure of scale and for a single variable, $\boldsymbol{s}$, it corresponds to the first quartile of the sorted pair-wise differences between all variable elements. It can be expressed as follows:

$$Q_n = 2.2219c\{|s_i - s_j|;\ i < j\}_{(k)} \quad (5)$$

where

$$k = \binom{h}{2} \approx \binom{m}{2}/4,$$

$h = [m/2] + 1$ and c is a correction factor, which depends on the number of objects, m. When the number of objects increases c tends to 1.

For comparison purpose, the same type of plots can be constructed in classical PCA. The classical distance is then calculated using the classical PCA score vectors. The classical z-scores are obtained by the use of classical measures of location and scale, i.e. mean and standard deviation.

Although there are many robust PCA approaches available (e.g. [1–11]), the main advantage of the spherical PCA approach is its conceptual simplicity and computational efficiency since the most time demanding step is PCA. By the use of the so-called kernel approaches to PCA for wide data matrices, the algorithm can be considerably speeded up [23], and thus, the construction of the SPCA model is fast. Since the EM method is computationally time-consuming, a fast robust PCA procedure should be applied. Therefore, SPCA will be considered here.

Despite its attractive properties, SPCA has a shortcoming: it fails to perform appropriate outlier's identification when data contain clusters. An outlier is an object, which is far from data majority. When data contain groups, the objects belonging to a whole group might be undesirably considered outliers [22].

2.3. EM-PCA and EM-SPCA

Classical PCA and spherical PCA, described above, can be implemented relatively easy in the expectation maximization framework. The EM procedure starts with initialization of the missing elements using for instance, the corresponding row's and column's means. In the follow-up iterations, the missing elements are re-filled in with their predicted values according to the currently constructed PCA or SPCA model. In general, the PCA model can be presented as a product of two matrices:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}}, \quad (6)$$

where $\boldsymbol{X}$ is the original data matrix of m objects and n variables, $\boldsymbol{T}$ of dimension $m \times f$ holds the first f score vectors and $\boldsymbol{P}$ $(n \times f)$ is the matrix containing the PCA loadings.

The EM algorithm iterates while the convergence criterion is not satisfied, which means while small changes in the model's residuals in two consecutive iterations are not observed. Once the convergence of the algorithm is reached, the elements replacing the missing values have residuals equal to zero. This means that the filled in values perfectly fit the PCA model of definite complexity and therefore, the differences between the original $\boldsymbol{X}$ (with filled in missing values according to PCA model) and predicted $\boldsymbol{X}$ are equal to zero. For the remaining elements (observed elements) these differences are not zeros. The missing elements are not taken into account in the model construction. The model is determined only based on the observed elements.

The steps of EM-PCA can be schematically described in the following way:

1. Initialize the missing elements by column's and row's means.
2. Preprocess the data.
3. Perform singular value decomposition (SVD) of the complete data set or other method of data decomposition.
4. Predict $\boldsymbol{X}$ according to the PCA model described by Eq. (6) using the predefined number of factors.
5. Re-fill in the missing elements with their predicted values according to the currently constructed model and go to step 2 while the convergence criterion is not fulfilled.

The algorithm of EM-SPCA follows almost the same scheme, but steps 2, and 3 are combined together as one, which performs spherical PCA and missing elements are initialized using the column's and row's medians in step 1. Like classical PCA, the prediction of $\boldsymbol{X}$ is also done according to the model described by Eq. (6), but using the robust scores and loadings.

In the above-described algorithms, the complexity of the final EM-PCA or EM-SPCA is assumed to be known beforehand. When this is not the case, a variant of a cross validation procedure should be involved. However, for a large data set, this can be a very time-consuming step. In EM-PCA, the selection of an optimal number of factors can be performed based on the minimum observed on the curve displaying eigenvalues versus number of factors. A similar way of a factor selection procedure can be adopted in EM-SPCA using robust eigenvalues. Another possibility can be to use the number of factors, which explain about 80% of data information [22]. This is also the rule used in our study.

Another important issue is how to preprocess data with missing values. Usually in classical PCA, the data are first centered, which involves the subtraction of the column's means from each corresponding data element. Centering or any other transformation of the data has to be performed within the iterative steps of the EM algorithm. This implies a proper estimation of the column's means or/and standard deviations together with the process of re-filling in the missing information in the data. Concerning the EM robust PCA approach, when there are large

differences in the variables' range and units, a standardization procedure in spherical PCA is considered.

3. Results and discussion

3.1. A simulation study

The goal of the simulation study is to show the performance of the proposed method, EM-SPCA, in exploring data containing missing elements and outliers simultaneously. For this purpose, a data set of a definite complexity was generated from the normal distribution, N (0,1). This means that each element of $\boldsymbol{X}$ ($m \times n$) comes from the population with zero mean and unit standard deviation. In order to obtain data of definite complexity, after data decomposition by means of PCA, $\boldsymbol{X}$ is reconstructed according to the PCA model defined by Eq. (6), using score and loading matrices ($\boldsymbol{T}$ of dimension $m \times f$ and $\boldsymbol{P}$ of dimension $n \times f$) with selected number of latent factors, f. The choice of latent factors depends entirely on the user and does not follow any special requirements. Furthermore, a normally distributed noise is additionally added to the data. The final simulation model is $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\mathrm{T}} + \boldsymbol{E}$, where $\boldsymbol{E}$ is a $m \times n$ matrix, the elements of which are generated from multivariate normal distribution, N (0,1). To keep the level of noise low, the generated values are multiplied by 0.2. In this way, a data set without missing elements and outliers, but with known complexity is created. In our study, $\boldsymbol{X}$ is of dimension 98×20 with complexity four, i.e. the original data matrix, $\boldsymbol{X}$, was reconstructed using four PCs.

In the next step of the study, different percentages (5, 10, 15, 20, 25, 30, 35, 40, 45 and 50, respectively) of elements were deleted from the data. The deleting mechanism is done completely at random, which means that the missingness is not related to variables of the data in terms defined by Rubin [24]. In order to simulate different patterns of a particular amount of missing elements, the deleting procedure was repeated 50 times. Moreover, to obtain information about the variability in the data depending on the different percentages of missing elements, the changes in the model's residuals evaluated only for observed elements can be traced. As it was already mentioned above, the residuals for the values replacing the missing elements are equal to zero. The root mean squared (RMS) error is evaluated as an average of 50 repetitions.

Fig. 1 presents the results from a comparative study of two algorithms, i.e. EM-PCA and EM-Spherical PCA, applied to normally distributed data with missing elements. Such a comparative study aims to show that the robust estimator performs similarly to the classical estimator at normal models. The models are indeed of complexity equal to four explaining above 90% of data information.

In general, the averaged RMS values decrease with the increasing percentage of missing elements (see Fig. 1a). This trend is expected since RMS is evaluated only for the observed elements, the number of which decreases with the increasing number of missing elements. Another important observation is that all the averaged RMS values obtained from the EM-SPCA algorithm are slightly higher than the RMS values obtained from EM-PCA. This is not also a surprising observation, because the robust estimators by definition show higher variance at normal models. There are no objects exceeding the cut-off values in the classical and robust distance–distance plots drawn for data with missing elements. Such plots are presented, for example, for data containing 10% of missing elements in Fig. 1b and c.

A more interesting case is to compare the performances of the methods when data contain both outliers and missing elements. Therefore, 10% of outliers were included in the original data. For this purpose, the error term $\boldsymbol{E}$ added to $\boldsymbol{X}$ is chosen such that 90% of objects have the normal distribution N (0,1) multiplied by factor 0.1 to keep the level of noise low, and 10% of objects, being outliers, have multivariate distribution N (10,1). Higher percentages of outliers were not considered in our study, because the breakdown point of SPCA is the highest possible one. In other words, the method will provide correct estimates even for highly contaminated data.

Fig. 1d shows clearly that the robust technique outperforms the classical EM-PCA method. The mean trimmed RMS values are much smaller for EM-SPCA than for the EM-PCA approach. The trimming operation has a sense only for the robust approach in order to honestly represent the model fit in the data. The outliers have high residuals, i.e. high orthogonal distances, when using a robust approach and their influence on RMS has to be discarded. In our study, the trimming procedure is done by weighting the objects exceeding the cut-off value of orthogonal distance. The orthogonal outliers and bad leverage observations receive weights proportional to the inverses of their distances to the cut-off line. The object, which is farther from the cut-off line, receives the smaller weight and consequently, it has the smaller impact on the RMS value estimated. In the simulated data set, the outliers are objects with nos. 1–10. Furthermore, the classical distance–distance plot (see Fig. 1e) fails to identify the ten outliers (nos. 1–10), whereas they are correctly found in the robust distance–distance plot (see Fig. 1f).

3.2. Convergence properties of the proposed method

Concerning the issues of the algorithm's convergence, one can monitor the stability of the replaced missing values in consecutive iterations or the changes in the objective function defined for the kth iteration as:

$$SS_k = \sum_p \sum_q (x_{pq})^2, \quad p, q \in \text{missing elements} \tag{7}$$

The model obtained will be the same in both situations. In our study, the differences in replaced values were traced. The calculations were considered converged when the sum of squared differences estimated for the re-filled values of the missing elements in two consecutive iterations were smaller than 10^{-8}. In general, the algorithm's convergence depends upon the complexity of the PCA model, percentage of missing elements and their distribution and most importantly upon the correlation structure of the data. If the data correlation structure is well defined, the convergence is achieved very rapidly. Otherwise, the convergence is slow and eventually not reached. For illustration, the number of iterations necessary to satisfy the convergence crite-

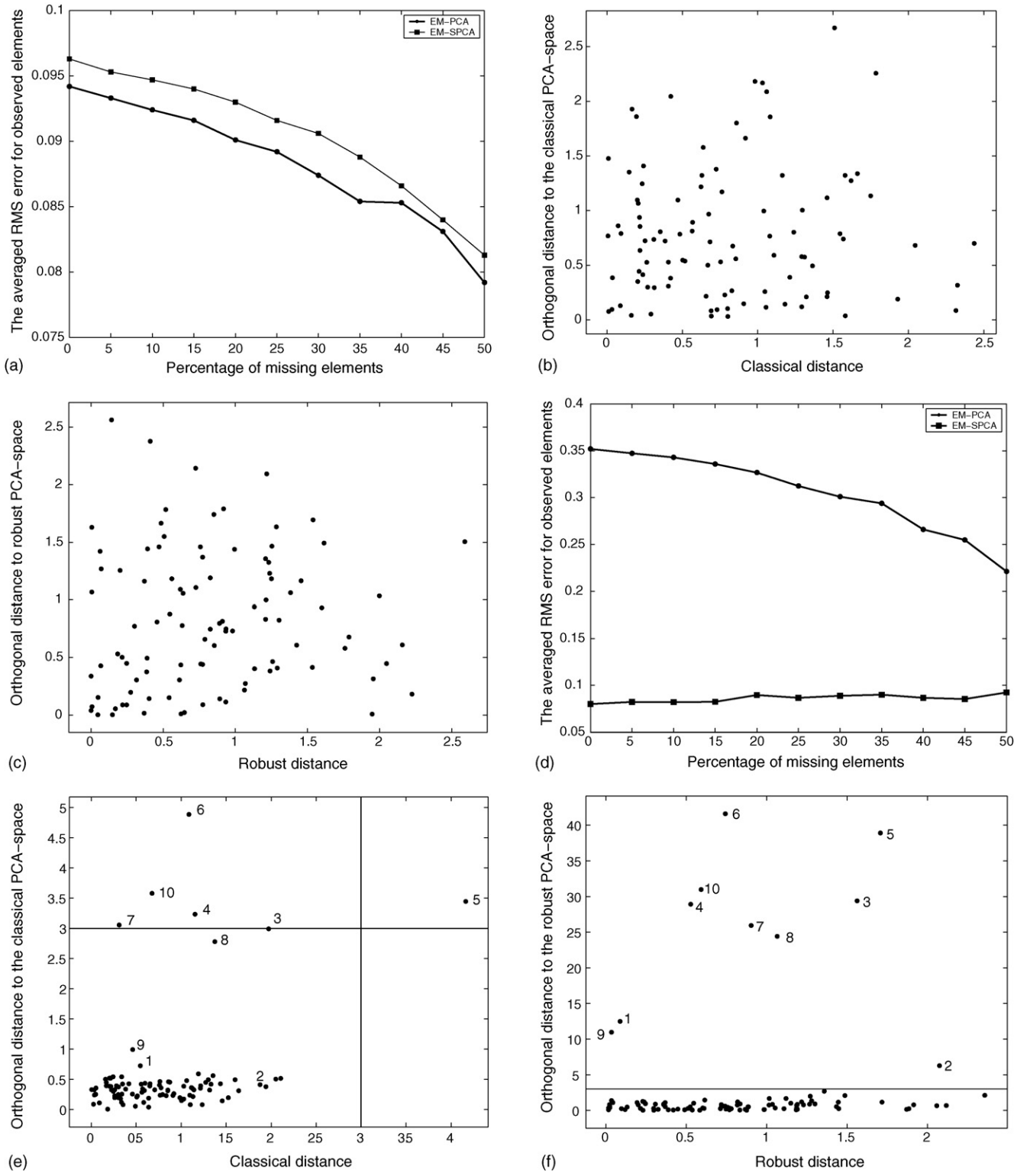

Fig. 1. Results of a simulation study: (a) the averaged RMS error calculated for observed elements as a function of percentage of missing elements for EM-PCA and EM-SPCA applied to normally distributed data, (b) the classical distance–distance plot constructed via EM-PCA applied to normally distributed data with 10% of missing elements, (c) the robust distance–distance plot constructed via EM-SPCA applied to normally distributed data with 10% of missing elements, (d) the averaged RMS error calculated for observed elements as a function of percentage of missing elements for EM-PCA and EM-SPCA applied to contaminated data, (e) the classical distance-distance plot constructed via EM-PCA applied to contaminated data with 10% of missing elements and (f) the robust distance–distance plot constructed via EM-SPCA applied to contaminated data with 10% of missing elements.

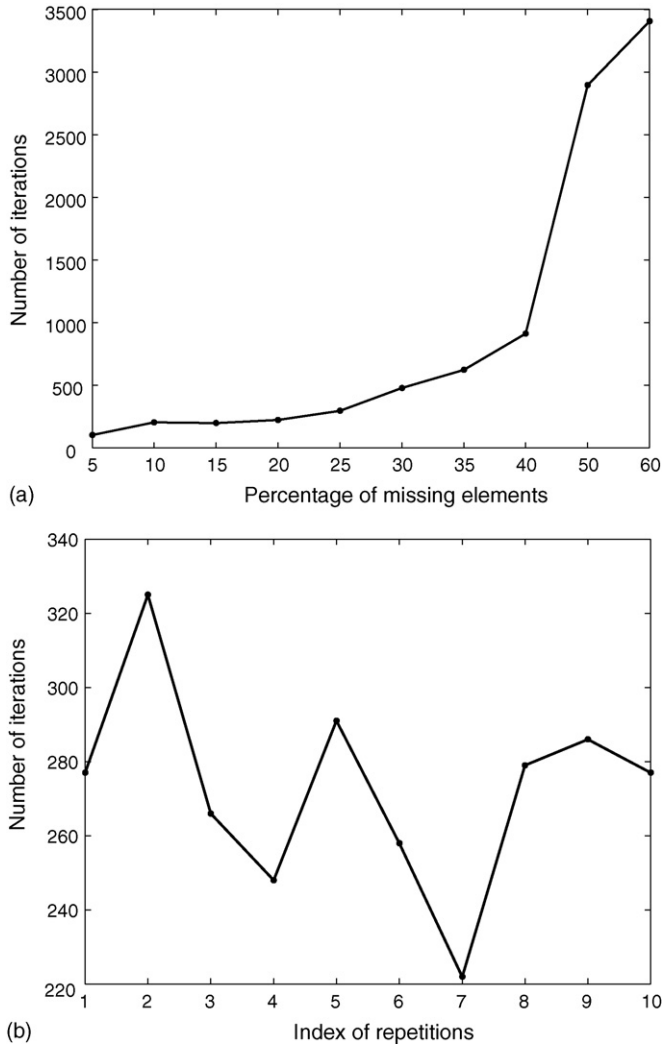

Fig. 2. Convergence properties of EM-SPCA: (a) number of iterations as a function of percentage of missing elements (number of iterations is an average of 50 runs of the algorithm applied to data with different random patterns of missing elements) and (b) number of iterations needed to achieve convergence in 10 consecutive runs of the algorithm.

rion is shown as a function of percentage of missing elements in the data in Fig. 2. The number of iterations is a mean value of 50 repetitions.

The number of iterations increases with the increasing number of missing elements (see Fig. 2a). Moreover, different number of iterations is needed depending upon the distribution of missing elements. This is shown for data containing outliers and 20% of missing elements in Fig. 2b. It is important to stress that the EM algorithm converges to the optimal solution no matter the number of iterations or the pattern of missing elements.

3.3. Application of EM-PCA and EM-SPCA to a real data set

The environmental data set contains the annual mean levels of nine major ions (H^+, NH_4^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, NO_3^- and SO_4^{2-}) measured in various sampling sites (Haunsberg, Innervillgraten, Kufstein, Reutte, Litschau, Luntz, Nassfeld,

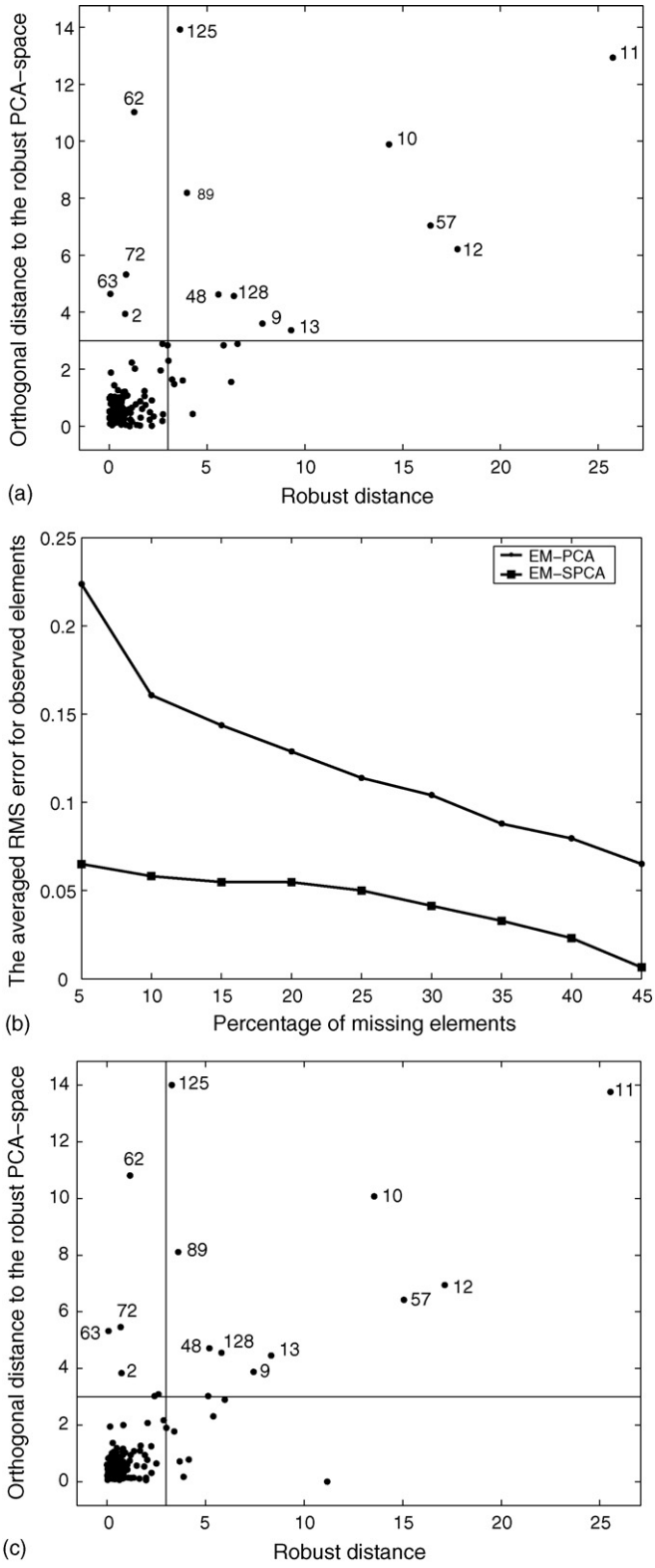

Fig. 3. Results of the study on a real data set: (a) the robust distance–distance plot constructed via SPCA (with standardization) applied to data without missing elements, (b) the averaged RMS error calculated for observed elements as a function of percentage of missing elements for EM-PCA and EM-SPCA applied to the studied environmental data and (c) the robust distance–distance plot constructed via EM-SPCA applied to the environmental data set containing 5% of missing elements.

Nasswald, Lobau, Sonnblick and Werfenweng) in Austria, over different periods of time (1988–1999). The data set is known to have several outlying observations [22], but there are no missing elements. The distance–distance plot obtained from SPCA (with standardization) applied to data without missing elements is shown in Fig. 3a. The diagnostic plot is constructed using four robust PCs explaining around 84% of data information. All objects exceeding the cut-off line of orthogonal distance are considered outliers. The largest orthogonal distances are observed for objects nos. 10, 11, 62 and 125 (see Fig. 3a). Unlike objects 62 and 125, objects 10 and 11 exhibit large robust distances as well.

The measured variables are in different ranges and units and the autoscaling procedure lends them the same importance in the data analysis. In EM-PCA, autoscaling is performed using column's means and standard deviations, whereas in EM-Spherical PCA this is done using column's medians and robust scales, Q_n. As already mentioned, the data transformation should be performed within the steps of the EM approach. In this way, the mean, median, standard deviation and robust scale are updated together with the missing information in the data.

The results from the comparison study of the EM-PCA and EM-SPCA approaches are presented in Fig. 3b for environmental data where different percentages of elements were deleted. The complexity of all the models is four. It can be seen that the EM-SPCA method clearly outperforms the EM-PCA in terms of lower trimmed RMS values. Every RMS value is a mean of 50 repetitions in Fig. 3b. Similarly to the results obtained from the simulation study, RMS decreases with the increasing percentage of missing elements. Looking at the distance–distance plot obtained for data with 5% nondetects, one can observe (see Fig. 3c) the same outliers identified like in the case without missing elements in the data.

4. Conclusions

The presented study proposes a combined approach for dealing successfully with missing elements and outliers present simultaneously in the data. It basically consists of robust version of PCA embedded within the expectation maximization framework. The robust version of PCA used in our study has been shown to have good robust properties. Only slightly worse performance at normal data models has been observed. Another attractive property is its computational efficiency.

The simulation study showed that EM-Spherical PCA is an efficient tool for processing data containing missing elements and outliers. When the experimental parameters measured has to be scaled in a robust way then a standardization step within the EM-SPCA algorithm should be concerned. This situation is demonstrated on a real environmental data set known to contain outlying observations. The EM-SPCA method (with standardization) was applied to these data after deleting different amounts of elements. The results lead to the same general conclusion as the one of the simulation study and namely that expectation maximization robust PCA is the preferred method over the classical approach when exploring data sets with nondetects and outliers. In particular, it is shown that EM-SPCA with implemented robust scaling outperforms EM-PCA with non-robust standardization.

Acknowledgements

I. Stanimirova and B. Walczak are grateful for financial support concerning scientific activities within the Sixth Framework Programme of the European Union, project TRACE—"TRAcing food Commodities in Europe" (project no. FOOD-CT-2005-006942). The publication reflects only the author's views and the community is not liable for any use that may be made of the information contained therein.

M. Daszykowski is grateful to Foundation for Polish Science for the financial support.

References

[1] G. Li, Z. Chen, J. Am. Stat. Assoc. 80 (1985) 759–766.
[2] L. Ammann, J. Am. Stat. Assoc. 88 (1993) 505–514.
[3] C. Croux, A. Ruiz-Gazen, COMPSTAT: Proceedings in Computational Statistics 1996, Physica-Verlag, Heidelberg, Germany, 1996, pp. 211–217.
[4] C. Croux, A. Ruiz-Gazen, High breakdown estimators for principal components: The Projection Pursuit approach revisited, vol. 29, The IMS Bulletin, 2000, p. 270.
[5] M. Hubert, P. Rousseeuw, S. Verboven, Chemometr. Intell. Lab. Syst. 60 (2002) 101–111.
[6] R.A. Maronna, Ann. Stat. 4 (1976) 51–67.
[7] C. Croux, G. Haesbroeck, Biometrika 87 (2000) 603–618.
[8] M. Hubert, S. Engelen, Bioinformatics 20 (2004) 1728–1736.
[9] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, Technometrics 47 (2005) 64–79.
[10] H. Hove, Y.-Z. Liang, O.M. Kvalheim, Chemometr. Intell. Lab. Syst. 27 (1995) 33–40.
[11] R.A. Maronna, Technometrics 47 (2005) 264–273.
[12] B. Grung, R. Manne, Chemometr. Intell. Lab. Syst. 42 (1998) 125–139.
[13] S. Roweis, IEEE 2001 Int. Conf. on Computer Vision (ICCV 2001), Vancouver, Canada, July 2001.
[14] B. Walczak, D.L. Massart, Part I, Chemometr. Intell. Lab. Syst. 58 (2001) 15–27.
[15] A. Smoliński, B. Walczak, J.W. Einax, Chemosphere 49 (2002) 233–245.
[16] P.R.C. Nelson, P.A. Taylor, J.F. McGregor, Chemometr. Intell. Lab. Syst. 35 (1996) 45–65.
[17] S. Wold, K. Esbensen, P. Geladi, Chemometr. Intell. Lab. Syst. 2 (1987) 37–52.
[18] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, Sociedad de Estadistica e Investigacion Operativa Test 8 (1999) 1–74.
[19] P.J. Huber, Robust Statistics, Wiley, New York, The USA, 1981.
[20] P.J. Rousseeuw, C. Croux, J. Am. Stat. Assoc. 88 (1993) 1273–1283.
[21] G. Boente, R. Fraiman, Test 8 (1999) 28–35.
[22] I. Stanimirova, B. Walczak, D.L. Massart, V. Simeonov, Chemometr. Intell. Lab. Syst. 71 (2004) 83–95.
[23] W. Wu, D.L. Massart, S. de Jong, Chemometr. Intell. Lab. Sys. 36 (1997) 165–172.
[24] D.B. Rubin, Biometrika 63 (1976) 581–592.