R TOOLS FOR ROBUST STATISTICAL ANALYSIS OF HIGH–DIMENSIONAL DATA

V. Todorov, P. Filzmoser
United Nations Industrial Development Organization (UNIDO)
Vienna, AUSTRIA
e-mail: v.todorov@unido.org

Abstract

The present work discusses robust multivariate methods specifically designed for high dimensions. Their implementation in R is presented and their application is illustrated on examples. The first group of classes are algorithms for outlier detection, already introduced elsewhere and implemented in other packages. The value added of the new package is that all methods follow the same pattern and thus can use the same graphical and diagnostic tools. The next topic covered is sparse principal components including an object oriented interface to the standard method proposed by Zou et al [14] and the robust one proposed by Croux et al [2]. Robust partial least squares (Hubert and Vanden Branden [6]) as well as partial least squares for discriminant analysis conclude the scope of the new package.

1 Introduction

High-dimensional data are typical in many contemporary applications in scientific areas like genetics, spectral analysis, data mining, image processing, etc. and introduce new challenges to the traditional analytical methods. First of all, the computational effort for the anyway computationally intensive robust algorithms increases with increasing number of observations n and number of variables p towards the limits of feasibility. Some of the robust multivariate methods available in R (see Todorov and Filzmoser [11]) are known to deteriorate rapidly when the dimensionality of data increases and others are not applicable at all when p is larger than n.

The present work discusses robust multivariate methods specifically designed for high dimensions. Their implementation in R is presented and their application is illustrated on examples. A key feature of this extension of the framework is the object model which follows the one already introduced by **rrcov** and based on statistical design patterns. The first group of classes are algorithms for outlier detection, already introduced elsewhere and implemented in other packages. The value added of the new package is that all methods follow the same pattern and thus can use the same graphical and diagnostic tools. The next topic covered is sparse principal components including an object oriented interface to the standard method proposed by Zou et al [14] and the robust one proposed by Croux et al [2]. These are presented and illustrated in Section 2. Robust partial least squares ([6], [10]) as well as partial least squares for discriminant analysis are presented in Section 3. Section 4 concludes.

2 Robust sparse Principal Component Analysis

Principal component analysis (PCA) is a widely used technique for dimension reduction achieved by finding a smaller number q of linear combinations of the originally observed p variables and retaining most of the variability of the data. It is important to be able to interpret these new variables, referred to as *principal components*, especially when the original variables have physical meaning. The link between the original variables and the principal components is given by the so called *loadings matrix* used for transforming the data and thus it should serve as a means for interpreting the PCs. Dimension reduction by PCA is mainly used for: (i) visualization of multivariate data by scatter plots (in a lower dimensional space); (ii) transformation of highly correlated variables into a smaller set of uncorrelated variables which can be used by other methods (e.g. multiple or multivariate regression); (iii) combination of several variables characterizing a given process into a single or a few *characteristic* variables or *indicators*.

The classical approach to PCA measures the variability through the empirical variance and is essentially based on computation of eigenvalues and eigenvectors of the sample covariance or correlation matrix. Therefore the results may be extremely sensitive to the presence of even a few atypical observations in the data. The outliers could artificially increase the variance in an otherwise uninformative direction and this direction will be determined as a PC direction. These discrepancies will carry over to any subsequent analysis and to any graphical display related to the principal components such as the biplot.

PCA was probably the first multivariate technique subjected to robustification, either by simply computing the eigenvalues and eigenvectors of a robust estimate of the covariance matrix or directly by estimating each principal component in a robust manner. Different approaches to robust PCA are discussed in many review papers, see for example [11] and [5], and examples are given how these robust analyses can be carried out in R. Details about the methods and algorithms can be found in the corresponding references. However, PCA usually tends to provide PCs which are linear combinations of all the original variables (by giving them non-zero loadings). Regarding the interpretability of the results it would be very helpful to reduce not only the dimensionality but also the number of used variables (ideally to relate each PC to only a few variables). It is not surprising that vast research effort was devoted to this issue and various proposals have been introduced in the literature. A straightforward informal method is to set to zeros those PC loadings which have absolute values below a given threshold (*simple thresholding*). Jolliffe et al [9] proposed SCoTLASS which applies a *lasso* penalty on the loadings in a PCA optimization problem, and recently Zou et al [14] reformulated PCA as a regression problem and used the *elastic net* to obtain a sparse version - SPCA.

Despite of being more or less successful in achieving sparsity, all these methods suffer a common drawback - all are based on the classical approach to PCA which measures the variability through the empirical variance, and is essentially based on computation of eigenvalues and eigenvectors of the sample covariance or correlation matrix. Therefore the results may be very sensitive to the presence of even a few

atypical observations in the data. The outliers could artificially increase the variance in an otherwise uninformative direction and this direction will be determined as a PC direction. To cope with the possible presence of outliers in the data, recently Croux et al [2] proposed a method which is sparse and robust at the same time. It utilizes the *projection pursuit* approach where the PCs are extracted from the data by searching the directions that maximize a robust measure of variance of data projected on it. An efficient computational algorithm was proposed by Croux et al [1].

Example We will use a real data example to illustrate the standard and robust sparse methods—the low-dimensional `cars` data set, which is available in the package **rrcovHD** as the data frame `cars`. For $n = 111$ cars, $p = 11$ characteristics were measured, including the length, width, and height of the car. After looking at pairwise scatterplots of the variables, and computing pairwise Spearman rank correlations $\rho(X_i, X_j)$ we see that there are high correlations among the variables, for example, $\rho(X_1, X_2) = .83$ and $\rho(X_3, X_9) = .87$. We conclude that PCA could be an appropriate method for finding the most important sources of variation in this data set (see also Hubert et al [8]). The first four classical PCs explain more than 96% of the total variance and the first four robust PCs explain more than 95%, therefore we decide to retain four components in both cases. Next we need to choose the degree of sparseness which is controlled by the regularization parameter λ. Since the sparse PCs have to provide a good trade-off between sparseness and achieved percentage of explained variance we can proceed similarly as in the selection of the number of principal components with the scree plot - we compute the sparse PCA for many different values of λ and plot the percent of explained variance against λ. We choose $\lambda = 0.78$ for classical PCA and $\lambda = 2.27$ for robust PCA, thus attaining 83 and 82 percent of explained variance, respectively, which is only an acceptable reduction compared to the non-sparse PCA. Retaining $k = 4$ principal components as above and using the selected parameters λ, we can construct the so called *diagnostic plots* which are especially useful for identifying outlying observations. The diagnostic plot is based on the *score distances* and *orthogonal distances* computed for each observation.

The *diagnostic plot* shows the orthogonal distances versus the score distance, and indicates with a horizontal and vertical line the cut-off values that allow to distinguish regular observations (those with small score and small orthogonal distance) from the different types of outliers: *bad leverage points* with large score and large orthogonal distance, *good leverage points* with large score and small orthogonal distance and *orthogonal outliers* with small score and large orthogonal distance (for detailed description see [8]). In Figure 1 the classical and robust diagnostic plot as well as their sparse alternatives are presented. The diagnostic plots for the standard PCA reveals only several orthogonal outliers and identifies two observations as bad leverage points. Three more observations are identified as bad leverage points by the sparse standard PCA which is already an improvement, but only the robust methods identify a large cluster of outliers. These outliers are masked by the non-robust score and orthogonal distances and cannot be identified by the classical methods. It is important to note that the sparsity feature added to the robust PCA did not influence its ability to detect

properly the outliers.

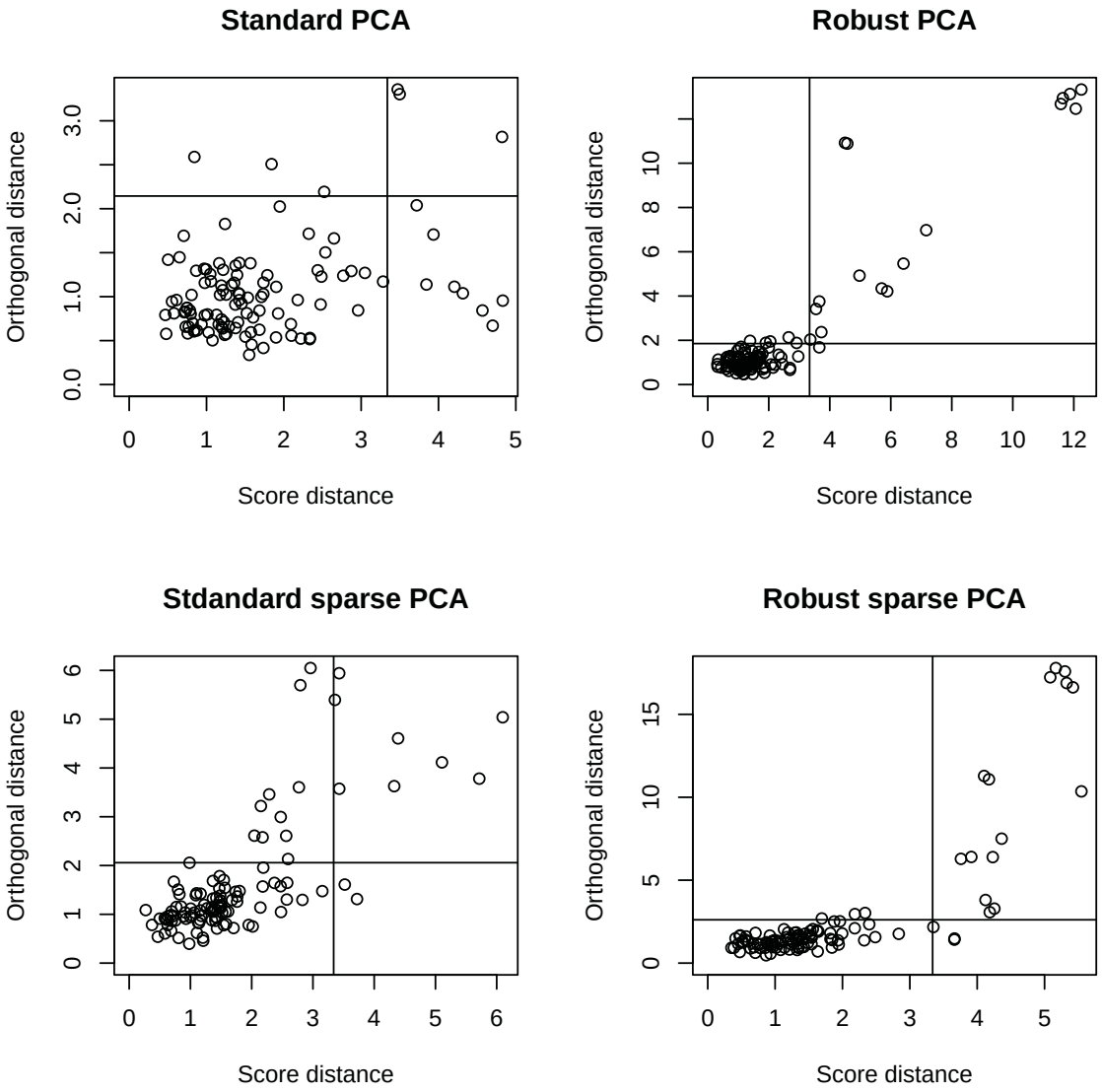

Figure 1: Distance-distance plots for standard and sparse PCA and their robust versions for the `cars` data.

3 Robust linear regression and classification in high dimensions

Regression problems become challenging when the number of explanatory variables p exceeds the number of observations n. The standard tool to use in these situations is partial least squares regression (PLS). PLS was developed by [13] in the 1960s in the

context of econometric path modeling but some twenty years later it was successfully adopted for regression problems in chemometrics and spectroscopy. It performs a dimensionality reduction to the original regressor variables $\boldsymbol{X}$ by searching for directions $\boldsymbol{w}$. The objective is to maximize the covariance between the scores $\boldsymbol{Xw}$ and a linear projection of the responses $\boldsymbol{Y}$. This ensures that the newly derived regressor variables contain relevant information for the prediction of the responses. There are different models and estimators for the PLS regression problem. The idea of PLS regression is a decomposition of the predictor matrix $\boldsymbol{X}$ and the response matrix $\boldsymbol{Y}$:

$$\boldsymbol{X} = \boldsymbol{TP}^\top + \boldsymbol{E_X} \tag{1}$$

$$\boldsymbol{Y} = \boldsymbol{TQ}^\top + \boldsymbol{E_Y} \tag{2}$$

where $\boldsymbol{T} = \boldsymbol{XW} \in \mathbb{R}^{n \times K}$ is a score matrix, and $\boldsymbol{W} = (\boldsymbol{w_1}, \ldots, \boldsymbol{w_K}) \in \mathbb{R}^{p \times K}$ is a matrix of direction (loading) vectors. The equations (1) and (2) can be regarded as ordinary least squares problems, so $\boldsymbol{P} \in \mathbb{R}^{p \times K}$ and $\boldsymbol{Q} \in \mathbb{R}^{q \times K}$ are matrices of coefficients, whereas $\boldsymbol{E_X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{E_Y} \in \mathbb{R}^{n \times q}$ are matrices of random errors. Again, K denotes the number of components, with $K \leq \min\{n, p, q\}$.

If we rewrite equation (2),

$$\boldsymbol{Y} = \boldsymbol{TQ}^\top + \boldsymbol{E_Y} = \boldsymbol{XWQ}^\top + \boldsymbol{E_Y}, \tag{3}$$

we see that $\boldsymbol{WQ}^\top \in \mathbb{R}^{p \times q}$ is a matrix of coefficients that relates $\boldsymbol{Y}$ to the original data $\boldsymbol{X}$ according to the original model:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}, \tag{4}$$

where the response variable is an $n \times q$ matrix $\boldsymbol{Y}$ of univariate responses. Accordingly, the matrix of regression coefficients $\boldsymbol{B}$ is of dimension $p \times q$, and the error matrix $\boldsymbol{E}$ has the same dimension as $\boldsymbol{Y}$.

In order to successively find direction vectors $\boldsymbol{w}$ that maximize the covariance between the explanatory variables and the responses, the SIMPLS criterion of [3] is used. The first normalized weight vectors $\boldsymbol{r_1}$ and $\boldsymbol{q_1}$ are obtained as linear combinations of $\boldsymbol{X}$ and $\boldsymbol{Y}$ that maximize

$$cov(\boldsymbol{X}\boldsymbol{r}_1; \boldsymbol{Y}\boldsymbol{q}_1). \tag{5}$$

The solution of this maximization problem is found by taking $\boldsymbol{r_1}$ and $\boldsymbol{q_1}$ as the first left and right singular eigenvectors of $\boldsymbol{S}_{xy} = \boldsymbol{X}^T\boldsymbol{Y}/(n-1)$, the cross-covariance matrix of the explanatory variables and response variables. For each observation the first coordinate of the score $\boldsymbol{t}_{i1}$ is computed as $\boldsymbol{t}_{i1} = \boldsymbol{x}_i\boldsymbol{r}_1$. The other weight vectors $\boldsymbol{r_a}$ and $\boldsymbol{q_a}$ for $a = 2, \ldots, K$ are obtained by imposing an orthogonality constraint to the elements of the scores.

In general, the weight vectors $\boldsymbol{r_a}$ and $\boldsymbol{q_a}$ are obtained as the left and right singular vector of $\boldsymbol{S}^a_{xy}$ where $\boldsymbol{S}^a_{xy}$ is the deflated covariance matrix:

$$\boldsymbol{S}^a_{xy} = \boldsymbol{S}^{a-1}_{xy} - \boldsymbol{P}_{[a-1]}(\boldsymbol{P}^T_{[a-1]}\boldsymbol{P}_{[a-1]})^{-1}\boldsymbol{P}^T_{[a-1]}\boldsymbol{S}^{a-1}_{xy} \tag{6}$$

Finally, when the scores are K-dimensional, multivariate linear regression is performed of the responses $\boldsymbol{y}_i$ on these scores $\boldsymbol{t}_i$:

$$\begin{aligned} \boldsymbol{B}_{p\times q} &= \boldsymbol{R}_{p\times K}(\boldsymbol{T}_{K\times n}^T \boldsymbol{T}_{n\times K})^{-1}\boldsymbol{T}_{K\times n}^T \boldsymbol{Y}_{n\times q} \\ \boldsymbol{B}_0 &= \bar{y} - \boldsymbol{B}_{q\times p}^T \bar{x}, \end{aligned} \tag{7}$$

where $\boldsymbol{R}_{p\times K} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K]$.

A robust alternative to PLS was proposed by Hubert and Vanden Branden [6]. It modifies the SIMPLS algorithm in only two steps. First, robust estimates of $\boldsymbol{S}_{xy}$ and $\boldsymbol{S}_x$ will be obtained by using robust PCA in order to compute a robust covariance matrix in high dimensions. Then a robust multivariate regression method is performed in the second stage. Another approach to robust PLS is the method called partial robust M (PRM) regression [10]. The main idea is to use an M estimator for regression not on the complete but only for a partial information of the explanatory variables. This partial information is obtained via latent variables that need to be extracted in a robust manner (see also [4]).

Classification in high dimensions The prediction of group membership and/or describing group separation on the basis of a data set with known group labels (training data set) is a common task in many applications and linear discriminant analysis (LDA) has often been shown to perform best in such classification problems. However, very often the data are characterized by far more variables than objects and/or the variables are highly correlated which renders LDA (and the other similar standard methods) unusable due to their numerical limitations. Let us assume that $\boldsymbol{Y}$ is univariate and categorical, i.e. $\forall 1 \leq i \leq n : y_i \in \{1, \ldots, G\}$ where G is the number of groups. For high dimensional data sets, classical linear discriminant analysis cannot be performed due to the singularity of the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$, as it requires the inverse of $\hat{\boldsymbol{\Sigma}}$. To overcome the high dimensionality problem in classification context one can reduce the dimensionality by either selecting a subset of "interesting" variables (variable selection) or construct K new components, $K \ll p$ which represent the original data with minimal loss of information (feature extraction, dimension reduction). Many methods for dimension reduction were considered in the literature but the two most popular are principal component analysis (PCA) and partial least squares (PLS). It is intuitively clear that a supervised method (which uses the group information while constructing the new components) like PLS should be preferred to unsupervised methods like PCA.

PLS was not originally designed to be used in the context of statistical discrimination but nevertheless was routinely applied with evident success by practitioners for this purpose. Taking into account the grouping variable(s) when decomposing the data one would intuitively expect an improved performance for group separation. Since the response variable in case of a classification problem is a categorical variable, none of the robust PLS methods proposed above can be used. Therefore, in order to obtain a robust PLS-DA we proposed to apply any of the outlier detection methods described in Filzmoser and Todorov [5], which are implemented in package **rrcovHD**, and then

use classical PLS on the already cleaned data set. Hubert and Van Driessen [7] used a

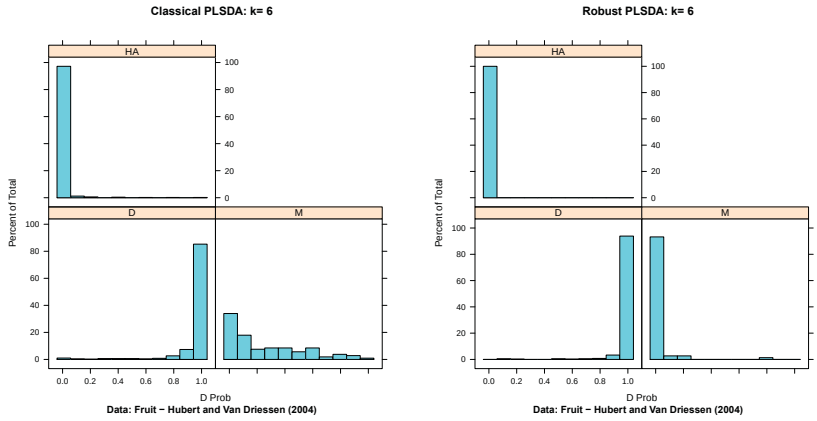

Figure 2: Prediction histograms for class D for the `fruit` data using classical and robust PLS-DA.

data set containing the spectra of three different cultivars of the same fruit. The three cultivars (groups) are named D, M and HA, and their sample sizes are 490, 106 and 500 observations, respectively. The spectra are measured at 256 wavelengths. The fruit data is thus a high-dimensional data set which was used to illustrate a new approach for robust linear discriminant analysis, and it was studied again by Vanden Branden and Hubert [12]. From these studies it is known that the first two cultivars D and M are relatively homogenous and do not contain atypical observations, but the third group HA contains a subgroup of 180 observations which were obtained with a different illumination system. In Figure 2 are shown the prediction histograms for class D for the `fruit` data using classical and robust PLS-DA.

4 Summary and conclusions

An object oriented framework for robust multivariate analysis developed in the S4 class system of the programming environment R already exists implemented in the package **rrcov** and is described in [11]. The main goal of this framework is to support the usage, experimentation, development and testing of robust multivariate methods as well as simplifying comparisons with related methods. In this article we investigated several robust multivariate methods specifically designed for high dimensions. All considered methods and data sets are available in the R package **rrcovHD**. A key feature of this extension of the framework is that the object model follows the one already introduced by **rrcov** which is based on statistical design patterns.

Acknowledgements

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization.

References

[1] Croux C, Filzmoser P, Oliveira M (2007) Algorithms for projection-pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems 87(218):218–225

[2] Croux C, Filzmoser P, Fritz H (2011) Robust sparse principal component analysis. Research report sm-2011-2, Vienna University of Technology

[3] de Jong S (1993) SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 18:251–263

[4] Filzmoser P, Todorov V (2011) Review of robust multivariate statistical methods in high dimension. Analytica Chimica Acta 705:2–14

[5] Filzmoser P, Todorov V (2012) Robust tools for the imperfect world. Information Sciences In press

[6] Hubert M, Vanden Branden K (2003) Robust methods for partial least squares regression. Journal of Chemometrics 17(10):537–549

[7] Hubert M, Van Driessen K (2004) Fast and robust discriminant analysis. Computational Statistics & Data Analysis 45:301–320

[8] Hubert M, Rousseeuw P, Vanden Branden K (2005) ROBPCA: A new approach to robust principal component analysis. Technometrics 47:64–79

[9] Jolliffe IT, Trendafilov NT, Uddin M (2003) A modified principal component technique based on the LASSO. J Comput Graph Statist 12(3):531–547

[10] Sernels S, Croux C, Filzmoser P, van Espen P (2005) Partial robust M-reression. Chemometrics and Intellegent Laboratory Systems 79:55–64

[11] Todorov V, Filzmoser P (2009) An object oriented framework for robust multivariate analysis. Journal of Statistical Software 32(3):1–47

[12] Vanden Branden K, Hubert M (2005) Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intellegent Laboratory Systems 79:10–21

[13] Wold H (1975) Soft modeling by latent variables: the non-linear iterative partial least squares approach. In: Giani J (ed) Perspectives in probability and statistics, papers in honor of M.S. Bartlett, Academic Press, London, pp 117–142

[14] Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. Journal of Computational and Graphical Statistics 15(2):265–286