Robust Multivariate Methods: The Projection Pursuit Approach

P. Filzmoser[1], S. Serneels[2], C. Croux[3], and P.J. Van Espen[2]

[1] Department of Statistics and Probability Theory,
Vienna University of Technology, A-1040 Vienna, Austria
[2] Department of Chemistry,
University of Antwerp, B-2610 Antwerp, Belgium
[3] Department of Applied Economics,
K.U. Leuven, B-3000 Leuven, Belgium

Abstract. Projection pursuit was originally introduced to identify structures in multivariate data clouds (Huber, 1985). The idea of projecting data to a low-dimensional subspace can also be applied to multivariate statistical methods. The robustness of the methods can be achieved by applying robust estimators to the lower-dimensional space. Robust estimation in high dimensions can thus be avoided which usually results in a faster computation. Moreover, flat data sets where the number of variables is much higher than the number of observations can be easier analyzed in a robust way.

We will focus on the projection pursuit approach for robust continuum regression (Serneels et al., 2005). A new algorithm is introduced and compared with the reference algorithm as well as with classical continuum regression.

1 Introduction

Multivariate statistical methods are often based on analyzing covariance structures. Principal Component Analysis (PCA) for example corresponds to a transformation of the data to a new coordinate system where the directions of the new axes are determined by the eigenvectors of the covariance matrix of the data. In factor analysis the covariance or correlation matrix of the data is the basis for determining the new factors, where usually the diagonal of this scatter matrix is reduced by a variance part that is unique for each variable ("uniqueness"). In Canonical Correlation Analysis (CCA) one is concerned with two sets of variables that have been observed on the same objects, and the goal is to determine new directions in each of the sets with maximal correlation. The problem comes down to an eigenvector decomposition of a matrix that uses information of the joint covariance matrix of the two variable sets. In discriminant analysis the group centers and group covariance matrices are used for finding discriminant rules that are able to separate two or more groups of data coming from different populations.

Traditionally, the population covariance matrix is estimated by the empirical sample covariance matrix. However, it is well known that outliers in the data can have severe influence to this estimator (see, e.g., Hampel et

al., 1986). For this reason, more robust scatter estimators have been introduced in the literature, for a review see Maronna and Yohai (1998). Although robustness is paid for by lower efficiency of the estimator and a higher computational effort, the resulting estimation will usually be more reliable for the data at hand. Plugging in robust covariance matrices into the before mentioned methods leads to robust counterparts of the multivariate methods. The robustness properties of the resulting estimators have been studied, e.g. Croux and Haesbroeck (2000) for PCA, or Pison et al. (2003) for factor analysis.

There exists another approach to robustify multivariate methods, without passing by a robust estimate of the covariance structure. This so-called Projection Pursuit (PP) approach uses the idea to project the multivariate data onto a lower dimensional space where robust estimation is much easier. PP was initially proposed by Friedman and Tukey (1974), and the original goal was to pursue directions that show the structure of the multivariate data if projected on these directions. This is done by maximizing a PP index, and the direction(s) resulting in a (local) maximum of the index are considered to reveal interesting data structures. Huber (1985) pointed out that PCA is a special case of PP, where the PP index is the variance of the projected data, and where orthogonality constraints have to be included in the maximization procedure. Li and Chen (1985) used this approach to robustify PCA by taking a robust scale estimator. Croux and Ruiz-Gazen (2005) investigated the robustness properties of this robust PCA approach, and they introduced an algorithm for fast computation. Robust estimation using PP was also considered for canonical correlation analysis (Branco et al., 2005), and this approach was compared with the method of robustly estimating the joint covariance matrix and with a robust alternating regression method.

The PP approach has several advantages, including the following:

(a) As mentioned earlier, robust estimation in lower dimension is computationally easier and faster, although on the other hand the search for "interesting" projection directions is again time consuming.
(b) Robust covariance estimation is limited to data sets where the number of observations is larger than the number of variables. Thus, for many problems–like in chemometrics–PP based methods are the methods of choice for a robust data analysis.
(c) The search for projection directions is sequential. Thus, the user can determine a certain number of directions he/she is interested in, and is not forced to perform a complete eigenanalysis of the covariance matrix. Especially for high dimensional problems the computation time can be reduced drastically by PP based methods as the number of interesting directions to be considered is often small.

In this article we will focus on Continuum Regression (CR), a multivariate method introduced by Stone and Brooks (1990) that combines ordinary least squares, partial least squares and principal components regression. Serneels

et al. (2005) introduced robust CR using the PP approach. In the next section we will describe CR and outline how the parameters can be estimated in a robust way. A new algorithm for computation will be introduced in Section 3, and the precision of this algorithm will be compared with the proposed algorithm of Serneels et al. (2005). Section 4 underlines the robustness of this method by presenting simulation results for the case of outliers in the space of the regressor variables. The final section provides a summary.

2 Robust Continuum Regression by Projection Pursuit

CR is a regression technique that was designed for problems with high dimensional regressors and few observations. Therefore, let $\boldsymbol{X}$ be the $n \times p$ matrix of regressors where typically $n \ll p$. Let $\boldsymbol{y}$ be a vector with n observations of the response variable. Like in the regression setting, the model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

with the error term $\boldsymbol{\varepsilon}$ is considered and the focus is on estimating the regression coefficients $\boldsymbol{\beta}$. Since the regressors are usually highly collinear, the coefficients $\boldsymbol{\beta}$ are not directly estimated, but a so-called latent variable model

$$\boldsymbol{y} = \boldsymbol{T}_h\boldsymbol{\xi} + \boldsymbol{\varepsilon} \tag{2}$$

with new regression coefficients $\boldsymbol{\xi}$ is considered. The score matrix $\boldsymbol{T}_h$ is of size $n \times h$ and h, the number of latent variables, is taken much smaller than p. The score matrix is related with the original regressors through $\boldsymbol{T}_h = \boldsymbol{X}\boldsymbol{W}_h$ with $\boldsymbol{W}_h = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_h)$ being a matrix with weights. The weight vectors are defined by

$$\boldsymbol{w}_i = \underset{\boldsymbol{a}}{\operatorname{argmax}}\{\operatorname{Cov}(\boldsymbol{X}\boldsymbol{a}, \boldsymbol{y})^2 \operatorname{Var}(\boldsymbol{X}\boldsymbol{a})^{\frac{\delta}{1-\delta}-1}\} \tag{3}$$

$(i = 1, \ldots, h)$ under the constraints

$$\|\boldsymbol{w}_i\| = 1 \quad \text{and} \quad \operatorname{Cov}(\boldsymbol{X}\boldsymbol{w}_i, \boldsymbol{X}\boldsymbol{w}_j) = 0 \text{ for } j < i. \tag{4}$$

The tuning parameter δ can be chosen in the interval $[0, 1]$. By taking $\delta = 0$ the criterion corresponds to ordinary least squares, $\delta = 0.5$ is the Partial Least Squares (PLS) criterion, and $\delta = 1$ results in principal component regression (see Stone and Brooks, 1990).

The definition (3) of the weight vectors can be understood as PP index that has to be maximized for a projection direction $\boldsymbol{a}$, and for subsequent projection directions the constraints (4) have to be fulfilled. The typically high dimensional regressor matrix $\boldsymbol{X}$ is projected to one dimension, namely $\boldsymbol{X}\boldsymbol{a}$, and the variance "Var" of the projected data as well as the covariance "Cov" between two univariate variables are the basis for finding the weight

vectors. "Var" and "Cov" are usually taken as sample variance and covariance estimators, respectively. By using more robust estimators instead, the influence of outliers will be reduced and the projection directions will be determined in a robust manner, resulting in a robust CR method. Serneels et al. (2005) suggested to take the α-trimmed variance and covariance because these estimators are easy to understand and fast to compute.

The algorithm for (robust) CR based on PP can be summarized as follows:

(a) Fix the number h of latent variables and the tuning parameter δ. The appropriate choice of h and δ is described in Serneels et al. (2005).
(b) Define $\boldsymbol{E}_1$ as the mean centered data matrix $\boldsymbol{X}$. For robust CR, robust mean centering can be achieved by using the L_1-median (for an efficient algorithm see Hössjer and Croux, 1995).
(c) Suppose that the weight vectors $\boldsymbol{W}_{i-1} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{i-1})$ have already been computed.
  (i) The i-th weight vector $\boldsymbol{w}_i$ is determined according to criterion (3) by scanning the projection directions $\boldsymbol{a}$. In Section 3 we will provide more details on this. Multiplying the matrix $\boldsymbol{E}_i$ (see below) with these weights gives the i-th score vector $\boldsymbol{t}_i$.
  (ii) The parameter vector $\boldsymbol{\xi}$ in model (2) is estimated by ordinary least squares in the classical case and in the robust case by any robust regression method, like Huber M-regression (Huber, 1981). Premultiplication with $\boldsymbol{W}_{i-1}$ gives the estimation of the coefficients $\boldsymbol{\beta}$ in the original model (1).
  (iii) Carry out a deflation in order to fulfill the model constraints (4):

$$\boldsymbol{E}_{i+1} = \Big(\boldsymbol{I}_n - \sum_{j=1}^{i-1} \frac{\boldsymbol{t}_j \boldsymbol{t}_j^\top}{\boldsymbol{t}_j^\top \boldsymbol{t}_j}\Big)\boldsymbol{X}. \tag{5}$$

3 Algorithms for Finding the PP Directions

A crucial point of CR is the maximization of the criterion (3) for the weights. In principle, all possible projection directions $\boldsymbol{a} \in \mathbb{R}^p$ have to be scanned, which is impossible especially in situations where p is large. For this reason, the number of candidate directions is limited to a set that is still computable in reasonable time. Serneels et al. (2005) suggested to construct k directions that are arbitrary linear combinations of the n data points at hand, the first n directions being directly the n observations. The computation time as well as the precision of this algorithms will thus strongly depend on the number k of candidate directions.

Here a new algorithm will be introduced and compared with the other proposal. This so-called grid algorithm works as follows. Let $\boldsymbol{x}_i$ $(i = 1, \ldots, p)$ be the columns, or "variables" of the data matrix $\boldsymbol{X}$.

(a) If $p = 2$:

(i) A first approximation $\boldsymbol{a}^1$ of the projection direction $\boldsymbol{a}$ is obtained by maximizing

$$C(\gamma_{1j}\boldsymbol{x}_1 + \gamma_{2j}\boldsymbol{x}_2) = \mathrm{Cov}(\gamma_{1j}\boldsymbol{x}_1 + \gamma_{2j}\boldsymbol{x}_2, \boldsymbol{y})^2 \mathrm{Var}(\gamma_{1j}\boldsymbol{x}_1 + \gamma_{2j}\boldsymbol{x}_2)^{\frac{\delta}{1-\delta}-1} \tag{6}$$

under the constraints $\gamma_{1j}^2 + \gamma_{2j}^2 = 1$ for $j = 1, \ldots, N$. The unknowns γ_{1j} and γ_{2j} are the coordinates of G grid points regularly chosen on the unit circle in the interval $[-\pi/2, \pi/2)$, and the maximum is taken among these G candidate directions.

(ii) The second approximation $\boldsymbol{a}^2$ is searched like before, but in a smaller interval $[-\pi/(2^f), \pi/(2^f))$ with $f = 2$. In each new iteration f is increased by 1, until after F interval halving steps the grid is fine enough to leave the solution essentially unchanged (marginal improvement smaller than a tolerance bound).

(b) If $p > 2$:

(i) Compute for each regressor variable $i = 1, \ldots, p$ the value of the objective function

$$C(\boldsymbol{x}_i) = \mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{y})^2 \mathrm{Var}(\boldsymbol{x}_i)^{\frac{\delta}{1-\delta}-1} \tag{7}$$

and sort the variables $\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(p)}$, being in the columns of $\boldsymbol{X}$, according to $C(\boldsymbol{x}_{(1)}) \geq C(\boldsymbol{x}_{(2)}) \geq \ldots \geq C(\boldsymbol{x}_{(p)})$.

(ii) The maximization is done now in the plane like in (a): Maximizing $C(\gamma_{1j}\boldsymbol{x}_{(1)} + \gamma_{2j}\boldsymbol{x}_{(2)})$ results in the approximation $\boldsymbol{a}^{(1)}$. A next approximation $\boldsymbol{a}^{(2)}$ is obtained by maximizing $C(\gamma_{1j}\boldsymbol{X}\boldsymbol{a}^{(1)} + \gamma_{2j}\boldsymbol{x}_{(3)})$. This procedure is repeated until the last variable has entered the optimization. In a next cycle each variable is considered again for improving the value of the objective function. The algorithm terminates when the improvement is considered to be marginal.

The precision of both algorithms is computed using the "Fearn" data (Fearn, 1983) which consists of 24 observations and 6 regressor variables. For $\delta = 0.5$ we compute all $h = 6$ latent variables. Since $\delta = 0.5$ corresponds to PLS, the solutions of both algorithms can be compared with the exact solution resulting from the SIMPLS algorithm (de Jong, 1993) in the case when the empirical sample variance and covariance are used in the criterion (3). The resulting regression coefficients are compared by computing the sum of all elementwise squared differences to the exact regression coefficients. This can be considered as measure of precision of the algorithm, which needs to be as small as possible. Since the precision measure could depend on the specifically generated directions for the algorithm described in Serneels et al. (2005), we average the precision measure over 100 runs. In Figure 1 the resulting precisions are presented for different parameter choices of the algorithms. For the algorithm of Serneels et al. (2005) different numbers of directions k are considered (scale on top), and for the grid algorithm different

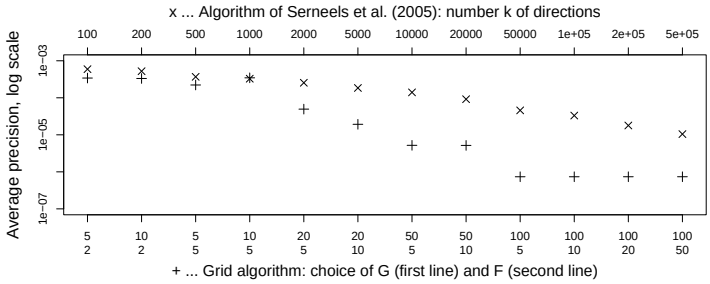

Fig. 1. Average precision for the regression coefficients of the Fearn data resulting from two different algorithms.

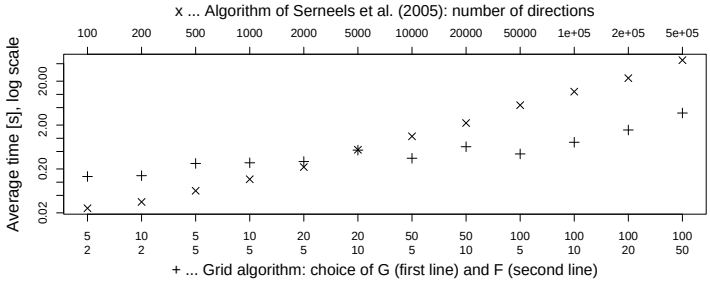

Fig. 2. Average computation time (in seconds) for both algorithms, see Figure 1.

numbers of grid points G and interval halving steps F are used (scale on bottom). From Figure 1 we see that the precision is comparable for $k = 1000$ directions and the choice $G = 10$ and $F = 5$. By taking more computational effort, the precision is getting much better for the grid algorithm.

It is also interesting to compare the algorithms with respect to computation time. Figure 2 presents the average computation time corresponding to the results of Figure 1. While the precision is about the same for $k = 1000$ and $G = 10$ and $F = 5$, the grid algorithm needs roughly twice as much time. On the other hand, the time for both algorithms is about the same for the parameters $k = 5000$ and $G = 20$, $F = 10$, but the precision of the grid algorithm is about $2 \cdot 10^{-5}$ compared to $2 \cdot 10^{-4}$ for the other algorithm. In general, if higher precision is needed, the grid algorithm will be much faster and at the same time more precise. On the other hand, if moderate precision is sufficient, the Serneels et al. (2005) algorithm is to be preferred.

4 Simulation

The advantage of robust CR over classical CR in presence of contamination was already demonstrated in Serneels et al. (2005) by simulations and an example. In the simulations different distributions of the error term $\boldsymbol{\varepsilon}$ in the

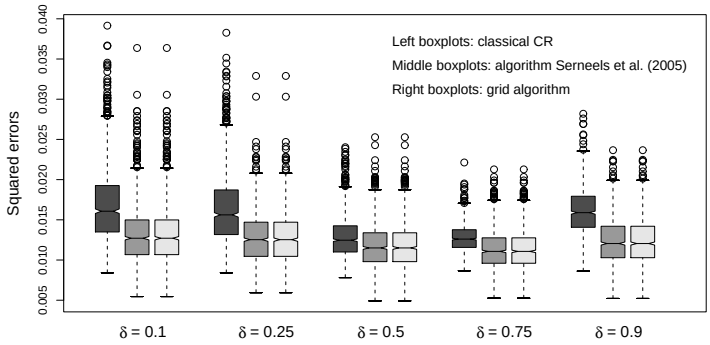

Fig. 3. Squared Errors from the simulation with outliers in the regressor variables.

model (1) were considered. We recomputed the simulations for the grid algorithm and obtained similar results as for the previously proposed algorithm.

Here we will consider the situation of outliers in the regressor variables. The matrix $\boldsymbol{X}$ of size $n \times p$ with $n = 100$ and $p = 10$ is generated from $N_p(\mathbf{0}, \boldsymbol{C})$, a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{C} = \mathrm{diag}(1, 1/2, \ldots, 1/p)$. $\boldsymbol{W}_h$ is constructed to fulfill the constraints (4) with $h = 3$, and $\boldsymbol{\xi}$ is generated from a uniform random distribution in $[0.5, 1]$. These matrices are fixed for a particular simulation setup. Hence, the true regression parameter $\boldsymbol{\beta} = \boldsymbol{W}_h\boldsymbol{\xi}$ is known. Then the error term is generated according to $\boldsymbol{\varepsilon} \sim N(0, 1/10)$ and 10% of the rows of $\boldsymbol{X}$ are replaced by outliers coming from $N_p(5 \cdot \mathbf{0}, \boldsymbol{I}_p)$. For several values of the tuning parameter δ the classical CR algorithm, the algorithm of Serneels et al. (2005) and the grid algorithm was applied in 1000 simulation replications. the resulting Squared Errors $\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(i)}_{\delta,h}\right)^\top \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(i)}_{\delta,h}\right)$ were computed for the estimated regression coefficients $\hat{\boldsymbol{\beta}}^{(i)}_{\delta,h}$ in the i-th simulation obtained from the different algorithms, and the results are presented by parallel boxplots in Figure 3. Each group of three boxplots corresponds to a different value of δ. Both algorithms for robust CR lead to comparable results, at least for the choice $k = 1000$ directions, $G = 10$ grid points and $F = 2$, and $\alpha = 10\%$ trimmed variance and covariance estimators. For all choices of δ the notches of the classical boxplots do not overlap with the robust ones, which is a strong evidence that the Squared Errors of the classical procedure is higher as for the robust ones, due to the presence of contamination.

5 Summary

The robustification of multivariate methods by plugging in robust covariance matrix estimates is limited to the case $n > p$. This limitation does not hold for methods based on PP, and the robustness can be achieved by applying

robust estimators to the projected data. Here we outlined the procedure for robust CR, and a new algorithm was introduced. Robust CR turns out to be robust against outliers in the error terms, but also robust with respect to outliers in the regressor variables, as was shown by the simulations in this paper. Programs for computation are available in the Matlab programming environment from the first author.

References

BRANCO, J.A., CROUX, C., FILZMOSER, P., and OLIVEIRA, M.R. (2005): Robust Canonical Correlations: A Comparative Study. *Computational Statistics, 2*. To appear.

CROUX, C. and HAESBROECK, G. (2000): Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika, 87, 603–618*.

CROUX, C. and RUIZ-GAZEN, A. (2005): High Breakdown Estimators for Principal Components: The Projection-pursuit Approach Revisited. *Journal of Multivariate Analysis*. To appear.

DE JONG, S. (1993): SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems, 18, 251–263*.

FEARN, T. (1983): A Misuse of Ridge Regression in the Calibration of a Near Infrared Reflectance Instrument. *Applied Statistics, 32, 73–79*.

FRIEDMAN, J.H., and TUKEY, J.W. (1974): A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers, 9, 881–890*.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W. (1986): *Robust Statistics. The Approach Based on Influence Functions*. John Wiley & Sons, New York.

HÖSSJER, O. and CROUX, C. (1995): Generalizing Univariate Signed Rank Statistics for Testing and Estimating a Multivariate Location Parameter. *Nonparametric Statistics, 4, 293–308*.

HUBER, P.J. (1981): *Robust Statistics*. John Wiley & Sons, New York.

HUBER, P.J. (1985): Projection Pursuit. *The Annals of Statistics, 13, 435–525*.

LI, G., and CHEN, Z. (1985): Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *Journal of the American Statistical Association, 80, 391, 759–766*.

MARONNA, R.A. and YOHAI, V.J. (1998): Robust Estimation of Multivariate Location and Scatter. In: S. Kotz, C. Read and D. Banks (Eds.): *Encyclopedia of Statistical Sciences*. John Wiley & Sons, New York, 589–596.

PISON, G., ROUSSEEUW, P.J., FILZMOSER, P., and CROUX, C. (2003): Robust Factor Analysis. *Journal of Multivariate Analysis, 84, 145–172*.

SERNEELS, S., FILZMOSER, P., CROUX, C. and VAN ESPEN, P.J. (2005): Robust Continuum Regression. *Chemometrics and Intelligent Laboratory Systems, 76, 197–204*.

STONE, M. and BROOKS, R.J. (1990): Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society B, 52, 237–269*.