Robust Principal Component Regression

P. Filzmoser
Dept. of Statistics, Prob. Theory, and Actuarial Maths.
Vienna University of Technology, AUSTRIA
e-mail: `P.Filzmoser@tuwien.ac.at`

Abstract: In this note we introduce a method for robust principal component regression. Robust principal components are computed from the predictor variables, and they are used afterwards for estimating a response variable by performing robust linear multiple regression. The performance of the method is evaluated at a test data set from geochemistry. Then it is used for the prediction of censored values of gold.

1 Introduction

The purpose of principal component regression (PCR) is to estimate the values of a response variable at the basis of selected principal components (PCs) of the explanatory variables. There are two main reasons for regressing the response variable on the PCs rather than directly on the explanatory variables. Firstly, the explanatory variables are often highly correlated (multicollinearity) which may cause inaccurate estimations of the least squares (LS) regression coefficients. This can be avoided by using the PCs in place of the original variables since the PCs are uncorrelated. Secondly, the dimensionality of the regressors is reduced by taking only a subset of PCs for prediction.

LS regression as well as classical principal component analysis (PCA) are vulnerable with respect to outlying observations. Even one massive outlier can heavily influence the parameter estimates of these methods. It is thus important to robustify PCR which in fact means to robustify both PCA and linear multiple regression. Walczak and Massart (1995) introduced a robust

PCR method for identifying outliers. Their method uses robust PCA based on a robust covariance matrix (ellipsoidal multivariate trimming), followed by least median of squares (LMS) regression (Rousseeuw and Leroy, 1987).

2 Robust PCR

For robustifying PCR we will use a robust PCA method proposed by Li and Chen (1985) which is based on the idea of projection pursuit. In contrary to most other robust PCA methods, the number of variables can exceed the number of observations. The resulting PCs are highly robust. Croux and Ruiz-Gazen (1996) introduced a fast algorithm which works as follows. For given observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$, which are collected in the rows of the data matrix $\boldsymbol{X}$, a coefficient vector $\boldsymbol{b} \in \mathbb{R}^p$ is defined for 1-dimensional projection of the data. Suppose that the first $k-1$ projection directions $\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_{k-1}$ (eigenvectors) ($k > 1$) have already been found. For finding the kth eigenvalue, a projection matrix is defined as $\boldsymbol{P}_k = \boldsymbol{I}_p - \sum_{j=1}^{k-1} \hat{\boldsymbol{\gamma}}_j \hat{\boldsymbol{\gamma}}_j^\top$ for projection on the orthogonal complement of the space spanned by the first $k-1$ eigenvectors (for $k = 1$ we can take $\boldsymbol{P}_k = \boldsymbol{I}_p$). The kth eigenvector is then defined by maximizing the function $\boldsymbol{b} \longrightarrow S(\boldsymbol{X}\boldsymbol{P}_k\boldsymbol{b})$ under the conditions $\boldsymbol{b}^\top\boldsymbol{b} = 1$ and $\boldsymbol{P}_k\boldsymbol{b} = \boldsymbol{b}$. S is a univariate scale estimator. For S we can use the classical standard deviation which results in classical PCA, or a robust measure of spread like the median absolute deviation (MAD) to obtain a robust PCA method. Note that the PCs are computed sequentially, and thus one can stop at a desired number of components to save computation time.

At the basis of the robust PCs we have to select a subset $k < p$ of components for predicting the response variable. One could simply take those first k PCs which include the main variation of the data set (sequential selection). However, it is known that PCs with the largest variances are not necessarily the best predictors. Hence we make a stepwise selection of PCs. We start with that PC resulting in the best prediction of the response variable (according to an appropriate association measure) and increase the number of PCs in each step until the quality of prediction cannot be essentially increased.

The prediction is done by robust regression. We prefer to use least trimmed squares (LTS) regression since it is very robust, has good statistical properties, and a fast algorithm exists (Rousseeuw and Van Driessen, 2000).

3 Example

We consider a data set from geochemistry which is available in form of a geochemical atlas (Reimann et al., 1998). It can be downloaded at *http://www.pangaea.de/Projects/Kola-Atlas/*. An area of 188000 km^2 in the so-called Kola region at the boundary of Norway, Finland, and Russia was sampled. One of the 5 sample media was the C-horizon of podzol profiles, developed on glacial drift. Although more than 50 chemical elements have been analyzed for all 606 samples, some of the most interesting elements like As, Au, Cs, Hg, Mo, Sb, Ta, U, and Zn, are censored data, i.e. a high proportion of samples returned values below detection. When strongly censored data are used for geochemical mapping, noisy maps, showing little or no regional structure, will be the result, even when smoothing methods like kriging are used to produce a smoothed surface map. We will apply our method to predict the censored data.

To test the prediction quality of robust PCR, several elements (Ag, As, Bi, Cr, Pb) without any censored data were used. The detection limit problem is simulated by deleting first the lower 50% and then the lower 60%, 70% and 80% of the data values. As an example let us consider Cr with 80% of the lower values deleted. The number of PCs for predicting the upper 20% of the values of Cr can be selected due to the sequential or stepwise criterion mentioned above. The criterion for the sequential selection of PCs suggests 30 components which result in a robust coefficient of determination (Rousseeuw and Leroy, 1987) of about 63%. The same percentage is achieved by 16 components selected by the stepwise criterion. 23 components explain even 78% of the variation of the upper part of Cr. Figure 1 shows the residual plots using LS regression (left) and LTS regression (right) of the upper 20% of the Cr values on the 23 selected robust PCs. The plots present the predicted values against the

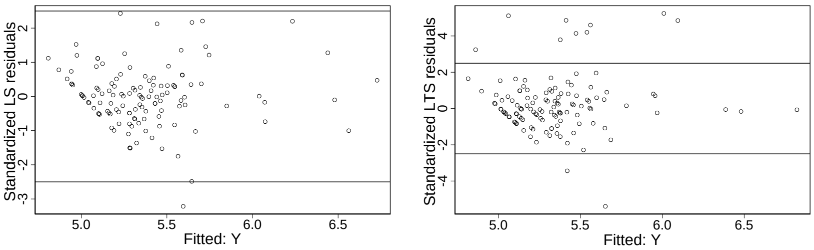
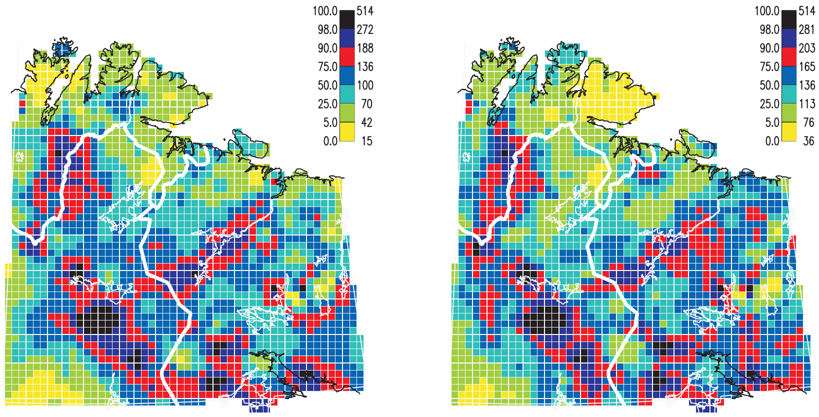


Figure 1: Residual plots for LS and LTS regression.

standardized residuals. The horizontal lines at ± 2.5 indicate outliers which are clearly visible in the plot for LTS regression, and masked by the LS regression. Figure 2 shows smoothed maps based on the krige estimation for Cr. The left map presents the original data, and the right map the concentrations predicted by robust PCR, combined with the original upper 20% of the values. This map demonstrates that the regional structure of Cr was successfully predicted. All other elements tested (Ag, As, Bi, and Pb) showed comparable results. It turned out that the method tends to overestimate the values at the lower end of the distribution. A possible explanation is the inhomogeneity of the data which is caused by several processes (e.g. lithology).

Following the successful test, we want to predict the regional structure of gold (Au) for the C-horizon where 72.4% of the values are below the detection limit. 21 PCs, selected according to the stepwise criterion, explain about 48% of the variation of the



Original Cr in C–horizon

Predicted Cr (80% missing)

Figure 2: Original and predicted Cr.

upper values of Au. Figure 3 shows the original (left) and the predicted (right) values (the upper part is replaced by the original values) by smoothed surface maps. The map of the predicted data show a very clear regional structure, and it could be used e.g. for developing gold exploration concepts for this area.

References

1. Croux C., Ruiz-Gazen A. (1996). A fast algorithm for robust principal components based on projection pursuit. In Prat A. (ed.) *Computational Statistics*. Physica-Verlag, Heidelberg, pp. 211-216.

2. Li G., Chen Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Amer. Statist. Assoc.*. Vol. **80**(391), pp. 759-766.

3. Reimann C., Äyräs M., Chekushin V., Bogatyrev I., Boyd R., Caritat P. de, Dutter R., Finne T.E., Halleraker J.H.,

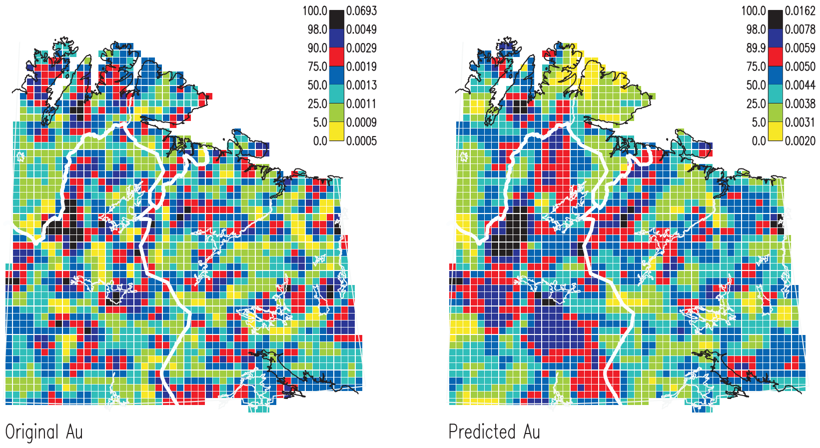


Figure 3: Original and predicted Au.

Jæger Ø., Kashulina G., Lehto O., Niskavaara H., Pavlov V., Räisänen M.L., Strand T., Volden T. (1998). *Environmental Geochemical Atlas of the Central Barents Region*. NGU-GTK-CKE special publication. Geological Survey of Norway, Trondheim.

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

5. Rousseeuw P.J., Van Driessen K. (2000). A fast algorithm for highly robust regression in data mining. In Bethlehem J.G., van der Heijden P.G.M. (eds.) *COMPSTAT: Proceedings in Computational Statistics*. Physica-Verlag, Heidelberg, pp. 421-426.

6. Walczak B., Massart D.L. (1995). Robust principal components regression as a detection tool for outliers. *Chemometrics and Intelligent Laboratory Systems*. Vol. **27**, pp. 41-54.