Robust Multivariate Methods in Geostatistics

Peter Filzmoser[1], Clemens Reimann[2]

[1]Department of Statistics, Probability Theory, and Actuarial Mathematics,
Vienna University of Technology, A-1040 Vienna, Austria

[2]Geological Survey of Norway, N-7491 Trondheim, Norway

Abstract: Two robust approaches to principal component analysis and factor analysis are presented. The different methods are compared, and properties are discussed. As an application we use a large geochemical data set which was analyzed in detail by univariate (geo-)statistical methods. We explain the advantages of applying robust multivariate methods.

1 Introduction

In regional geochemistry an advantage could be that instead of presenting maps for 50 (or more) chemical elements only a few maps of the principal components or factors may have to be presented, containing a high percentage of the information of the single element maps. Additionally, it might be possible to find effects which are not visible in the single element maps. Especially factor analysis is used in different kinds of applications to detect hidden structures in the data.

Geochemical data sets usually include outliers which are caused by a multitude of different processes. It is well known that outliers can heavily influence classical statistical methods, including multivariate statistical methods. Even one single (huge) outlier can completely determine the result of principal component analysis. For that reason it is advisable to use robust multivariate methods for detecting the multivariate structure. Section 2 treats two methods of robust principal component analysis. Two different versions of robust factor analysis which have recently been proposed, are considered in Sections 3 and 4. Section 5 gives an example with a real geochemical data set.

2 Robust Principal Component Analysis

Let $\boldsymbol{x}$ be a p-dimensional random vector with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. The covariance matrix can be decomposed as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^\top$, where the columns of $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_{.1}, \ldots, \boldsymbol{\gamma}_{.p})$ are the eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues (arranged in descending order) of $\boldsymbol{\Sigma}$. The principal components of $\boldsymbol{x}$ are defined by $\boldsymbol{z} = \boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})$. Classically, $\boldsymbol{\mu}$ is estimated by the sample mean $\bar{\boldsymbol{x}}$, and

$\boldsymbol{\Sigma}$ by the sample covariance matrix $\boldsymbol{S}$, which is decomposed into eigenvectors and -values. $\bar{\boldsymbol{x}}$ as well as $\boldsymbol{S}$ are highly sensitive with respect to outlying observations. Hence, for seriously analyzing geochemical data, a robust version of principal component analysis (PCA) has to be applied.

PCA can easily be robustified by estimating the covariance matrix $\boldsymbol{\Sigma}$ in a robust way, e.g. by taking the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985). The robustly estimated covariance matrix is not influenced by outliers, and hence the eigenvector/eigenvalue decomposition is also robust. Since the MCD additionally gives a robust estimation of $\boldsymbol{\mu}$, the whole PCA procedure is robust. We will discuss the usage of the MCD estimator in more detail in the context of factor analysis (Section 3).

Another way for robustifying PCA was introduced by Li and Chen (1985). The method is based on the projection pursuit technique. PCA can be seen as a special case of projection pursuit, where the variance of the projected data points is to be maximized. Let $\boldsymbol{X} = (\boldsymbol{x}_{1.}^\top, \ldots, \boldsymbol{x}_{n.}^\top)^\top$ be a data matrix with observation vectors $\boldsymbol{x}_{i.} \in \mathbb{R}^p$ $(i = 1, \ldots, n)$. Now, let us assume that the first $(k-1)$ projection directions $\widehat{\boldsymbol{\gamma}}_{.1}, \ldots, \widehat{\boldsymbol{\gamma}}_{.(k-1)}$ are already known. We define a projection matrix

$$\boldsymbol{P}_1 = \boldsymbol{I}_p \ , \qquad \boldsymbol{P}_k = \boldsymbol{I}_p - \sum_{j=1}^{k-1} \widehat{\boldsymbol{\gamma}}_{.j} \widehat{\boldsymbol{\gamma}}_{.j}^\top \ . \tag{1}$$

$\boldsymbol{P}_k$ corresponds to a projection onto the space spanned by the first $(k-1)$ projection directions. We are interested in finding a projection direction $\boldsymbol{a}$ which maximizes the function

$$\boldsymbol{a} \quad \longrightarrow \quad S(\boldsymbol{X}\boldsymbol{P}_k\boldsymbol{a}) \tag{2}$$

under the restrictions $\boldsymbol{a}^\top\boldsymbol{a} = 1$ and $\boldsymbol{P}_k\boldsymbol{a} = \boldsymbol{a}$ (orthogonality to previously found projection directions). Defining S in (2) as the classical sample standard deviation would result in classical PCA. The method can easily be robustified by taking a robust measure of spread, e.g. the *median absolute deviation* (MAD)

$$\mathrm{MAD}(\boldsymbol{y}) = \operatorname*{med}_i |y_i - \operatorname*{med}_j (y_j)| \ . \tag{3}$$

Since the number of possible projection directions is infinite, an approximative solution for maximizing (2) is as follows. The k-th projection direction is only searched in the set

$$A_{n,k} = \left\{ \frac{\boldsymbol{P}_k(\boldsymbol{x}_{1.} - \widehat{\boldsymbol{\mu}}_n)}{\|\boldsymbol{P}_k(\boldsymbol{x}_{1.} - \widehat{\boldsymbol{\mu}}_n)\|}, \ldots, \frac{\boldsymbol{P}_k(\boldsymbol{x}_{n.} - \widehat{\boldsymbol{\mu}}_n)}{\|\boldsymbol{P}_k(\boldsymbol{x}_{n.} - \widehat{\boldsymbol{\mu}}_n)\|} \right\} \tag{4}$$

where $\widehat{\boldsymbol{\mu}}_n$ denotes a robust estimation of the mean, like the L_1-median or the component-wise median.

The algorithm outlined above was suggested by Croux and Ruiz-Gazen (1996). It is easy to implement and fast to compute which makes the method quite attractive to use in practice. Furthermore, this robust PCA method has a big advantage for high-dimensional data (large p) because it allows to stop at a desired number $k < p$ of components, whereas usually all p components are to be extracted by the eigenvector/eigenvalue decomposition of the (robust) covariance matrix. Moreover, the method still gives reliable results for $n < p$, which is important for a variety of applications. The computation of the MCD estimator requires at least $n > p$.

3 Robust Factor Analysis using the MCD

The aim of factor analysis (FA) is to summarize the correlation structure of observed variables $x_1, \ldots, x_p$. For this purpose one constructs $k < p$ unobservable or latent variables $f_1, \ldots, f_k$, which are called the *factors*, and which are linked with the original variables through the equation

$$x_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \ldots + \lambda_{jk} f_k + \varepsilon_j, \tag{5}$$

for each $1 \leq j \leq p$. The error variables $\varepsilon_1, \ldots, \varepsilon_p$ are supposed to be independent, but they have *specific variances* $\psi_1, \ldots, \psi_p$. The coefficients λ_{jl} are called the factor *loadings*, and they are collected into the matrix of loadings $\boldsymbol{\Lambda}$.

Using the vector notations $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$, $\boldsymbol{f} = (f_1, \ldots, f_k)^\top$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^\top$, the usual conditions on factors and error terms can be written as $E(\boldsymbol{f}) = E(\boldsymbol{\varepsilon}) = 0$, $\mathrm{Cov}(\boldsymbol{f}) = \boldsymbol{I}_k$, and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$, with $\boldsymbol{\Psi}$ a diagonal matrix containing on its diagonal the specific variances. Furthermore, $\boldsymbol{\varepsilon}$ and $\boldsymbol{f}$ are assumed to be independent.

From the above conditions it follows that the covariance matrix of $\boldsymbol{x}$ can be expressed by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}. \tag{6}$$

In classical FA the matrix $\boldsymbol{\Sigma}$ is estimated by the sample covariance matrix. Afterwards, decomposition (6) is used to obtain the estimators for $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. Many methods have been proposed for this decomposition, of which maximum likelihood (ML) and the principal factor analysis (PFA) method are the most frequently used.

Similar to the previous section, the parameter estimates can heavily be influenced when using a classical estimation of the scatter matrix. The problem can be avoided when $\boldsymbol{\Sigma}$ is estimated by the MCD estimator, which looks for the subset of h out of all n observations having the smallest determinant of its covariance matrix. Typically, $h \approx 3n/4$.

Pison et al. (1999) used the MCD for robustifying FA. They have shown that PFA based on MCD results in a resistant FA method with bounded influence function. It has better robustness properties than the ML-based counterpart. The empirical influence function can be used as a data-analytic tool. The method is also attractive for computational reasons since a fast algorithm for the MCD estimator has recently been developed (Rousseeuw and Van Driessen (1999)).

4 FA using Robust Alternating Regressions

A limitation of the MCD-based approach is that the sample size n needs to be bigger than the number of variables p. For samples with $n \leq p$ (which occur quite frequently in the practice of FA), a robust FA technique based on alternating regressions, originating from Croux et al. (1999), can be used.

For this we consider the sample version of model (5):

$$x_{ij} = \sum_{l=1}^{k} \lambda_{jl} f_{il} + \varepsilon_{ij} \tag{7}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Suppose that preliminary estimates for the factor scores f_{il} are known, and consider them as constants for a moment. The loadings λ_{jl} can now be estimated by linear regressions of the x_j's on the factors. Moreover, by applying a robust scale estimator on the computed residuals, estimates $\hat{\psi}_j$ for ψ_j can easily be obtained (for example by computing the MAD of the residuals).

On the other hand, if preliminary estimates of the loadings are available, linear regression estimators can again be used for estimating the factor scores. Indeed, if we take i fixed in (7) and suppose that the λ_{jl} are fixed, a regression of x_{ij} on the loadings λ_{jl} yields updated estimates for the factor scores. Since there is heteroscedasticity, weights proportional to $(\hat{\psi}_j)^{-1/2}$ should be included.

Using robust principal components (Section 2) as appropriate starting values for the factor scores, an iterative process (called alternating or interlocking regressions) can be carried out to estimate the unknown parameters of the factor model. To ensure robustness of the procedure we use a weighted L_1-regression estimator since it is fast to compute and very robust. More details about the method and the choice of the weights can be found in Croux et al. (1999). Note that in contrast to the method described in Section 3, the factor scores are estimated *directly*.

5 Example

We consider a data set described and analyzed by univariate methods in Reimann et al. (1998). From 1992-1998 the Geological Surveys of Finland (GTK), and Norway (NGU) and the Central Kola Expedition (CKE), Russia, carried out a large multi-element geochemical mapping project, covering an area of 188,000 km^2 between 24° and 35.5°E up to the Barents Sea coast. One of the sample media was the C-horizon of podzol profiles, developed on glacial drift. C-horizon samples were taken at 605 sites, and the contents of more than 50 chemical elements was measured for all samples. Although the project was mainly designed to reveal the environmental conditions in the area, the C-horizon was sampled to reflect the geogenic background.

In the following we will apply the alternating regression-based FA approach (Section 4). Robust PCA and MCD-based FA was used in Filzmoser (1999) for the upper layer, humus, of the complete data set.

For the investigation of the C-horizon data we only considered the elements Ag, Al, As, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, K , Mg, Mn, Na, Ni, P , Pb, S, Si, Sr, Th, V and Zn. These variables have been transformed to a logarithmic scale to give a better approximation to the normal distribution. In order to put everything to a common scale we first standardized (robustly) the variables to mean zero and variance one. We want to analyze the data by using non-robust least squares (LS) regression and robust weighted L_1-regression in the alternating regression scheme. We decided to extract 6 factors which results in a proportion of total variance of 75% for both cases. The loadings of factors $F1$ to $F6$ are shown in Figure 1. We just printed the elements with an absolute value of the loadings larger than 0.3 to avoid confusion. The percentage of explained variance is printed at the top of the plots. Figure 1 shows that for the first factor $F1$ there is just a slight difference between the non-robust (a) and the robust (b) method. However, for the subsequent factors this difference grows. Especially the loadings of factors $F4$ and $F6$ are strongly changing.

It is also interesting to inspect the factor scores which are directly estimated by our method. Because of space limitations we only show the scores of the second factor $F2$ (Figure 2), which is interesting because it nicely reflects the distribution of alkaline intrusions in the survey area. Figure 2 shows the whole region under consideration. The dark lines are the borders of the countries Russia (east), Norway (north-west), and Finland (south-west). The gray lines show rivers and the coast.

At a first glance the two results presented in Figure 2, the non-robust (a) and the robust (b) scores of factor $F2$ seem to be very similar. But already the ranges of the estimated scores are different ($[-3.08, 4.98]$ for the non-robust and $[-3.82, 5.13]$ for the robust method (in the maps we used the same scaling). The smaller range is typical for LS-based methods because *all* data points, including the outliers, are tried to be fitted. Robust methods fit the majority of "good" data points which leads to a reliable estimation. As a consequence, the regions with high and low outliers are presented more reliable by the robust method. In the map the two uppermost classes (crosses) mark areas which are underlain by alkaline bedrocks. The anomalies in the factor maps are much more prominent than the intrusions themselves in a geological map. The reason is that the emplacement of the intrusions was accompanied by the movement of large amounts of hydrothermal fluids. These changed the chemical composition of the intruded bedrocks. The map thus reflects the alteration haloes of these intrusions and demonstrates the importance of the geological process for a very large region.

References

CROUX, C., FILZMOSER, P., PISON, G., and ROUSSEEUW, P. J. (1999): Fitting Factor Models by Robust Interlocking Regression. Preprint, Vienna University of Technology.

CROUX, C. and RUIZ-GAZEN, A. (1996): A Fast Algorithm for Robust Principal Components based on Projection Pursuit, in Prat (Ed.): Proceedings in Computational Statistics, Physika-Verlag, Heidelberg.

FILZMOSER, P. (1999): Robust Principal Component and Factor Analysis in the Geostatistical Treatment of Environmental Data. *Environmetrics, 10, 363-375.*

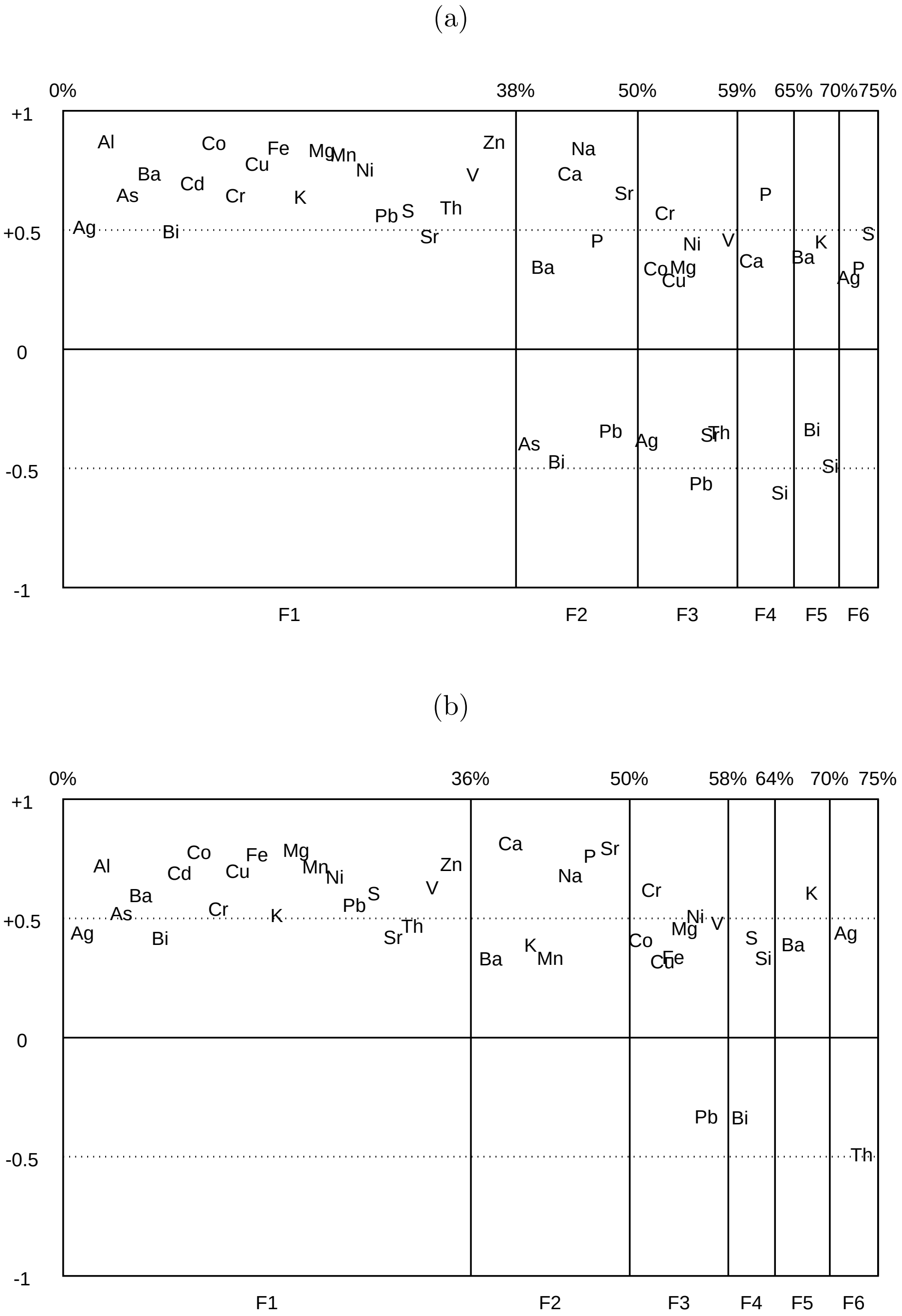
LI, G. and CHEN, Z. (1985): Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *J. Amer. Statist. Assoc., 80, 759-766.*

PISON, G., ROUSSEEUW, P. J., FILZMOSER, P., and CROUX, C. (1999): Robust Factor Analysis. Preprint, Vienna University of Technology.

REIMANN, C., ÄYRÄS, M., CHEKUSHIN, V., BOGATYREV, I., BOYD, R., CARITAT, P. DE, DUTTER, R., FINNE, T. E., HALLERAKER, J. H., JÆGER, Ø., KASHULINA, G., LEHTO, O., NISKAVAARA, H., PAVLOV, V., RÄISÄNEN, M. L., STRAND, T., and VOLDEN, T. (1998): Environmental Geochemical Atlas of the Central Barents Region. Geological Survey of Norway (NGU), Geological Survey of Finland (GTK), and Central Kola Expedition (CKE), Special Publication, Trondheim, Espoo, Monchegorsk.

ROUSSEEUW, P. J.(1985): Multivariate Estimation with High Breakdown Point, in Grossmann et al. (Eds.): Mathematical Statistics and Applications, Vol. B, Akadémiai Kiadó, Budapest.

ROUSSEEUW, P. J. and VAN DRIESSEN, K. (1999): A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics, 41, 212-223.*


(a)

(b)

Figure 1: Loadings of the alternating regression based FA method using (a) LS-regression and (b) weighted L_1-regression.

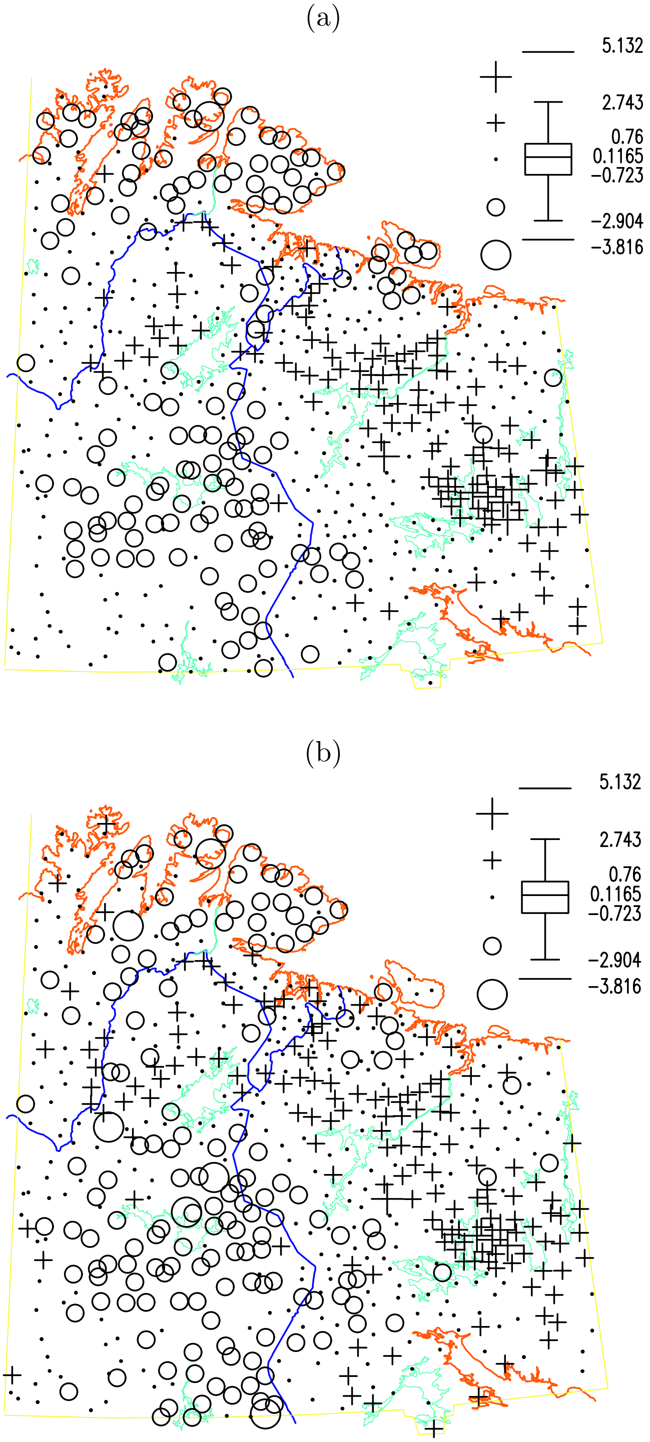


Figure 2: Scores of the second factor of the alternating regression based FA method using (a) LS-regression and (b) weighted L_1-regression.