Robust Factorization of a Data Matrix

Christophe Croux[1] and Peter Filzmoser[2]

[1] ECARE and Institut de Statistique, Université Libre de Bruxelles, Av. F.D. Roosevelt 50, B-1050 Bruxelles, Belgium
[2] Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria

Abstract. In this note we show how the entries of a data matrix can be approximated by a sum of row effects, column effects and interaction terms in a robust way using a weighted L_1 estimator. We discuss an algorithm to compute this fit, and show by a simulation experiment and an example that the proposed method can be a useful tool in exploring data matrices. Moreover, a robust biplot is produced as a byproduct.

Keywords. Alternating regressions, biplot, factor model, robustness

1 Introduction

Multivariate data can often be represented in the form of a data matrix whose elements will be denoted by y_{ij}, where $1 \leq i \leq n$ denotes the row index, and $1 \leq j \leq p$ the column index. Each entry in the data matrix is supposed to be the realization of a random variable

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \tag{1}$$

where μ_{ij} is the median value of each variable Y_{ij} and the residuals ε_{ij} are supposed to form a white noise. It is assumed that the values μ_{ij} can be decomposed as a sum of four terms:

$$\mu_{ij} = c + a_i + b_j + \sum_{l=1}^{k} \lambda_{jl} f_{il}, \tag{2}$$

with $k \leq p$. The constant c can be interpreted as an overall median, a_i as a row effect and b_j as a column effect. The last term represents the interaction between rows and columns and is factorized as the scalar product between a vector of *loadings* $\lambda_{j\cdot} = (\lambda_{j1}, \ldots, \lambda_{jk})^\top$ and a vector of *scores* $f_{i\cdot} = (f_{i1}, \ldots, f_{ik})^\top$. The above model is like the FANOVA model introduced by Gollob (1968), which combines aspects of analysis of variance and factor analysis. We are mainly interested in data matrices in which the rows represent individuals and the columns variables, possibly representing different types of measurement. Therefore we will not continue to pursue symmetry between rows and columns. To identify uniquely the parameters a_i, b_j, and c, the following restrictions are imposed:

$$\operatorname*{med}_i(a_i) = \operatorname*{med}_j(b_j) = 0 \quad \text{and} \quad \operatorname*{med}_i(f_{il}) = \operatorname*{med}_j(\lambda_{jl}) = 0, \tag{3}$$

for $l = 1, \ldots, k$. Furthermore, the scores are standardized by imposing $f_{1l}^2 + \cdots + f_{nl}^2 = 1$ for $l = 1, \ldots, k$. Note that there is no orthogonality condition for the factors, implying that the vectors of loadings $\lambda_{j\cdot}$ and scores $f_{i\cdot}$ are not uniquely determined, as is common in factor models.

By taking $k = 2$, and representing in the same two-dimensional plot the rows by (f_{i1}, f_{i2}) and the columns by $(\lambda_{j1}, \lambda_{j2})$, a biplot is obtained. The biplot allows us to investigate the row and column interaction by visual inspection of a two-dimensional graphical display.

Among others, Gabriel (1978) considered models like (2) and estimated the unknown parameters using a least squares fit. It is however well known that an LS-based method is very vulnerable in the presence of outliers. In this paper, we will propose a robust approach to fit model (2), show by a simulation experiment its merits and illustrate it with an example.

2 A robust fit

A first suggestion is to use the L_1-criterion to fit the model. If we denote by θ the vector of all unknown parameters in the model, and by $\hat{y}_{ij}(\theta) = \hat{\mu}_{ij}(\theta)$ the corresponding fit, then this procedure minimizes the objective function

$$\sum_{i=1}^{n}\sum_{j=1}^{p} |y_{ij} - \hat{y}_{ij}(\theta)|. \tag{4}$$

For the computation of the estimator we use an iterative procedure known as *alternating regressions*, which was originally proposed by Wold (1966) and used in the context of generalized bilinear models by de Falguerolles and Francis (1992). The idea is very simple: if we take the row index i in the model equation (2) fixed and consider the parameters b_j and $\lambda_{j\cdot}$ as known for all j, then we see that a regression with intercept of the ith row of the two-way table on the k vectors of loadings yields estimates for a_i and the vector of scores $f_{i\cdot}$. Reversely, if we take j fixed and suppose that a_i and $f_{i\cdot}$ are known for all i, and regress the jth column of the data matrix on the k vectors of scores, then we can update the estimates for b_j and $\lambda_{j\cdot}$. To make things robust, we will of course use a robust regression method, as was already proposed by Ukkelberg and Borgen (1993). Minimizing the criterion (4) results in performing alternating L_1 regressions.

Unfortunately, L_1-regression is sensitive to leverage points. Therefore we propose a weighted L_1-regression, corresponding to minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{p} |y_{ij} - \hat{y}_{ij}(\theta)| w_i(\theta) w_j(\theta). \tag{5}$$

These weights will downweight outlying vectors of loadings or scores. The row weights are defined by

$$w_i = \min(1, \chi^2_{k,0.95}/\mathrm{RD}_i^2) \quad \text{for } i = 1, \ldots, n,$$

where $\mathrm{RD}_1, \ldots, \mathrm{RD}_n$ are robust Mahalanobis distances computed from the collection of score vectors $\{f_{i\cdot} | 1 \leq i \leq n\}$ and based on the Minimum Volume Ellipsoid (Rousseeuw and van Zomeren, 1990). Analogously, we define the set of column weights w_j using the vectors of loadings. Since the true loadings and scores are unobserved, w_i and w_j depend on the unknown parameters, and will be updated at each iteration step in the alternating regression procedure. To start the iterative procedure one can take initial values obtained by robust principal component analysis (Croux and Ruiz-Gazen, 1996). It is recommended to orthogonalize the vectors of scores at the end of the iteration procedure.

It was shown by many simulations and experiments, that the above method works well, is highly robust and converges. As a byproduct of the algorithm, robust biplots can be produced. An S-plus program of the proposed algorithm is available at *http://www.statistik.tuwien.ac.at/public/filz/research.html.*

3 Simulation experiment

In this section we study the performance of the proposed method by a modest simulation study. We generated data sets with $n = 25$ rows and $p = 15$ columns according to a model with two factors:

$$Y_{ij} = c + a_i + b_j + \sum_{l=1}^{2} \lambda_{jl} f_{il} + \varepsilon_{ij}$$

$(i = 1, \ldots, n;\ j = 1, \ldots, p)$. Values for c, a_i, b_j, f_{il}, and λ_{jl} were randomly generated and fulfilled the restrictions discussed in Section 1. The noise term ε_{ij} was quite small (distributed according to a $N(0, 0.05)$) for $n \times p - n_{out}$ of the entries in the data matrix. However, for n_{out} entries, randomly placed in the data matrix, the noise term followed a $N(0, 10)$, which induced n_{out} outlying cells.

Fitting the model gave estimated parameters $\hat{c}^s$, $\hat{a}_i^s$, $\hat{b}_j^s$, $\hat{f}_{il}^s$, and $\hat{\lambda}_{jl}^s$, for $s = 1, \ldots, nsim = 150$ simulated samples. As a measure of deviation of the estimated parameters from the true ones we took the mean squared error (MSE):

$$\text{MSE}(c) = \frac{1}{nsim} \sum_{s=1}^{nsim} \|\hat{c}^s - c\|^2, \quad \text{MSE}(a) = \frac{1}{nsim} \sum_{s=1}^{nsim} \|\hat{a}^s - a\|^2,$$

where a^s is a vector of length n with components a_i^s and $\|\cdot\|$ is the Euclidean norm. (The expression for MSE(b) is obtained analogously.) It is also possible to compute proximity indices between the sets of estimated and true vectors of loadings, resp. scores, using e.g. angles between subspaces. We preferred, however, to compute an overall measure of the quality of the estimation procedure :

$$\frac{1}{nsim} \sum_{s=1}^{nsim} \sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{\mu}_{ij}^s - \mu_{ij})^2, \tag{6}$$

with μ and $\hat{\mu}^s$ defined according to (2).

This simulation experiment was repeated for a percentage of outliers in the data set varying from 1 to 27. Figure 1 displays the summary measures as a function of the percentage of outliers when using the algorithm based on LS, L_1 and weighted L_1 regression. We clearly see that the approach based on LS is highly non-robust: even for a small percentage of outliers, we observe huge MSEs and a bad quality of the fit. For the estimation of the overall median, row and column effects, L_1 and weighted L_1 behave similarly. But the overall fit is much better for weighted L_1 than for L_1, since the latter approach is not capable of extracting the factor structure in the interaction terms when outliers are present.

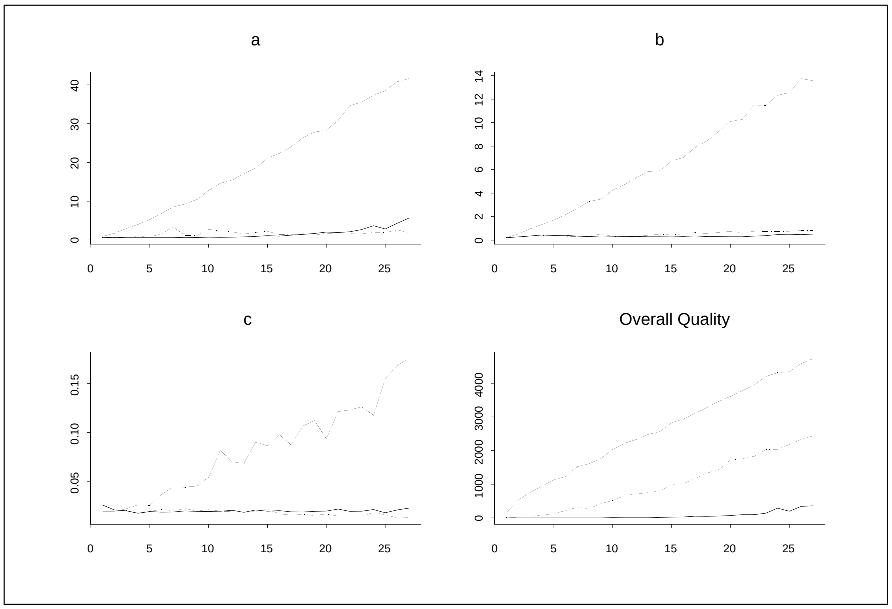

Fig. 1. MSE of the estimates for the row effects, column effects and for the overall median, and a general measure for the quality of the fit using the Least Squares (−−), the L_1 (− · −) and the weighted L_1 (solid line) estimators, in function of the percentage of outliers.

4 Example

We measured $p = 13$ variables for the 17 Styrian political districts (Styria is part of Austria). One district is the capital Graz (G). The typical rural districts are Feldbach (FB), Hartberg (HB), Murau (MU), Radkersburg (RA), and Weiz (WZ), while typical industrial regions are Bruck/Mur (BM), Judenburg (JU), Knittelfeld (KN), and Mürzzuschlag (MZ). Graz-Umgebung (GU) is the surroundings of Graz. Liezen (LI) is a touristic region with beautiful nature. The remaining districts are Deutschlandsberg (DL), Fürstenfeld (FF), Leibnitz (LB), Leoben (LE), and Voitsberg (VO). As variables were considered: the proportion of children ($<$ 15 years) (`chi`) and old people ($>$ 60 years) (`old`) in each district. Furthermore, the proportion of employed people in the industry (`ind`), trade (`tra`), tourism (`tou`), service (`ser`), and agriculture (`agr`), and the total proportion of unemployed people (`une`) was measured. Other variables are the proportion of mountain farms (`mou`), of people with university education (`uni`), of people which just attended primary school (`pri`), of employed people not commuting daily (`cnd`), and the proportion of employed people commuting to another district (`cdi`). The origin of these measurements is the Austrian census of 1991, and the data are available at the before mentioned web page.

We fitted the model, using weighted L_1 regression, with $k = 2$ to the raw data, although that it may be more appropriate to apply the logit transformation first. In Table 1, we displayed the estimated row effect $\hat{a}_i$ and column effect $\hat{b}_j$, together with the residual matrix $y_{ij} - \hat{\mu}_{ij}$. We see that Graz (G) appears as an outlier for a lot of variables, indicating that it is clearly distinct from most other districts. The district GU has a high residual for commuting to another district (namely to Graz), which is also true for VO, and for employed people in the industry (it is a quite and refined district). District RA has an atypical row effect, and an outlying residual for the cell corresponding with employed people in agriculture.

The biplot (Figure 2) pictures the estimates $(\hat{f}_{i1}, \hat{f}_{i2})$ and $(\hat{\lambda}_{j1}, \hat{\lambda}_{j2})$. The typical agricultural districts (FB, HB, MU, RA, WZ) have high loadings on the variable representing the employed people in agriculture, but they also have high values for commuting to another district (the latter is also true for GU, the surroundings of Graz). Additionally, the districts FB, HB, RA, and MU have high loadings for the variable "commuting not daily" (`cnd`). The industrial regions (BM, JU, KN, MZ) have high values at the vector "industry" (`ind`), but also GU and LE have high values there. LE additionally has a high value for employed people in service.

Graz appears as an outlier again. Fortunately the biplot is robust, implying that Graz will not influence too much the estimates of loadings and scores. A classical biplot would also reveal Graz as an outlier, but then the estimated loadings and scores are heavily influenced by this outlier, making their interpretation subject to a lot of doubt.

Acknowledgment

We wish to thank A. de Falguerolles for bringing the technique of alternating regressions under our attention and the referee for his very helpful remarks.

References

Croux, C. and Ruiz-Gazen, A. (1996). A fast algorithm for robust principal components based on projection pursuit. In: *Proceedings in Computational Statistics, COMPSTAT 1996*, Prat, A. (Ed.), 211-216. Heidelberg: Physica-Verlag.

de Falguerolles, A., and Francis, B. (1992). Algorithmic approaches for fitting bilinear models. In: *Computational Statistics, COMPSTAT 1992*, Dodge, Y. and Whittaker, J. (Eds.), Vol. 1, 77-82. Heidelberg: Physica-Verlag.

Gabriel, K.R. (1978). Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society B*, **40**(2), 186-196.

Gollob, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques, *Psychometrika*, 33, 73-116.

Rousseeuw, P.J. and van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633-639.

Ukkelberg, Å. and Borgen, O. (1993). Outlier detection by robust alternating regressions. *Analytica Chimica Acta*, 277, 489-494.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In: *A Festschrift for F. Neyman*, David, F. N. (Ed.), 411-444. New York: Wiley and Sons.

WL_1	chi	old	ind	tra	tou	ser	agr	mou	une	uni	pri	cnd	cdi	$\hat{\mathbf{a}}$
BM	-0.1	-0.1	0.2	1.1	0.7	-1.6	0.0	0.0	1.5	0.0	0.0	1.4	-2.8	0.1
DL	-0.7	0.4	0.1	-0.5	-0.1	0.0	0.0	0.0	0.1	0.2	-0.1	3.1	1.5	0.6
FB	-0.5	1.5	0.0	0.0	-0.2	0.0	4.1	-3.1	0.2	0.0	1.1	-0.9	-1.2	-0.3
FF	0.0	1.6	-2.2	4.3	-0.1	0.0	0.0	-2.0	0.0	1.0	-0.2	0.0	-0.7	0.0
G	2.0	0.0	0.0	12.8	0.0	8.1	0.0	-0.6	-0.2	12.8	-15.0	-13.5	-12.4	-4.8
GU	0.0	-1.4	-7.2	0.0	0.9	3.1	0.0	0.0	-0.9	0.8	-1.5	0.0	8.6	1.1
HB	0.4	-0.2	0.1	0.1	2.0	0.0	-0.2	0.3	0.0	0.0	0.0	6.3	-0.4	-0.1
JU	0.1	0.0	7.3	-0.2	-0.7	-2.5	-2.9	0.0	0.0	0.3	0.0	1.8	0.0	-0.3
KN	0.6	0.0	0.0	-1.4	-1.9	2.8	0.0	-0.1	-0.7	0.0	0.4	-0.4	0.0	0.4
LB	0.0	-1.3	-2.6	1.5	0.0	0.9	0.0	-2.5	0.3	-0.9	0.9	0.0	6.5	0.3
LE	-0.7	0.0	-3.8	0.0	-0.7	0.0	2.7	-1.0	1.1	-0.1	0.0	-0.5	4.8	-0.2
LI	1.2	-1.7	-3.3	1.3	4.1	0.0	-1.0	0.1	-0.1	0.0	-0.5	0.0	0.0	-0.2
MZ	0.0	1.2	5.9	-2.4	0.8	-5.1	0.0	0.2	-0.5	-1.0	0.8	0.0	-2.6	0.3
MU	0.0	-1.5	0.7	-2.6	0.6	0.0	0.0	0.1	-1.9	0.3	-3.8	6.2	7.5	0.3
RA	-1.7	3.5	-3.4	0.0	0.0	1.9	8.2	-3.4	0.0	0.8	0.0	-3.4	0.0	-1.1
VO	-0.4	0.3	0.0	-1.1	0.0	-3.6	0.0	1.2	2.0	-0.6	0.4	0.0	9.7	-0.2
WZ	0.5	0.0	3.2	-0.6	0.1	-2.9	-0.2	0.2	0.0	0.0	0.0	0.0	0.0	0.3
$\hat{\mathbf{b}}$	5.7	8.8	27.9	-0.3	-7.1	16.7	0.0	-10.0	-6.1	-7.3	76.0	-1.1	40.7	$\hat{c} = 12.2$

Table 1. Estimates for the row effects and column effects together with the residuals for the Styrian districts data set using the weighted L_1 approach (rounded values, in %).

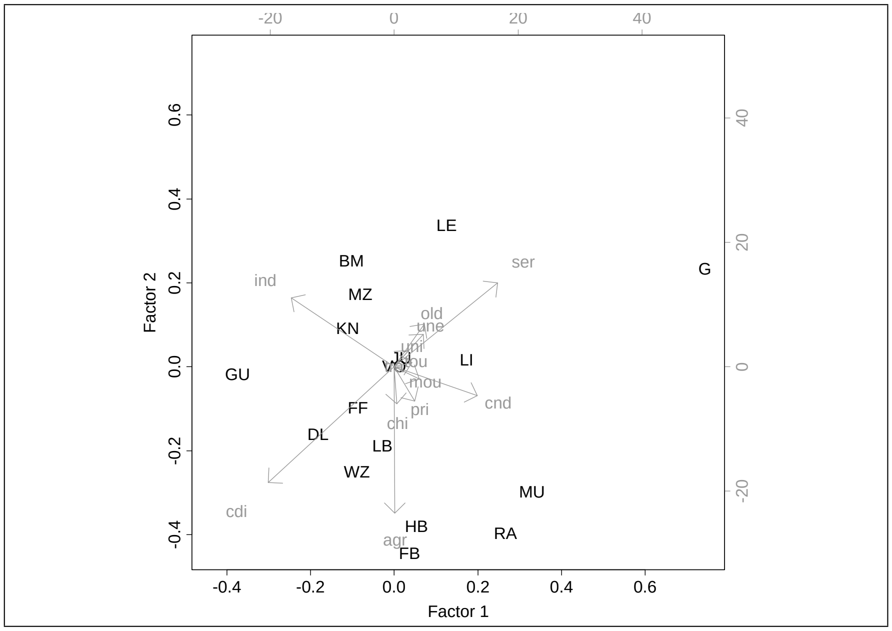

Fig. 2. Robust biplot representation of the interactions between rows and columns for the Styrian districts data set.