ABSTRACT FOR CODAWORK'08

Robust Factor Analysis for Compositional Data

Peter Filzmoser,[1] Karel Hron,[2] Clemens Reimann,[3] and Robert G. Garrett[4]

Factor analysis as frequent technique for multivariate data inspection is widely used also for compositional data analysis. The usual way is to use a centered logratio (clr) transformation to obtain the random vector $\boldsymbol{y}$ of dimension D. The factor model is then

$$\boldsymbol{y} = \boldsymbol{\Lambda f} + \boldsymbol{e} \tag{1}$$

with the factors $\boldsymbol{f}$ of dimension $k < D$, the error term $\boldsymbol{e}$, and the loadings matrix $\boldsymbol{\Lambda}$. Using the usual model assumptions (see, e.g., Basilevsky, 1994), the factor analysis model (1) can be written as

$$\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{\Lambda\Lambda}^T + \boldsymbol{\Psi} \tag{2}$$

where $\boldsymbol{\Psi} = \mathrm{Cov}(\boldsymbol{e})$ has a diagonal form. The diagonal elements of $\boldsymbol{\Psi}$ as well as the loadings matrix $\boldsymbol{\Lambda}$ are estimated from an estimation of $\mathrm{Cov}(\boldsymbol{y})$.

Given observed clr transformed data $\boldsymbol{Y}$ as realizations of the random vector $\boldsymbol{y}$. Outliers or deviations from the idealized model assumptions of factor analysis can severely effect the parameter estimation. As a way out, robust estimation of the covariance matrix of $\boldsymbol{Y}$ will lead to robust estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ in (2), see Pison et al. (2003). Well known robust covariance estimators with good statistical properties, like the MCD or the S-estimators (see, e.g. Maronna et al., 2006), rely on a full-rank data matrix $\boldsymbol{Y}$ which is not the case for clr transformed data (see, e.g., Aitchison, 1986).

The isometric logratio (ilr) transformation (Egozcue et al., 2003) solves this singularity problem. The data matrix $\boldsymbol{Y}$ is transformed to a matrix $\boldsymbol{Z}$ by using an orthonormal basis of lower dimension. Using the ilr transformed data, a robust

[1]Dept. of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria; e-mail: P.Filzmoser@tuwien.ac.at

[2]Dept. of Mathematical Analysis and Applications of Mathematics, Palacký University Olomouc, Tomkova 40, CZ-77100 Olomouc, Czech Rep.; e-mail: hronk@seznam.cz

[3]Geological Survey of Norway (NGU), N-7491 Trondheim, Norway; e-mail: Clemens.Reimann@ngu.no

[4]Geological Survey of Canada (GSC), Natural Resources Canada, 601 Booth Street, Ottawa, Ont., Canada K1A 0E8; e-mail: garrett@NRCan.gc.ca

covariance matrix $C(\boldsymbol{Z})$ can be estimated. The result can be back-transformed to the clr space by

$$C(\boldsymbol{Y}) = \boldsymbol{V} C(\boldsymbol{Z}) \boldsymbol{V}^T$$

where the matrix $\boldsymbol{V}$ with orthonormal columns comes from the relation between the clr and the ilr transformation. Now the parameters in the model (2) can be estimated (Basilevsky, 1994) and the results have a direct interpretation since the links to the original variables are still preserved.

The above procedure will be applied to data from geochemistry. Our special interest is on comparing the results with those of Reimann et al. (2002) for the Kola project data.

References

J. Aitchison. *The Statistical Analysis of Compositional Data.* Chapman and Hall, London, U.K., 1986. 416 pp.

A. Basilevsky. *Statistical Factor Analysis and Related Methods. Theory and Applications.* John Wiley & Sons, Inc., New York, USA, 1994. 737 pp.

J.J. Egozcue, V. Pawlowsky-Glahn, F. Mateu-Figueros, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300, 2003.

R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods.* John Wiley & Sons Canada Ltd., Toronto, ON, 2006. 436 pp.

G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux. Robust factor analysis. *Journal of Multivariate Analysis*, 84:145–172, 2003.

C. Reimann, P. Filzmoser, and R.G. Garrett. Factor analysis applied to regional geochemical data: problems and possibilities. *Applied Geochemistry*, 17:185–206, 2002.