Robust orthogonal regression for compositional data in R

V. Todorov[1,*], K. Hrůzová[2], K. Hron[2] and P. Filzmoser[3]

[1] *United Nations Industrial Development Organization, Vienna, Austria; v.todorov@unido.org*
[2] *Palacky University, Olomouc, Czech Republic; klara.hruzova@gmail.com, hronk@seznam.cz*
[3] *Vienna University of Technology, Vienna, Austria; p.filzmoser@tuwien.ac.at*
[*] *Presenting author*

Keywords. *Compositional data; Orthogonal regression; Isometric logratio coordinates; MM-estimates; Bootstrap inference*

In the context of building a regression model on compositional data, orthogonal regression (as a special case of errors-in-variable models) is appropriate since all compositional parts - also the explanatory variables - are measured with errors. The classical approach to estimate the model is based on an eigenvector analysis of the joint covariance matrix of the observations. However, in the presence of outlying observations in compositional data the orthogonal regression (that is able to handle the regression problem statistically) should be replaced by its robust counterpart. Therefore, we consider also a robust version of orthogonal regression. In **?**, M- and S-estimators for robust orthogonal regression are presented. However, S-estimators are computed using inefficient algorithms and M-estimators have low breakdown point. Another possibility can be found in **?**, where the projection-pursuit approach is used, which is also suitable for more than one response variable. In order to benefit from the better statistical properties of the MM-estimates, we decided to follow the above mentioned (classical) approach and develop robust orthogonal regression in orthonormal coordinates using robust PCA, which is obtained through a robust estimation of the covariance matrix. Among other possibilities, the MM-estimators are employed for this purpose. The reason for choosing MM-estimators is that they are highly efficient when the errors have a normal distribution, their breakdown point is 50% and they have bounded influence function.

In order to perform statistical inference, like deriving confidence intervals or testing hypotheses, bootstrap techniques for classical and robust orthogonal regression are proposed. Although bootstrap is a very useful tool, in case of robust estimators there are two problems: computational complexity of robust estimators and the instability of the bootstrap in case of outliers. Therefore we used fast and robust bootstrap **?** which is based on the fact that the robust estimators (namely S- and MM-estimators) can be represented by smooth fixed point equations which allow to

calculate a fast approximation of the estimates in each bootstrap sample.

As the estimation of parameters and statistical inferences is performed in real (unconstrained) coordinates, the resulting orthogonal regression model is not exclusively designed for compositional data, but it could be used also with any non-compositional data.

The R package *oreg* provides functions for classical and robust orthogonal regression. These functions can be applied on both compositional and non-compositional data. In case of compositional data, all regression models are estimated, one for each orthonormal basis. The regression parameters are estimated using (classical or robust) principal components and the MM-estimates are computed by a call to the implementation in the *rrcov* package **?**. The results can be viewed by standard `print()` and `plot()` functions, while a `summary()` function presents the parameter estimates and also the corresponding statistical inference (confidence intervals and p-values for significance testing) obtained through bootstrap. In the robust version, fast and robust bootstrap from the package *FRB* **?** is used.

The robustness, the efficiency and the computational performance of the procedure are studied through simulation and are illustrated with a data set from macroeconomics representing the structure of gross value added and the relation between its components. The data set comes from the World Bank database (http://data.worldbank.org) and includes observations for 131 countries in 2010 at constant 2005 USD.

References

Croux C, Fekri M & Ruiz-Gazen A. Fast and robust estimation of the multivariate errors in variables model. *Test* **19** 286–303.

Salibian-Barrera M, Van Aelst S & Willems G. (2006). PCA based on multivariate MM-estimators with fast and robust bootstrap. *J Am Stat Assoc* **101** 1198–1211.

Todorov V & Filzmoser Filzmoser P. An object oriented framework for robust multivariate analysis. *Journal of Statistical Software* **2009** 32/3.

Van Aelst S & Willems G. Fast and robust bootstrap for multivariate inference: The R package FRB. *Journal of Statistical Software* **2013** 53/3.

Zamar, RH. Robust estimation in the errors-in-variables model. *Biometrika* **76** (1) 149–160.