Robust compositional data analysis

P. FILZMOSER[1], K. HRON[2], and M. TEMPL[1,3]

[1]Department of Statistics and Probability Theory - Vienna University of Technology, Austria P.Filzmoser@tuwien.ac.at
[2]Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
[3]Statistics Austria, Vienna, Austria

Abstract

Many practical data sets contain outliers or other forms of data inhomogeneities. Robust statistics offers concepts how to deal with these situations where the data do not follow strict model assumptions. These concepts are designed for the usual Euclidean space, and they can be easily applied to compositional data if they are represented in this space as well. It turns out that the isometric logratio (ilr) transformation is best suitable in the context of robust estimation. Depending on the method applied, an interpretation of result is usually done in a back-transformed space.

1 Introduction

Statistical estimators rely on more or less strict assumptions, typically distributional assumptions. If these are fulfilled, it is usually possible to derive optimality properties of the estimator. On the other hand, a violation of the assumptions may lead to uncontrollable effects. In such cases an estimator can even deliver arbitrary results. Using this undesired outcome of an estimator, it is usually impossible to check the validity of model assumptions.

To be more specific, an example in the context of simple linear regression is shown in Figure 1. The majority of data points follows a linear trend, and some points are strongly deviating from this trend. The left plot shows the least squares regression line using all data points (solid), and two lines indicating a residual distance of two standard errors (dashed). It is obvious that the outliers have attracted the regression line. They also have increased the estimation of the residual variance, and the resulting band for outlier diagnostics identifies not the outliers but regular data points. Thus, neither the regression estimates nor the diagnostics is useful. In contrast, the right plot is based on a robust regression estimator. The outliers seem to have no effect on the estimation. The diagnostic band is based on a robust estimation of the residual variance. It flags the correct data points as outliers, and reveals also some further data points as slightly deviating from the linear trend.

The identification of outliers as shown in Figure 1 is trivial, but it is not so straightforward in the context of multiple linear regression with several explanatory variables, or if the position of the deviating data points is close to the data majority forming the linear trend. Robust regression estimation and diagnostics is still reliable in such cases.

The main idea of robust statistics is to allow for certain deviations from idealized model assumptions (Maronna et al., 2006). The estimators are supposed to still give meaningful results in the "surroundings" of an ideal model. The concept of robust statistics has also been formalized into the "theory of robust statistics" (Huber, 1981). The most prominent tools to characterize a robust estimator are the influence function and the breakdown point. While the influence function measures the effect of an infinitesimal contamination on an estimator, the breakdown point studies the behavior of an estimator under higher amounts of contamination. Robust estimators are characterized by a bounded influence function, because any small contamination, even if it is placed on an arbitrary position, should only have a bounded influence on the estimator. Moreover, a robust estimator with a high breakdown point should not give arbitrary results if a small or even moderate fraction of the observations is replaced by any outlying values.

2 Consequences for compositional data analysis

Just as any "usual" statistical data, also compositional data may contain outlying observation. Since compositional data are by definition multivariate data, outlier identification or, more generally, the

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

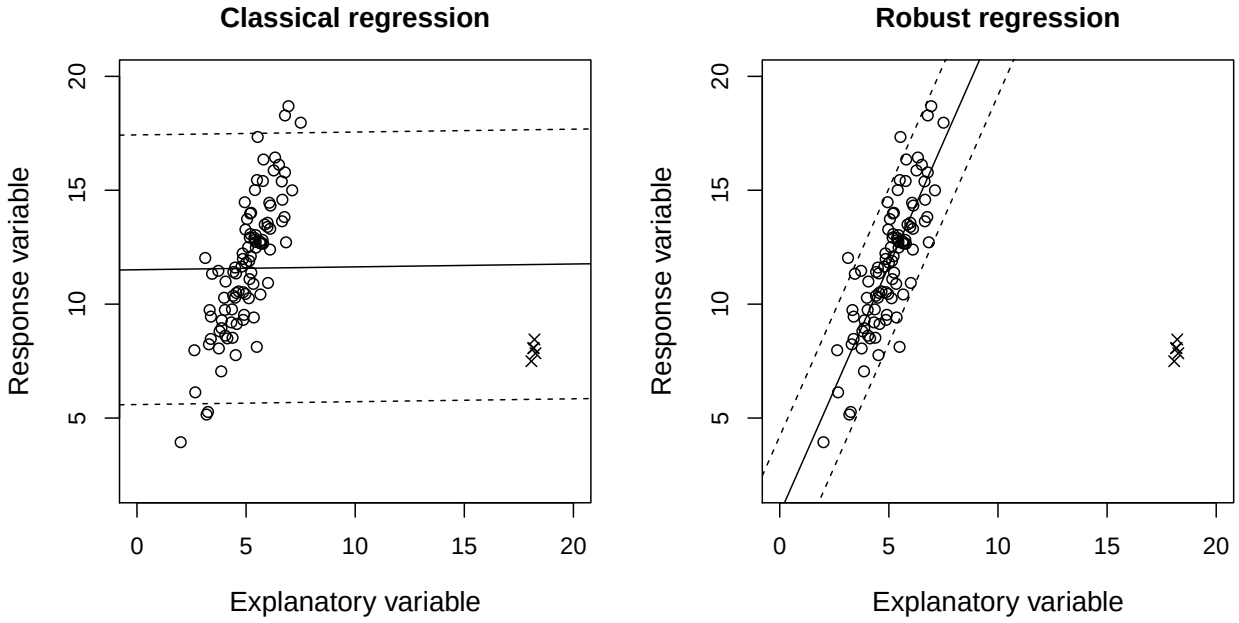

Figure 1: Classical and robust regression for data containing outliers. The dashed lines are at a residual distance of two standard errors, and they are usually used for outlier diagnostics.

identification of observations that are deviating from an underlying model, becomes more and more difficult with increasing dimension (number of compositional parts). Even more, "deviation" has another meaning for compositional data, because it needs to refer to the simplex, the sample space of compositional data with the Aitchison geometry (Aitchison, 1986; Egozcue et al., 2003). The problem of robust regression, for example, becomes extremely difficult now when applying this method directly to the compositional raw data. Firstly, linear regression would no longer be appropriate because of the geometry of the simplex, and secondly, distances need to be measured in terms of the Aitchison distance. The natural way is thus to transfer the problem to the usual Euclidean space, where all the concepts of robust regression are valid, and where robustness is designed for. The isometric logratio (ilr) transformation is best suited for this purpose, but it introduces for some methods the difficulty of interpreting the results (Egozcue et al., 2003). In that case it is necessary to back-transform them, usually to the space obtained by the centered logratio (clr) transformation.

Most multivariate methods are based on an estimation of the covariance structure. Robust counterparts to classical covariance estimation have been developed, and efficient algorithms for their computation exist (Maronna et al., 2006). Figure 2 shows a demonstration of classical and robust covariance estimation of compositional data. The ternary diagram of the original compositions (left plot) already indicates that some data points marked by $\times$ are outlying. However, the magnitude of their deviation from the bulk of the data seems rather small, and accordingly their effect on the classical covariance estimation might be negligible. The right plot shows the data transformed to the ilr space. The majority of the data points shows an elliptical structure. The outlying data points are indeed not extreme along the coordinates; it would even be impossible to identify them when inspecting the single coordinates. In this sense, these are multivariate outliers. Still, they have quite an effect on the covariance estimation: The classical sample covariance estimation is determining the structure of the dashed tolerance ellipse, an ellipse covering 97.5% of the data points in case of bivariate normal distribution. The tolerance ellipse drawn with the solid line is based on a robust covariance estimation, and it corresponds much better to the structure of the data majority. In fact, the tolerance ellipses can also be used for multivariate outlier detection: multivariate outliers are placed outside the ellipse. However, the outliers themselves inflate the classical ellipse, leading to the so-called masking effect. The robust ellipse clearly flags the outliers. The tolerance ellipses can be back-transformed to the simplex, where the inflation of the classical ellipse is also visible. However, in the simplex it is sometimes not so clear which of the data points are deviating, and if the majority of the data points

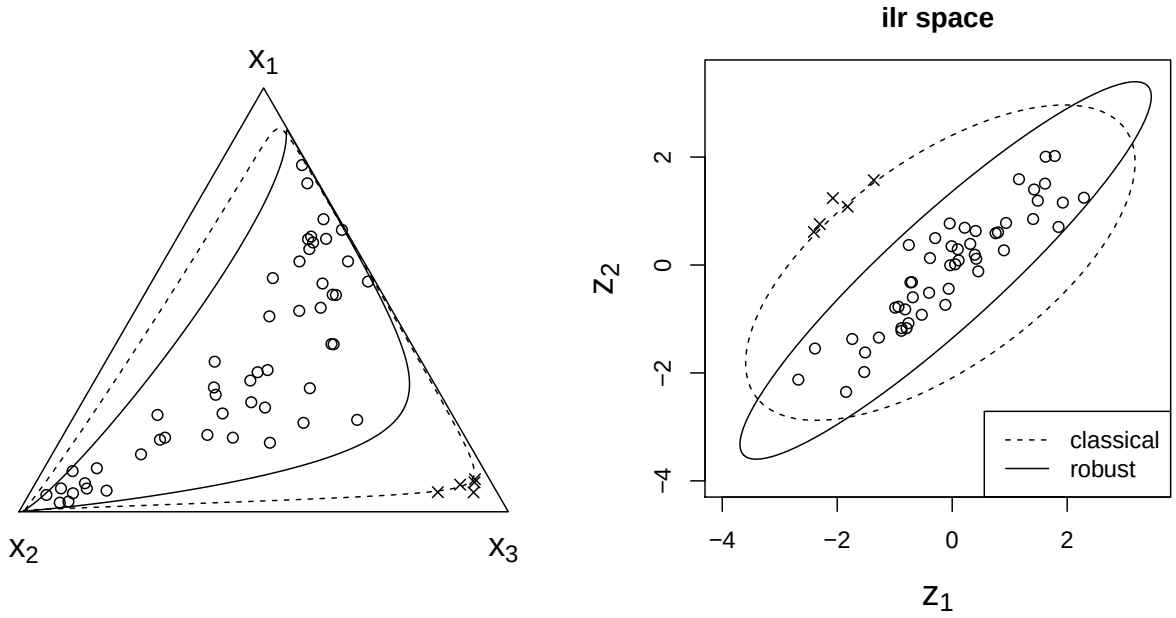

Figure 2: Classical and robust covariance estimation, visualized by tolerance ellipses. Covariance estimation is applied in the ilr space, the results are back-transformed to the simplex.

is indeed generated from one distribution.

Multivariate outlier detection is usually based on the computation of Mahalanobis distances. The Mahalanobis distances are invariant to the logratio transformations alr (additive logratio), clr, and ilr (Filzmoser and Hron, 2008). The Mahalanobis distances are based on an estimation of multivariate location and covariance, and for reliable outlier detection a robust estimation is necessary. This excludes the clr transformation, because robust covariance estimation is only possible for a non-singular data matrix. But also the invariance with respect to the different versions of the alr transformation and the ilr transformation is only valid if the robust estimators are affine equivariant, i.e. if they transform accordingly under affine transformations.

At the basis of robust covariance estimation, it is straightforward to robustify principal component analysis (Filzmoser et al., 2009a), factor analysis (Filzmoser et al., 2009b), or discriminant analysis (Filzmoser et al., 2009c). The resulting estimates are indeed robust in the sense of bounded influence and positive breakdown point, because these properties can be derived in the usual Euclidean geometry (see, e.g., Pison et al., 2003). Robust versions of these multivariate methods may not always seem preferable to classical ones:

- *Robust principal component analysis:* The directions of the robust principal components are not determined by the outliers, and thus they express the main data variability formed by the data majority. Classical principal components, on the other hand, can be attracted by the outliers, because their (classical) variance can be very high. This, however, allows to immediately "see" large outliers in plots of the principal components. Outliers may not be well visible when plotting the scores of robust principal components. For this reason, a diagnostic tool has been developed that shows the influence of outlying data points on the classical estimation (Hubert et al., 2005).

- *Robust discriminant analysis:* The goal of discriminant analysis is usually a minimization of the misclassification rate for new test data, hereby using the discriminant rules established from the training data. Rules based on robust estimates of location and covariance (like for linear or quadratic discriminant analysis) may not necessarily lead to smaller misclassification errors. Even more, applying robust discriminant analysis in the transformed space does not necessarily lead to a smaller misclassification rate than applying the method to the raw compositional data. The performance of the discriminant rules depends strongly on the location of the samples in the data space, on the data structure, and on the position of potential outliers. Applying the

rules in the usual Euclidean geometry has the advantage that the method works as expected, because it is designed for this geometry; an application in the simplex can cause unexpected behavior. Finally, robust estimators are based on the points that are drawn from underlying model distributions. If we assume that test data are also drawn from these distributions, the rules are appropriate. A possibly larger error rate compared to a classical rule may then only be caused by an unfavorable configuration of the outliers.

3 Conclusions

Robust estimation for compositional data is straightforward after transforming the compositions to an appropriate space, usually the ilr space. In this space, the properties of the robust estimators are valid and well-known. The concept of influence function or breakdown point, for example, can be determined only in the usual Euclidean geometry. A further argument for the ilr transformation against the clr transformation is that clr results in singularity, making robust estimation difficult or even impossible.

It is not always straightforward to interpret results of robust statistical methods in this transformed space. An example is factor analysis, where the meaning of the factors is based on the loading matrix. Loadings from ilr coordinates are thus usually not helpful for an interpretation and thus results need to be back-transformed to an appropriate space (Filzmoser et al., 2009b).

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Filzmoser, P., and K. Hron (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences 40*(3), 233–248.

Filzmoser, P., K. Hron, and C. Reimann (2009a). Principal component analysis for compositional data with outliers. *Environmetrics 20*(6), 621–632.

Filzmoser, P., K. Hron, C. Reimann, and R.G. Garrett (2009b). Robust factor analysis for compositional data. *Computers and Geosciences 35*, 1854–1861.

Filzmoser, P., K. Hron, and M. Templ (2009c). Discriminant analysis for compositional data and robust parameter estimation. Technical Report SM-2009-3, Vienna University of Technology, Austria. Submitted for publication.

Huber, P.J. (1981). *Robust Statistics*. John Wiley & Sons, New York.

Hubert, M., P.J. Rousseeuw, and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal components analysis. *Technometrics 47*, 64–79.

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

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