Principal component analysis for compositional data with outliers

Peter Filzmoser[1], Karel Hron[2] and Clemens Reimann[3]

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

[2] *Department 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*

SUMMARY

Compositional data (almost all data in geochemistry) are closed data, i.e. they sum up to a constant (e.g. 100 weight percent). Thus the correlation structure of compositional data is strongly biased and results of many multivariate techniques become doubtful without a proper transformation of the data. The centered logratio transformation (clr) is often used to open closed data. However the transformed data do not have full rank following a logratio transformation and cannot be used for robust multivariate techniques like principal component analysis (PCA). Here we propose to use the isometric logratio transformation (ilr) instead. However, the ilr transformation has the disadvantage that the resulting new variables are no longer directly interpretable in terms of the originally entered variables. Here we propose a technique how the resulting scores and loadings of a robust PCA on ilr transformed data can be back-transformed and interpreted. The procedure is demonstrated using a real data set from regional geochemistry and compared to results from non-transformed and non-robust versions of PCA. It turns out that the procedure using ilr transformed data and robust PCA delivers superior results to all other approaches. The examples demonstrate that due to the compositional nature of geochemical data PCA should not be carried out without an appropriate transformation. Furthermore a robust approach is preferable if the dataset contains outliers.

KEY WORDS: robust statistics; compositional data; isometric logratio transformation; principal component analysis

1. INTRODUCTION

The statistical analysis of compositional multivariate data is a much discussed topic in the field of multivariate statistics. The data values of compositional data consist of proportions that sum up to a constant (e.g. to 100%) for each sample. If not all variables or components have been analyzed, this constant sum property is not directly visible in the data, but the relation between the variables is still not the real relation but a forced one. For example, if the concentrations of chemical elements are measured in soil samples, and if a variable like SiO_2 has a big proportion of, say, 70%, then automatically the sum of the remaining concentrations can at most be 30%. Increasing values of SiO_2 automatically lead to decreasing values of the other compounds, and even if not all constituents of the soil have been measured, the correlations will be mainly driven by the constant sum constraint. This restriction also leads to a geometrical artefact of the multivariate data because the data are in fact in a sub-space, the so called simplex sample space. The new view of this problem, as stated in Aitchison (1986), allowed a possibility to use standard statistical methods for the inspection of compositional data. It is based on transformations (from the family of so-called logratio transformations) of compositional data from the simplex to the usual real space. The statistical methods are applied to the transformed data and the results are back-transformed to the original space.

Beginning with papers by Aitchison (1983, 1984), a lot of research was devoted to finding a useful transformation for compositional data in the context of principal component analysis (PCA). The centered logratio (clr) transformation turned out to be a preferable option (Aitchison and Greenacre, 2002). It is based on dividing each sample by the geometric mean of its values, and taking the logarithm. The principal components (PCs) are then aimed at summarizing the multivariate data structure, and subsequently they can be used for dimension reduction. The goal of keeping the most important data information with only few PCs can fail for data containing outliers because these can spoil the estimation of the PCs (see, e.g., Maronna et al., 2006). This artefact arises for classical PCA where the estimation of the PCs is based on the classical sample covariance matrix. As a solution, robust PCA uses a robust estimation of the covariance matrix, and the PCs will still point in directions of the main variability of the bulk of data (see, e.g., Filzmoser, 1999). This procedure, however, does not work with clr transformed data because robust covariance estimators usually need a full rank data matrix.

In this paper we will solve this problem by taking the isometric logratio (ilr) transformation rather than the clr transformation (Section 2) where robust covariance estimation is again possible (Section 3). The resulting scores and loadings have to be back-transformed, and it is another purpose of this article to show how this is done (Section 4). Moreover, we demonstrate at a real data example from geochemistry how the results of classical and robust PCA, as well as appropriately transformed or just log-transformed data can differ (Section 5).

2. LOGRATIO TRANSFORMATIONS OF COMPOSITIONAL DATA

As stated in Aitchison (1986), compositional or closed data are multivariate data with positive values that sum up to a constant, usually chosen as 1. The sample space of compositional data is thus the simplex

$$\mathcal{S}^D = \{\boldsymbol{x} = (x_1, \ldots, x_D)', \, x_i > 0, \, \sum_{i=1}^{D} x_i = 1\}$$

where the prime stands for transpose and the simplex sample space is a $D-1$ dimensional subset of $\mathbb{R}^D$.

Standard statistical methods can lead to useless results if they are directly applied to original closed data. For this reason, the family of logratio transformations was introduced. It includes the additive logratio (alr) and the centered logratio (clr) transformation (Aitchison, 1986), as well as the isometric logratio (ilr) transformation (Egozcue et al., 2003). Since the alr transformation divides the data values by a reference variable (and uses the logarithm thereof), the choice of this reference variable will mainly determine the results, and thus this transformation is rather subjective.

The clr transformation is a transformation from $\mathcal{S}^D$ to $\mathbb{R}^D$, and the result for an observation $\boldsymbol{x} \in \mathcal{S}^D$ are the transformed data $\boldsymbol{y} \in \mathbb{R}^D$ with

$$\boldsymbol{y} = (y_1, \ldots, y_D)' = \left(\log \frac{x_1}{\sqrt[D]{\prod_{i=1}^D x_i}}, \ldots, \log \frac{x_D}{\sqrt[D]{\prod_{i=1}^D x_i}} \right)', \tag{1}$$

or, written in matrix notation,

$$\boldsymbol{y} = \boldsymbol{F} \log(\boldsymbol{x}), \tag{2}$$

where

$$\boldsymbol{F} = \boldsymbol{I}_D - \frac{1}{D}\boldsymbol{J}_D, \quad \text{with } \boldsymbol{I}_D = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}, \boldsymbol{J}_D = \begin{pmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{pmatrix}. \tag{3}$$

The matrices $\boldsymbol{F}$, $\boldsymbol{I}_D$, and $\boldsymbol{J}_D$ are all of dimension $D \times D$. The clr transformation treats all components symmetrically by dividing by the geometric mean. Thus it is possible to use the original variable names for the interpretation of statistical results based on clr transformed data. The main disadvantage of this transformation is that the resulting data are collinear because $\sum_{i=1}^D y_i = 0$. Methods that rely on full rank data matrices, like standard robust covariance estimators (Maronna et al., 2006), will thus not be applicable.

The isometric logratio (ilr) transformation solves the problem of data collinearity resulting from the clr transformation, while preserving all its advantageous properties like isometry between the simplex and the real space (Egozcue et al., 2003). It is based on the choice of an orthonormal basis (in the well known Euclidean sense) on the hyperplane $\mathcal{H} : y_1 + \ldots + y_D = 0$ in $\mathbb{R}^D$ that is formed by the clr transformation so that the compositions $\boldsymbol{x} \in \mathcal{S}^D$ result in non-collinear ilr transformed data $\boldsymbol{z} \in \mathbb{R}^{D-1}$. Egozcue et al. (2003) suggested to use the basis

$$\boldsymbol{v}_i = \sqrt{\frac{i}{i+1}} \left(\frac{1}{i}, \ldots, \frac{1}{i}, -1, 0, \ldots, 0 \right)' \quad \text{for } i = 1, \ldots, D-1, \tag{4}$$

resulting in the ilr transformed data $\boldsymbol{z} = (z_1, \ldots, z_{D-1})'$ with

$$z_i = \sqrt{\frac{i}{i+1}} \log \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}} \quad \text{for } i = 1, \ldots, D-1. \tag{5}$$

Clearly, another chosen orthonormal basis of $\mathcal{H}$ leads to orthogonal transformation of both the resulting data and the original data on the simplex (here in the sense of the corresponding simplicial, so called Aitchison inner product), see Egozcue et al. (2003).

Equations (1) and (5) can be used to express the relation between clr and ilr transformed data in matrix notation by

$$\boldsymbol{y} = \boldsymbol{V}\boldsymbol{z} \tag{6}$$

where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{D-1})$ is the $D \times (D-1)$ matrix with orthonormal basis vectors from (4) on the hyperplane $\mathcal{H}$. Multiplying equation (6) from the left-hand side with $\boldsymbol{V}'$ and utilizing that $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_{D-1}$ results in the inverse relation

$$\boldsymbol{z} = \boldsymbol{V}'\boldsymbol{y}, \tag{7}$$

see also Egozcue et al. (2003). This notation will be useful later on in the context of PCA.

It is easy to see that the interpretation of the ilr transformed data is not possible because the new $D-1$ variables have no direct connection to the original variables but they are only combinations thereof. Hence, for an interpretation results like loadings and scores from PCA based on ilr transformed data have to be back-transformed to the clr space.

3. PRINCIPAL COMPONENT ANALYSIS AND ITS ROBUSTIFICATION

Principal component analysis (PCA) is one of the most important multivariate statistical methods. It is widely applied for data pre-processing and dimension reduction, and the resulting PCs are then used for plotting or for subsequent multivariate analyses (see, e.g., Johnson and Wichern, 2007).

The PCs are usually derived from the composition of the covariance matrix of the $n \times D$ data matrix $\boldsymbol{X}$ with multivariate observations $\boldsymbol{x}_i'$, $i = 1, \ldots, n$ in its rows. Hence, for the PCA transformation the location estimator $T(\boldsymbol{X})$ and the scatter estimator $C(\boldsymbol{X})$ are needed. After singular value decomposition $C(\boldsymbol{X}) = \boldsymbol{G_x L G_x'}$ with the diagonal matrix $\boldsymbol{L}$ of eigenvalues and the matrix $\boldsymbol{G_x}$ of eigenvectors of $C(\boldsymbol{X})$ we can define the PCA transformation as

$$\boldsymbol{X}^* = (\boldsymbol{X} - \mathbf{1}T(\boldsymbol{X})')\boldsymbol{G_x}, \tag{8}$$

where $\mathbf{1}$ denotes a vector of n ones. The matrix $\boldsymbol{X}^*$ has obviously the same dimension as $\boldsymbol{X}$. Its columns are called the *scores* of the j-th PC ($j = 1, \ldots, D$). The columns of $\boldsymbol{G_x}$ are called *loadings* of the j-th PC, and they represent the influence of the original variables on the new PCs. For dimension reduction only the first few PCs are considered that cover the most important data information.

It is crucial which estimators are used for the PCA transformation (8). For classical PCA the location estimator $T(\boldsymbol{X})$ is the arithmetic mean vector, and the scatter estimator $C(\boldsymbol{X})$ is the sample covariance matrix. Both estimators are sensible with respect to outliers, and thus more robust counterparts can be used, like the MCD or S estimators (see Maronna et al., 2006). In case of the MCD (minimum covariance determinant) estimator, the location and scatter estimators are obtained by looking for a subset of at least h observations with the smallest determinant of their sample covariance matrix. The robust location estimator is then the arithmetic mean of this subset, and the scatter estimator is the sample covariance matrix of the subset, multiplied by a factor for consistency (Rousseeuw and Van Driessen, 1999). The choice of the number h determines both the robustness and the efficiency of the estimators. h should at least be taken as half of the total sample size n which results in the best resistance to outlying observations, but in a poorer efficiency. On the other hand, if h is large, e.g. close to n, the robustness of the MCD location and scatter estimators is poor, but the efficiency increases. A compromise is thus to take h approximately as $\frac{3}{4}n$. This choice would tolerate an outlier fraction of about $\frac{n-h}{n} = \frac{1}{4}$ of the observations. Thus, practically one has to take care that $\frac{n-h}{n}$ is larger than the fraction of outliers in the data, since otherwise the estimators can become unreliable.

Since the MCD estimator is based on minimizing the determinant of the covariance matrix of subsets of observations, it is only computable for non-singular data with rank equal to the number of variables. There are similar problems with other robust estimators for location and covariance. Thus, if non-classical PCA should be undertaken for compositional data, the clr transformation is not appropriate but the ilr transformation can be considered.

It should be noted that MCD or S estimators are based on elliptical symmetry of the data. Usually one even assumes that the data majority follows a multivariate normal distribution. Hence, prior to applying the ilr transformation for computing robust principal components, the raw data are supposed to follow a multivariate normal distribution on the simplex sample space (for details, see Pawlowsky-Glahn et al., 2007).

4. ROBUST PCA FOR ISOMETRIC LOGRATIO TRANSFORMED DATA

Given an $n \times D$ data matrix $\boldsymbol{X}_{n,D}$ with n compositions $\boldsymbol{x}_i'$, $i = 1, \ldots, n$, in its rows. Applying (2) to each row results in the clr transformed matrix

$$\boldsymbol{Y} = \log(\boldsymbol{X})\, \boldsymbol{F}'.$$

The relation

$$\boldsymbol{Z} = \boldsymbol{Y}\boldsymbol{V} \tag{9}$$

for the ilr transformed data matrix $\boldsymbol{Z}$ of dimension $n \times (D-1)$ follows immediately from (6) using $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_{D-1}$ (identity matrix of order $D-1$) and basic properties of matrix transposition, see e.g. Harville (1997). Using the location estimator $T(\boldsymbol{Z})$ and the covariance estimator $C(\boldsymbol{Z})$ for the ilr transformed data, the PCA transformation is defined as

$$\boldsymbol{Z}^* = [\boldsymbol{Z} - \boldsymbol{1}\,T(\boldsymbol{Z})']\boldsymbol{G}_{\boldsymbol{z}} \tag{10}$$

(compare to (8)). The $(D-1)\times(D-1)$ matrix $\boldsymbol{G_z}$ results from the singular value decomposition of

$$C(\boldsymbol{Z}) = \boldsymbol{G_z L_z G'_z}. \tag{11}$$

If the original data matrix has full rank D, the matrix $\boldsymbol{Z}$ will also have full rank $D-1$, and the MCD estimator can be used for $T(\boldsymbol{Z})$ and $C(\boldsymbol{Z})$, resulting in robust principal component scores $\boldsymbol{Z}^*$ and loadings $\boldsymbol{G_z}$. However, since these are no longer interpretable, we have to back-transform the results to the clr space. Using (9) we obtain the back-transformed scores

$$\boldsymbol{Y}^* = \boldsymbol{Z}^*\boldsymbol{V}'. \tag{12}$$

For obtaining the back-transformed loading matrix we can use again relation (9). For an affine equivariant scatter estimator we have

$$C(\boldsymbol{Y}) = C(\boldsymbol{Z}\boldsymbol{V}') = \boldsymbol{V}\, C(\boldsymbol{Z})\, \boldsymbol{V}' = \boldsymbol{V}\, \boldsymbol{G_z L_z G'_z}\, \boldsymbol{V}'.$$

The MCD scatter estimator has the property of affine equivariance (see, e.g., Maronna et al., 2006), and thus the matrix

$$\boldsymbol{G_y} = \boldsymbol{V}\boldsymbol{G_z} \tag{13}$$

represents the matrix of eigenvectors to the *nonzero* eigenvalues of $C(\boldsymbol{Y})$ (with the property $\boldsymbol{G'_y G_y} = \boldsymbol{I}_{D-1}$). The nonzero eigenvalues of $C(\boldsymbol{Y})$ are the same as for $C(\boldsymbol{Z})$ and consequently the explained variance with the chosen number of principal components remains unchanged.

It is useful to display the loadings and scores together in *biplots* (Gabriel, 1971). The interpretation of the biplot depends on the chosen scale for loadings and scores. For the special interpretation of biplots for compositional data in the clr space we refer to the results of Aitchison and Greenacre (2002) and Pawlowsky-Glahn et al. (2007).

5. EXAMPLE

As an example for robust PCA with compositional data we use the so-called Baltic Soil Survey (BSS) data (Reimann et al., 2003). This data set originates from a large-scale geochemistry project carried out in northern Europe, in an area of about 1 800 000 km^2. On an irregular grid 769 samples of agricultural soils have been collected. The samples came from two different layers, the top layer (0-25 cm) and the bottom layer (50-75 cm). All samples were analyzed for the concentration of more than 40 chemical compounds. The data sets of the top and bottom layer are available in the R package *mvoutlier* (R development core team, 2008) as data files *bsstop* and *bssbot*, respectively, both including 44 variables as well as x and y coordinates of the survey area.

This project was carried out to document element concentrations and spatial variation in agricultural soils from northern Europe. The element distributions will not only be influenced by the underlying lithology but also by other factors like climate, the input of marine aerosols, agricultural practice and contamination. For our example we use the major elements (Al_2O_3, Fe_2O_3, K_2O, MgO, MnO, CaO, TiO_2, Na_2O, P_2O_5 and SiO_2), plus LOI (Loss on ignition). A PCA inspection of this data set should allow to better understand the relations between the variables and thus the geochemical processes dominating the element distribution in the survey area. A visualization of the results in biplots should allow an interpretation of the relations among the compounds, and maps of the first view PCs should show the regions where certain concentrations are higher or lower due to some key geochemical processes.

In geochemistry it is most often wrongly argued that the variables follow a lognormal distribution and thus they are simply log-transformed (Reimann and Filzmoser, 2000). Here we want to compare the PCA results of the log-transformed data with those of the ilr transformed data, back-transformed to the clr space (in order to be able to use compositional biplots). Moreover, a comparison is made for classical and robust PCA. The results of these combinations will be denoted by *log-classical*, *ilr-classical*, *log-robust*, and *ilr-robust*, respectively.

Figure 1 shows the biplots of the first 2 PCs for the four considered combinations classical and robust PCA for log-transformed and ilr transformed data. The biplot for log-transformed data and classical PCA (upper left) shows clearly the typical data closure problem because either SiO_2 is high in quartz-rich samples or LOI is very high in the organic samples from northern Finland. Thus all other concentrations must decline and plot towards negative loadings of PC1. Also the curve-shape configuration of the scores is typical for data closure problems. The plot for the log-transformed robust PCA (upper right) shows again the closure problem. Samples with very high SiO_2 must have low concentrations of the other compounds. The influence of the organic samples is downweighted because there are much fewer samples with high LOI (only in northern Finland) than with high SiO_2. Although the plot is dominated by many outliers in the scores, the PCA axes are derived in a robust way and not essentially influenced by these outliers caused by the organic samples. The biplot for classical PCA for the ilr transformed data (lower left) shows that the data are now opened because the bias due to data closure has disappeared. However, outliers can still play an important role and spoil the correlation structure. This is no longer the case when robust PCA is applied to the ilr transformed data (Figure 1, lower right) where the true geochemical correlations become clearly visible. For example, K_2O and SiO_2 indicate the coarse grained sediments in the southern project area with high amounts of quartz and potassium feldspar, LOI and P_2O_5 an organic association (samples from northern Finland), Al_2O_3, CaO, and Na_2O a plagioglace association, MgO and Fe_2O_3 a mafic association.

In contrast to the biplots presented in Figure 1, the maps of the corresponding first PCs do not differ drastically (Figure 2). The map for classical PCA of the log-transformed data (upper left) combines the samples with high SiO_2 and the samples with high LOI against the samples that have high concentrations with all the other compounds. In the robust version (upper right) the organic-rich samples play a less dominant role, they are downweighted. The map for classical PCA on the ilr transformed data (lower left) is rather noisy, but the robust version (lower right) shows a clear separation between

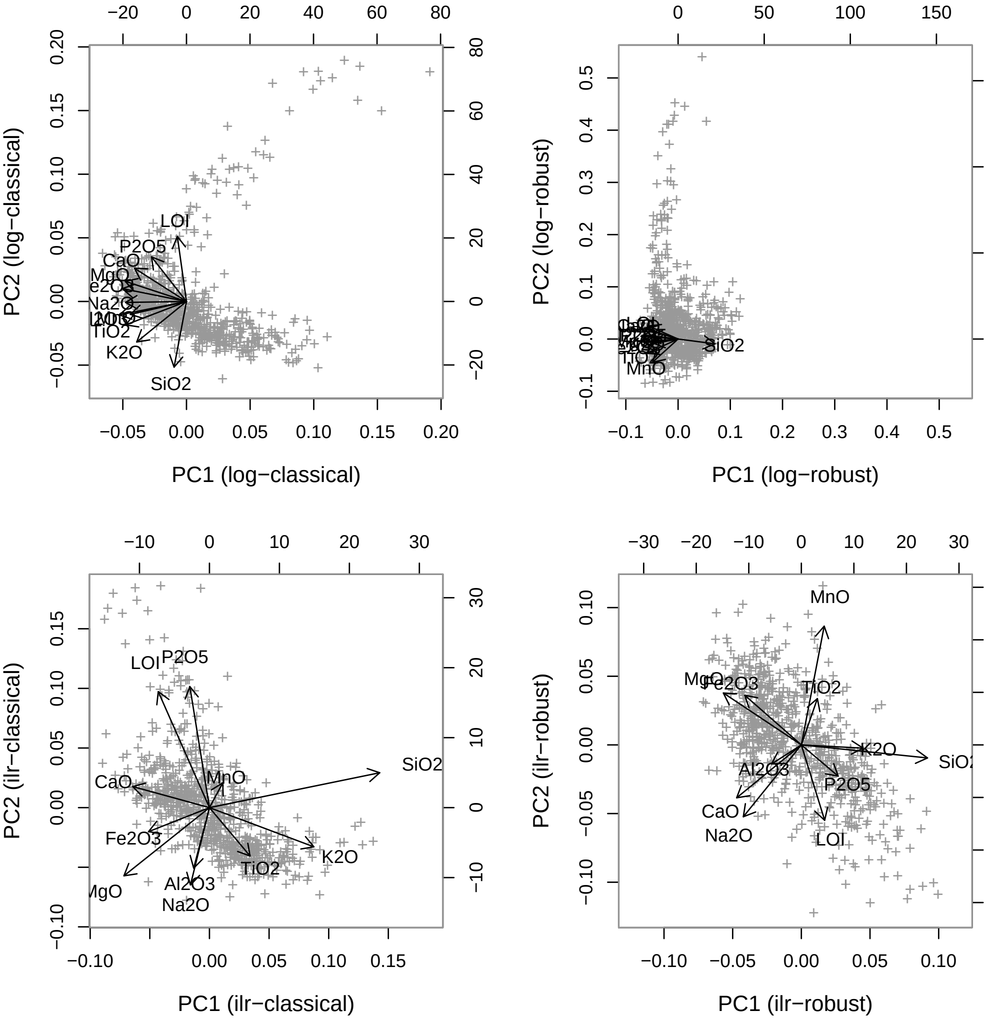

Figure 1: Biplots of the first two PCs for the log-transformed (upper row) and ilr transformed (lower row) BSS data using classical (left column) and robust (right column) PCA.

the samples from the southern and northern half of the survey area.

The maps of the second principal components (see Figure 3) show a much clearer difference. They are straightforward to interpret using the biplots from Figure 1. The map for PC2 (log-classical, upper left) is dominated by the outliers in LOI (see biplot) and shows in general the location of the samples with much organic material due to climatic (wet and cold) conditions. The information as such is interesting, but would not need a PCA to detect. The map for the log-robust case (upper right) shows even less structure, because the information from the "outliers", that provides interesting information in the non-robust map is now also lost, only the samples in Finland with very high organic content are still clearly marked, the coastal pattern in Norway is, however, almost lost due to the down-weighing of the outliers. The map for the ilr-classical case (lower left) contains much more detailed information on "organic rich samples" in the northern countries versus coarser grained glacial sediments in the southern project area. Finally, the map for the ilr-robust case (lower right) shows more detail (e.g. S-tip of Norway, southern border of Poland) than any of the other maps and actually contains really new, valuable information for the geochemist.

The above analysis has demonstrated that robust PCA for ilr transformed data gave the most useful and interpretable results. It is often argued that instead of a robust statistical analysis the outliers could be removed and the classical procedure could be applied. However, multivariate outlier detection in case of compositional data is again not trivial (see Filzmoser and Hron, 2008), and thus an approach that is robust by itself is preferable.

Geochemical data are by definition compositional data because if all constituents of a soil sample have been analyzed they must sum up to 100%. Also in the above example the sum of the 11 compounds considered was nearly 100% for each sample (Figure 4, left). However, even if the row sums of the data are not constant the same problems with closure are present. Suppose that the variable SiO_2 had not been measured or included in the above example data set. SiO_2 has a median of about 70 wt.-% (weight-percent) and is thus very

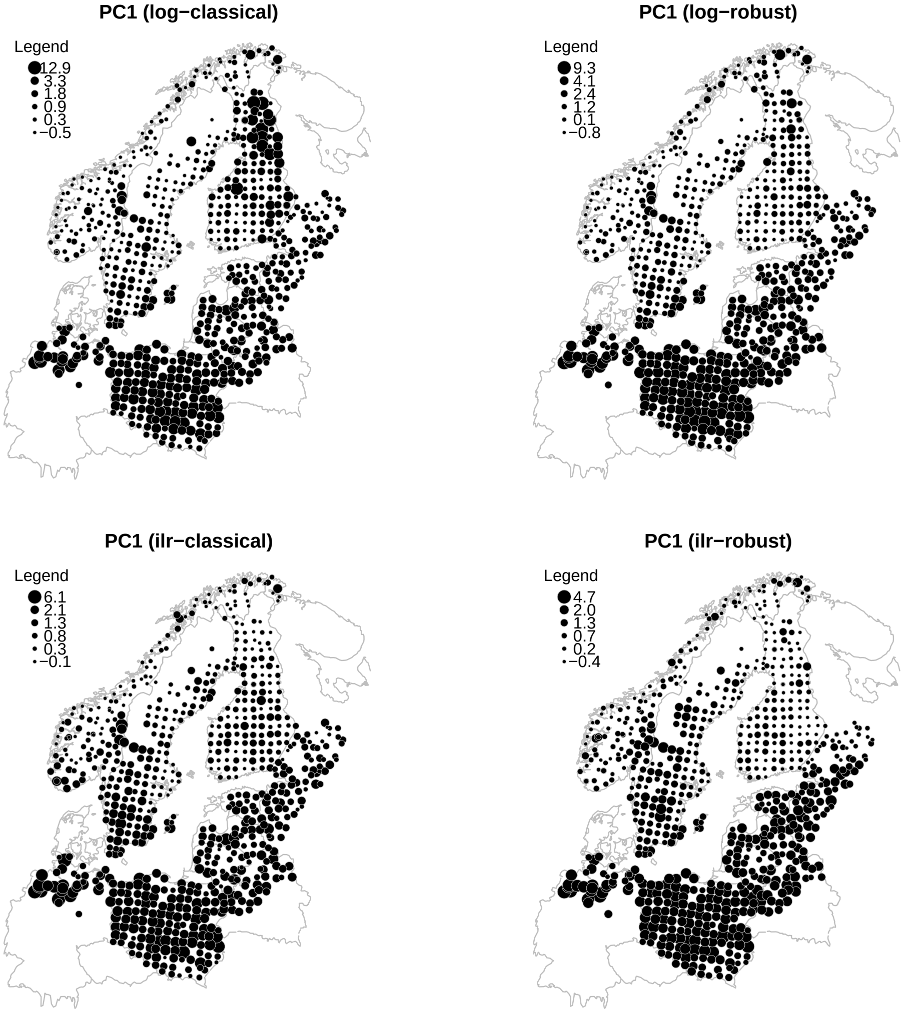

Figure 2: Maps for the first principal component scores for the log-transformed (upper row) and ilr transformed (lower row) BSS data using classical (left column) and robust (right column) PCA.

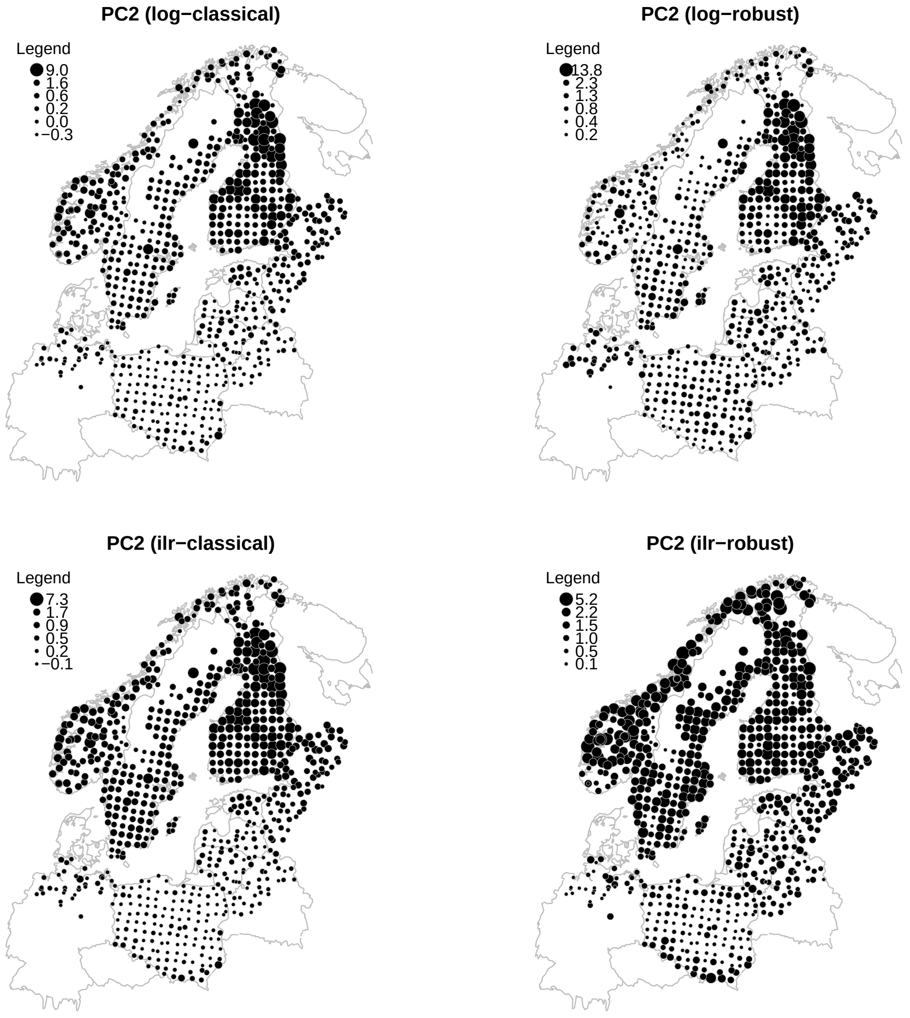

Figure 3: Maps for the second principal component scores for the log-transformed (upper row) and ilr transformed (lower row) BSS data using classical (left column) and robust (right column) PCA.

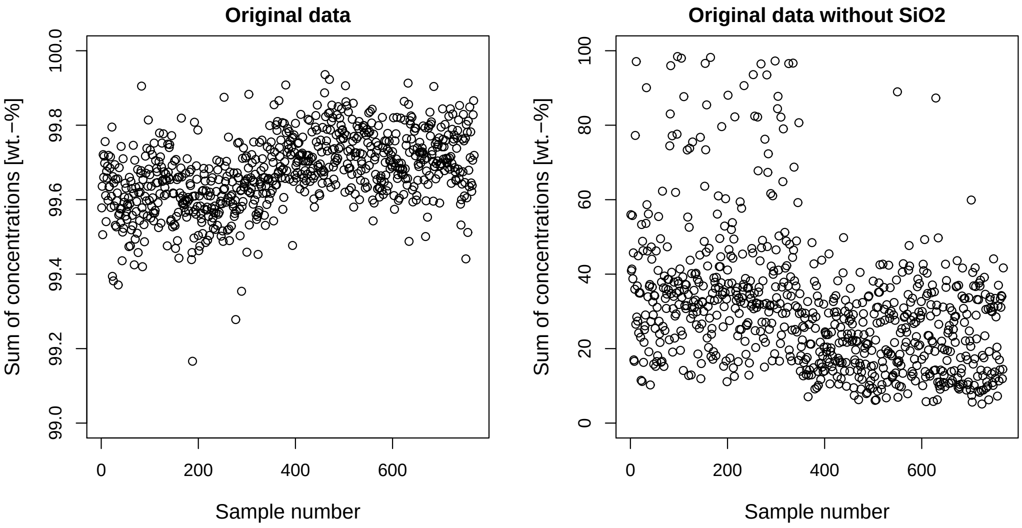

Figure 4: Row sums for each observation of the original data (left) and the original data without SiO_2 (right).

dominant in most samples. When taking the reduced data matrix without SiO_2, the new row sums vary in the range 5-99 wt.-% and are far from being constant or nearly constant (Figure 4, right). Here the same problems arise when multivariate methods like PCA are applied. Figure 5 shows the biplots of the first two robust PCs for the log-transformed (left) and the ilr transformed (right) data. When comparing them with Figure 1 (right column) where SiO_2 was still included, more or less the same picture appears. The closure effect is clearly visible for the analysis based on the log-transformed data whereas the ilr transformation allows a meaningful variable configuration. This demonstrates the problem of so-called subcompositions (Aitchison, 1986) which still require an appropriate transformation but where the closure effect is not visible when inspecting row sums. This situation where the user is not able to *check* (using e.g. row sums) whether the data are closed or not is rather unfortunate. Geochemists often hoped that the problem of closure just disappears if the major elements – or at least compounds like SiO2 (or

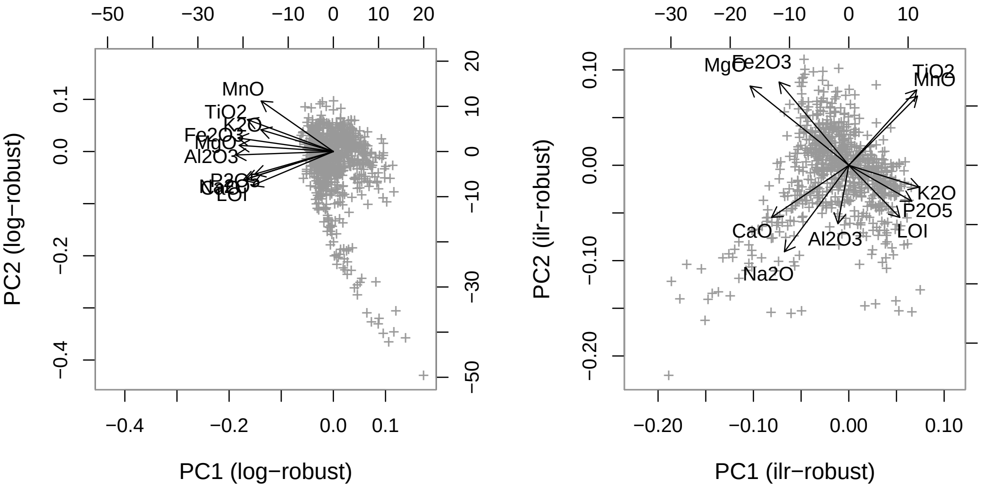

Figure 5: Biplots of the first two PCs for the log-transformed (left) and ilr transformed (right) reduced (without SiO_2) BSS data using robust PCA.

LOI), that can completely dominate the composition are simply not analyzed or not used in data analysis. Our example demonstrates that closure is an inherent problem of geochemical data that cannot be simply overcome, it is more a philosophical than a geochemical problem. Closure will always influence compositional data no matter how many of the elements are used or not and independent of high or low element concentrations. Thus whenever the data are of compositional nature (i.e. practically always in geochemistry and environmental sciences) appropriate data transformations are recommended prior to multivariate analysis.

6. CONCLUSIONS

Robust PCA for compositional data is not possible for the clr transformed data if robust PCA is based on a robust covariance estimator like the MCD. A solution is to take ilr transformed data which do not result in singularity problems for robust covariance estimation. The resulting loadings and scores have to be back-transformed to the clr space in order to allow for an interpretation in terms of the original variable names.

We need to conclude that the closure problem is inherent in geochemical data, it cannot be overcome by not including some major element in the data analysis (or not analyzing them). It is thus suggested to always use appropriate transformations prior to any multivariate analysis of compositional data.

ACKNOWLEDGMENTS

The authors are grateful to the referee for helpful comments and suggestions. We also thank Prof. Horova for handling the paper.

REFERENCES

[1] Aitchison J. 1983. Principal component analysis of compositional data. *Biometrika* **1**:57-65.

[2] Aitchison J. 1984. Reducing the dimensionality of compositional data sets. *Math. Geol.* **16**:617-635.

[3] Aitchison J. 1986. *The statistical analysis of compositional data.* Chapman and Hall: London.

[4] Aitchison J, Greenacre M. 2002. Biplots of compositional data. *Applied Statistics* **51**:375 - 392.

[5] Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueraz G, Barceló-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. *Math. Geol.* **35**:279-300.

[6] Filzmoser P. 1999. Robust principal components and factor analysis in the geostatistical treatment of environmental data. *Environmetrics* **10**:363-375.

[7] Filzmoser P, Hron K. 2008. Outlier detection for compositional data using robust methods. *Math. Geol* **40(3)**:233 - 248.

[8] Gabriel KR. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**: 453-467.

[9] Harville DA. 1997. *Matrix algebra from a statistician's perspective.* Springer-Verlag: New York.

[10] Johnson R, Wichern D. 2007. *Applied multivariate statistical analysis.* Sixth edition. Prentice-Hall: London.

[11] Maronna R, Martin RD, Yohai VJ. 2006. *Robust statistics: Theory and methods.* John Wiley: New York.

[12] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado J (2007) Lecture notes on compositional data analysis. http://diobma.udg.edu/handle/10256/297/

[13] R development core team. 2008. R: A language and environment for statistical computing. Vienna, `http://www.r-project.org`.

[14] Reimann C, Filzmoser P. 2000. Normal and lognormal data distribution in geochemistry: Death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology* **39**: 1001-1014.

[15] Reimann C, Siewers U, Tarvainen T, Bityukova L, Eriksson J, Gilucis A, Gregorauskiene V, Lukashev VK, Matinian NN, Pasieczna A. 2003. Agricultural Soils in Northern Europe: A Geochemical Atlas. In *Geologisches Jahrbuch.* Schweizerbart'sche Verlagsbuchhandlung: Stuttgart.

[16] Rousseeuw PJ, Van Driessen K. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**:212-223.