Correlation analysis for compositional data

Peter Filzmoser[1] **and Karel Hron**[2]

Abstract

Compositional data need a special treatment prior to correlation analysis. In this paper we argue why standard transformations for compositional data are not suitable for computing correlations, and why the use of raw or log-transformed data is neither meaningful. As a solution, a procedure based on balances is outlined, leading to sensible correlation measures. The construction of the balances is demonstrated using a real data example from geochemistry. It is shown that the considered correlation measures are invariant with respect to the choice of the binary partitions forming the balances. Robust counterparts to the classical, non-robust correlation measures are introduced and applied. By using appropriate graphical representations it is shown how the resulting correlation coefficients can be interpreted.

Keywords Correlation analysis; ilr transformation; log-ratio transformation; Compositional data; Balances; Subcompositions; Amalgamation; Robust statistics

1 Introduction

Correlation analysis is popular in many applications because it is a quantitative way to evaluate whether two or more variables are related or not. Thus, correlation analysis allows to reduce the information contained in n observations that have been measured on pairs or groups of data to a single number falling into a normed interval. It is then convenient to proceed with the derived correlation coefficients for interpreting the relations. On the other hand, depending on the data structure and data quality, the correlation measure can be quite misleading because it can be influenced by the skewness of the data distributions or by outliers. Transformations of the variables, or nonparametric (Conover 1998) or robust (Maronna, Martin and Yohai 2006) correlation measures can avoid such problems.

Great care is necessary when attempting any correlation analysis with compositional data (Aitchison 1986). Compositional data are data that carry only relative

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

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

information (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2007), and in the most common situation they sum up to a constant. For example, if soil samples are completely analyzed, the concentrations of the chemical elements per sample sum up to 1,000,000 mg/kg. More formally, a compositional vector that describes a sample consisting of D compositions or *parts* is defined as

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

where the relevant information is contained only in the ratios between the parts (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2007). From this definition it follows that $(x_1, \ldots, x_n)^t$ and $(ax_1, \ldots, ax_n)^t$ include essentially the same information, for any non-zero number a. A way to simplify the use of compositions is to represent them as positive vectors, the parts of which sum up to a positive constant value κ (usually chosen as 1 or 100 when dealing with percentages), namely as

$$\boldsymbol{x} = (x_1, \ldots, x_D)^t,\ x_i > 0,\ i = 1, \ldots, D,\ \sum_{i=1}^{D} x_i = \kappa.$$

Due to this constraint, the set of all D-parts form a simplex sample space $\mathcal{S}^D$, a $(D-1)$-dimensional subset of the real space $\mathbb{R}^D$.

The above mentioned indicates that the closure constant κ is not the key aspect, but that the scale is important, what makes the problem when using standard correlation measures. These are based on variances and covariances that are defined for the Euclidean space and not for the simplex. For instance, the mean minimizes the expected squared distance and the variance the expected squared distance from the mean. A further problem is the presence of negative bias in the covariance structure on the simplex (Aitchison 1986), represented by the relation

$$\mathrm{cov}(x_1, x_2) + \mathrm{cov}(x_1, x_3) + \ldots + \mathrm{cov}(x_1, x_D) = -\mathrm{var}(x_1).$$

This artifact was already noted by Pearson (1897). Using standard correlation analysis for compositional data thus leads to such undesirable properties like scale dependence and subcompositional incoherence. The latter means that if not all parts of the compositional data are available but only a subcomposition, the correlation between compositional parts depends on the subcomposition used. So, if two teams have measured only some parts out of all available compositions, then the correlation matrices computed from the common parts of the data are in general different (Aitchison 1986).

It can be concluded at this point that prior to applying correlation analysis, compositional data first need to be transformed into an appropriate sample space. In geochemistry the distribution of element concentrations are often very skewed, and it is argued that a logarithmic transformation symmetrizes or even normalizes

the distribution, a requirement that is not a "must" but a recommendation for computing and interpreting Pearson correlation coefficients (Reimann and Filzmoser 2000). However, even after log-transformation the compositional nature remains inherent in the data, and the derived correlation coefficients could thus be severely misleading.

A family of transformations, the so-called logratio transformations (= logarithms of ratios) has been introduced to transform compositional data to an unconstrained real space (Aitchison 1986). The alr (additive logratio) transformation builds on logratios to a single reference variable. Unfortunately, the alr transformation is not isometric, which means that distances are not preserved. The clr (centered logratio) transformation is an isometric transformation and is defined by the logratio to the geometric mean of all variables. This avoids the selection of a ratio variable like for alr, and simplifies the interpretation of the transformed variables, because one could think in terms of the original variables. However, also here problems arise in the context of correlation analysis, because correlations between them cannot be interpreted as correlations between the original variables. There are further arguments against the use of the clr transformation: the transformed data are singular and subcompositionally incoherent. The former property is a result from the definition of clr (Aitchison 1986).

The ilr (isometric logratio) transformation (Egozcue et al. 2003) solves the problem of data collinearity resulting from the clr transformation, while preserving all its advantageous properties like isometry. It is based on the choice of an orthonormal basis (in the well known Euclidean sense) in the hyperplane formed by the clr transformation. For an appropriate choice of the basis also the problem of subcompositional incoherence can be avoided (Egozcue and Pawlowsky-Glahn 2005). Correlations computed from the ilr transformed data can, however, not be interpreted in the sense of the original variables, because the ilr variables (basis vectors) are related to the original variables only through non-linear functions (Egozcue et al. 2003). In general, there is also no way to transform the correlations back to the original space. One can, however, choose different bases by considering non-overlapping groups of the original variables (Egozcue and Pawlowsky-Glahn 2005). The results of the procedure to construct such a new basis are called *sequential binary partitions*, and the constructed basis vectors are called *balances*. The balances can be viewed as new variables with the property that they represent both groups of parts and relations between the groups. Additionally, the relative information contained in the non-overlapping groups is separated from the relations between the groups. Correlation analysis can then be applied to the balances representing the separated groups.

In this paper we will follow the ideas of Egozcue and Pawlowsky-Glahn (2005) by

using balances for computing correlations, which seems to be the only sensible way of applying correlation analysis to compositional data. For choosing the balances, either expert knowledge or the compositional biplot (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2007) can be used. We extend the use of correlations for pairs of balances to multiple correlations, relating single balances with groups of balances, and to group correlations, relating two groups of balances. Moreover, we will present robust versions of these correlation measures.

The paper is organized as follows. In Section 2 the idea of sequential binary partitioning to construct balances is explained in more detail by using a real data example. Section 3 reminds the reader about the different correlation measures and robustified versions thereof. A result is presented that proofs the invariance of the correlation measures to different choices of the sequential binary partitions. In Section 4 applications to real data are presented, and the final Section 5 concludes.

2 Balances and compositional data

We want to define correlation coefficients between the balances by using a well-known data set from geochemistry, the so-called Kola data (Reimann et al. 1998). These data contain the concentrations of more than 50 chemical elements in about 600 soil samples taken at the Peninsula Kola. At each sample site four different layers of soil have been analyzed. The complete data set is available in the R library *StatDA* (R Development Core Team 2008). Here we use the analytical results from the O-horizon.

Considering the arguments of Section 1, correlation analysis is only possible for groups of at least two parts. Interesting groups are parts reflecting effects like pollution, seaspray, bioproductivity, etc. Various statistical analyses of these data (see, e.g., Reimann et al. (2008)) have been indicative for the group assignments shown in Table 1. The group reflecting bioproductivity (B) consists of the elements forming contamination (C) and mineralization (M), all other groups consist of non-overlapping elements.

Table 1 about here.

Table 1 contains 12 different elements or parts in the simplex, and this information can be expressed with 11 dimensions that will form the balances. Additionally, since the groups of parts should be separated for correlation analysis, we need to construct balances describing each group. A group consisting of k parts can be described by $k-1$ balances. So, for group P we need two balances, for group S two balances, for C three balances, and for group M one balance is required. The remaining three balances contain the information that are linking the groups, like one balance that links groups C and M to form group B.

The above procedure to construct the balances is described in detail in Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2007). It is possible to present the sequential binary partitions and the resulting balances in table form (see Table 2).

Table 2 about here.

The main idea for the construction of sequential binary partitions is as follows. We start with all parts of the composition (here 12 parts), and at each order of the partition a group in the previous level is subdivided into two subgroups: those with label + and the other ones with label −. In Table 2 the balance z_{11} separates all parts into two groups. The group with labels + is separated in the next level by balance z_{10} into two further subgroups, and so on. The empty entries in Table 2 have the meaning that these parts are not involved in the partition at this order. The resulting balances z_1 and z_2 describe the relative information within group P, z_3 and z_4 are for group S, group C is described by the balances z_6, z_7, and z_8, and balance z_9 is for mineralization (group M). Bioproductivity (B) which combines groups C and M is formed by balances z_5 to z_9. The remaining balances z_{10} and z_{11} contain relative information about the relation between the groups, and they will not be of interest for the correlation analysis of the groups.

The symbols used in Table 2 refer to the rules for computing the balances. The general formula is

$$z_i = \sqrt{\frac{rs}{r+s}} \ln \frac{(\prod_+ x_j)^{\frac{1}{r}}}{(\prod_- x_k)^{\frac{1}{s}}} \quad \text{for} \quad i = 1, \ldots, D-1, \tag{1}$$

where the products $\prod_+$ and $\prod_-$ only include parts coded with + and −, and r and s are the numbers of positive and negative signs (parts) in the i-th order partition, respectively (see Egozcue and Pawlowsky-Glahn (2006) for details). Using this general formula, the formulas for our balances z_1 to z_9 of interest can be derived directly from Table 2 and are shown in Table 3.

Table 3 about here.

The sequential binary partitions always describe separated groups of parts. However, the construction procedure suggests the use of parts rather than groups of parts and to find a correlation-like measure for two chosen parts x_i and x_j, $1 \leq i, j \leq D$, $i \neq j$. According to (1) the resulting balance is $\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}$. Since the balance is univariate, a sensible "correlation measure" is the variance, called *normalized variation* (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2007),

$$\tau_{ij} = \operatorname{var}\left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}\right). \tag{2}$$

This measure does not depend on the scale of the data. A small value of (2) is obtained if the ratio x_i/x_j is nearly constant. This indicates a strong relationship between i-th and j-th part, because samples with high values at x_i will also have high values at x_j and vice versa (for regional data this will result in very similar maps for the two parts). On the other hand, the measure τ_{ij} gets larger the more dissimilar x_i and x_j are (maps show different patterns). The measure in general increases even more if there is an opposite behavior of the parts, meaning that samples with high values at x_i will have low values at x_j and vice versa (one map is the "negative" of the other map). For this reason, τ_{ij} is not working like a usual correlation measure. Also the normalized version $\exp(-\tau_{ij})$ which transforms the values to the interval [0,1] (Buccianti and Pawlowsky-Glahn, 2005) does not give the results we would expect from "correlation analysis".

3 Correlation measures and robust estimation

With the procedure of Section 2 we obtain G separated groups of parts, where each group consists of g_1, $g_2, \ldots, g_G$ ilr variables. Let p denote the number of these balances, i.e. $p = g_1 + \ldots + g_G$, and $\boldsymbol{z} = (z_1, \ldots, z_p)^t$ the corresponding random variables. Moreover, the groups of parts are denoted by the random variables $\boldsymbol{z}_k = (z_{g_1}, \ldots, z_{g_k})^t$, for $k = 1, \ldots, G$.

The population covariance matrix of all balances describing the groups is the $p \times p$ matrix $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{z})$. The covariance between two balances z_i and z_j is the value $\mathrm{cov}(z_i, z_j)$, and the covariance between a balance z_i and a group $\boldsymbol{z}_k$ is the vector $\mathrm{cov}(z_i, \boldsymbol{z}_k)$ of length g_k.

For correlation analysis we distinguish among three different cases (see, e.g., Johnson and Wichern (2007)):

Correlation between two balances: A measure of linear dependency of two random variables z_i and z_j (for $1 \leq i, j \leq p$) is the well known *correlation coefficient*, defined as

$$\rho_{z_i,z_j} = \frac{\mathrm{cov}(z_i, z_j)}{\sqrt{\mathrm{var}(z_i)\,\mathrm{var}(z_j)}}.$$

This measure is normed to the interval [-1,1], with 0 indicating no linear relation, and 1 (-1) for perfect positive (negative) linear relation. We can express the squared correlation coefficient as

$$\rho^2_{z_i,z_j} = 1 - \frac{|\boldsymbol{\Sigma}^*|}{\mathrm{var}(z_i)\,\mathrm{var}(z_j)} \quad \text{with} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} \mathrm{var}(z_i) & \mathrm{cov}(z_i, z_j) \\ \mathrm{cov}(z_j, z_i) & \mathrm{var}(z_j) \end{pmatrix}$$

where $|\boldsymbol{\Sigma}^*|$ denotes the determinant of $\boldsymbol{\Sigma}^*$.

Correlation between a balance and a group: A measure of linear relationship between a random variable z_i and a group of random variables $\boldsymbol{z}_k$ is the *multiple correlation coefficient* $\rho_{z_i,\boldsymbol{z}_k}$. This measure falls into the range [0,1] where 0 indicates no linear relationship and 1 perfect linear relation. The square of the multiple correlation coefficient is defined as

$$\rho^2_{z_i,\boldsymbol{z}_k} = \frac{\mathrm{cov}(z_i, \boldsymbol{z}_k)\boldsymbol{\Sigma}_k^{-1}\mathrm{cov}(\boldsymbol{z}_k, z_i)}{\mathrm{var}(z_i)},$$

where $\boldsymbol{\Sigma}_k = \mathrm{cov}(\boldsymbol{z}_k)$. An equivalent formulation is

$$\rho^2_{z_i,\boldsymbol{z}_k} = 1 - \frac{|\boldsymbol{\Sigma}^*|}{|\boldsymbol{\Sigma}_k|\,\mathrm{var}(z_i)} \quad \text{with} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} \mathrm{var}(z_i) & \mathrm{cov}(z_i, \boldsymbol{z}_k) \\ \mathrm{cov}(\boldsymbol{z}_k, z_i) & \mathrm{cov}(\boldsymbol{z}_k) \end{pmatrix}.$$

Correlation between two groups: The definition of the multiple correlation coefficient can be extended to the *group correlation coefficient* $\rho_{\boldsymbol{z}_k,\boldsymbol{z}_l}$ for two random vectors $\boldsymbol{z}_k$ and $\boldsymbol{z}_l$ by defining its square as

$$\rho^2_{\boldsymbol{z}_k,\boldsymbol{z}_l} = 1 - \frac{|\boldsymbol{\Sigma}^*|}{|\boldsymbol{\Sigma}_k|\,|\boldsymbol{\Sigma}_l|} \quad \text{with} \quad \boldsymbol{\Sigma}^* = \begin{pmatrix} \mathrm{cov}(\boldsymbol{z}_k) & \mathrm{cov}(\boldsymbol{z}_k, \boldsymbol{z}_l) \\ \mathrm{cov}(\boldsymbol{z}_l, \boldsymbol{z}_k) & \mathrm{cov}(\boldsymbol{z}_l) \end{pmatrix},$$

where $\boldsymbol{\Sigma}_k = \mathrm{cov}(\boldsymbol{z}_k)$ and $\boldsymbol{\Sigma}_l = \mathrm{cov}(\boldsymbol{z}_l)$ (see Anděl (1978), p. 309, and Anderson (1958)). This straightforward extension is less well-known, but there is a link to the more frequently used canonical correlation analysis (CCA). CCA not only measures the linear relation between two multivariate data sets, but searches for latent variables – so-called *canonical variates* – in each of the data groups such that the scores on the latent variables have maximal correlation (see, e.g., Johnson and Wichern (2007)). There exists a subspace of solutions which has dimension $r = \min(g_k, g_l)$. The results are r pairs of uncorrelated score vectors for both groups, each leading to a maximal correlation ρ_f, for $f = 1, \ldots, r$. These correlations are called *canonical correlation coefficients*, and there exists the relation

$$\rho^2_{\boldsymbol{z}_1,\boldsymbol{z}_2} = 1 - (1-\rho_1^2)(1-\rho_2^2)\cdots(1-\rho_r^2).$$

So, the group correlation summarizes all canonical correlation coefficients by one number in the interval [0,1]. Note that this summary measure is also closely related to the form of the test statistic used for testing uncorrelatedness between two groups of data (Johnson and Wichern 2007).

An important question in this context is whether the choice of the sequential binary partition will alter the resulting correlation coefficients or not. For example, in Table 2 the groups of parts could have been defined differently by choosing the ilr vectors in a different way (e.g. exchange "−" and "+"). Consequently, the formulas

for computing the balances (Table 3) would change. The following theorems state that the resulting correlations will remain the same, and the proofs are given in the Appendix.

Theorem 1: The correlation coefficient, the multiple correlation coefficient, and the group correlation coefficient are invariant with respect to the choice of the sequential binary partition for the representation of given groups of parts.

Theorem 2: The canonical correlation coefficients are invariant with respect to the choice of the sequential binary partition for the representation of given groups of parts.

For given data, all three correlation measures can immediately be computed once the covariance matrix $\mathbf{\Sigma}^*$ has been estimated, because the correlation measures use sub-matrices of $\mathbf{\Sigma}^*$ in their definition. The classical way of estimating the covariance matrix of q-dimensional observations $\boldsymbol{y}_i$, $i = 1, \ldots, n$, is the sample covariance matrix, given by $\frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^t$, where $\bar{\boldsymbol{y}}$ is the arithmetic mean vector of the observations $\boldsymbol{y}_i$. For many practical situations, the quality of this estimation can be poor, especially in case of outlying observations or inhomogeneous data. There are several proposals for a robust estimation of the covariance with the common idea that outliers are downweighted in the estimation procedure (Maronna, Martin and Yohai 2006). A frequently used method for this purpose is the minimum covariance determinant (MCD) estimator (Rousseeuw and Van Driessen 1999). The MCD estimator looks for a subset h out of n observations with the smallest determinant of their sample covariance matrix. A robust estimator of covariance is the sample covariance matrix of these h observations, multiplied by a factor for consistency at normal distribution. The subset size h can vary between half the sample size and n, and it will determine the robustness of the estimator, but also its efficiency.

From the robust estimator of $\mathbf{\Sigma}^*$ the sub-matrices that are needed in the definition of the correlation measures can be extracted. The resulting correlation measures are robust in a sense that they can resist a certain amount of outlying observations.

4 Application and results

In the following we will use the groups from the Kola O-horizon data (see Table 1) to illustrate the use of the correlations measures and their robust versions. The groups, resulting from expert knowledge, will be represented by the balances constructed in Table 2 and computed according to the formulas of Table 3. It is possible to choose the balances for the same groups differently, however, this would leave the resulting correlation measures unchanged (see Theorems 1 and 2). The balances represent

- z_1, $z_2 \longrightarrow$ pollution P,
- z_3, $z_4 \longrightarrow$ seaspray S,
- $z_6, z_7, z_8 \longrightarrow$ contamination C,
- $z_9 \longrightarrow$ mineralization M

and the groups C and M together form bioproductivity B. The results for the classical (non-robust) and robust group correlations are displayed in Table 4 as well as the first and second (classical and robust) canonical correlation coefficients. The multiple correlation coefficient between the groups C and M is 0.17 if classically estimated, and 0.09 if robustly estimated.

Table 4 about here.

Although it is convenient to obtain a single number that expresses the linear relation, it is difficult to interpret these relations. For example, there is a relatively strong linear relationship between the groups seaspray S and bioproductivity B, and there is also an interesting difference between classical and robust estimation (see Table 4). What is the meaning of the relation, and what is the reason for this difference? The first question can be answered by inspecting the canonical variates for both groups, i.e. the projection directions used within the first and within the second group that were responsible for finding the best possible linear relations between the groups. The first canonical variates are the following linear combinations:

Group S:	$2.6z_3$	$-1.3z_4$				(classical)
	$2.9z_3$	$+2.0z_4$				(robust)
Group B:	$-1.2z_5$	$-0.5z_6$	$+2.4z_7$	$+1.4z_8$	$-1.5z_9$	(classical)
	$-1.7z_5$	$-1.8z_6$	$+3.2z_7$	$+0.7z_8$	$-1.9z_9$	(robust)

Some balances have stronger, some have weaker influence to the projection directions, some have positive, and some have negative influence. There is also a certain change from classical to robust estimation. However, since the balances are only mathematical constructions, it is impossible to find any interpretation for the canonical variates.

Another possibility for gaining more insight into the results is the use of appropriate plots. It is natural to visualize the projection of the data onto the first canonical variates. This corresponds to one score vector in the first group and one in the second group, which result in the maximum correlation among all possible projection directions. This plot for the groups S and B is shown in Figure 1. The left plot is for the classical analysis, the right for robust canonical correlation analysis based

on robust covariance estimation using the MCD estimator. In the right plot we also used different symbols (+) for the multivariate outliers that were identified with the MCD estimator and downweighted for the analysis. Obviously, those observations are downweighted that are not following the joint covariance structure of both groups, and this has the effect of a higher robust canonical correlation coefficient (0.54 for the robust analysis compared to 0.44 for the classical analysis).

Figure 1 about here.

A sensible interpretation of the results can be obtained by presenting the results in a geographical map. For this reason, we use different symbols for the plots shown in Figure 1 in order to highlight the observations that are responsible for the linear relation. The first classical canonical correlation coefficient between the groups is the same as the usual Pearson correlation between the scores shown in Figure 1 (left) (and similar for the robust counterparts with the outliers downweighted). We can thus partition the data points shown in the plots at the cooordinate-wise medians into 4 quadrants: points in the upper right and lower left quadrant will increase the correlation measure, and points in the other 2 quadrants will decrease the measure. This fact comes from the definition of the covariance $\mathrm{cov}(x, y) = \mathrm{E}[(x - \mathrm{E}(x))(y - \mathrm{E}(y))]$. We are interested in the points that allow for a high relation, and thus these points obtain a special symbol in Figure 2, with the symbol size according to the Mahalanobis distance from the center. In contrast to the Euclidean distance, the Mahalanobis distance (Mahalanobis 1936) is a distance measure that accounts for the covariance structure (see Filzmoser and Hron (2008) for its usage in the context of compositional data), and points with high Mahalanobis distance will be most influential to an increase of the correlation coefficient. Figure 2 (left) shows the same picture as Figure 1, but with the modified symbols. These symbols are also used in the right plot of Figure 2 representing the geographical map of the Kola project area. Figure 3 shows the corresponding plots for the robust analysis, with the outliers highlighted by the symbol "+". In both analyses we can see high influence for an increase of the correlation measure of observations located at the coast in the north of the area. It is known that seaspray decreases from north to south, and that bioproductivity increases from south to north. The industrial centers around Monchegorsk and Zapoljarnij in the east, and around Nikel in the north are disturbing this relation. The robust analysis finds many outliers in these regions (see Figure 3, right) and thus leads to a more stable analysis. Also in the south-west we can find (grey) points with high influence on the increase of the correlation measure. This is an area where both factors, seaspray and bioproductivity, are very low.

Figure 2 about here.

Figure 3 about here.

In a similar way we could present the results for the other groups in order to gain more insight into their relations.

In the remaining part of this section we will compare the above approach based on the balances with the results when the compositional nature of the considered data would simply be ignored. In this case one would probably check the distribution of the variables and apply a log-transformation because the variables are all right-skewed. Canonical correlation analysis (classical and robust) can then be applied to the log-transformed data. As a result one can expect a different value of the correlation coefficient, and the practitioner has to decide which of the results makes more sense. Besides a theoretical/mathematical argumentation it may be convincing to inspect and compare biplots (Gabriel 1971) of the log-transformed data and of the balances. Figure 4 (left) shows the biplot of the considered log-transformed Kola O-horizon data. Although the relations between the variables reflect the groups summarized in Table 1, the overall correlation structure is distorted, because all arrows representing the variables in Figure 4 (left) are arranged in a half-plane. This is the typical outcome when ignoring the compositional nature of the data, and the correlation measures will be unrealistic in general. The biplot shown in Figure 4 (right) is based on a robust covariance estimation based on the MCD (see, e.g., Reimann et al. 2008), and it shows the same effect. Thus, robust estimation is not helpful in this case, and one has to use an appropriate approach like the construction of balances. Figure 5 presents the biplots (left plot for classical estimation, right plot for robust estimation) for the balances constructed according to Table 2 and 3. The effect of distorted variable relations is no longer visible. Because of the presence of outliers, a robust treatment of the data will be more reliable.

Figure 4 about here.

Figure 5 about here.

5 Conclusions

Compositional data need a special treatment for correlation analysis. The application of correlation measures to the raw data or to log-transformed data is inappropriate because of the geometry of compositional data. In this paper we presented an approach based on balances. It was shown that the considered correlation measures are invariant with respect to the choice of the sequential binary partitions for defining the balances. It is, however, important to use all balances of a group rather than single balances for the computation of the correlation.

One might argue that the groups that are described by the balances can not always be clearly defined and separated from each other. For the Kola data example used in this paper the groups were based on expert knowledge, but there are probably also other influential elements for the groups, and there could even be elements affecting two or more groups. This seems to be a weak point in the concept of balances, and even if balances are constructed with compositional biplots (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2007), it will in general not be possible to define the groups in a unique way.

Although the procedure for constructing the balances is more demanding than other standard transformations, it is strictly defined and easy to program. The example in Section 4 has shown that the use of log-transformed data for correlation analysis – which is frequently done in geochemistry and other fields dealing with compositional data – can lead to biased variable relations and thus to unrealistic correlation measures. In contrast to the above mentioned procedure based on balances, a log-transformation does not respect the nature of compositional data. Thus a statistical analysis for log-transformed data has no theoretical background and can be completely misleading. But even an appropriate analysis based on balances can be misleading in case of outliers. In this case a robust procedure will give a reliable answer.

Acknowledgements

The authors are grateful to the referees for helpful comments and suggestions. Moreover, Dr. Clemens Reimann from the Geological Survey of Norway is thanked for his help with the Kola data example.

References

[1] Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman & Hall, London, 416 p

[2] Anděl J (1978) Mathematical statistics. SNTL/ALFA, Prague, 346 p (in Czech)

[3] Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York, 374 p

[4] Buccianti A, Pawlowsky-Glahn V (2005) New perspectives on water chemistry and compositional data analysis. Math Geol 37(7):703-727

[5] Conover WJ (1998) Practical nonparametric statistics. 3rd edition, John Wiley & Sons, Inc., New York, 584 p

[6] Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueraz G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Math Geol 35(3):279-300

[7] Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Math Geol 37(7):795-828

[8] Egozcue JJ, Pawlowsky-Glahn V (2006) Simplicial geometry for compositional data. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V (eds) Compositional data analysis in the geosciences: From theory to practice. Geological Society, London, Special Publications 264:145-160

[9] Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Math Geosci 40(3):233–248

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

[11] Harville DA (1997) Matrix algebra from a statistican's perspective. Springer-Verlag, New York, 630 p

[12] Johnson R, Wichern D (2007) Applied multivariate statistical analysis. Sixth edition. Prentice-Hall, London, 816 p

[13] Mahalanobis P (1936) On the generalized distance in statistics. Proceedings of the National Institute of Science of India 12: 49-55

[14] Maronna R, Martin RD, Yohai VJ (2006) Robust statistics: Theory and methods. John Wiley, New York, 436 p

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

[16] Pearson K (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London LX: 489-502

[17] R Development Core Team (2008) R: A language and environment for statistical computing. Vienna, http://www.r-project.org

[18] Reimann C, Äyräs M, Chekushin V, Bogatyrev I, Boyd R, Caritat P d, Dutter R, Finne T, Halleraker J, Jæger O, Kashulina G, Lehto O, Niskavaara H, Pavlov V, Räisänen M, Strand T, Volden T, (1998) Environmental Geochemical Atlas of the Central Barents Region. Geological Survey of Norway (NGU), Geological Survey of Finland (GTK), and Central Kola Expedition (CKE), Special Publication, Trondheim, Espoo, Monchegorsk, 745 p

[19] 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

[20] Reimann C, Filzmoser P, Garrett RG, Dutter R (2008) Statistical data analysis explained. Applied environmental statistics with R. John Wiley, Chichester, 362 p

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

Appendix

Proof of Theorem 1: A different choice of the sequential binary partition corresponds to orthogonal transformations of the balances representing nonoverlapping groups. Let $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ be the random variables that represent two different nonoverlapping groups with g_1 and g_2 balances, respectively. Denote their covariance matrices by $\boldsymbol{\Sigma}_{\boldsymbol{z}_1} = \mathrm{cov}(\boldsymbol{z}_1)$ and $\boldsymbol{\Sigma}_{\boldsymbol{z}_2} = \mathrm{cov}(\boldsymbol{z}_2)$. Furthermore, let $\boldsymbol{P}_i$ $(i = 1, 2)$ be an orthogonal matrix with g_i, such that $\boldsymbol{P}_i\boldsymbol{P}_i^t = \boldsymbol{P}_i^t\boldsymbol{P}_i = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. An orthogonal transformation of $\boldsymbol{z}_i$ is then given by $\boldsymbol{w}_i = \boldsymbol{P}_i\boldsymbol{z}_i$, and the resulting covariance matrix is $\boldsymbol{\Sigma}_{\boldsymbol{w}_i} = \mathrm{cov}(\boldsymbol{w}_i)$. The determinant of this covariance matrix is

$$|\boldsymbol{\Sigma}_{\boldsymbol{w}_i}| = |\boldsymbol{P}_i\boldsymbol{\Sigma}_{\boldsymbol{z}_i}\boldsymbol{P}_i^t| = |\boldsymbol{P}_i|\,|\boldsymbol{\Sigma}_{\boldsymbol{z}_i}|\,|\boldsymbol{P}_i^t| = |\boldsymbol{\Sigma}_{\boldsymbol{z}_i}|,$$

using the property that $|\boldsymbol{P}_i| = \pm 1$ and $|\boldsymbol{P}_i^t| = \mp 1$ which comes from the orthogonality of the matrix $\boldsymbol{P}_i$ (see, e.g., Harville (1997), Corollary 13.3.5). Now consider the joint random vectors $\boldsymbol{z} = (\boldsymbol{z}_1^t, \boldsymbol{z}_2^t)^t$ and $\boldsymbol{w} = (\boldsymbol{w}_1^t, \boldsymbol{w}_2^t)^t$ and their covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{z}} = \mathrm{cov}(\boldsymbol{z})$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \mathrm{cov}(\boldsymbol{w})$, respectively. From the sequential binary partition procedure which constructs nonoverlapping groups it is clear that the covariance matrix of $\boldsymbol{w}$ equals

$$\boldsymbol{\Sigma}_{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2 \end{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{z}} \begin{pmatrix} \boldsymbol{P}_1^t & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{P}_2^t \end{pmatrix},$$

with $\boldsymbol{0}$ being a matrix of zeros of the corresponding order. Since the matrix

$$\begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2 \end{pmatrix}$$

is orthogonal, we have again $|\boldsymbol{\Sigma}_{\boldsymbol{w}}| = |\boldsymbol{\Sigma}_{\boldsymbol{z}}|$. So, all quantities needed for the group correlation coefficient remain unchanged.

The proof for the correlation coefficient and for the multiple correlation coefficient is analogous. □

Proof of Theorem 2: We use the same notation as for the proof of Theorem 1. Additionally, the covariance between the two random vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ is denoted by $\boldsymbol{\Sigma}_{\boldsymbol{z}_1,\boldsymbol{z}_2} = \mathrm{cov}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \boldsymbol{\Sigma}_{\boldsymbol{z}_2,\boldsymbol{z}_1}^t$. Then the covariance matrix of $\boldsymbol{w}$ can be written as

$$\boldsymbol{\Sigma}_{\boldsymbol{w}} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{w}_1} & \boldsymbol{\Sigma}_{\boldsymbol{w}_1,\boldsymbol{w}_2} \\ \boldsymbol{\Sigma}_{\boldsymbol{w}_2,\boldsymbol{w}_1} & \boldsymbol{\Sigma}_{\boldsymbol{w}_2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2 \end{pmatrix} \times$$

$$\times \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{z}_1} & \boldsymbol{\Sigma}_{\boldsymbol{z}_1,\boldsymbol{z}_2} \\ \boldsymbol{\Sigma}_{\boldsymbol{z}_2,\boldsymbol{z}_1} & \boldsymbol{\Sigma}_{\boldsymbol{z}_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{P}_1^t & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2^t \end{pmatrix} = \begin{pmatrix} \boldsymbol{P}_1\boldsymbol{\Sigma}_{\boldsymbol{z}_1}\boldsymbol{P}_1^t & \boldsymbol{P}_1\boldsymbol{\Sigma}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\boldsymbol{P}_2^t \\ \boldsymbol{P}_2\boldsymbol{\Sigma}_{\boldsymbol{z}_2,\boldsymbol{z}_1}\boldsymbol{P}_1^t & \boldsymbol{P}_2\boldsymbol{\Sigma}_{\boldsymbol{z}_2}\boldsymbol{P}_2^t \end{pmatrix}.$$

Since the canonical correlation coefficients are the square roots of the eigenvalues of the matrix product $\boldsymbol{\Sigma}_{\boldsymbol{z}_1}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\boldsymbol{\Sigma}_{\boldsymbol{z}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{z}_2,\boldsymbol{z}_1}\boldsymbol{\Sigma}_{\boldsymbol{z}_1}^{-1/2}$, we have to show that the matrix

product $\boldsymbol{\Sigma}_{\boldsymbol{w}_1}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{w}_1,\boldsymbol{w}_2}\boldsymbol{\Sigma}_{\boldsymbol{w}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w}_2,\boldsymbol{w}_1}\boldsymbol{\Sigma}_{\boldsymbol{w}_1}^{-1/2}$ has the same eigenvalues (see, e.g., Johnson and Wichern (2007)). Using the properties $\boldsymbol{P}_i^t\boldsymbol{P}_i = \boldsymbol{I}$ and $\boldsymbol{\Sigma}_{\boldsymbol{z}_i}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{z}_i}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{z}_i}^{-1/2}$ for $i = 1, 2$, the inverse of the covariance matrix of $\boldsymbol{w}_i$ can be written as

$$\boldsymbol{\Sigma}_{\boldsymbol{w}_i}^{-1} = (\boldsymbol{P}_i\boldsymbol{\Sigma}_{\boldsymbol{z}_i}\boldsymbol{P}_i^t)^{-1} = \boldsymbol{P}_i\boldsymbol{\Sigma}_{\boldsymbol{z}_i}^{-1}\boldsymbol{P}_i^t = (\boldsymbol{P}_i\boldsymbol{\Sigma}_{\boldsymbol{z}_i}^{-1/2}\boldsymbol{P}_i^t)(\boldsymbol{P}_i\boldsymbol{\Sigma}_{\boldsymbol{z}_i}^{-1/2}\boldsymbol{P}_i^t) = \boldsymbol{\Sigma}_{\boldsymbol{w}_i}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{w}_i}^{-1/2}.$$

It follows that

$$|\boldsymbol{\Sigma}_{\boldsymbol{w}_1}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{w}_1,\boldsymbol{w}_2}\boldsymbol{\Sigma}_{\boldsymbol{w}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w}_2,\boldsymbol{w}_1}\boldsymbol{\Sigma}_{\boldsymbol{w}_1}^{-1/2} - \lambda\boldsymbol{I}| =$$

$$|\boldsymbol{P}_1\boldsymbol{\Sigma}_{\boldsymbol{z}_1}^{-1/2}\boldsymbol{P}_1^t\boldsymbol{P}_1\boldsymbol{\Sigma}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\boldsymbol{P}_2^t\boldsymbol{P}_2\boldsymbol{\Sigma}_{\boldsymbol{z}_2}^{-1}\boldsymbol{P}_2^t\boldsymbol{P}_2\boldsymbol{\Sigma}_{\boldsymbol{z}_2,\boldsymbol{z}_1}\boldsymbol{P}_1^t\boldsymbol{P}_1\boldsymbol{\Sigma}_{\boldsymbol{z}_1}^{-1/2}\boldsymbol{P}_1^t - \boldsymbol{P}_1\lambda\boldsymbol{I}\boldsymbol{P}_1^t| =$$

$$= |\boldsymbol{\Sigma}_{\boldsymbol{z}_1}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{z}_1,\boldsymbol{z}_2}\boldsymbol{\Sigma}_{\boldsymbol{z}_2}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{z}_2,\boldsymbol{z}_1}\boldsymbol{\Sigma}_{\boldsymbol{z}_1}^{-1/2} - \lambda\boldsymbol{I}|$$

which completes the proof. □

Figure Captions

Figure 1 Data projected on the first canonical variates: horizontal axes for seaspray S, vertical axes for bioproductivity B; left plot for classical, and right plot for robust canonical correlation analysis. The symbols "+" in the right plot refer to multivariate outliers.

Figure 2 Data projected on the first classical canonical variates shown with special symbols that emphasize the increase of the correlation coefficient (compare Figure 1) (left), and map using the same symbols (right).

Figure 3 Data projected on the first robust canonical variates shown with special symbols that emphasize the increase of the correlation coefficient (compare Figure 1) (left), and map using the same symbols (right).

Figure 4 Biplots of the log-transformed Kola O-horizon data. Left: biplot based of classical covariance estimation; right: biplot based on robust covariance estimation.

Figure 5 Biplots of the constructed balances for Kola O-horizon data. Left: biplot based of classical covariance estimation; right: biplot based on robust covariance estimation.

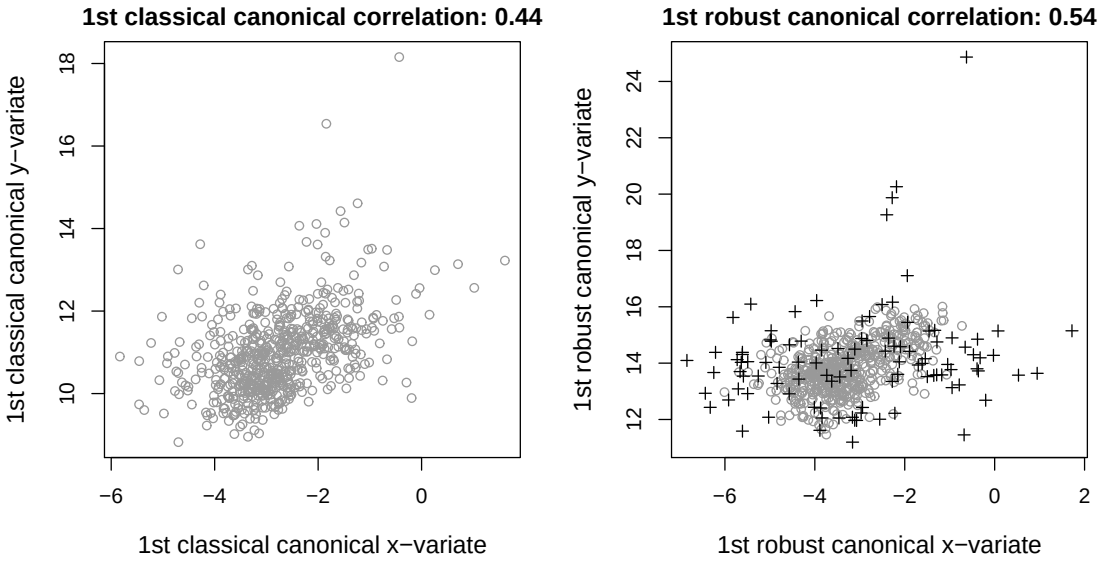

Figure 1: Data projected on the first canonical variates: horizontal axes for seaspray S, vertical axes for bioproductivity B; left plot for classical, and right plot for robust canonical correlation analysis. The symbols "+" in the right plot refer to multivariate outliers.

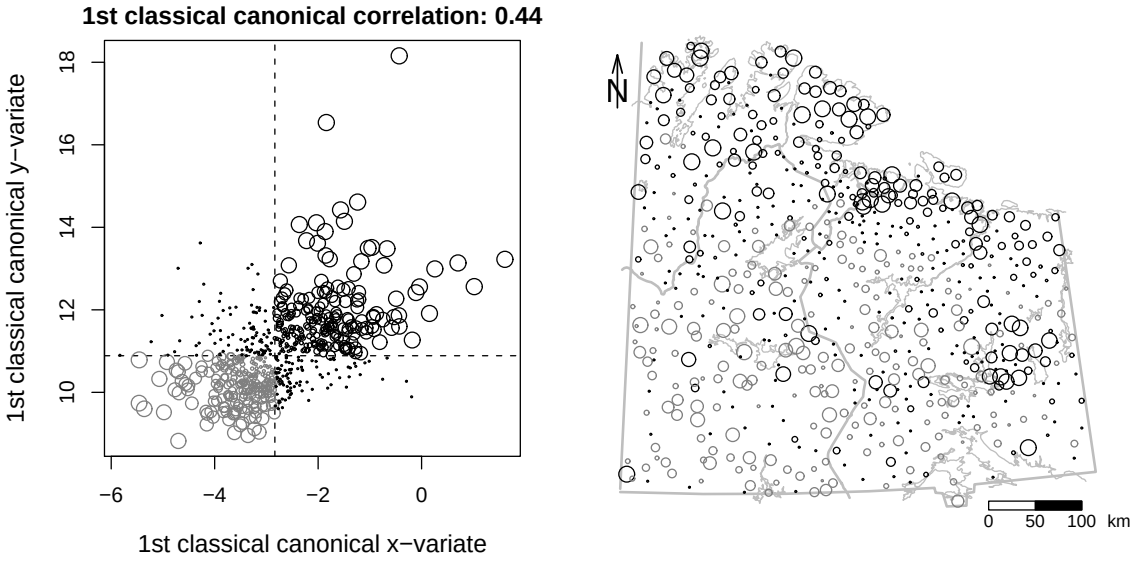

Figure 2: Data projected on the first classical canonical variates shown with special symbols that emphasize the increase of the correlation coefficient (compare Figure 1) (left), and map using the same symbols (right).

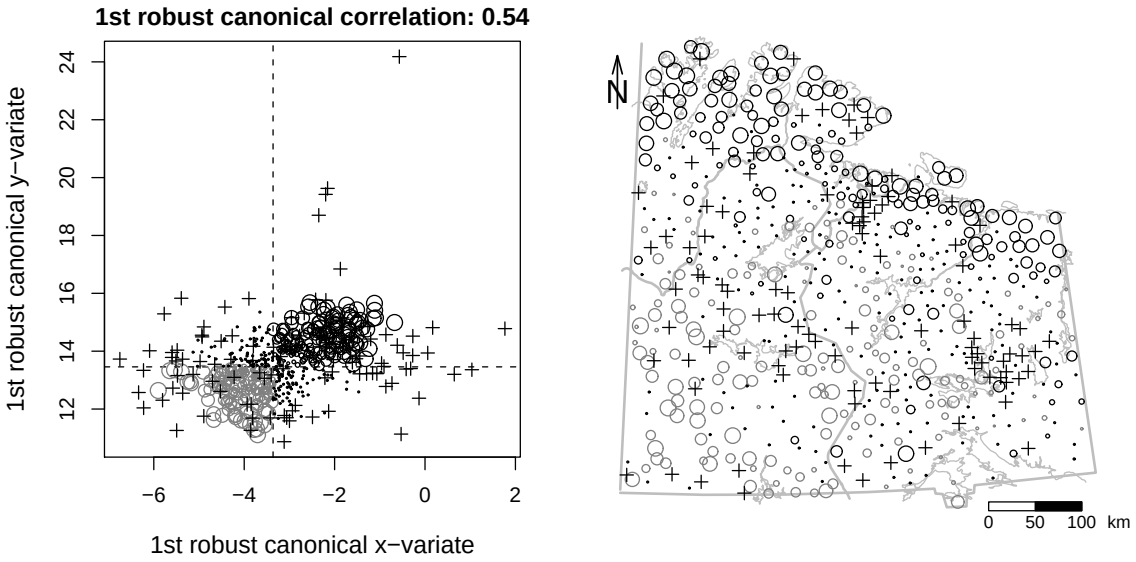

Figure 3: Data projected on the first robust canonical variates shown with special symbols that emphasize the increase of the correlation coefficient (compare Figure 1) (left), and map using the same symbols (right).

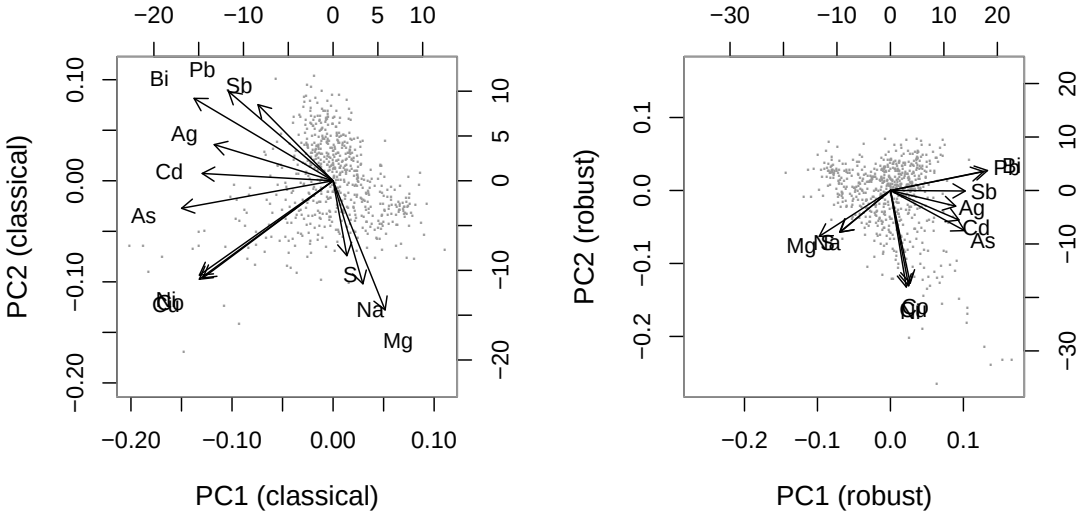

Figure 4: Biplots of the log-transformed Kola O-horizon data. Left: biplot based of classical covariance estimation; right: biplot based on robust covariance estimation.

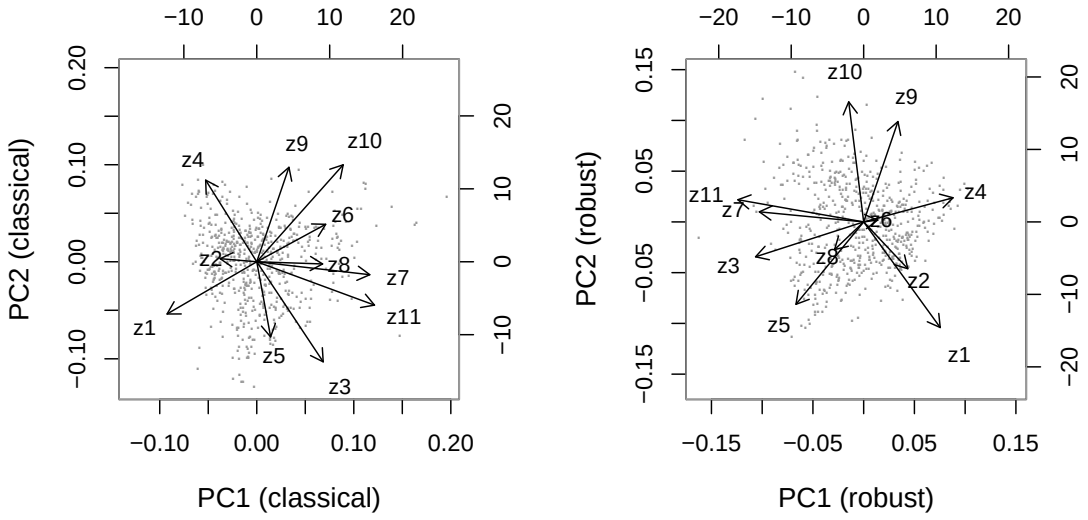

Figure 5: Biplots of the constructed balances for Kola O-horizon data. Left: biplot based of classical covariance estimation; right: biplot based on robust covariance estimation.

Table 1: Group assignments for elements of the Kola O-horizon data

Group	Elements
Pollution (P)	Co, Cu, Ni
Seaspray (S)	Mg, Na, S
Contamination (C)	As, Bi, Cd, Sb
Mineralization (M)	Ag, Pb
Bioproductivity (B)	As, Bi, Cd, Sb, Ag, Pb

Table 2: Sequential binary partitions and resulting balances of the elements of the Kola O-horizon data.

balance	Co	Cu	Ni	Mg	Na	S	As	Bi	Cd	Sb	Ag	Pb
z_1	+	+	−									
z_2	+	−										
z_3				+	+	−						
z_4				+	−							
z_5							+	+	+	+	−	−
z_6							+	+	−	−		
z_7							+	−				
z_8									+	−		
z_9											+	−
z_{10}	+	+	+	−	−	−						
z_{11}	+	+	+	+	+	+	−	−	−	−	−	−

Table 3: Formulas for computing the balances describing the groups of the Kola O-horizon data. These new "variables" will be used for correlation analysis.

balance	z_1	z_2	z_3	z_4
formula	$\frac{\sqrt{2}}{\sqrt{3}} \ln \frac{(x_1 x_2)^{\frac{1}{2}}}{x_3}$	$\frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}$	$\frac{\sqrt{2}}{\sqrt{3}} \ln \frac{(x_4 x_5)^{\frac{1}{2}}}{x_6}$	$\frac{1}{\sqrt{2}} \ln \frac{x_4}{x_5}$
balance	z_5	z_6	z_7	z_8
formula	$\frac{2}{\sqrt{3}} \ln \frac{(x_7 x_8 x_9 x_{10})^{\frac{1}{4}}}{(x_{11} x_{12})^{\frac{1}{2}}}$	$\ln \frac{(x_7 x_8)^{\frac{1}{2}}}{(x_9 x_{10})^{\frac{1}{2}}}$	$\frac{1}{\sqrt{2}} \ln \frac{x_7}{x_8}$	$\frac{1}{\sqrt{2}} \ln \frac{x_9}{x_{10}}$
balance	z_9	z_{10}	z_{11}	
formula	$\frac{1}{\sqrt{2}} \ln \frac{x_{11}}{x_{12}}$	$\frac{\sqrt{3}}{\sqrt{2}} \ln \frac{(x_1 x_2 x_3)^{\frac{1}{3}}}{(x_4 x_5 x_6)^{\frac{1}{3}}}$	$\sqrt{3} \ln \frac{\prod_{i=1}^{6} x_i^{\frac{1}{6}}}{\prod_{j=7}^{12} x_j^{\frac{1}{6}}}$	

Table 4: Group correlations, and first and second canonical correlation coefficients for the groups of the Kola O-horizon data; robust correlations are provided in the lower left parts, and classical correlations in the upper right parts.

	Group corr.			1st can. corr.			2nd can. corr.		
Group	P	S	B	P	S	B	P	S	B
P	–	0.24	0.59	–	0.24	0.50	–	0.05	0.37
S	0.32	–	0.46	0.34	–	0.44	0.18	–	0.14
B	0.67	0.56	–	0.55	0.54	–	0.48	0.24	–