**Covariance-based variable selection for compositional data**[1]

by Karel Hron[23], Peter Filzmoser[4], Sandra Donevska[23] and Eva Fišerová[23]

[1] Received .....; accepted .....

[2] Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, 771 46 Olomouc, Czech Republic; e-mail: hronk@seznam.cz, sdonevska@seznam.cz, eva.fiserova@upol.cz

[3] Department of Geoinformatics, Faculty of Science, Palacký University, tř. Svobody 26, 771 46 Olomouc, Czech Republic

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

Corresponding author:
Karel Hron
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
17. listopadu 12, 771 46 Olomouc, Czech Republic
Phone +420 585 634 605
e-mail: hronk@seznam.cz

**Abstract**

Omitting variables in compositional data analysis may lead to a substantial change of the results of a multivariate statistical analysis. In particular, this is the case for principal component analysis and the compositional biplot, where both the interpretation of loadings and scores of the remaining subcomposition is affected. A stepwise procedure is introduced that allows for a reduction of the original composition to a subcomposition by avoiding a substantial change of the information, like those carried by the compositional biplot. The subcomposition is easier to handle and interpret. Numerical results give evidence of the usefulness of the procedure.

KEY WORDS: Aitchison geometry on the simplex; Centered log-ratio transformation; Isometric log-ratio transformation; Variable selection.

# 1 Introduction

Compositional data occur frequently in geosciences. They are defined as quantitative descriptions of parts of some whole, thus as data carrying only relative information (Aitchison, 1986; Egozcue, 2009). As a consequence, only ratios between the parts of a composition are informative, but not the single measurements. This data type needs a different treatment in a statistical analysis, because the sample space of $D$-part compositional data is the simplex,

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

with a special geometry, nowadays called the Aitchison geometry, with the Euclidean space structure. Since only ratios between the parts are of interest, the constant sum $\kappa$ can be chosen arbitrarily and it leads just to a proper representation of compositions (as proportions or percentages, for example).

Most statistical methods are designed for the usual Euclidean geometry, and thus compositional data first need to be transformed from the simplex to the real space, using so called log-ratio transformations. In the following, we will consider the centered log-ratio (clr) and the isometric log-ratio (ilr) transformations, both of them carrying the property of isometry (Aitchison, 1986; Egozcue et al., 2003). The clr transformation is an isometric mapping between $\mathcal{S}^D$ and a hyperplane of $\mathbb{R}^D$,

$$\mathbf{y} = clr(\mathbf{x}) = (y_1, y_2, ..., y_D)' = \left( \ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, ..., \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)' . \qquad (1)$$

Due to their construction, the clr variables lead to collinear data, because $y_1 + \cdots + y_D = 0$. This has consequences if the statistical methods require data with full rank, e.g. for computing an inverse covariance matrix. Despite this problem, clr variables are still frequently in use because of an intuitive interpretation. A prominent example is the compositional biplot (Aitchison and Greenacre, 2002), which is constructed with the clr variables. Here, the single clr variables are usually interpreted in terms of the original compositional parts (Filzmoser et al., 2009; Tolosana-Delgado et al., 2005).

Another important use of clr variables is due to the relation to a special choice of the ilr transformation, as it will be outlined in the following. The ilr transformation represents an isometric mapping from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$, and it has an additional advantageous feature: it represents compositions in coordinates of an orthonormal basis on the simplex. In contrast, the clr transformation results in coordinates with respect to a generating system (note that the dimension of the simplex equals $D-1$). Consequently, the resulting ilr data matrix has full rank and possible numerical problems are avoided. There are several possibilities how to construct the ilr coordinates. A reasonable choice is to use sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005), where in each of $D-1$ steps of the procedure the compositional parts are split into two non-overlapping groups; the resulting $D-1$ ilr variables represent balances between these groups. An alternative interpretation of ilr coordinates (named in this context also balances) is based on a decomposition of their covariance structure (Fišerová and Hron, 2011): each balance explains ratios (log-ratios) between compositional parts in both groups that arise from the corresponding step of a sequential binary partition. As a consequence, the resulting balances form a unique representation of all log-ratios in the composition. Even more, by a proper choice of the balances, like proposed below, it is possible to proceed from the full composition to a subcomposition, i.e. to assign directly coordinates to the subcomposition without the necessity of forming a new orthonormal basis. Note that this is not possible with clr transformed data.

As a special case, such an orthonormal basis can be constructed where the first balance explains all information concerning the log-ratios about a component $x_l$, $l = 1, \ldots, D$, of the original composition (the mentioned balance thus "represents" the part $x_l$). Consequently, the remaining $D-2$ balances can be chosen arbitrarily in the sequential binary partition, so that they explain information concerning the remaining log-ratios in the composition. Such a basis is frequently useful in applications as well as in theoretical considerations (Egozcue et al., 2003; Fišerová and Hron, 2011; Filzmoser et al., 2012; Hron et al., 2010). This special ilr transfor-

mation for a chosen part $x_l$ is defined as

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}}\ \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}},\ i = 1, \ldots, D-1,\ l = 1, \ldots, D. \qquad (2)$$

Here, $(x_1^{(l)}, x_2^{(l)}, ..., x_l^{(l)}, x_{l+1}^{(l)}, ..., x_D^{(l)})'$ stands for the re-ordered composition with the part $x_l$ in the first position, $(x_l, x_1, ..., x_{l-1}, x_{l+1}, ..., x_D)'$, in order to keep the above interpretation of the variable $z_1^{(l)}$. Unfortunately, there is no way in general to find a similar interpretation for more than one part simultaneously using this construction of an orthonormal basis.

There is a linear relation between the first coordinate of $D$ bases corresponding to (2) and the clr variables (Egozcue and Pawlowsky-Glahn, 2006),

$$y_i = \sqrt{\frac{D-1}{D}} z_1^{(i)},\ i = 1, \ldots, D. \qquad (3)$$

Consequently, the clr variables can be interpreted like $z_1^{(l)}$, $l = 1, \ldots, D$, the values differ by a constant. This relation has advantages from a computational point of view, when the single variables $z_1^{(l)}$ are of interest.

The above relation of the clr transformation with this special choice of the ilr transformation is a main motivation to investigate the change of the multivariate data structure when moving from the full composition to a subcomposition. The goal of the paper is to derive a stepwise procedure to obtain such a subcomposition of the original composition, where the effect of the change of the information (concerning the resulting subcomposition) is rather negligible. In this way it is possible to use a subset of variables that does not lead to substantial changes when a statistical analysis (like the compositional biplot of clr transformed compositions) is used to investigate the multivariate compositional data structure. The subset usually allows for a simplified interpretation of a statistical analysis.

In the next section we investigate the change of the variances of clr variables when coming from the original composition to a subcomposition. The theoretical considerations on clr variables will be used in Section 3 to construct a covariance-based stepwise procedure to recognize the main contributing parts in a composition,

i.e., parts, whose log-ratios to the other parts in the composition contain most of the total variability of a compositional data set. In other words, we search for such a subcomposition of the original composition, where the changes of a multivariate analysis (with respect to the original composition) are rather negligible. The procedure is completed by a stepwise testing whether there is a significant loss of the total variability when reducing the composition from the previous step by one part. Section 4 contains a real world example from the field of geochemistry where the properties of the proposed stepwise procedure are demonstrated. The final section comments on other possible ways to variable selection in compositional data.

## 2 Covariance structure and coordinates

The basic measure of variability of a random composition $\mathbf{x} = (x_1, \ldots, x_D)'$ is the variation matrix (Aitchison, 1986), defined as

$$\mathbf{T} = \left\{ \mathrm{var}\left( \ln \frac{x_i}{x_j} \right) \right\}_{i,j=1}^{D}.$$

The elements of the variation matrix describe the variability of the random log-ratio $\ln \frac{x_i}{x_j}$: the smaller the value of this variance, the more the log-ratio tends to be a constant. The (normed) sum of the elements of the variation matrix is called total variance,

$$\mathrm{totvar}(\mathbf{x}) = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \mathrm{var}\left( \ln \frac{x_i}{x_j} \right),$$

expressing the total variability of the compositional data set. Note that

$$\mathrm{totvar}(\mathbf{x}) = \sum_{i=1}^{D} \mathrm{var}(y_i) = \sum_{i=1}^{D-1} \mathrm{var}(z_i^{(l)}),\ l = 1, \ldots, D,$$

i.e. the total variance can also be computed using the variability of the clr variables or the ilr coordinates, respectively (Pawlowsky-Glahn and Egozcue, 2001).

Fišerová and Hron (2011) mentioned that it is possible to express the covariance structure of balances using linear combinations of variances of log-ratios. Taking

the balances $z_i = z_i^{(1)}$ from the previous section,

$$z_i = \sqrt{\frac{D-i}{D-i+1}}\, \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}},\ i = 1, \ldots, D-1, \tag{4}$$

we obtain (Fišerová and Hron, 2011)

$$\mathrm{var}(z_i) = \frac{1}{D-i+1} \sum_{p=i+1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_p}\right) - \frac{1}{2(D-i)(D-i+1)} \sum_{p=i+1}^{D} \sum_{q=i+1}^{D} \mathrm{var}\left(\ln \frac{x_p}{x_q}\right). \tag{5}$$

As a consequence of (3), the variance of the clr variable $y_i$ corresponds (up to a constant) to the variance of $z_1^{(i)}$. Also the covariance structure of the clr variables can be analyzed, which has been done thoroughly in Aitchison (1986).

From (5), multiplied by $(D-1)/D$ to obtain clr variances, we can also expect quite a strong relation between $\mathrm{var}(y_i)$ and the sum of the $i$-th row (column) of the corresponding variation matrix $\mathbf{T}$. This finding induces a useful property, mentioned in the next theorem. Particularly, it shows that ordered variances of different clr variables (or, alternatively, of the first ilr coordinates from (2)) correspond to the same order of the sums in the variation matrix connected with the related compositional parts.

**Theorem 1:** *Consider the clr variables $y_i$ and $y_j$, $i \neq j$, $i, j \in \{1, \ldots, D\}$ (or, equivalently, balances $z_1^{(i)}$ and $z_1^{(j)}$ from (2), corresponding to two different orthonormal bases). Then* $\mathrm{var}(y_i) \geq \mathrm{var}(y_j)$, *if and only if*

$$\sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_p}\right) \geq \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_p}\right).$$

The proof is in the Appendix. This theorem can be used to identify compositional parts ("markers") that are responsible for larger clr variances. In particular, it allows to detect possible sources of changes in the multivariate analysis of compositional data, like those resulting from the compositional biplot (Aitchison and Greenacre,

2002). Theorem 1 makes it possible to identify the ordered contribution of the single compositional parts to the overall variance with the corresponding clr variables. Using this fact, a stepwise algorithm is introduced in the following that helps to derive a subcomposition with a minimal loss concerning the total variance of the original composition.

## 3 Proposed stepwise procedure

Let us consider a composition $\mathbf{x} = (x_1, \ldots, x_D)'$. Without loss of generality, let

$$\operatorname{var}(y_1) \geq \cdots \geq \operatorname{var}(y_D), \tag{6}$$

which is, according to Theorem 1, equivalent to

$$\sum_{p=1}^{D} \operatorname{var}\left(\ln \frac{x_1}{x_p}\right) \geq \sum_{p=1}^{D} \operatorname{var}\left(\ln \frac{x_2}{x_p}\right) \geq \cdots \geq \sum_{p=1}^{D} \operatorname{var}\left(\ln \frac{x_D}{x_p}\right).$$

Since $y_D$ has the smallest variance, its contribution to the overall variance of the compositional data set, totvar($\mathbf{x}$), is minimal. This is equivalent to the statement that the aggregated variances of the log-ratios with the part $x_D$ have the smallest contribution to the overall variance. Consequently, the part $x_D$ is not determining the multivariate data structure and it can be omitted from the composition. Hence, we arrive at a subcomposition $\mathbf{x}_1 = (x_1, \ldots, x_{D-1})'$. In the next step we perform a clr transformation on $\mathbf{x}_1$, calculate variances of the clr transformed variables and again omit the part corresponding to the clr variable with the smallest variance. So we continue until a certain number of parts is obtained, and we stop at latest after $D - 2$ steps.

The order of the variances of the clr variables is generally not maintained after omitting the part of composition $\mathbf{x}$ corresponding to the clr variable with the smallest variance. In fact, as a simple consequence of Theorem 1, the order of the clr variances when moving from a $D$-part to a $(D-1)$-part composition is maintained only under the assumption

$$\operatorname{var}\left(\ln \frac{x_1}{x_D}\right) \geq \operatorname{var}\left(\ln \frac{x_2}{x_D}\right) \geq \cdots \geq \operatorname{var}\left(\ln \frac{x_{D-1}}{x_D}\right).$$

Nevertheless, from simulations using real geochemical data (see next section) it follows that the ordering of the clr variables of the original composition according to their variances is a relatively accurate indicator whether the corresponding part of the original composition will be included in the final subcomposition or not.

The prescribed number of parts of the target subcomposition is usually not provided. Thus, an important question is when the selection of compositional parts should be stopped. It is easy to see that the main idea of the above algorithm is to select a subcomposition such that the loss in total variance of the composition from the previous step is minimal. This inspires to find such a criterion that compares the total variance of the subcomposition, obtained in the $i$-th step of the algorithm, $i = 1, \ldots, D-2$, with the total variance of the composition from the previous step. In more detail, denote $\widehat{\text{totvar}}(\mathbf{x}_i)$ the total variance of the $i$-th subcomposition, estimated from the data. We want to test whether its difference to the total variance $\text{totvar}(\mathbf{x}_{i-1})$ can be considered as negligible (i.e., the null hypothesis is $\text{totvar}(\mathbf{x}_i) = \text{totvar}(\mathbf{x}_{i-1})$) or rather as a result of a systematic pattern (alternative hypothesis $\text{totvar}(\mathbf{x}_i) < \text{totvar}(\mathbf{x}_{i-1})$). Obviously, $\text{totvar}(\mathbf{x}_{i-1})$ is not known and also needs to be estimated from the data (in the previous step of the algorithm), as it is the case with $\widehat{\text{totvar}}(\mathbf{x}_i)$. Here we assume that $\text{totvar}(\mathbf{x}_{i-1})$ is fixed from the previous step of the algorithm, and thus it can be considered as a given (non-random) number in the current step. The following test statistic from Hron and Kubáček (2011),

$$U_i^+ = \frac{\widehat{\text{totvar}}(\mathbf{x}_i) - \text{totvar}(\mathbf{x}_{i-1})}{\sqrt{\frac{2}{n-1}\text{tr}(\widehat{\mathbf{\Sigma}}_i^2)}}, \tag{7}$$

is used for this purpose; the matrix $\widehat{\mathbf{\Sigma}}_i$ stands for the sample covariance matrix of the composition $\mathbf{x}_i$ in (arbitrarily chosen) ilr coordinates. Small values of $U_i^+$ favor the alternative, so we reject the null hypothesis, if $U_i^+$ realizes in the critical region $\mathsf{W} = (-\infty, u_\alpha\rangle$, where $u_\alpha$ denotes the $\alpha$-quantile (preferably $\alpha = 0{,}05$) of the standard normal distribution (being inspired by the asymptotic distribution of $U_i^+$, see Hron and Kubáček (2011) for details). Thus, in each step of the algorithm we compute the statistic $U_i^+$, and the procedure is stopped when $U_i^+$ realizes for the first time in $\mathsf{W}$.

Practical properties of the proposed iterative procedure will be demonstrated on real-world examples in the next section.

# 4 Illustrative examples

The behavior of the proposed stepwise algorithm is demonstrated at the well-known Kola data (Reimann et al., 1998). This data set is the result of a large geochemical mapping project, carried out from 1992 to 1998 by the Geological Surveys of Finland and Norway, and the Central Kola Expedition, Russia. An area covering 188 000 $km^2$ at the peninsula Kola in northern Europe was sampled. In total, around 600 samples of soil were taken in 4 different layers (moss, humus, B-horizon, C-horizon), and subsequently analyzed by a number of different techniques for more than 50 chemical elements. The project was primarily designed to reveal the environmental conditions in the area. The data are available in the package `StatDA` of the software environment R (R Development Core Team, 2012).

In the first experiment we are interested in observing the reduction in total variance by the stepwise procedure. For this purpose we use all the 31 elements of the moss layer and select randomly 15 variables. Then the stepwise algorithm is applied until a two-part subcomposition is reached (i.e. here we are not using the proposed stopping criterion). After each step the total variance is computed. The whole procedure is repeated 1000 times, and the results are shown in Figure 1. Each boxplot summarizes the total variances achieved by the given size of the subcomposition. A decreasing sequence of the total variance (and its variability among subcompositions of the given size, see whiskers of the boxplots) is clearly visible. Subsequent steps of the algorithm result in increasing relative differences of the median total variances. This is important for obtaining significance at a certain step using the proposed test statistic. Note that this feature as well as the below mentioned properties are characteristic for the stepwise procedure even in general, independent on the concrete data set chosen.

In a next experiment we want to check if the test statistic is able to select appropriate compositional parts. For this we again select randomly 15 parts of the

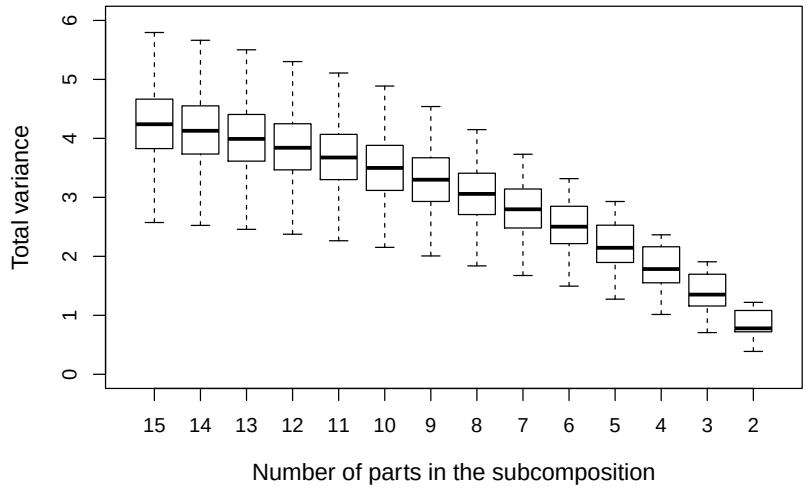


Figure 1: Total variances of subcompositions obtained from the stepwise algorithm.

Kola moss data as a starting composition. Then the stepwise procedure is applied until the test statistic suggests to stop the process. Repeating this experiment 1000 times results in a distribution of the number of remaining parts, which is visualized by a barplot in Figure 2 (left). The algorithm arrives typically at subcompositions with 10 to 12 parts, i.e. around two thirds of the starting number of parts. The important question is whether the resulting target compositions are indeed consisting of parts with large clr variances of the initial compositions. Therefore we sort the parts of all 1000 initial subcompositions according to decreasing values of their clr variances, and count how often the top $k$ clr variables were included in the target compositions, where $k = 1, \ldots, 15$. Figure 2 (right) shows the result. The counts have to decrease for larger values of $k$ because of the possibly smaller total number of parts in the target compositions, see Figure 2 (left). We can see that the initial clr variable with largest variance ($k = 1$) was selected in all 1000 cases. This also holds for $k$ up to 5, i.e., the top 5 clr variables were always selected. Then

the counts drop, partly because the resulting subcompositions were smaller than $k$, and partly because not all considered $k$ clr variables were selected. The figure, however, clearly indicates that the important clr variables were included in the target subcompositions, although Theorem 1 provides here no theoretical guarantee.

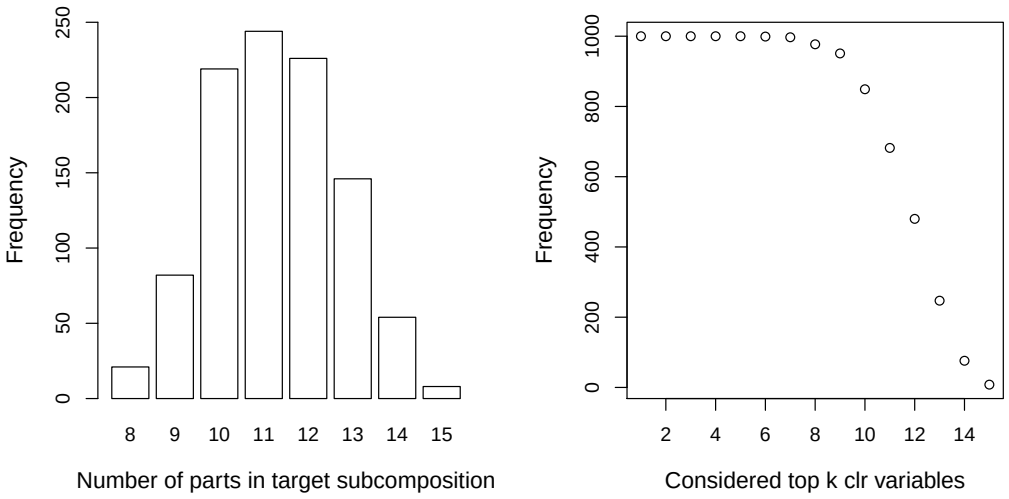


Figure 2: Barplot of the number of parts of the subcomposition resulting from the stepwise procedure using the stopping-criterion (left); clr variables of the initial composition, sorted according to decreasing variance, versus number of times the corresponding compositional parts were included in the resulting subcomposition (right).

In a further simulation experiment we analyze the behavior of the stepwise procedure for different sizes of the starting composition. We use the same simulation setting as before, but select as initial composition 5, 10, 20, and 25 parts of the Kola moss data, respectively. For each case 1000 simulations are performed, and the distributions of the resulting numbers of parts in the target compositions are shown in Figure 3. If the starting composition has only $D = 5$ parts (upper left), the procedure usually arrives at a target composition again with 5 parts. On the other hand, if one starts with $D = 25$ parts (lower right), the number of parts will

be reduced to about a half. This behavior of shrinking larger compositions more and more is very desirable for practice.

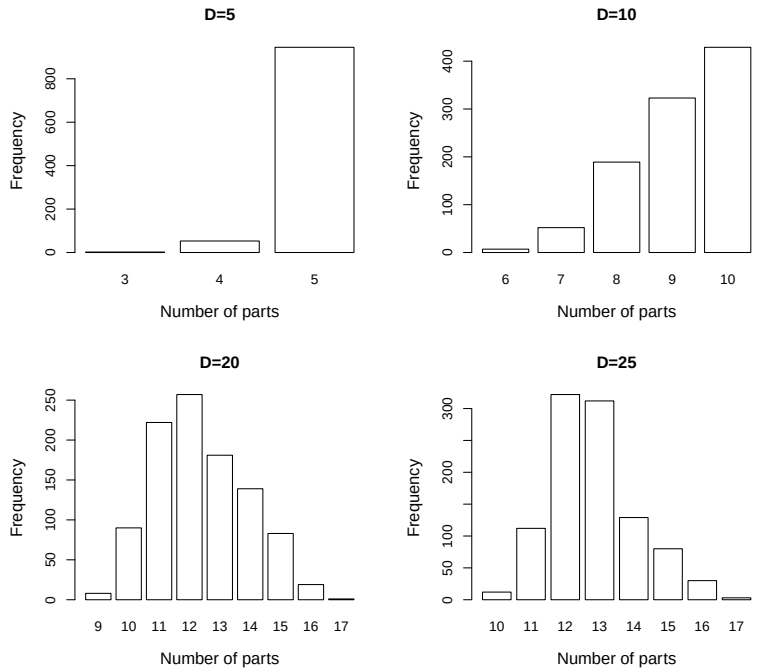


Figure 3: Barplots of the number of parts of the subcomposition resulting from the stepwise procedure using the stop-criterion with 5-part (upper left), 10-part (upper right), 20-part (lower left) and 25-part (lower right) original compositions.

In the next illustration we apply the stepwise procedure to the whole moss layer data set consisting of 31 compositional parts. The total variances of the resulting subcompositions are plotted in Figure 4. They quite nicely correspond to the trend as indicated in Figure 1 (left). The right picture shows the values of the test statistic $U_i^+$, for each of the $i = 1, \ldots, 29$ possible steps, and they reflect the same trend. The algorithm will stop after the value of the test statistic falls below $u_{0.05} = -1.64$ (horizontal line). This happens at step 19 of the algorithm, and thus a 13-part subcompositions is remaining.

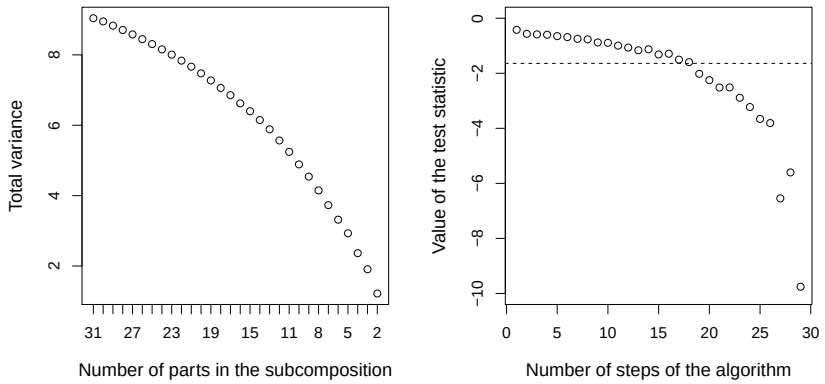


Figure 4: Total variances of subcompositions obtained from the stepwise algorithm for the whole moss layer data set (left), corresponding values of the test statistic $U_i^+$ together with the cut-off value (right).

Finally, we demonstrate the effect of the stepwise procedure using the compositional biplot (Aitchison and Greenacre, 2002) as a visualization tool. Here we employ the Baltic Soil Survey (BSS) data (Reimann et al., 2003), which originate from a large-scale geochemistry project carried out in northern Europe, in an area of about $1\,800\,000\,\mathrm{km}^2$. 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`. Here 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) of the top layer, i.e. an 11-part composition. Note that the same elements were used also in Filzmoser et al. (2009), where classical and robust biplots of both log- and clr-transformed compositions were compared.

Figure 5 (left) shows the classical compositional biplot of the initial 11-part composition. If we apply the stepwise procedure, we arrive at a 9-part subcomposition.

The elements $Al_2O_3$ and $Fe_2O_3$ were subsequently excluded with the corresponding values of the test statistics $U_1^+ = -0.7185$ and $U_2^+ = -1.4753$, respectively. The next step with an exclusion of $TiO_2$ would already lead to significance with a value of $U_3^+ = -2.2712$. The resulting biplot (Figure 5, right) shows that there is nearly no difference visible in the relations among the remaining compositional parts (arrows in the biplot) compared to the original biplot. Thus, the multivariate data structure is widely preserved and the information of the excluded elements is still contained in the remaining subcomposition.

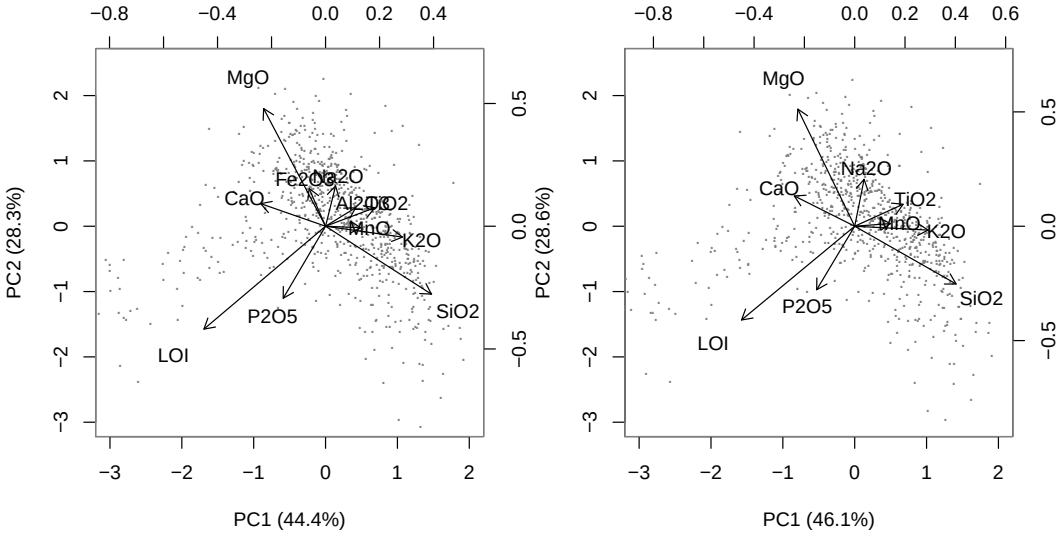


Figure 5: Biplots of the BSS data with all major elements (left) and after exclusion of $Al_2O_3$ and $Fe_2O_3$ (right).

## 5 Discussion

Practitioners are often interested in reducing the number of compositional parts for the statistical analysis, because this simplifies the analysis, and also the interpretation of the results is simplified. An intuitive selection of parts based on expert knowledge of subject matter specialists may lead to major changes of the multivari-

ate statistical analysis results. For example, experts may be interested in analyzing certain geochemical processes, and select elements for the statistical analysis which are somehow related to these processes. In this selection they may miss variables that are responsible for substantial information to the multivariate information, and their omission changes the statement about the resulting subcomposition. Note that the presented approach differs from the known problem of subcompositional incoherence of the original composition with a (prescribed) unit-sum representation of the compositional vector (Chayes, 1960; Aitchison, 1986), that is against the general definition of compositional data as multivariate observations where the only relevant information is contained in the ratios between the parts (Egozcue, 2009).

A stepwise procedure for excluding compositional parts allows to arrive at a subcomposition that still retains the important information contained in the multivariate data structure. The goal of this procedure is to retain the total variance from one step to the next, and it is stopped before a significant reduction would occur. The larger the original composition, the more reduction of the number of parts is made. Examples have demonstrated that indeed those "marker" variables are selected, and an omission of these variables would have resulted in substantial changes of multivariate statistical analyses of compositional data.

**Acknowledgments:** The authors are grateful to the referee for helpful comments and suggestions. The authors also gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic) and the Grant No. PrF-2012-017 of the Internal Grant Agency of the Palacký University in Olomouc.

## References

[1] Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London.

[2] Aitchison J, Greenacre M (2002) Biplots of compositional data. Applied Statistics 51:375–392.

[3] Chayes F (1960) On correlation between variables of constant sum. Journal of Geophysical Research 65, 4185–4193.

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

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

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

[7] Egozcue JJ (2009) Reply to "On the Harker Variation Diagrams; ..." by J. A. Cortés. Math Geosci 41:829–834.

[8] Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics 20:621–632.

[9] Filzmoser P, Hron K, Reimann C (2012) Interpretation of multivariate outliers for compositional data. Computers & Geosciences 39: 77–85.

[10] Fišerová E, Hron K (2011) On interpretation of orthonormal coordinates for compositional data. Math Geosci 43:455–468.

[11] Hron K, Kubáček L (2011) Statistical properties of the total variation estimator for compositional data. Metrika 74: 221–230.

[12] Hron K, Templ M, Filzmoser P (2010) Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis 54:3095–3107.

[13] Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment 15:384–398.

[14] R development core team (2012) R: A language and environment for statistical computing. Vienna, `http://www.r-project.org`.

[15] Reimann C, Äyräs M, Chekushin V, Bogatyrev I, Boyd R, Caritat P, 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.

[16] 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.

[17] Tolosana-Delgado R, Otero N, Pawlowsky-Glahn V, Soler A (2005) Latent compositional factors in the Llobregat river basin (Spain) hydrogeochemistry. Math Geol 37:681-702.

# Appendix

**Proof of Theorem 1:**

Let $\mathrm{var}(y_i) \geq \mathrm{var}(y_j)$, $i, j = 1, \ldots, D$. According to (5), this is equivalent to

$$\frac{D-1}{D^2} \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_i}{x_p}\right) - \frac{1}{2D^2} \sum_{\substack{p=1 \\ p \neq i}}^{D} \sum_{\substack{s=1 \\ s \neq i}}^{D} \mathrm{var}\left(\ln \frac{x_p}{x_s}\right) \geq$$

$$\frac{D-1}{D^2} \sum_{p=1}^{D} \mathrm{var}\left(\ln \frac{x_j}{x_p}\right) - \frac{1}{2D^2} \sum_{\substack{p=1 \\ p \neq j}}^{D} \sum_{\substack{s=1 \\ s \neq j}}^{D} \mathrm{var}\left(\ln \frac{x_p}{x_s}\right).$$

Extending the left-hand side of this inequality by the term $\pm(1/D^2)\sum_{p=1}^{D}\mathrm{var}\left(\ln(x_i/x_p)\right)$ and using the relationship $\mathrm{var}\left(\ln(x_p/x_s)\right) = \mathrm{var}\left(\ln(x_s/x_p)\right)$, the left-hand side can be rewritten in the form

$$\frac{1}{D}\sum_{p=1}^{D}\mathrm{var}\left(\ln\frac{x_i}{x_p}\right) - \frac{1}{2D^2}\sum_{p=1}^{D}\sum_{s=1}^{D}\mathrm{var}\left(\ln\frac{x_p}{x_s}\right).$$

Similarly we can adjust the right-hand side of the inequality, and thus $\mathrm{var}(y_i) \geq \mathrm{var}(y_j)$ if and only if

$$\sum_{p=1}^{D}\mathrm{var}\left(\ln\frac{x_i}{x_p}\right) \geq \sum_{p=1}^{D}\mathrm{var}\left(\ln\frac{x_j}{x_p}\right).$$

□

**Figure captions:**

Figure 1: Total variances of subcompositions obtained from the stepwise algorithm.

Figure 2: Barplot of the number of parts of the subcomposition resulting from the stepwise procedure using the stop-criterion (left); clr variables of the initial composition, sorted according to decreasing variance, versus number of times the corresponding compositional parts were included in the resulting subcomposition (right).

Figure 3: Barplots of the number of parts of the subcomposition resulting from the stepwise procedure using the stop-criterion with 5-part (upper left), 10-part (upper right), 20-part (lower left) and 25-part (lower right) original compositions.

Figure 4: Total variances of subcompositions obtained from the stepwise algorithm for the whole moss layer data set (left), corresponding values of the test statistic $U_i^+$ together with the cut-off value (right).

Figure 5: Biplots of the BSS data with all major elements (left) and after exclusion of $Al_2O_3$ and $Fe_2O_3$ (right).