Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **52**, 2 (2013) 61–69

Covariance Structure of Principal Components for Three-Part Compositional Data*

Klára HRŮZOVÁ [1a], Karel HRON [2a], Miroslav RYPKA [1b],
Eva FIŠEROVÁ [2b]

[1] *Department of Geoinformatics, Faculty of Science, Palacký University*
17. listopadu 50, 779 Olomouc, Czech Republic
[a] *e-mail: klara.hruzova@gmail.com*
[b] *e-mail: miroslav.rypka01@upol.cz*

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

(Received June 21, 2013)

Abstract

Statistical analysis of compositional data, multivariate observations carrying only relative information (proportions, percentages), should be performed only in orthonormal coordinates with respect to the Aitchison geometry on the simplex. In case of three-part compositions it is possible to decompose the covariance structure of the well-known principal components using variances of log-ratios of the original parts. They seem to be helpful for the interpretation of these special orthonormal coordinates. Theoretical results are applied to real-world data containing relative structure of landscape use in German regions.

Key words: compositional data, covariance structure, principal components, log-contrasts

2010 Mathematics Subject Classification: 15A18, 62H25, 62H99, 62J10

1 Introduction

Compositional data as multivariate observations carrying only relative information (specially percentages, proportions, etc.) frequently occur in practice

*Supported by the ESF project Operational Program Education for Competitiveness, CZ.1.07/2.3.00/20.0170.

[1, 8]. The simplex $\mathcal{S}^D$, the sample space of their representations with a prescribed constant sum constraint κ, can be expressed for D-part compositional data as

$$\mathcal{S}^D = \{\mathbf{x} = (x_1, \ldots, x_D)', \, x_i > 0, \, x_1 + \cdots + x_D = \kappa\}.$$

Compositional data naturally follow the so called Aitchison geometry with Euclidean vector space structure (of dimension $D-1$). The standard statistical methods cannot be applied to compositional data until they are expressed in coordinates with respect to an orthonormal basis on the simplex (in the Aitchison geometry sense). The corresponding mapping is called the isometric logratio (ilr) transformation [2] due to isometry property between the Aitchison geometry on the simplex and the real space associated with the Euclidean geometry.

Special case of compositional data is represented by three-part compositions, $\mathbf{x} = (x_1, x_2, x_3)'$ that are of interest since of the possibility to represent them in ternary diagrams graphically. Consequently, there are three possible choices of the isometric logratio transformation according to [3] (up to orientation of coordinates), which differ only in permutation of the parts x_1, x_2, x_3,

$$z_{11} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \quad z_{12} = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}, \tag{1.1}$$

$$z_{21} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_2}{\sqrt{x_1 x_3}}, \quad z_{22} = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_3}, \tag{1.2}$$

$$z_{31} = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_3}{\sqrt{x_1 x_2}}, \quad z_{32} = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}. \tag{1.3}$$

The interpretation of orthonormal coordinates can be obtained from their covariance structure, expressed using variances of log-ratios [5, 6]. In case of (1.1), the variances of z_{11} and z_{12} are given by

$$\begin{aligned} \operatorname{var}(z_{11}) &= \frac{1}{3} \operatorname{var}\left(\ln \frac{x_1}{x_2}\right) + \frac{1}{3} \operatorname{var}\left(\ln \frac{x_1}{x_3}\right) - \frac{1}{6} \operatorname{var}\left(\ln \frac{x_2}{x_3}\right), \\ \operatorname{var}(z_{12}) &= \frac{1}{2} \operatorname{var}\left(\ln \frac{x_2}{x_3}\right). \end{aligned} \tag{1.4}$$

Thus the first coordinate captures all relative information about the first compositional part (expressed by log-ratios between x_1, x_2 and x_1, x_3, respectively). The second coordinate stands for the remaining log-ratio between x_2, x_3. The variance of z_{11} consists of variances of the first two mentioned log-ratios including x_1 in the nominator and it is reduced by the variance of log-ratio of remaining two compositional parts. This is a consequence of the fact that each ilr variable forms a log-contrast, i.e. term of the form $\mathbf{h}' \ln \mathbf{x}$, where $\mathbf{h}'\mathbf{1} = h_1 + h_2 + h_3 = 0$. Furthermore, the total variance, which represents the sum of variances of both coordinates, results in

$$\operatorname{totvar}(\mathbf{x}) = \operatorname{var}(z_{11}) + \operatorname{var}(z_{12}) = \frac{1}{3}\left[\operatorname{var}\left(\ln \frac{x_1}{x_2}\right) + \operatorname{var}\left(\ln \frac{x_1}{x_3}\right) + \operatorname{var}\left(\ln \frac{x_2}{x_3}\right)\right]. \tag{1.5}$$

Analogous relations would be obtained also for (1.2) and (1.3) by permutation of parts of the original composition.

The main goal of this paper is to analyze the variance structure of the well-known principal components as a popular tool for dimension reduction and its impact to interpretation of these orthonormal coordinates. Note that principal components are obtained from such rotation of the original variables that maximizes variance of the resulting coordinates. Although in case of standard real data the covariance structure of principal components can be also expressed using elements of the original covariance matrix [11], we will follow an alternative way of its derivation that enables a deeper insight into covariance structure of three-part compositional data.

2 Variance structure of principal components

Principal component analysis represents a popular tool for dimension reduction in multivariate data sets. Consequently, only few new variables (principal components) are able to capture most of the overall variability of the original data. In case of compositional data, principal component analysis needs to be performed in (preferably) orthonormal coordinates. From the theoretical point of view, principal components of a random composition $\mathbf{x} = (x_1, \dots, x_D)'$ are formed by orthogonal rotation of centred ilr coordinates $\mathbf{z} = (z_1, \dots, z_{D-1})'$,

$$\mathbf{z}^* = \mathbf{G}_z'(\mathbf{z} - \mathsf{E}(\mathbf{z})),$$

where $\mathbf{G}_z$ comes from spectral decomposition of the covariance matrix $\mathrm{var}(\mathbf{z}) = \mathbf{G}_z\mathbf{L}\mathbf{G}_z'$ as matrix of eigenvectors (the diagonal matrix $\mathbf{L}$ contains eigenvalues of $\mathrm{var}(\mathbf{z})$). Of course principal component are orthonormal coordinates as well.

At the beginning of this section we introduce a general constrained problem of finding stationary values [7] that will be used consequently to derive the main theorem concerning covariance structure of principal components for three-part compositional data, denoted in the following as z_1^*, z_2^*. Taking the main idea of principal component analysis into account, we search for maximal difference between variances of both variables.

Let $\mathbf{A}$ be a real symmetric matrix of order D, and $\mathbf{c}$ a given real vector that fulfills the condition $\mathbf{c}'\mathbf{c} = 1$. The goal is to find the stationary values of $\mathbf{h}'\mathbf{A}\mathbf{h}$, taking constraints $\mathbf{h}'\mathbf{h} = 1$, $\mathbf{c}'\mathbf{h} = 0$ into account. Denote

$$\varphi(\mathbf{h}, \nu, \mu) = \mathbf{h}'\mathbf{A}\mathbf{h} - \nu(\mathbf{h}'\mathbf{h} - 1) + 2\mu\mathbf{h}'\mathbf{c}, \tag{2.1}$$

where ν, μ are Lagrange multipliers. Differentiating (2.1) with respect to $\mathbf{h}$ leads to

$$\mathbf{A}\mathbf{h} - \nu\mathbf{h} + \mu\mathbf{c} = \mathbf{0}. \tag{2.2}$$

Multiplying (2.2) from left by $\mathbf{c}'$ and using the condition $\mathbf{c}'\mathbf{c} = 1$, we have

$$\mu = -\mathbf{c}'\mathbf{A}\mathbf{h}. \tag{2.3}$$

Then substituting (2.3) into (2.2) we obtain

$$\mathbf{PAh} = \nu\mathbf{h}, \tag{2.4}$$

where $\mathbf{P} = \mathbf{I} - \mathbf{cc}'$. Although $\mathbf{P}$ and $\mathbf{A}$ are symmetric, $\mathbf{PA}$ is not necessarily so. Note that $\mathbf{P}^2 = \mathbf{P}$, so that $\mathbf{P}$ is a projection matrix.

It is well-known that for two arbitrary square matrices $\mathbf{G}$ and $\mathbf{H}$, the eigenvalues of $\mathbf{GH}$ equal the eigenvalues of $\mathbf{HG}$. Thus we can write

$$\lambda(\mathbf{PA}) = \lambda(\mathbf{P}^2\mathbf{A}) = \lambda(\mathbf{PAP}),$$

where λ corresponds to any (fixed) eigenvalue of the matrix in brackets.

The matrix $\mathbf{PAP}$ is symmetric and hence one can use the standard algorithms for finding its eigenvalues. Then if we denote $\mathbf{K} = \mathbf{PAP}$ and if $\mathbf{Kz}_i = \lambda_i\mathbf{z}_i$, it follows that $\mathbf{h}_i = \mathbf{Pz}_i$, where $\mathbf{h}_i$ is the eigenvector which satisfies (2.4) and also the initial problem. At least one eigenvalue of $\mathbf{K}$ will be equal to zero, and $\mathbf{c}$ will be an eigenvector associated with a zero eigenvalue.

The following lemma (see [1, p. 93]) establishes a relation between log-contrasts, corresponding to orthonormal coordinates and their covariance structure.

Lemma 1 *Variances and covariances for log-contrasts* $\mathbf{h}_1' \ln \mathbf{x}$ *and* $\mathbf{h}_2' \ln \mathbf{x}$ *of a D-part composition* $\mathbf{x}$ *are*

$$\operatorname{var}(\mathbf{h}_1' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{h}_1'\mathbf{Th}_1, \quad \operatorname{var}(\mathbf{h}_2' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{h}_2'\mathbf{Th}_2, \tag{2.5}$$

$$\operatorname{cov}(\mathbf{h}_1' \ln \mathbf{x}, \mathbf{h}_2' \ln \mathbf{x}) = -\frac{1}{2}\mathbf{h}_1'\mathbf{Th}_2, \tag{2.6}$$

where $\mathbf{T}$ *is the variation matrix defined by*

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

Theorem 1 *The covariance structure of principal components (orthonormal coordinates)* z_1^*, z_2^* *for three-part composition* $\mathbf{x} = (x_1, x_2, x_3)'$ *can be expressed as*

$$\begin{aligned} \operatorname{var}(z_1^*) &= \frac{a+b+c}{6} + \frac{\sqrt{(a-b)^2 + (b-c)^2 + (c-a)^2}}{3\sqrt{2}}, \\ \operatorname{var}(z_2^*) &= \frac{a+b+c}{6} - \frac{\sqrt{(a-b)^2 + (b-c)^2 + (c-a)^2}}{3\sqrt{2}}, \end{aligned} \tag{2.7}$$

where a, b, c *correspond to* $\operatorname{var}\left(\ln \frac{x_1}{x_2}\right)$, $\operatorname{var}\left(\ln \frac{x_1}{x_3}\right)$, $\operatorname{var}\left(\ln \frac{x_2}{x_3}\right)$, *respectively.*

Proof Taking properties of the variation matrix into account [1], the general problem of finding stationary values can be replaced by maximizing $\mathbf{h}'\mathbf{Th}$ with respect to constraints $\mathbf{h}'\mathbf{c} = 0$, $\mathbf{h}'\mathbf{h} = 1$. Here $\mathbf{c} = \frac{1}{\sqrt{3}}(1,1,1)'$ and

$$\mathbf{T} = -\frac{1}{2}\begin{pmatrix} 0 & a & b \\ a & 0 & c \\ b & c & 0 \end{pmatrix}.$$

Consequently, by solving the equation $\mathbf{Kh} = \lambda\mathbf{h}$ ($\mathbf{K} = \mathbf{PTP}$, $\mathbf{P} = \mathbf{I} - \mathbf{c}'\mathbf{c}$), the resulting non-zero eigenvalues correspond to variances of principal components and eigenvectors to their log-contrasts. □

Note that in context of compositional data analysis, the matrix $\mathbf{K}$ represents covariance matrix of centred log-ratio transformed compositions [1]. It is easy to see that principal components and their variances, resulting as log-contrasts of eigenvectors and (non-zero) eigenvalues of the clr covariance matrix, respectively, correspond to those coming from ilr transformed compositional data [4]. Log-contrasts, corresponding to coordinates z_1^*, z_2^*, thus can be expressed as

$$\mathbf{h}_1 = \Bigg(-\frac{a-c+S}{2(b-c)\sqrt{S^2+(a-c)(a-b)S}}, \frac{a-b+S}{2(b-c)\sqrt{S^2+(a-c)(a-b)S}}, \frac{1}{2\sqrt{S^2+(a-c)(a-b)S}}\Bigg)'$$

and

$$\mathbf{h}_2 = \Bigg(-\frac{a-c-S}{2(b-c)\sqrt{S^2-(a-c)(a-b)S}}, \frac{a-b-S}{2(b-c)\sqrt{S^2-(a-c)(a-b)S}}, \frac{1}{2\sqrt{S^2-(a-c)(a-b)S}}\Bigg)',$$

respectively, where

$$S = \sqrt{\frac{1}{2}[(a-b)^2+(b-c)^2+(c-a)^2]}.$$

Because z_1^*, z_2^* are orthonormal coordinates, $\mathbf{h}_1$, $\mathbf{h}_2$ are standard and orthogonal log-contrasts, i.e. $\mathbf{h}_1'\mathbf{h}_1 = \mathbf{h}_2'\mathbf{h}_2 = 1$, $\mathbf{h}_1'\mathbf{h}_2 = 0$ (see [1, p. 85] for details). The latter property as well as zero covariance between z_1^* and z_2^* results from construction of principal components [10].

Note that big differences between variances of logratios contribute for maximalization of the first principal component at the expense of the second one. This is obvious from the second part of (2.7)—in variance of z_1^* we add square root of the sum of squared differences of these variances while in $\mathrm{var}(z_2^*)$ we substract it. Further it is not neccesary to consider the covariance because principal components are uncorrelated [10]. Obviously, the interpretation of principal components seems to be not straightforward even with the above decomposition of the covariance structure using variance of log-ratios of compositional parts. It will strongly depend on the concrete analyzed problem. On the other hand, some features of variances of these coordinates are now easily detectable. As already mentioned, the first part of both variances is formed by half of the total variance. Particularly, for higher difference between variances of both principal components high differences between variances of log-ratios are crucial (see the term contained in the square root).

3 Illustrative example

We demonstrate the above theoretical results using real-world data set from [12] containing the relative structure of landscape use (habitation, x_1; agriculture, x_2; wood and water areas, x_3) in 415 German regions (2009). Note that original data were expressed in percentages and the third part results from amalgamation of two trace parts.

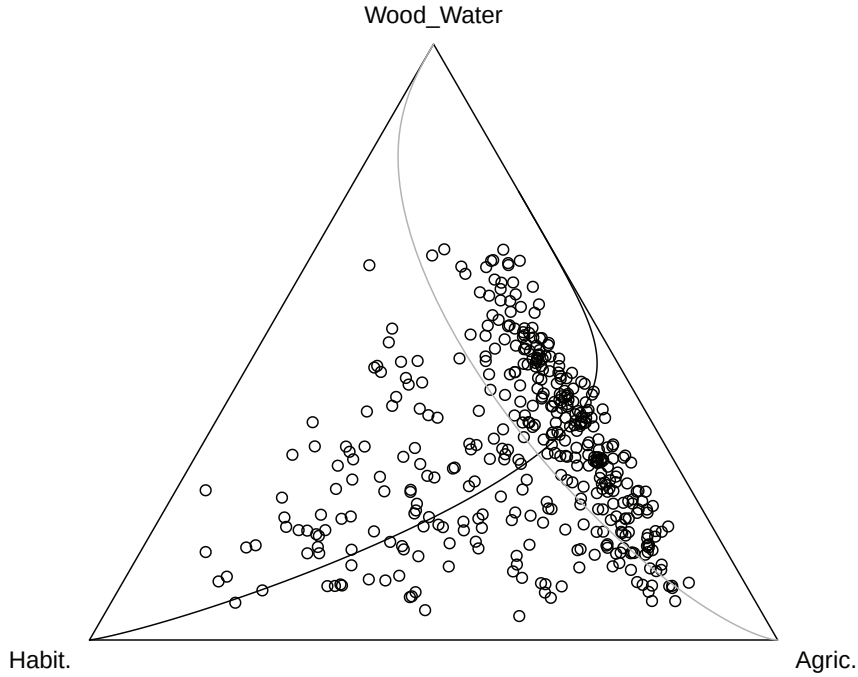

Figure 1: Relative structure of landscape use, original data in ternary diagram. Black line corresponds to first principal component and gray line to second principal component.

All computations and plots were performed using package 'compositions' of statistical software R [9].

From ternary diagram (Figure 1) we can see that data are concentrated mainly between parts correponding to agriculture and nature areas. Thus it might seem that the ratio between these two parts contributes at most to the total variance of the whole composition. Nevertheless, the variability concerning the part x_1 has small relative values except for big cities where it plays a dominating role. This turns out to be crutial and it becomes visible when the original data are expressed in orthonormal coordinates.

The preliminary considerations concerning variability are justified by variation matrix

$$\mathbf{T} = \begin{pmatrix} 0 & 0.876 & 1.006 \\ 0.876 & 0 & 0.597 \\ 1.006 & 0.597 & 0 \end{pmatrix}; \tag{3.1}$$

indeed, the highest variability is contained in log-ratio between the first and third part, followed closely by log-ratio between the first and second part of the composition. On the other hand, the effect of log-ratio between parts x_2 and x_3 is substantially smaller.

Figure 2 displays scatterplots of the above introduced three ilr coordinate systems together with principal components.

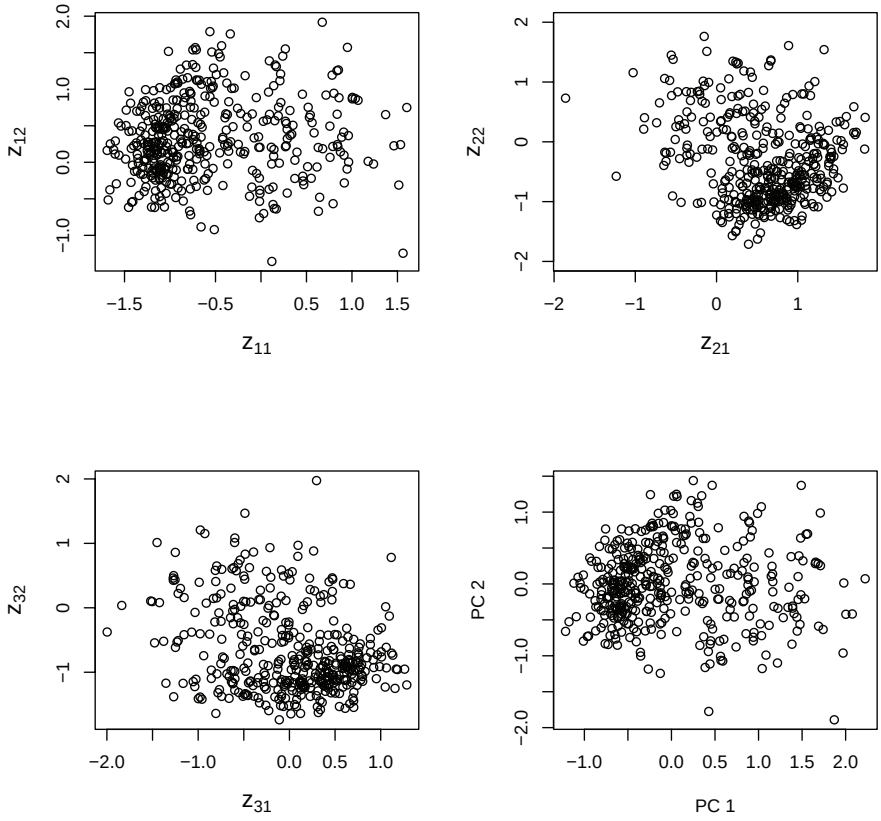

Figure 2: Relative structure of landscape use, plots of orthonormal coordinates. The upper left plot corresponds to formula (1.1), upper right plot to (1.2), lower left plot to (1.3) and lower right plot to principal components.

Note that these coordinates are rotations of each other. The upper left plot corresponds to coordinates resulting from (1.1). We can observe that the main data cloud contains negative values of z_{11} and positive values of z_{12}. It means that the second part (agriculture) is dominating in the composition, followed by nature area and habitation parts. It is also easy to see that the first coordinate captures more variability of the data set which is confirmed by Table 1. From this figure we can also see that the main data cloud in the other two coordinates systems is located in the fourth quadrant. This means that ratios x_1/x_3 and x_1/x_2 are mostly below zero. This confirms the fact that the part x_1 contributes at least to the relative structure of landscape use. Consequently, it is not suprising that the scatterplot for principal components is quite close just to coordinates (1.1).

i	$\mathrm{var}(z_{i1})$	$\mathrm{var}(z_{i2})$
1	0.528	0.299
2	0.323	0.503
3	0.388	0.438

Table 1: Variances of ilr coordinates.

From the variation matrix (3.1) the variances of principal components using (2.7) can be easily computed, $\mathrm{var}(z_1^*) = 0.5337$, $\mathrm{var}(z_2^*) = 0.2926$, where the first part of both variance terms, half of the total variance $\mathrm{totvar}(\mathbf{x})$, equals 0.4131. Difference between both variances results from the sum of squared differences between variances of log-ratios. Variances of log-ratios with x_1 differ substantially from variance of $\ln(x_2/x_3)$ that once more confirm the exceptional role of the habitation part for the overall variability of the compositional data set. This is also reflected by Figure 1 where both principal directions are displayed.

4 Conclusions

The case of three-part compositional data enables to decompose the covariance structure of principal components of ilr transformed compositions and allows for their better interpretation in sense of log-ratios of the original compositional parts. This is advantageous in the practice because interpretation of principal components usually strongly depends on the concrete data set and any additional support in this direction is warmly welcome. Although the maximization problem can be solved also in general for D-part compositions, from our experience it seems to be not possible to arrive to an interpretable decomposition of the covariance structure as for the case of three-part compositional data. Moreover, generalization for D-part compositions leads to demanding computation of solution of an algebraic equation.

References

[1] Aitchison, J.: The Statistical Analysis of Compositional Data. *Chapman and Hall*, London, 1986.

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

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

[4] Filzmoser, P., Hron, K.: *Robustness for compositional data*. In: Becker C., Fried R., Kuhnt S. (eds.): Robustness and complex data structures, *Springer*, Heidelberg, 2013, 117–131.

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

[6] Fišerová, E., Hron, K.: *Statistical inference in orthogonal regression for three-part compositional data using a linear model with type-II constraints*. Communications in Statistics – Theory and Methods **41** (2012), 2367–2385.

[7] Golub, G. H.: *Modified matrix eigenvalue problems*. SIAM Review **15** (1973), 318–334.

[8] Pawlowsky-Glahn, V., Buccianti, A. (eds.): *Compositional data analysis: Theory and applications*. *Wiley*, Chichester, 2011.

[9] R Development Core Team: *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing*, Vienna, 2013.

[10] Härdle, W. K., Simar, L.: Applied Multivariate Statistical Analysis. *Springer-Verlag*, Berlin, Heidelberg, 2012.

[11] Jackson, J. D., Dunlevy, J. A.: *Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences*. J. Roy. Statist. Soc. Ser. D (The Statistician) **37**, 1 (1988), 7–14.

[12] Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen: Kreiszahlen: Ausgewählte Regional Daten für Deutschland. *Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen*, Hannover, 2012.