To appear in *Statistics: A Journal of Theoretical and Applied Statistics*
Vol. 00, No. 00, Month 20XX, 1–18

Compositional biplots including external non-compositional variables

P. Kynčlová[a,c*], P. Filzmoser[a] and K. Hron[b,c]

[a] *Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria* [b] *Department of Mathematical Analysis and Applications of Mathematics, Palacký University, Olomouc, Czech Republic* [c] *Department of Geoinformatics, Palacký University, Olomouc, Czech Republic*

(*v4.1 released December 2013*)

Biplots represent a widely used statistical tool for visualizing the resulting loadings and scores of a dimension reduction technique applied to multivariate data. If the underlying data carry only relative information (i.e. compositional data expressed in proportions, mg/kg, etc.) they have to be pre-processed with a logratio transformation before the dimension reduction is carried out. In the context of principal component analysis, the resulting biplot is called compositional biplot. We introduce an alternative, the *ilr biplot*, which is based on a special choice of orthonormal coordinates resulting from an isometric logratio (ilr) transformation. This allows to incorporate also external non-compositional variables, and to study the relations to the compositional variables. The methodology is demonstrated on real data sets.

Keywords: compositional data; logratio transformations; principal component analysis; singular value decomposition; compositional biplot; ilr biplot

1. Introduction

Compositional data represent multivariate observations where the relevant information is contained in the ratios between the variables. Usually, already the measurement unit of such data (proportions, percentages, mg/kg, ppm, etc.) reflects their relative character. Since the interest is only in the ratios, the chosen unit is irrelevant, and it forms just a proper representation of the variables, called compositional parts [1]. Geometrically, compositional data follow the Aitchison geometry on the simplex [2]. Consequently, standard statistical methods that rely on the standard Euclidean geometry in real space usually fail when they attempt to capture the multivariate structure of compositional data.

In the last two decades, several papers related to the proper statistical treatment of compositional data have appeared, employing the logratio methodology to compositional data analysis [1, 3–7]. This is also the case in the context of principal component analysis (PCA) for compositional data [8–12]. Nevertheless, the recent developments concern just the case of PCA working only with compositional parts [8, 10, 12] or when supplementary variables are projected into a PCA biplot of compositional data [9]. A concise methodology on how to incorporate also additional non-compositional variables into one PCA is still not available, despite the fact that these cases frequently occur in practice. Examples are chemical concentration data of air quality measurements with external in-

[*] Corresponding author. Email: kynclova.petra@gmail.com

formation like wind-speed or solar radiation, or election data with external information characterizing the districts or regions.

The goal of this paper is to introduce an approach, based on the isometric logratio transformation for compositional data [5], for exploring the relations between compositional parts and external non-compositional variables using biplots of principal components. In the next section, some basics on biplots are recalled (Section 2). Section 3 treats biplots from a compositional data analysis point of view. Section 4 provides a detailed description of the methodology to include additional variables to compositional data in this context. Its usefulness for practical applications is demonstrated on two examples (Sections 5): for a data set from the German federal election, and for employment data in the European Union. The final Section 6 discusses possible problems and extension of the new analytical tool.

2. The PCA biplot: construction and interpretation

Consider a given data matrix $\mathbf{X}$ of dimension $n \times D$. The n rows are formed by the observation vectors $\mathbf{x}_{i.}$, for $i = 1, \ldots, n$, and the D columns by the variable vectors $\mathbf{x}_j$, for $j = 1, \ldots, D$. Throughout the manuscript, a "." in the index of a vector will refer to the corresponding row of a matrix, and a vector will always be a column-vector. Thus, $\mathbf{X} = (\mathbf{x}_{1.}^\top, \ldots, \mathbf{x}_{n.}^\top)^\top = (\mathbf{x}_1, \ldots, \mathbf{x}_D)$. We further assume that $\mathbf{X}$ is mean-centered, i.e. the column-wise arithmetic mean is subtracted from each column.

A PCA biplot can be constructed using singular value decomposition (SVD) of $\mathbf{X}$, given by

$$\mathbf{X} = \mathbf{UDV}^\top, \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{n\times n}$ and $\mathbf{V} \in \mathbb{R}^{D\times D}$ represent orthogonal matrices and $\mathbf{D} \in \mathbb{R}^{n\times D}$ is a (rectangular) diagonal matrix, where the diagonal consists of non-negative values, the *singular values*, which are arranged in descending order ($d_{11} \geq d_{22} \geq \cdots \geq d_{kk} \geq 0$). Here, $k \leq \min(n, D)$ denotes the rank of $\mathbf{X}$. With this decomposition, $\mathbf{X}$ can be expressed as

$$\mathbf{X} = \sum_{i=1}^{k} d_{ii}\mathbf{u}_i\mathbf{v}_i{}^\top,$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$, respectively, represent the i-th column of the matrix $\mathbf{U}$ and $\mathbf{V}$, respectively. Due to the orthogonality of $\mathbf{U}$ and $\mathbf{V}$ the following equations hold:

$$\mathbf{XX}^\top\mathbf{u}_i = d_{ii}^2\mathbf{u}_i,$$

$$\mathbf{X}^\top\mathbf{X}\mathbf{v}_i = d_{ii}^2\mathbf{v}_i.$$

Thus, $\mathbf{u}_i$ is the i-th eigenvector of $\mathbf{XX}^\top$ to the eigenvalue d_{ii}^2, and $\mathbf{v}_i$ is the i-th eigenvector of $\mathbf{X}^\top\mathbf{X}$ to the same eigenvalue d_{ii}^2. From the latter equation it is immediate that $\mathbf{v}_i$ is also an eigenvector of the sample covariance matrix

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^\top\mathbf{X},$$

which thus corresponds to the i-th loading vector of a classical PCA. Accordingly, the PCA scores information is contained in the matrix $\mathbf{V}$ [13].

The goal of the biplot is to plot information of the observations (PCA scores) as well as information of the variables (PCA loadings) in one plot [14]. For this purpose we define the decomposition $\mathbf{X} = \mathbf{G}\mathbf{H}^\top$, where the rows of the matrix

$$\underset{n\times k}{\mathbf{G}} = (\mathbf{g}_{1\cdot}, \ldots, \mathbf{g}_{n\cdot})^\top = \sqrt{n-1}\mathbf{U} \tag{2}$$

contain the information of the observations, and the rows of the matrix

$$\underset{D\times k}{\mathbf{H}} = (\mathbf{h}_{1\cdot}, \ldots, \mathbf{h}_{D\cdot})^\top = \frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{D}, \tag{3}$$

contain information of the variables. The scores information is usually shown by points in the biplot, while the loadings information is drawn by rays. Since a biplot is usually two-dimensional, the information contained in $\mathbf{X}$ is exactly reproduced if the rank k of $\mathbf{X}$ is two (or less). Otherwise, the descriptive ability of the biplot relies on the amount of variability explained by the first two principal components, and we only obtain $\mathbf{G}\mathbf{H}^\top \approx \mathbf{X}$.

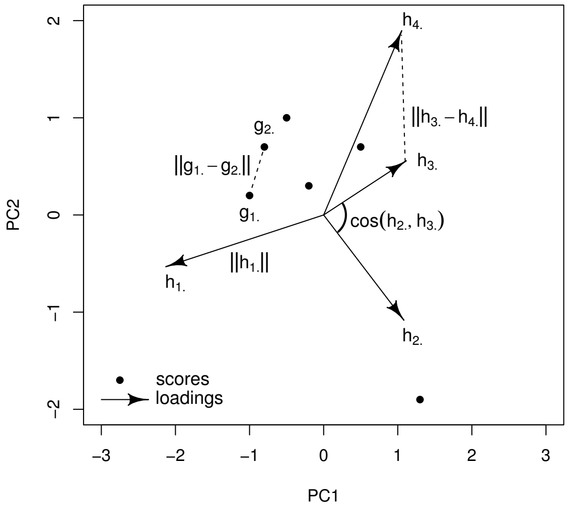

Figure 1. Graphical illustration of standard biplot properties.

With the above choices of the matrices $\mathbf{G}$ and $\mathbf{H}$, the following properties are obtained [14] and visually explained in Figure 1:

- The inner product between the rows of $\mathbf{G}$ and the rows of $\mathbf{H}$ estimates the original matrix of observations $\mathbf{X}$, i.e. $\mathbf{g}_{i\cdot}^\top \mathbf{h}_{j\cdot} \approx x_{ij}$.
- Since $\mathbf{H}\mathbf{H}^\top \approx \frac{1}{n-1}\mathbf{X}^\top\mathbf{X} = \mathbf{S}$, a biplot constructed in this way is called *covariance biplot*.
- The length of a ray estimates the standard deviation of the respective variable, $\|\mathbf{h}_{j\cdot}\|^2 = \mathbf{h}_{j\cdot}^\top \mathbf{h}_{j\cdot} \approx \frac{1}{n-1}\mathbf{x}_j^\top \mathbf{x}_j$.
- Consequently, the cosine of the angle between two rays expresses the approximated correlation coefficients between the corresponding variables, $\cos(\mathbf{h}_{i\cdot}, \mathbf{h}_{j\cdot}) = \frac{\mathbf{h}_{i\cdot}^\top \mathbf{h}_{j\cdot}}{\|\mathbf{h}_{i\cdot}\|\|\mathbf{h}_{j\cdot}\|} \approx$

$\frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$.

- The squared distances between the rows of $\mathbf{H}$ approximate the mean squared difference between the variables, $\|\mathbf{h}_{i\cdot} - \mathbf{h}_{j\cdot}\|^2 \approx \frac{1}{n-1}\|\mathbf{x}_i - \mathbf{x}_j\|^2$.
- The squared distances between the rows of $\mathbf{G}$ approximate the squared Mahalanobis distance between the observations, $\|\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}\|^2 \approx (\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot})^\top \mathbf{S}^{-1}(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot})$.

The above well-known properties for the covariance biplot will be explored in the following for compositional data.

3. Biplots for compositional data

3.1. *The clr transformation and corresponding biplot properties*

Compositional data follow the Aitchison geometry on the simplex. Before applying PCA and constructing a biplot, the data need to be transformed to the usual Euclidean geometry. A popular transformation for this purpose [8] is the centered logratio (clr) transformation [1], defined for a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)^\top$ as

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

The expression in the denominator, $\sqrt[D]{\prod_{i=1}^D x_i}$, represents the geometric mean of the given composition $\mathbf{x}$, denoted as $g(\mathbf{x})$.

Let us assume the $n \times D$ matrix $\mathbf{Y}$ as a matrix of clr coefficients of $\mathbf{X}$, the original uncentered compositional data matrix. The elements of $\mathbf{Y}$ are denoted by y_{ij}, the rows by $\mathbf{y}_{i\cdot}$, and the columns by $\mathbf{y}_j$. Since the clr transformation preserves the distances between the objects [5], the standard procedures can be applied for the newly constructed matrix $\mathbf{Y}$. For the sake of convenience, we will use the same notation as in the last section. Accordingly, in analogy to (1), the SVD decomposition of $\mathbf{Y}$ is given by

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top. \tag{4}$$

Further, the matrices $\mathbf{G}$ and $\mathbf{H}$ are defined according to (2) and (3), respectively. Using only the first two components of these matrices for the biplot construction, the relation $\mathbf{G}\mathbf{H}^\top = \mathbf{Y}$ holds if the rank of $\mathbf{Y}$ is not larger than two–otherwise this relation is only approximately valid, and the quality of the approximation depends on the rank of $\mathbf{Y}$. The rows of the matrix $\mathbf{G}$ contain the object information, and the rows of the matrix $\mathbf{H}$ contain the information of the clr variables. Both sources of information are used to construct the so-called *compositional biplot* [8].

The essential difference between the standard and the compositional biplot is that $\mathbf{H}$ does not directly represent the original variables but transformed versions thereof. It is possible to interpret the single clr variables as those capturing all the relative information (ratios) about the corresponding compositional parts (in the numerator of the logratio) [15]. Nevertheless, from a numerical perspective, one should be aware of the fact that the geometric mean in the denominator can be driven by possible distortion (like rounding errors) of the involved parts. For this reason, the interpretation of clr variables in the sense of the original compositional parts (in terms of (sub)dominance of the part of interest to the "mean" part in the composition) requires a careful selection of parts, included in the parent composition. As a consequence, the interpretation of the relations in the compositional biplot has to be adapted (Figure 1):

- Similar to the standard biplot, the inner product between the rows of $\mathbf{G}$ and the rows of $\mathbf{H}$ estimates the matrix of clr coefficients $\mathbf{Y}$,
$$\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{j\cdot} = \sqrt{n-1}\mathbf{u}_{i\cdot}^{\top}\frac{1}{\sqrt{n-1}}(\mathbf{v}_{j\cdot}\mathbf{D}) = \mathbf{u}_{i\cdot}^{\top}\mathbf{D}\mathbf{v}_{j\cdot} \approx y_{ij} = \ln\frac{x_{ij}}{g(\mathbf{x})}, \tag{5}$$
where $\mathbf{u}_{i\cdot}$ and $\mathbf{v}_{j\cdot}$ are i-th and j-th row of $\mathbf{U}$ and $\mathbf{V}$, respectively.
- The lengths of the rays estimate the standard deviations of clr transformed variables (clr coefficients),
$$\|\mathbf{h}_{j\cdot}\|^2 = \mathbf{h}_{j\cdot}^{\top}\mathbf{h}_{j\cdot} = \frac{1}{n-1}(\mathbf{v}_{j\cdot}\mathbf{D})^{\top}(\mathbf{v}_{j\cdot}\mathbf{D}) \approx \frac{1}{n-1}\mathbf{y}_j^{\top}\mathbf{y}_j = \operatorname{var}\left(\ln\frac{\mathbf{x}_j}{g(\mathbf{x})}\right). \tag{6}$$
- The links between the vertices of the rays estimate the standard deviation of the logratio between the corresponding compositional parts, hence
$$\begin{aligned}\|\mathbf{h}_{i\cdot} - \mathbf{h}_{j\cdot}\|^2 &\approx \frac{1}{n-1}(\mathbf{y}_i - \mathbf{y}_j)^{\top}(\mathbf{y}_i - \mathbf{y}_j) = \frac{1}{n-1}\sum_{l=1}^{n}(\mathbf{y}_{li} - \mathbf{y}_{lj})^2 \\ &= \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{li}}{g(\mathbf{x})} - \ln\frac{x_{lj}}{g(\mathbf{x})}\right)^2 = \frac{1}{n-1}\sum_{l=1}^{n}\left(\ln\frac{x_{li}}{x_{lj}}\right)^2 = \operatorname{var}\left(\ln\frac{\mathbf{x}_i}{\mathbf{x}_j}\right).\end{aligned} \tag{7}$$
- The projection of a score onto a link represents an approximate difference between the two clr coordinates y_{ij} and y_{ik}, which is the logratio between the original values x_{ij} and x_{ik},
$$\begin{aligned}\mathbf{g}_{i\cdot}{}^{\top}(\mathbf{h}_{j\cdot} - \mathbf{h}_{k\cdot}) &= \sqrt{n-1}\mathbf{u}_{i\cdot}{}^{\top}\frac{1}{\sqrt{n-1}}(\mathbf{v}_{j\cdot} - \mathbf{v}_{k\cdot})\mathbf{D} \\ &\approx y_{ij} - y_{ik} = \ln\frac{x_{ij}}{g(\mathbf{x})} - \ln\frac{x_{ik}}{g(\mathbf{x})} = \ln\frac{x_{ij}}{x_{ik}}.\end{aligned} \tag{8}$$
- The Euclidean distance between the rows of $\mathbf{G}$ approximates the Mahalanobis distance between the clr coefficients in the full space with the estimated covariance matrix $\mathbf{S}_{\mathbf{Y}}$ of the clr-transformed variables,
$$\begin{aligned}\|\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}\|^2 &= (\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot})^{\top}(\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}) = (n-1)(\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot})^{\top}(\mathbf{u}_{i\cdot} - \mathbf{u}_{j\cdot}) \\ &\approx (\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot})^{\top}\mathbf{S}_{\mathbf{Y}}^{-1}(\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot}).\end{aligned} \tag{9}$$

Several further properties of the compositional biplot are listed in [8]. Although these are important for interpreting the relations among the compositional parts, they cannot be explored for relating compositional variables with external information.

The important difference between the standard and the compositional biplot is in the interpretation of the rays and of the links between the vertices of the rays. While in the standard biplot, rays and links represent variability among the variables, they represent *relative* variability in the compositional biplot. Specifically, the correlation measure expressed by the cosine of the angle between two rays (standard biplot) is replaced by the variance of a logratio, expressed as the (squared) length of a link in the compositional biplot [1]. Accordingly, when the vertices coincide, or nearly so, then the variance

$\mathrm{var}(\ln \frac{x_i}{x_j})$ is approximately equal to zero. Thus, the ratio between x_i and x_j is constant, or nearly so, and it could be stated that variables x_i and x_j are interchangeable.

In many situations, the clr coordinates themselves are not appropriate for a statistical analysis, because due to the constraint $y_1 + \cdots + y_D = 0$, resulting from the fact that clr variables represent coordinates with respect to a generating system, the corresponding covariance matrix is singular. A correlation coefficient between clr variables would thus result in biased values. The reason is that for the covariance structure of clr variables the following relations hold: $\sum_{i \neq j} \mathrm{cov}(y_i, y_j) = -\mathrm{var}(y_i)$, $i = 1, \ldots, D$. Consequently, the corresponding correlation coefficients loose their predicative value, because they cannot vary freely between -1 and 1. From this perspective, also for combining the clr variables with external non-compositional ones, the singularity constraint would result in problematic issues. For example, any clr variable cannot be principally taken separately without considering its relation to the other variables, expressed by the zero sum constraint. It thus complicates intepretability of the biplot in the sense of relative information on single compositional parts, discussed in the following. To sum up, this all makes the use of clr variables for the purpose of PCA and the compositional biplot with additional non-compositional variables not recommendable.

3.2. *The ilr transformation and biplot construction*

The isometric logratio (ilr) transformation results in orthonormal coordinates $\mathbf{z} = (z_1, \ldots, z_{D-1})^\top$ with respect to the Aitchison geometry, and it also leads to an orthonormal basis of the hyperplane $\mathcal{H} : y_1 + \cdots + y_D = 0$, formed by the clr transformation [5]. Consequently, there exists a linear relation between the clr variables and the orthonormal coordinates [5],

$$\mathbf{y} = \mathcal{V}\mathbf{z}. \tag{10}$$

The columns of the $D \times (D-1)$ matrix $\mathcal{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ are orthonormal basis vectors on the hyperplane $\mathcal{H}$,

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}}\left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i}\right)^\top, \qquad i = 1, \ldots, D-1, \tag{11}$$

resulting in the ilr coordinates $\mathbf{z}$. In particular, this means that PCA results in the same principal component scores with non-zero variances (the last principal component is formed by the normal vector on $\mathcal{H}$, thus having zero variability).

There are infinitely many possibilities to construct an orthonormal basis. A special choice of orthonormal coordinates that allows to interpret them in terms of the contributions of the single compositional parts is as follows [16]. Consider the compositions $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$, which are re-arranged such that the l-th part is in the first position. We will use the notation $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$, where each part with index $l = 1, \ldots, D$ could be placed on the first position, and the sequence of the other parts remains unchanged. The ilr transformation of $\mathbf{x}^{(l)}$ results in $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})^\top$, where the components are defined by

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \qquad i = 1, \ldots, D-1. \tag{12}$$

Then, the first ilr variable $z_1^{(l)}$ explains all the relative information (log-ratios) about the original compositional part x_l. The coordinates $z_2^{(l)}, \ldots, z_{D-1}^{(l)}$ explain the remaining log-ratios in the composition [15]. Note that the only important position is that of $x_1^{(l)}$, because it can be fully explained by $z_1^{(l)}$. The other parts can be chosen arbitrarily, because different ilr transformations are orthogonal rotations of each other [5]. Note that the relation

$$y_l = \sqrt{\frac{D-1}{D}} z_1^{(l)},\ l = 1, \ldots, D \tag{13}$$

confirms our preliminary requirement on interpretability of the resulting coordinates, for $D \to \infty$ both variables approach the same values. On the other hand, both y_l and $z_1^{(l)}$ thus share also interpretational doubts, mentioned by defining the clr variables.

The advantage of obtaining an interpretation for each compositional part is redeemed by the necessity of constructing D coordinate systems, where always just one variable is of primary interest (at the first position). It is obvious that always the first coordinate $z_1^{(l)}$ in each given system corresponds to the clr coordinate y_l, for $l = 1, \ldots, D$, differing by the constant $\sqrt{\frac{D}{D-1}}$.

Consider now an $n \times (D-1)$ matrix $\mathbf{Z}^{(l)}$ with ilr coefficients due to (12), for each of the n observations. Assuming D different coordinate systems, then D singular value decompositions are required to obtain scores and loadings for the biplot construction. For $l = 1, \ldots, D$, an SVD gives

$$\mathbf{Z}^{(l)} = \mathbf{U}^{(l)} \mathbf{D} \mathbf{V}^{(l)^\top}. \tag{14}$$

As it has been shown in [10], the diagonal matrix $\mathbf{D}$ is the same as in (4) for the clr-transformed data. Moreover, all matrices $\mathbf{U}^{(l)}$ are equal, and they correspond to the matrix $\mathbf{U}$ in (4). This means that the scores in the clr space are identical to the scores of the ilr space, apart from the last column of the clr score matrix that contains zeros. Due to the relationship (10) between clr and ilr coordinates by a matrix with orthonormal columns, and the fact that different ilr-transformations are orthogonally related, we get

$$\mathbf{V} = \mathcal{V}^{(l)} \mathbf{V}^{(l)}, \qquad \text{for } l = 1, \ldots, D, \tag{15}$$

where $\mathbf{V}$ are the loadings from an SVD of $\mathbf{Y}$, and the matrix $\mathcal{V}^{(l)}$ stands for corresponding permutations of the orthonormal basis matrix $\mathcal{V}$, see [5] and [10]. Considering relation (13) it is immediate that the l-th row of $\mathbf{V}$ is equivalent to the first row of $\mathbf{V}^{(l)}$, differing only by the constant $\sqrt{\frac{D}{D-1}}$.

For constructing the biplot, a decomposition of the form

$$\mathbf{Z}^{(l)} = \mathbf{G}^{(l)} \mathbf{H}^{(l)^\top}, \qquad l = 1, \ldots, D, \tag{16}$$

is required. With the above statements, and in analogy to the clr biplot, it is clear that

$$\mathbf{G}^{(l)} = \mathbf{G} = \sqrt{n-1}\mathbf{U}, \tag{17}$$

and

$$\mathbf{H}^{(l)} = \frac{1}{\sqrt{n-1}} \mathbf{V}^{(l)} \mathbf{D}. \tag{18}$$

Due to the relation between the matrices $\mathbf{V}$ and $\mathbf{V}^{(l)}$, the first row $\mathbf{h}_{1\cdot}^{(l)}$ of the ilr loadings information $\mathbf{H}^{(l)}$ is related to the l-th row $\mathbf{h}_{l\cdot}$ of the clr loadings information $\mathbf{H}$ by

$$\mathbf{h}_{1\cdot}^{(l)} = \sqrt{\frac{D}{D-1}}\mathbf{h}_{l\cdot}, \qquad l = 1, \ldots, D. \tag{19}$$

The relationships between the loadings of ilr and clr coefficients are leading to similar properties as in the compositional biplot, only differing by a constant. The properties of the ilr biplot are illustrated in Figure 2.

- The inner product between the rows of $\mathbf{G}$ and the rows of $\mathbf{H}^{(l)}$ gives

$$\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{1\cdot}^{(l)} = \sqrt{\frac{D}{D-1}}\mathbf{g}_{i\cdot}^{\top}\mathbf{h}_{l\cdot} = \sqrt{\frac{D}{D-1}}\mathbf{u}_{i\cdot}^{\top}\mathbf{D}\mathbf{v}_{l\cdot} \approx \sqrt{\frac{D}{D-1}}y_{il} = \sqrt{\frac{D}{D-1}}\ln\frac{x_{il}}{g(\mathbf{x})}. \tag{20}$$

- The lengths of the rays represent

$$\|\mathbf{h}_{1\cdot}^{(l)}\|^2 = \frac{D}{D-1}\mathbf{h}_{l\cdot}^{\top}\mathbf{h}_{l\cdot} \approx \frac{D}{D-1}\frac{1}{n-1}\mathbf{y}_l^{\top}\mathbf{y}_l = \frac{D}{D-1}\text{var}\left(\ln\frac{\mathbf{x}_l}{g(\mathbf{x})}\right). \tag{21}$$

- The links between the vertices are

$$\|\mathbf{h}_{1\cdot}^{(i)}-\mathbf{h}_{1\cdot}^{(j)}\|^2 = \frac{D}{D-1}\|\mathbf{h}_{i\cdot}-\mathbf{h}_{j\cdot}\|^2 \approx \frac{D}{D-1}\frac{1}{n-1}(\mathbf{y}_i-\mathbf{y}_j)^{\top}(\mathbf{y}_i-\mathbf{y}_j) = \frac{D}{D-1}\text{var}\left(\ln\frac{\mathbf{x}_i}{\mathbf{x}_j}\right). \tag{22}$$

- The projection of a score to the link yields

$$\mathbf{g}_{i\cdot}{}^{\top}(\mathbf{h}_{1\cdot}^{(j)} - \mathbf{h}_{1\cdot}^{(k)}) = \sqrt{\frac{D}{D-1}}\mathbf{g}_{i\cdot}{}^{\top}(\mathbf{h}_{j\cdot} - \mathbf{h}_{k\cdot}) \approx \sqrt{\frac{D}{D-1}}(y_{ij} - y_{ik}) = \sqrt{\frac{D}{D-1}}\ln\frac{x_{ij}}{x_{ik}}. \tag{23}$$

- As for the clr biplot, the Euclidean distance between the rows of $\mathbf{G}$ gives

$$\|\mathbf{g}_{i\cdot} - \mathbf{g}_{j\cdot}\|^2 \approx (\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot})^{\top}\mathbf{S}_{\mathbf{Y}}^{-1}(\mathbf{y}_{i\cdot} - \mathbf{y}_{j\cdot}). \tag{24}$$

- The angles between ilr coordinates and clr coefficients remain the same, despite the fact that they are not used for interpreting a correlation structure of a compositional biplot.

$$\cos(\mathbf{h}_{1\cdot}^{(i)}, \mathbf{h}_{1\cdot}^{(j)}) = \frac{\frac{D}{D-1}\mathbf{h}_{i\cdot}^{\top}\mathbf{h}_{j\cdot}}{\frac{D}{D-1}\|\mathbf{h}_{i\cdot}\|\|\mathbf{h}_{j\cdot}\|} \approx \frac{\mathbf{y}_i^{\top}\mathbf{y}_j}{\|\mathbf{y}_i\|\|\mathbf{y}_j\|}. \tag{25}$$

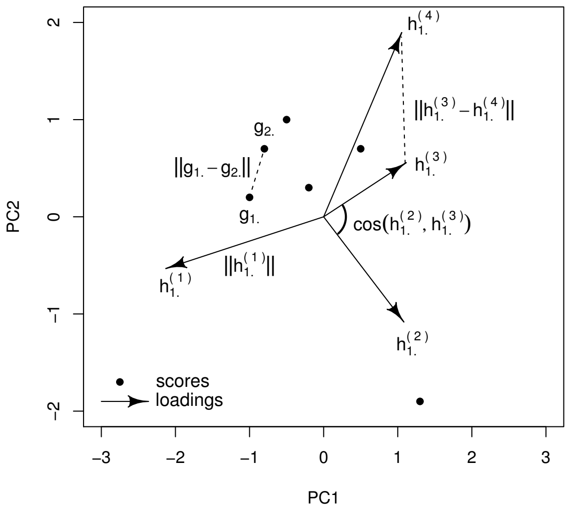

Figure 2. Graphical illustration of ilr biplot properties.

In the following, the biplot constructed by merging information from loadings of D orthonormal coordinate systems together into one planar graph, as described above, will be called *ilr biplot*. In order to avoid possible confusion, we should note that the ilr biplot as defined here thus corresponds to a scaled compositional biplot of clr variables; they both differ just in the interpretation of the loadings, coming from the employed orthonormal coordinate systems in the ilr biplot. This helps to consider the (scaled) clr variables separately (consequently also within the compositional biplot), and not as an inherent part of the coordinates with respect to a generating system. On the other hand, a *biplot of ilr coordinates* for an interpretable choice of balances [6], following the properties of a standard biplot, can be constructed as well. In the next step we describe how the ilr biplot can be extended by additional non-compositional variables.

4. Compositional biplots with additional variables

The next step to construct a meaningful biplot for both compositional data and external non-compositional variables is to analyze, whether the use of a clr transformation or ilr coordinate systems (12) for the compositional part of the data would yield the same results (up to a scaling constant) as in the previous section. Consider q additional non-compositional variables $\mathbf{X}^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_q^*)^\top$, which have already been preprocessed accordingly (e.g. scaled). In the following we have to distinguish different cases how to combine external and compositional variables.

Initially, let us assume only one composition and external variables. We could consider two joint matrices $(\mathbf{Y}\vdots\mathbf{X}^*) \in \mathrm{R}^{n\times(D+q)}$ and $(\mathbf{Z}^{(l)}\vdots\mathbf{X}^*) \in \mathrm{R}^{n\times(D+q-1)}$, where $\mathbf{Y}$ represents clr coordinates and $\mathbf{Z}^{(l)}, l = 1,\ldots,D$, are ilr coefficients for D different coordinate systems. Subsequently, it is required to apply the SVD for both matrices to compare scores and loadings of a compositional and ilr biplot, respectively. For $l = 1,\ldots,D$ the SVD gives

$$(\mathbf{Y}\vdots\mathbf{X}^*) = \mathbf{U}^*\mathbf{D}^*\mathbf{V}^{*\top} = \mathbf{G}^*\mathbf{H}^{*\top}, \tag{26}$$

$$(\mathbf{Z}^{(l)} \vdots \mathbf{X}^{*}) = \mathbf{U}^{*(l)}\mathbf{D}^{*(l)}\mathbf{V}^{*(l)^{\top}} = \mathbf{G}^{*(l)}\mathbf{H}^{*(l)^{\top}}. \tag{27}$$

The diagonal matrices $\mathbf{D}^{*(l)}, l = 1, \dots, D$, are the same and they are equal to $\mathbf{D}^{*}$ (up to its last zero row/column) corresponding to the SVD for clr coordinates with external variables. Similarly, it is straightforward to show that the scores for the compositional and ilr biplot, respectively, are identical,

$$\mathbf{G}^{*(l)} = \mathbf{G}^{*} = \sqrt{n-1}\mathbf{U}^{*}, \qquad l = 1, \dots, D. \tag{28}$$

The loadings of the SVD of (4) and (27) are related according to a linear relation between the clr and the ilr transformation (10) as

$$\mathbf{V}^{*} = \mathcal{V}^{(l)}\mathbf{V}^{*(l)}, \qquad l = 1, \dots, D, \tag{29}$$

where the matrix $\mathcal{V}^{(l)}$ represents the corresponding permutation of the orthonormal basis matrix $\mathcal{V}$ (11). Accordingly, a relation between the loadings using the ilr transformation and the clr transformation to construct a biplot including external non-compositional variables is obtained as

$$\mathbf{h}_{1.}^{*\,(l)} = \sqrt{\frac{D}{D-1}}\mathbf{h}_{l.}^{*} \quad \text{for} \quad l = 1, \dots, D. \tag{30}$$

Since we have stated the same relation for loadings without external variables (19), it is obvious that incorporating new non-compositional variables to the construction of a biplot does not influence the resulting loadings and scores of the compositional parts.

Consequently, a meaningful interpretation between compositional parts and external variables can be investigated. The representation of the relations among the compositional variables has been introduced in Section 3 and in the case of external variables the important role is played by the angles showing the approximate correlation coefficient between two external variables as in the standard biplot. Similarly, for the purpose of interpreting the relations between both types of variables only angles can be considered. Thus the angles can also approximate the correlation structure between the chosen external variable x_i^{*} $i = 1, \dots, q$ and an arbitrary compositional part x_l $(l = 1, \dots, D)$, since the compositional variable is expressed (in the above sense) using coordinate $z_1^{(l)}$, $l = 1, \dots, D$, being a standard real variable.

Furthermore, let us assume two different compositional variables to investigate their mutual relations among parts in a biplot (external variables are not considered for simplicity). Let $\mathbf{X}_1 = (\mathbf{x}_{11}, \dots, \mathbf{x}_{1D_1})^{\top}$ and $\mathbf{X}_2 = (\mathbf{x}_{21}, \dots, \mathbf{x}_{2D_2})^{\top}$ be two different compositions with D_1 respectively D_2 parts. To compare loadings and scores it is necessary to construct the SVD for the merged matrices of the clr and ilr coordinates for both compositional variables as follows

$$(\mathbf{Y}_1 \vdots \mathbf{Y}_2) = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^{\top} = \tilde{\mathbf{G}}\tilde{\mathbf{H}}^{\top}, \tag{31}$$

$$(\mathbf{Z}_1^{(l)} \vdots \mathbf{Z}_2^{(k)}) = \tilde{\mathbf{U}}^{(lk)}\tilde{\mathbf{D}}^{(lk)}\tilde{\mathbf{V}}^{(lk)} = \tilde{\mathbf{G}}^{(lk)}\tilde{\mathbf{H}}^{(lk)}, \tag{32}$$

where $l = 1, \dots, D_1$ and $k = 1, \dots, D_2$. Here $\mathbf{Y}_1$ and $\mathbf{Y}_2$, respectively, represent clr coefficients of $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively, and $\mathbf{Z}_1^{(l)}, \mathbf{Z}_2^{(k)}$ stand for their ilr coordinates according to (12). The relationships between scores and loadings for the compositional biplot and the ilr biplot correspond directly to the simple case of one composition in Section 3.2. Since, by omitting the last two rows and columns of $\tilde{\mathbf{D}}$, the diagonal matrices $\tilde{\mathbf{D}}^{(lk)}$ and $\tilde{\mathbf{D}}$ are the same, for $l = 1, \dots, D$, the corresponding scores are equal,

$$\tilde{\mathbf{G}}^{(lk)} = \tilde{\mathbf{G}} = \sqrt{n-1}\tilde{\mathbf{U}}. \tag{33}$$

We can derive an analogous relation also for the loadings,

$$\tilde{\mathbf{h}}_{1\cdot}^{(lk)} = \sqrt{\frac{D_1}{D_1 - 1}}\tilde{\mathbf{h}}_{l\cdot}, \qquad l = 1, \dots, D_1, \tag{34}$$

thus the ilr loadings concerning the first composition differ only by a constant $\sqrt{\frac{D_1}{D_1-1}}$, where D_1 is the number of parts of the first composition. A similar relation can also be derived for the loadings highlighting parts of the latter composition, i.e.

$$\tilde{\mathbf{h}}_{D_1\cdot}^{(lk)} = \sqrt{\frac{D_2}{D_2 - 1}}\tilde{\mathbf{h}}_{(D_1+k)\cdot}, \qquad k = 1, \dots, D_2. \tag{35}$$

Taking into account the mentioned relations between scores and loadings, an appropriate interpretation of the properties can be incorporated for the case of a biplot constructed for two different compositional variables. Because the ilr coordinates represent standard real variables, their relation for those coordinates resulting from different compositions can be analyzed using angles of the corresponding rays like in the standard biplot. Of course, for measuring the strength of the relative relation between the parts within one composition, the links between the rays still represent the preferred option.

Generally, it is feasible to construct a meaningful biplot for more compositions and external non-compositional variables simultaneously as a simple extension of two previously described cases. The main idea consists in applying a special choice of ilr coordinates (12) for each composition and preprocessing external non-compositional variables by the corresponding transformations. Consequently, the transformed variables are merged into one joint matrix followed by SVD to obtain scores and loadings for a biplot construction. Such a biplot representation reflects all possible combinations of the previously mentioned cases.

The main convenience is given by a simple relationship between the resulting SVD for clr and ilr coordinates. Obviously, it is not necessary to construct D coordinate systems when scores are always the same and loadings differ from using the clr transformation only by a scaling constant. It is possible to apply the clr transformation for the compositional parts followed by the same interpretation of the biplot as for a special choice of ilr coordinates (12). It is apparent that an appropriate interpretation of scores and loadings always depends also on the characteristic structure of the examined data. Selected cases are demonstrated on real-world data examples in Section 5.

5. Applications

5.1. *Election data*

The first example describes the results of a federal election in Germany in different federal states (Table A1) in September 2013 (data come from German Federal Statistical Office). The aim is to analyze the relations between the votes for the political parties in the elections (compositional variables), and their relation to the unemployment rate and the average monthly income (external non-compositional variables). We consider the votes for the Christian Democratic Union and Christian Social Union of Bavaria, also called The Union (CDU/CSU), Social Democratic Party (SDP), The Left (DIE LINKE), Alliance '90/The Greens (GRÜNE), Free Democratic Party (FDP) and the rest of the parties participated in the elections (other parties). The votes are examined in absolute values (number of valid votes). The unemployment in the federal states is reported in percentages, and the average monthly income in Euros.

As mentioned formerly, we are interested in relative information (ratios between the votes for the parties) contained in the data and also the influence of some additional effects. Initially, it is necessary to use appropriate transformations for all variables to obtain a meaningful biplot structure. For the numbers of valid votes, the ilr transformation (12) is used. The unemployment information, provided in percentages, is logit-transformed in order to change the relative scale of percentages (as a special case of compositional data) into the absolute one [17], and the average monthly income is scaled using its mean and its standard deviation. Subsequently, PCA is performed on these joint data to obtain scores and loadings for constructing the biplot.

Figure 3 (left) shows the resulting biplot. The explained variance is high, with 92.8%. It is obvious that the federal states are split into two groups. The right located group of states corresponds exactly to the states of former East Germany, except Berlin. The rest of them, left located, are states of former West Germany.

The lengths of the rays of the compositional variables represent the variability of respective ilr coordinates, and the lengths of the rays of the external variables stand for their own variability. The longest ray of the compositional variables represents the standard deviation of ilr variable DIE LINKE, which explains all the relative information of DIE LINKE to the rest of the considered parties. This means that the relative variability of the obtained votes differs a lot among all observed states. On the other hand, SDP and other parties show the smallest relative variability.

The important role in the interpretation of compositional variables in biplots is played by links between vertices of the rays. As the links stand for standard deviations of logratios, they can provide the information about relative variability of compositional parts. When the variance of the logratios $\mathrm{var}(\ln \frac{x_i}{x_j})$ is approximately zero or nearly so, we can say that the proportion of the variables is stable, thus x_i and x_j are interchangeable. This is the case for the pair GRÜNE and SDP, and to some extent also for the pair CDU/CSU and other parties. It means their proportion is almost equal among all observations. On the other hand, GRÜNE and DIE LINKE, FDP and DIE LINKE show the highest proportional variability.

The relation between external non-compositional variables can be examined as in the standard biplot. Accordingly, since the rays for income and unemployment are almost orthogonal, these variables seem to be nearly uncorrelated. The angles of the rays are also informative for investigating the relations between external variables and compositional ones, since the latter are ilr coordinates $z_1^{(l)}$ which explain all relative information about the original part x_l. Accordingly, the parties GRÜNE and SDP are strongly positively related to average monthly income. In contrast, the income variable is uncorrelated with voters of FDP and DIE LINKE, there is no essential relationship between income and votes for these political parties. The variable unemployment is strongly negatively related to FDP and CDU/CSU. The opposite relation seems to exist between unemployment and DIE LINKE, i.e. the rate of unemployment influences the proportional structure of people voting DIE LINKE.

Also the federal states can now be associated with the variables: The division of the states into the western and the eastern group is based on differences in the income (higher in the west) and unemployment (higher in the east), but also in the voting behavior. For example, in the eastern states DIE LINKE is much more dominant, and FDP is stronger in the west.

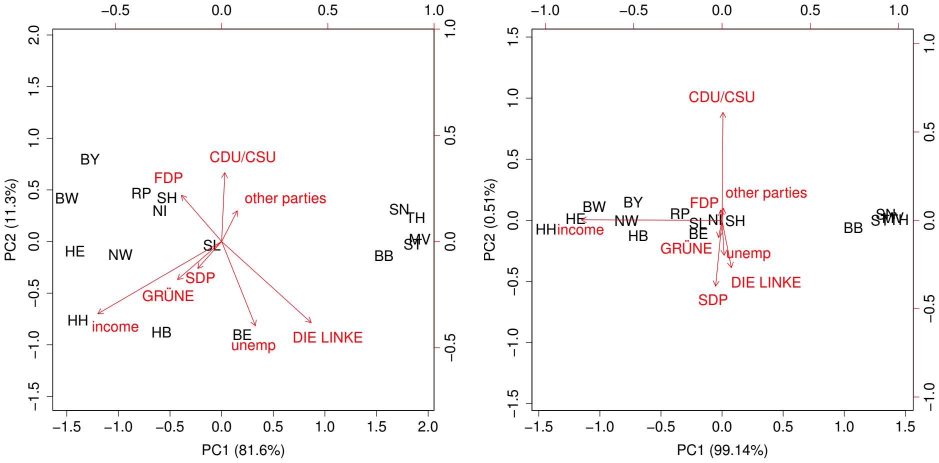

Figure 3. Biplots for the German federal elections including unemployment and average monthly income: ilr biplot (left) and standard biplot (right).

We also want to compare the results obtained using the ilr biplot with the case, where the compositional nature of the election data is not accounted for. Therefore, these raw percentage data are combined with the external variables unemployment (in percent) and income (in absolute numbers, scaled). Then, an SVD is carried out to the combined data, and the results are shown in a standard biplot in Figure 3 (right). Despite the high explained proportion of variance (99.65%), it is obvious that the resulting biplot differs a lot from the previous solution. We still have the separation of the states into the two groups, which are the result of different income. However, all other variables are essentially uncorrelated to this main direction. Also, this second PCA direction expresses not even 1% of the variability, and is thus rather irrelevant for an interpretation.

We also tried to use a logit-transformation for each of the compositional variables, and join this information with the external variables, i.e. with logit-transformed unemployment and scaled income. The resulting biplot is quite similar to the ilr biplot. There is, however, no guarantee for this phenomenon, as it will be shown in the next example.

5.2. *Employment data*

The aim of the second example is to show how it is possible to construct and interpret a biplot for two different compositions with external non-compositional variables. We consider a data set consisting of the number of employed people in the countries of the European Union (except of Ireland); the data come from EUROSTAT. The first composition describes the number of employed people in different fields of economic activity: agriculture, forestry and fishing (*agri*); industry and construction (*industry*); financial and insurance activities (*finance*); real estate activities (*real estate*); public administration, defense, education, human health and social work activities (*public*); arts, entertainment, recreation and other service activities (*arts*). The second composition illustrates employment in various age categories: from 15 to 24 years (15-24); from 25 to 64 years (25-64); and from 65 years and over (65+). The external variables are: shares of young people living with their parents (*young*), and people at the risk of poverty or social exclusion (*poverty*); both are given in percentages.

Each compositional data set is ilr-transformed with D coordinate systems (12) (instead, for simplicity, just the clr transformation can be taken), and afterwards joined together with the external variables. Figure 4 shows the resulting ilr biplot. The proportion of explained variance for these first two components is 79.1%. It is visible that many observations which are close to each other are also geographically in a neighborhood, for instance the Baltic states (Estonia, Latvia, Lithuania) or the Scandinavian countries (Denmark, Finland, Sweden). Close groups of observations have a similar proportional behavior of the considered variables. In general, richer countries are concentrated in the left part of the biplot, whereas less economically strong ones are in the right part. This is also supported by the external variables *poverty*, pointing to the right side, and *real estate*, pointing to the richer countries. On the top of graph we can recognize a group of countries whose gross domestic product (GDP) consists particularly from activities of the financial sector (Luxembourg, Malta and also Cyprus).

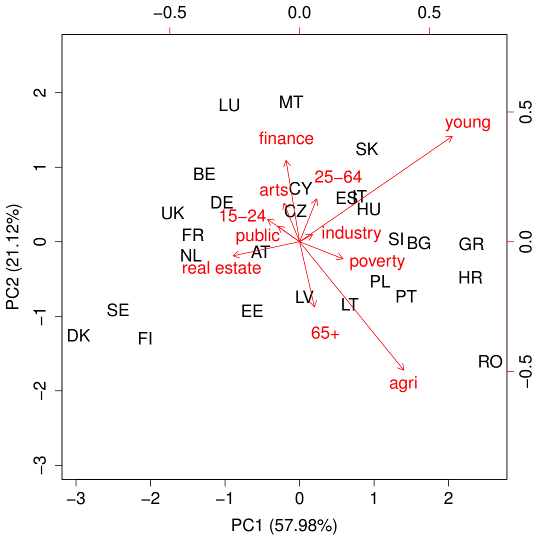

Figure 4. Ilr biplot of employed people by economic activity and age, including the risk of poverty or social exclusion and the share of young people living with their parents.

Initially, let us consider only the first composition (economic activity) and relations within these variables. The most significant ray is apparent for the agricultural sector, expressing a large standard deviation of all relative information of the variable *agri* to the remaining sectors. The links suggest that variables *public* and *industry* are proportionally almost equal, the proportion remains almost the same among all the observations. The same behavior can also be observed for the pairs *finance* and *arts*, *industry* and *arts*, *public* and *arts*, *finance* and *public*. On the contrary, the highest proportional variability is evident between agriculture and real estate activities, resp. financial activities. Within the second composition, the links seem to be very similar for all given parts.

Subsequently, we can also investigate relations between both compositions. Since the same ilr transformation (12) was used for both of them, the resulting conclusions (concerning the biplot interpretation) are made in the same way as dealing with standard real variables. We can see that the rays for *public* and young employees (15-24) nearly coincide, thus the behavior of these two variables within their parent compositions is positively correlated. The analogous relation can also be identified between 15-24 to *arts* and *finance*, then between 25-64 to *industry* and *finance*. Oppositely, the dominance of agriculture and young workers in their respective compositions is negatively correlated. It means that the agricultural sector is more important in countries with lower relative representation of young workers (this corresponds also to its positive correlation with employees over 65 years).

Considering now the external variables, we see that the percentage of young people living with their parents is uncorrelated to the proportion of employed people in the agricultural sector. The same conclusion can be stated also for the arts sector. On the contrary, the variable *young* is strongly related to the relative information of the industrial sector. On the other hand, the *young* is strongly negatively correlated with real estate activities since these variables lay approximately on the same line. The risk of poverty appears uncorrelated with employed people between 25 and 64 years. Moreover, the variable *poverty* is strongly negatively correlated with the relative amount of people employed in the public administration. Additionally, the risk of poverty seems to be related also to the variables *agri* and *industry*.

It is interesting to compare the compositional biplot also with a biplot constructed in the standard way, i.e. by ignoring the compositional nature of both compositions. Figure 5 shows two standard biplots with different data preprocessing transformations. The left graph represents data without scaling, since the data are expressed in percentages, scaling seems to be unreasonable in this case. Regardless, the resulting biplot does not look very meaningful for the purpose of interpretation. The non-compositional variables reflect significantly higher variability than other observed variables (much longer rays). For this reason, the logit transformation was used for all non-compositional variables and the biplot is shown in Figure 5 (right). The explained variance is much lower in this case (72.18%) and there are some significant differences to the ilr biplot. For instance, the age structure of the employed people is completely different. The ray of employed people in age of 25 to 64 is slightly visible and the variance of young people (15-24) has changed its direction. In the ilr biplot the ray coincides with the *public* variable, whereas here it seems to be correlated with real estate activities.

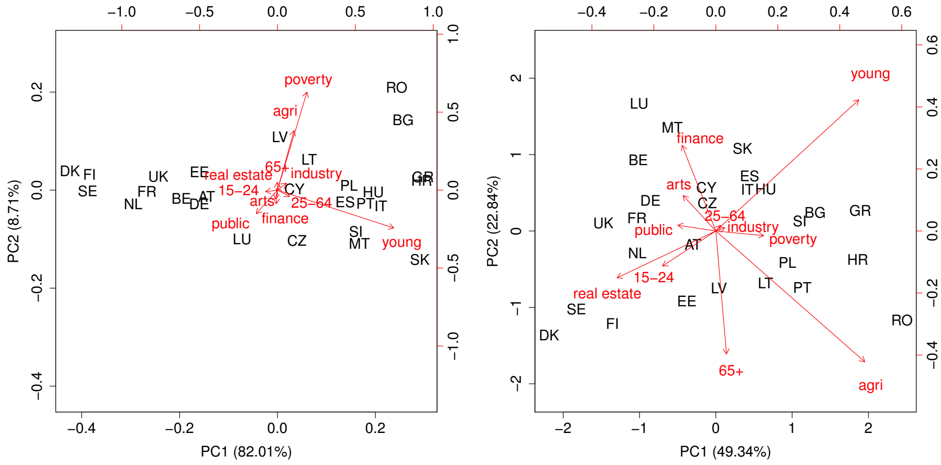

Figure 5. Standard biplot of employed people by economic activity and age with external non-compositional variables: no transformation used (left), using logit transformation (right).

In conclusion, the construction of the ilr biplot enhances the applicability of the compositional biplot, whereas they visualize the same scores and loadings (up to a scaling constant). Frequently, standard biplots result in misleading representations and their construction does not consider the natural geometric structure of compositional data. As it was shown in the examples, the ilr biplot usually yields more reasonable results.

6. Discussion

The multivariate data structure of compositional data can be analyzed with the clr biplot, i.e. a biplot based on singular value decomposition of the clr-transformed data. Instead of the clr transformation, we considered a specific ilr transformation for each compositional variable. Variable-wise, this yields the same information as clr, up to a scaling constant. However, from an interpretation point of view, the ilr version is more convenient, since each ray in the plot represents an individual orthonormal coordinate with a meaningful interpretation.

The ilr version of the biplot has the additional advantage that it is possible to reasonably combine compositional data with other compositions, and/or with non-compositional (external) variables. The idea is that each composition is ilr-transformed, the results are combined, and then external variables merged. We have shown how the relations between the variables of different compositions, relations to external variables, and relations to the observations can be interpreted.

As in the non-compositional case, a proper preprocessing of external variables should be considered. It has been shown on real examples that the most convenient transformations are logit transformation for percentage data and simple scaling for variables containing absolute values. Possibly also the log-transformation can be applied, when the effect of relative scale of the original variable needs to be suppressed [18]. A scaling of the compositions is not necessary since the logratio transformations are invariant with respect to scaling.

It has been shown on practical real-world examples that the ilr biplot provides a more reasonable representation of the data structure than standard biplots since it captures the different geometrical features of compositional data. As in the usual case, a proper interpretation depends also on the explained proportion of variance. The higher variance, the better the ilr biplot reveals the real multivariate data structure. It is of course possible to show an ilr biplot not only for the first two components, but also for higher-order pairs.

In further research, a robust version of the ilr biplot can be considered and constituted, based on robust PCA for compositional data [10]. A robustified version will be less sensitive to outlying observations.

Acknowledgment

This work was supported 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).

References

[1] Aitchison J. The statistical analysis of compositional data. London, UK: Chapman and Hall; 1986.

[2] Egozcue JJ, Pawlowsky-Glahn V. Simplicial geometry for compositional data. In: Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V, editors. Compositional data analysis in the geosciences: From theory to practice. London, UK: Geological Society Publishing House; 2006. p. 67–77.

[3] Buccianti A. Is compositional data analysis a way to see beyond the illusion? Computers & Geosciences. 2013;50:165–173.

[4] Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V, editors. Compositional data analysis in the geosciences: From theory to practice. Special publications ed. 264; Geological Society; 2006.

[5] Egozcue J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. Mathematical Geology. 2003;35:279–300.

[6] Egozcue J, Pawlowsky-Glahn V. Groups of parts and their balances in compositional data analysis. Mathematical Geology. 2005;37:795–828.

[7] Pawlowsky-Glahn V, Buccianti A. Compositional data analysis: Theory and applications. Chichester: Wiley; 2011.

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

[9] Daunis-i-Estadella J, Thió-Henestrosa S, Mateu-Figueras G. Including supplementary elements in a compositional biplot. Computers & Geosciences. 2011;37:696–701.

[10] Filzmoser P, Hron K, Reimann C. Principal component analysis for compositional data. Environmetrics. 2009;20:621–632.

[11] Filzmoser P, Hron K. Robustness for compositional data. In: Becker C, Fried R, Kuhnt S, editors. Robustness and complex data structures. Heidelberg, DE: Springer; 2013. p. 117–131.

[12] Hron K, Filzmoser P. Exploring compositional data with the robust compositional biplot. In: Carpita M, Brentari E, Qannari E, editors. Advances in latent variables. Heidelberg: Springer; 2014. p. 219–226.

[13] Jackson JE. A user's guide to principal components. New York: Wiley & Sons; 1991.

[14] Gabriel KR. The biplot graphic display of matrices with application to principal component analysis. Biometrika. 1971;58:453–467.

[15] Fišerová E, Hron K. On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences. 2011;43:455–468.

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

[17] Filzmoser P, Hron K, Reimann C. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Science of the Total Environment. 2009;407:6100–6108.

[18] Mateu-Figueras G, Pawlowsky-Glahn V. A critical approach to probability laws in geochemistry. Mathematical Geosciences. 2008;40(5):489–502.

Appendix A. Abbreviations

Table A1. Codes representing names of German states

Abbreviation	State
BB	Brandenburg
BE	Berlin
BY	Bavaria
BW	Baden-Württemberg
HB	Bremen
HE	Hesse
HH	Hamburg
MV	Mecklenburg-Vorpommern
NI	Lower Saxony
NW	North Rhine-Westphalia
RP	Rhineland-Palatinate
SH	Schleswig-Holstein
SL	Saarland
SN	Saxony
ST	Saxony-Anhalt
TH	Thuringia

Table A2. Codes representing names of European countries

Abbreviation	Country
AT	Austria
BE	Belgium
BG	Bulgaria
CY	Cyprus
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GR	Greece
HR	Croatia
HU	Hungary
IT	Italy
LT	Lithuania
LU	Luxemburg
LV	Latvia
MT	Malta
NL	Netherlands
PL	Poland
PT	Portugal
RO	Romania
SE	Sweden
SI	Slovenia
SK	Slovakia
UK	United Kingdom