HHS Public Access

Author manuscript

J Geochem Explor. Author manuscript; available in PMC 2021 January 27.

Published in final edited form as:

J Geochem Explor. 2018 November ; 194: 120–133. doi:10.1016/j.gexplo.2018.07.014.

Exploration of geochemical data with compositional canonical biplots

Jan Graffelman[a,b,*], **Vera Pawlowsky-Glahn**[c], **Juan José Egozcue**[d], **Antonella Buccianti**[e,f]

[a]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, Barcelona 08028, Spain

[b]Department of Biostatistics, University of Washington, UW Tower, 15th Floor, 4333 Brooklyn Avenue NE, Seattle 98105, WA, USA

[c]Department of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona Campus Montilivi, Edifici P4, Girona E-17003, Spain

[d]Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya, Jordi Girona Salgado 1-3, Edifici C2, Barcelona E-08034, Spain

[e]Department of Earth Sciences, University of Florence, Via G. La Pira 4, 50121 Firenze, Italy

[f]IGG - CNR Pisa, Italy

Abstract

The study of the relationships between two compositions is of paramount importance in geochemical data analysis. This paper develops a compositional version of canonical correlation analysis, called CoDA-CCO, for this purpose. We consider two approaches, using the centred log-ratio transformation and the calculation of all possible pairwise log-ratios within sets. The relationships between both approaches are pointed out, and their merits are discussed. The related covariance matrices are structurally singular, and this is efficiently dealt with by using generalized inverses. We develop compositional canonical biplots and detail their properties. The canonical biplots are shown to be powerful tools for discovering the most salient relationships between two compositions. Some guidelines for compositional canonical biplot construction are discussed. A geochemical data set with X-ray fluorescence spectrometry measurements on major oxides and trace elements of European floodplains is used to illustrate the proposed method. The relationships between an analysis based on centred log-ratios and on isometric log-ratios are also shown.

Keywords

Biplot; Biplot link; Biplot ray; Canonical loadings; Canonical weights; Floodplain sediment; Generalized least squares; Goodness-of-fit; Log-ratio transformation; Moore-Penrose inverse; X-ray fluorescence spectrometry

[*]Corresponding author at: Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, Barcelona 08028, Spain. jan.graffelman@upc.edu (J. Graffelman).

Appendix C. Supplementary Figures

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gexplo.2018.07.014.

1. Introduction

Many geological investigations concern compositional data sets, which are characterized by components that form part of a whole. Classical examples are the mineral composition of rocks, the oxide and trace composition of sediments and the chemistry of water and natural gases. The corresponding compositions typically contain more than two parts, and the data are therefore inherently of multivariate nature. In several cases, due to analytical requirements for detecting components with different properties, several compositional datasets are obtained from the same whole sample. Again, compositions can be associated with different portions of the same sample when the partition is expected to give some sense to the investigation. Particularly interesting in geochemistry is the development of graphical and numerical methods able to associate trace elements to major and minor components when considered as different (sub)compositions. In fact trace elements tend to follow the behaviour of major and minor components with coherent properties, and the identification of clear associations could help to point out the dynamics of natural processes for different concentration scales but characterized parallel paths. Log-ratio principal component analysis (Aitchison, 1983) has become a standard multivariate technique in compositional data analysis (CoDA), and is often one of the first tools used to explore a compositional data set (e.g. Otero et al., 2005; Tolosana-Delgado et al., 2005). Specific compositional biplots (Aitchison, 1990; Aitchison and Greenacre, 2002) have been proposed that allow efficient visualization of geochemical data sets.

Compositional data often go together with other variables that can appear as predictors of the compositions, or that can appear as responses explained by compositions. In this paper we address the situation where there are two sets of variables which are both geochemical compositions, and our goal is to study the relationships between the two sets by means of a canonical correlation analysis (CCO). The CCO of compositional data has been previously addressed by several authors (Aitchison, 1986a, Section 14.4; Reyment and Savazzi, 1999, Chapter 6; Mateu-Figueras et al., 2016), who used the additive log-ratio transformation. In this paper, we use the centred log-ratio transformation and deal with structural singularity by using a generalized inverse, the Moore-Penrose inverse. We extend the previous work with a detailed development of compositional canonical biplots and goodness-of-fit statistics.

Over the last decades, compositional data analysis (Aitchison, 1982, 1986a) has experienced a strong development. Scientists have become increasingly aware of the fact that compositional data are special data and this has to be taken into account in any statistical analysis. Recent books by Pawlowsky-Glahn and Buccianti (2011), Van den Boogaart and Tolosana-Delgado (2013), and Pawlowsky-Glahn et al. (2015) show that the analysis of compositional data is an active field of research. It is now clear that compositional data are multivariate data and that the only way to capture the complex dynamics of natural phenomena is to adopt adequate tools as the CoDA ones. However, since in Earth Sciences relationships between (sub)compositions with different types of compounds can be of interest, this item will be here developed from a theoretical and practical point of view. The structure of this paper is as follows. In Section 2 we provide the theory for our compositional version of CCO, hereafter called CoDA-CCO and develop the corresponding compositional biplots. In Section 3 we illustrate our methodology with an artificial example

and with the analysis of a geochemical data set of major oxides and trace elements measured in European floodplain sediments. Floodplain sediments are represented by a continuum of sediment types that range from clay- to gravel-size particles, including both terrigenous and organic deposits. Their importance is related to economically relevant reservoirs of oil, natural gas, and water, and as a fundamental tool to provide detailed records of past and present environments. Finally, a discussion completes the paper.

Canonical correlation analysis (CCO) is an important classical multivariate method developed by Hotelling (1935, 1936) dedicated to the study of relationships between two sets of multiple variables, an X- set and a Y-set. Statistics courses on multivariate analysis usually cover the method, and textbooks in the field typically dedicate a chapter to the technique (Anderson, 1984, Mardia et al., 1979, Johnson and Wichern, 2002, Dillon and Goldstein, 1984, Manly, 1989). The monograph by Gittins (1985) is entirely dedicated to canonical analysis. Canonical correlation analysis offers a unifying theoretical framework, since several multivariate techniques are particular cases of it. CCO is a generalization of multiple regression with more than one response variable (Mardia et al., 1979; Gittins, 1985), relates to multivariate analysis of variance (MANOVA) and discriminant analysis when one of the two sets of variables consists of indicator variables (Gittins, 1985, Section 4.6), and is also intricately related to correspondence analysis (Greenacre, 1984, Section 4.4) when both the X variables and the Y variables consist of indicator variables. CCO has been greatly enhanced by the development of biplots that efficiently depict the correlation structure of the variables. The method provides a generalized least squares approximation to the between-set correlation matrix. Haber and Gabriel (1976), Ter Braak (1990) and Graffelman (2005) have shown that canonical correlation analysis allows the construction of a biplot of the between-set correlation matrix. The biplot greatly helps the interpretation of the output of a canonical correlation analysis.

2. Theory

In this section we establish our notation, briefly summarize classical CCO and then develop a compositional version of CCO.

2.1. Classical CCO

We consider one set containing p predictor variables (X-variables) and a second set containing q criterion variables (Y-variables). Both sets are assumed real, that is, the sample space is the ordinary Euclidean space. The Y-variables can be thought of as response variables, though not necessarily so, as the analysis treats X and Y in a symmetric fashion. The main aim of a CCO is to search for linear combinations $\mathbf{U} = \mathbf{X}_c\mathbf{A}$ and $\mathbf{V} = \mathbf{Y}_c\mathbf{B}$ of the column-mean centred variables in $\mathbf{X}_c$ and $\mathbf{Y}_c$ that have maximal correlation. The coefficient matrices $\mathbf{A}$ and $\mathbf{B}$ are known as the *canonical weights* or the *canonical coefficients*, and the constructed linear combinations are known as the *canonical variables* (also termed *canonical variates* by some authors). The solution of a CCO is efficiently computed by using the singular value decomposition (s.v.d.) of the transformed between-set covariance matrix. In particular, the canonical coefficients and correlations can be obtained by the s.v.d. of

$$\mathbf{K} = \mathbf{S}_{xx}^{-1/2}\mathbf{S}_{xy}\mathbf{S}_{yy}^{-1/2} = \widetilde{\mathbf{A}}\mathbf{D}\widetilde{\mathbf{B}}', \tag{1}$$

where $\mathbf{S}_{xx}$, $\mathbf{S}_{yy}$ and $\mathbf{S}_{xy}$ are the sample covariance matrices of the X-variables, the Y-variables, and the between-set covariances, respectively. Matrix $\widetilde{\mathbf{A}}$ is a $p \times r$ orthonormal matrix of left singular vectors ($(\widetilde{\mathbf{A}}'\widetilde{\mathbf{A}} = \mathbf{I}_r)$) and matrix $\widetilde{\mathbf{B}}$ is a $q \times r$ orthonormal matrix of right singular vectors ($\widetilde{\mathbf{B}}'\widetilde{\mathbf{B}} = \mathbf{I}_r$). Diagonal matrix $\mathbf{D}$ is of rank r ($r = \min(p,q)$) and contains the canonical correlations in non-increasing order of magnitude (Gittins, 1985, Section 2.3.2). The canonical coefficients are related to the left and right singular vectors by

$$\mathbf{A} = \mathbf{S}_{xx}^{-1/2}\widetilde{\mathbf{A}}, \quad \mathbf{B} = \mathbf{S}_{yy}^{-1/2}\widetilde{\mathbf{B}}. \tag{2}$$

The canonical coefficients are normalized so that $\mathbf{A}'\mathbf{S}_{xx}\mathbf{A} = \mathbf{I}_r$ and $\mathbf{B}'\mathbf{S}_{yy}\mathbf{B} = \mathbf{I}_r$ and, consequently, the canonical variables are standardized variables,

$$(1/n)(\mathbf{X}_c\mathbf{A})'\mathbf{X}_c\mathbf{A} = \mathbf{A}'\mathbf{S}_{xx}\mathbf{A} = \mathbf{I}_r, \quad (1/n)(\mathbf{Y}_c\mathbf{B})'\mathbf{Y}_c\mathbf{B} = \mathbf{B}'\mathbf{S}_{yy}\mathbf{B} = \mathbf{I}_r.$$

The singular value decomposition in Eq. (1) shows that we do a weighted least squares approximation of given rank to the between-set covariance matrix $\mathbf{S}_{xy}$. Row markers ($\mathbf{F}$) and column markers ($\mathbf{G}$) for the biplot can be obtained by:

$$\mathbf{F}_p = \mathbf{S}_{xx}\mathbf{A}\mathbf{D}, \quad \mathbf{G}_s = \mathbf{S}_{yy}\mathbf{B}. \tag{3}$$

We use the subindices p and s to indicate “principal” and “standard” coordinates, respectively. This convenient terminology was proposed by Greenacre (1984) in the context of correspondence analysis, and was previously used in CCO by Graffelman (2005); it serves to distinguish the different biplot scalings. The principal coordinates are characterized by the presence of diagonal matrix $\mathbf{D}$ in the formula, whereas standard coordinates refer to coordinates without matrix $\mathbf{D}$ in their formula. An alternative scaling for the biplot is to have rows in standard coordinates, and columns in principal coordinates:

$$\mathbf{F}_s = \mathbf{S}_{xx}\mathbf{A}, \quad \mathbf{G}_p = \mathbf{S}_{yy}\mathbf{B}\mathbf{D}. \tag{4}$$

In CCO all these sets of coordinates for biplots can be interpreted as covariances. The principal coordinates $\mathbf{F}_p$ are cross covariances between X-variables and canonical Y-variables. The standard coordinates $\mathbf{G}_s$ are the covariances between canonical Y-variables and the original Y-variables. In the same manner, the standard coordinates $\mathbf{F}_s$ are intra-set covariances for the X-variables and the canonical X-variables, and the principal coordinates $\mathbf{G}_p$ are cross covariances between Y-variables and X-variables. This is shown by the following set of equations,

$$\mathbf{S}_{xu} = \frac{1}{n}\mathbf{X}_c'\mathbf{U} = \mathbf{S}_{xx}\mathbf{A} = \mathbf{F}_s, \tag{5}$$

$$\mathbf{S}_{xv} = \frac{1}{n}\mathbf{X}_c'\mathbf{V} = \mathbf{S}_{xy}\mathbf{B} = \mathbf{S}_{xx}\mathbf{A}\mathbf{D} = \mathbf{F}_p, \tag{6}$$

$$\mathbf{S}_{yu} = \frac{1}{n}\mathbf{Y}_c'\mathbf{U} = \mathbf{S}_{yx}\mathbf{A} = \mathbf{S}_{yy}\mathbf{B}\mathbf{D} = \mathbf{G}_p, \tag{7}$$

$$\mathbf{S}_{yv} = \frac{1}{n}\mathbf{Y}_c'\mathbf{V} = \mathbf{S}_{yy}\mathbf{B} = \mathbf{G}_s\,. \tag{8}$$

A biplot of the between-set covariance matrix $\mathbf{S}_{xy}$ can be obtained as $\mathbf{F}_p\mathbf{G}_s'$ in Eq. (3) or as $\mathbf{F}_s\mathbf{G}_p'$ in Eq. (4). Numerical output of a CCO typically also includes the *canonical loadings*. The canonical loadings are the correlations between the original variables and the canonical variables and can be used to interpret the canonical variables. In a correlation-based CCO the previous covariance expressions (Eqs. (5)-(8)) are in fact equal to the canonical loadings. If a covariance-based CCO is used, then the loadings are obtained by premultiplying the previous covariances with the inverse of a diagonal matrix containing the standard deviations ($\mathbf{D}_{sx}$,$\mathbf{D}_{sy}$), so that the loadings are obtained by:

$$\mathbf{R}_{xu} = \mathbf{D}_{sx}^{-1}\mathbf{S}_{xu} = \mathbf{D}_{sx}^{-1}\mathbf{F}_s, \tag{9}$$

$$\mathbf{R}_{xv} = \mathbf{D}_{sx}^{-1}\mathbf{S}_{xv} = \mathbf{D}_{sx}^{-1}\mathbf{F}_p, \tag{10}$$

$$\mathbf{R}_{yu} = \mathbf{D}_{sy}^{-1}\mathbf{S}_{yu} = \mathbf{D}_{sy}^{-1}\mathbf{G}_p, \tag{11}$$

$$\mathbf{R}_{yv} = \mathbf{D}_{sy}^{-1}\mathbf{S}_{yv} = \mathbf{D}_{sy}^{-1}\mathbf{G}_s\,. \tag{12}$$

Note that, in order to obtain the loadings, post-multiplication by the inverse of the standard deviation of the canonical variables is not needed, as the latter are already standardized by virtue of the normalization constraints on the singular vectors in Eq. (1). This shows that the correlation-based and covariance-based biplots are almost identical, and that the only difference is a rescaling of the variable vectors. In correlation-based CCO biplots all variable vectors will be within the unit circle. In covariance-based CCO biplots, variable vectors can be outside the unit circle. The angles between the variable vectors are the same in both types of analysis, and the goodness-of-fit of $\mathbf{S}_{xy}$ equals the goodness of fit of $\mathbf{R}_{xy}$.

We briefly summarize the main measures of goodness-of-fit in canonical analysis. The goodness-of-fit of the between-set covariance matrix in a k-dimensional biplot is given by

$$\frac{\sum_{i=1}^{k} d_i^2}{\sum_{i=1}^{\min(p,q)} d_i^2}\,. \tag{13}$$

Matrix $\mathbf{X}_c$ is approximated by the inner products between the rows of $\mathbf{U}$ and the columns of $\mathbf{F}_s$. If the X-set is the smaller set ($p \leq q$) then, in the full space of the solution, $\mathbf{X}_c$ is perfectly recovered, because

$$\mathbf{U}\mathbf{F}_s' = \mathbf{X}_c\mathbf{A}\mathbf{F}_s' = \mathbf{X}_c\mathbf{A}\mathbf{A}'\mathbf{S}_{xx} = \mathbf{X}_c\,. \tag{14}$$

In a k-dimensional biplot $\mathbf{X}_c$ is approximated by $\widehat{\mathbf{X}} = \mathbf{U}_{(k)}\mathbf{F}_{s(k)}{}'$. The total variance of the $\mathbf{X}$ variables accounted for by a given number of canonical $\mathbf{U}$ variables, called the *adequacy coefficient* (Thompson, 1984), is

$$\frac{\mathrm{tr}(\mathbf{S}_{xu(k)}\mathbf{S}_{xu(k)}{}')}{\mathrm{tr}(\mathbf{S}_{xx})}, \tag{15}$$

where $\mathbf{S}_{xu(k)}$ refers to the covariance matrix between X variables and the first k canonical X variables. Note that the adequacy coefficient is not scale-invariant under standardization of the original variables. With standardized variables, the adequacy coefficients are obtained by changing all covariances in Eq. (15) by correlations, and this reduces to $\frac{1}{p}\sum_{i=1}^{p}\sum_{j=1}^{k} r^2(x_i, u_j)$. The latter is also the average of the coefficients of determination (R^2) obtained by regressing all X variables onto k canonical variables. Likewise, the inner products of $\mathbf{U}$ with the Y-variables in principal scaling approximate the Y-measurements in the full space, and we have

$$\mathbf{U}\mathbf{G}_p{}' = \mathbf{U}\mathbf{U}'\mathbf{Y}_c(1\,/\,n) = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'\mathbf{Y}_c = \widehat{\mathbf{Y}}, \tag{16}$$

which can be interpreted as the fitted values obtained in a regression of $\mathbf{Y}_c$ onto the canonical X-variables. In general, it will not be possible to exactly recover the measurements of the variables in principal scaling, even if we use the full space of the solution. The amount of explained variation of the Y-variables in a k-dimensional solution, known as the *redundancy coefficient* (Stewart and Love, 1968), is

$$\frac{\mathrm{tr}(\mathbf{S}_{yu(k)}\mathbf{S}_{yu(k)}{}')}{\mathrm{tr}(\mathbf{S}_{yy})}\,. \tag{17}$$

The redundancy coefficients are neither scale-invariant under standardization of the original variables. With standardized variables, the redundancy coefficients are obtained by changing the covariances in Eq. (17) by correlations, and reduce to $q^{-1}\sum_{i=1}^{q}\sum_{j=1}^{k} r^2(y_i, u_j)$. Analogous adequacy and redundancy coefficients can be calculated for the canonical Y variables.

In conclusion, classical CCO basically provides a biplot of the between-set covariance or correlation structure, in which the original observations are absent. Classical biplots made by principal component analysis (Gabriel, 1971) provide more information, since they do not only represent the variables, but also the original samples. In previous work, Graffelman (2005) has shown that it is possible to represent the original samples in the CCO biplot by

using regression results for the representation of supplementary information (Graffelman and Aluja-Banet, 2003). If samples are fitted to the biplot by generalized least squares, it is particularly simple to represent them in the biplot: the $\mathbf{F}_s\mathbf{G}_p'$ biplot should be overplotted with the canonical X variables, and the $\mathbf{F}_p\mathbf{G}_s'$ biplot should be overplotted with the canonical Y variables. These results are of particular relevance for a compositional version of CCO, as they will allow the representation of the original compositions in the CCO biplot (see Section 2.2). The corresponding plots could be termed *triplots* because they represent three entities: X variables, Y variables and data points. The term triplot stems from ecological multivariate analysis, as triplots are commonly made in canonical correspondence analysis (Ter Braak, 1986; Ter Braak and Smilauer, 2002) and redundancy analysis (Ter Braak and Looman, 1994).

We finish this section with a few remarks on the scaling of the original data matrix, as this is also relevant for the compositional analysis that is to follow. One can decide to perform CCO using covariance matrices (as outlined above), or using correlation matrices. A correlation based analysis is possible by simply standardizing the data matrices prior to the analysis, e.g. dividing the columns of **X** and **Y** by their respective standard deviations. CCO is, to a large extent, invariant to such standardization. Canonical correlations, canonical variables, and canonical loadings will all be the same in a covariance-based and a correlation-based analysis. In this sense CCO differs from principal component analysis (PCA), since it is well known that a PCA of the centred data matrix is different from the PCA of the standardized data matrix, giving rise to two “variants” of PCA. The main difference between a covariance based CCO and a correlation based CCO concerns the biplot: the first produces a biplot of the between-set covariance matrix, whereas the latter produces a biplot of the between-set correlation matrix. The goodness-of-fit of these matrices will be the same in both approaches. Finally, the goodness-of-fits of the original data matrices, as expressed by the adequacy and redundancy coefficients, are different in a covariance-based and correlation-based analysis as explained above.

2.2. Compositional CCO

In the development in the previous section, **X** and **Y** typically stand for matrices of quantitative real variables. We now consider **X** and **Y** to be matrices with n compositions in their rows, and having D_x and D_y parts (columns) respectively. Recall that compositional data can be defined as strictly positive vectors for which the information of interest is in the ratios between components. There are several ways to perform a CoDA-CCO, depending on how the compositions are transformed. One can use the additive, the centred or the isometric log-ratio transformation, or one can also use the matrices with all pairwise log-ratios of the X-set and the Y-set. The different approaches are largely equivalent, though the biplots obtained will be different. We develop two approaches to CoDA-CCO in the corresponding subsections below, using the canonical analysis of the clr transformed compositions (Section 2.2.1), and, largely equivalently, the canonical analysis of all pairwise log-ratios of the X-set and the Y-set (Section 2.2.2). Both these transformations lead to a visualization of the pairwise log-ratios which form the most simple representation of the data, and from which more complex ratios can be build. The clr-based approach is also the usual approach taken in log-ratio principal component analysis (Pawlowsky-Glahn et al., 2015; Aitchison and

Greenacre, 2002). Some invariance properties for the isometric log-ratio transformation are derived in Appendix A.

2.2.1. The centred log-ratio (clr) approach—We consider the centred log-ratio transformation (clr) of a composition **x** given by

$$\mathrm{clr}(\mathrm{x}) = \left[\ln\left(\frac{x_1}{g_m(\mathbf{x})}\right), \ln\left(\frac{x_2}{g_m(\mathbf{x})}\right), \cdots, \ln\left(\frac{x_D}{g_m(\mathbf{x})}\right)\right], \tag{18}$$

where $g_m(\mathbf{x})$ is the geometric mean of the components of the composition **x.** Let $\mathbf{X}_\ell$ be the log transformed compositions, that is $\mathbf{X}_\ell = \ln(\mathbf{X})$ with the natural logarithmic transformation applied element-wise. The clr transformed data can be obtained by just centring the rows of this matrix, using the centring matrix $\mathbf{H}_r = \mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}'$, with D equal to D_x or D_y as corresponds. Then

$$\mathbf{X}_{\mathrm{clr}} = \mathbf{X}_\ell \mathbf{H}_r, \qquad \mathbf{Y}_{\mathrm{clr}} = \mathbf{Y}_\ell \mathbf{H}_r\,. \tag{19}$$

These clr transformed data matrices have the same dimensions as **X** and **Y**. The columns of $\mathbf{X}_{\mathrm{clr}}$ and $\mathbf{Y}_{\mathrm{clr}}$ are subject to a zero sum constraint because $\mathbf{H}_r\mathbf{1} = \mathbf{0}$. The column rank of these matrices is, in the absence of additional linear constraints, equal to $D_x - 1$ and $D_y - 1$, respectively. We now column-centre the clr transformed data, producing data matrices that have column means that are zero,

$$\mathbf{X}_{c\mathrm{clr}} = \mathbf{H}_c\mathbf{X}_{\mathrm{clr}} = \mathbf{H}_c\mathbf{X}_\ell\mathbf{H}_r, \qquad \mathbf{Y}_{c\mathrm{clr}} = \mathbf{H}_c\mathbf{Y}_{\mathrm{clr}} = \mathbf{H}_c\mathbf{Y}_\ell\mathbf{H}_r, \tag{20}$$

where $\mathbf{H}_c$ is the idempotent centring matrix $\mathbf{H}_c = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}'$. Thus, $\mathbf{X}_{c\mathrm{clr}}$ and $\mathbf{Y}_{c\mathrm{clr}}$ have zero row means due to the subtraction of the geometric means, and zero column means due to centring operation $\mathbf{H}_c$. We propose to use $\mathbf{X}_{c\mathrm{clr}}$ and $\mathbf{Y}_{c\mathrm{clr}}$ as the input matrices for a classical CCO described in Section 2.1. Due to the zero row sum constraint, the covariance matrices of $\mathbf{X}_{c\mathrm{clr}}$ and $\mathbf{Y}_{c\mathrm{clr}}$ are singular. In CCO the covariance (or correlation) matrices of the X and Y variables are inverted. In order to be able to deal with the structural singularity due to the compositional nature of the data, we use a generalized inverse, the Moore-Penrose inverse (Searle, 1982), in order to be able to proceed with the analysis. In CCO with non-singular covariance matrices, the inverse of the square roots of the covariance matrices are needed (Eq. (1)) and these can be obtained from the spectral decomposition of the covariance matrices, in particular

$$\mathbf{S}_{xx}^{-1/2} = \mathbf{W}\mathbf{\Lambda}^{-1/2}\mathbf{W}', \tag{21}$$

where **W** and $\mathbf{\Lambda}$ contain eigenvectors and eigenvalues obtained in the spectral decomposition of $\mathbf{S}_{XX} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}'$. Under singularity of $\mathbf{S}_{XX}$, the Moore-Penrose inverse denoted by $\mathbf{S}_{xx}^{+}$ is obtained by $\mathbf{W}\widetilde{\mathbf{\Lambda}}\mathbf{W}'$, with $\widetilde{\mathbf{\Lambda}} = \mathrm{diag}(1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_{D-1}, 0)$, which satisfies the four Moore-Penrose conditions. Compositional canonical correlation analysis (CoDA-CCO) can then be carried out using the singular value decomposition

$$\mathbf{K} = (\mathbf{S}_{xx}^{+})^{1/2}\mathbf{S}_{xy}(\mathbf{S}_{yy}^{+})^{1/2} = \widetilde{\mathbf{A}}\mathbf{D}\widetilde{\mathbf{B}}'. \quad (22)$$

Due to the compositional nature of the data, the number of dimensions in the solution, the rank of **D**, is now given by $r = \min(D_x - 1, D_y - 1)$. The canonical coefficients are now obtained as

$$\mathbf{A} = (\mathbf{S}_{xx}^{+})^{1/2}\widetilde{\mathbf{A}}, \qquad \mathbf{B} = (\mathbf{S}_{yy}^{+})^{1/2}\widetilde{\mathbf{B}}. \quad (23)$$

The biplot coordinates and the canonical loadings of a CoDA-CCO are now obtained by the same expressions given for the classical analysis in Eqs. (3) and (4) and (5) through (8). Note that the between-set covariance matrix of the clr coordinates $\mathbf{S}_{xy}$ has dimension $D_x \times D_y$, but that the generalized inverses have at most rank $D_x - 1$ and $D_y - 1$, respectively. Consequently, matrix **K** is not full rank, but has at most rank $\min(D_x - 1, D_y - 1)$. We note that computer programs typically produce an s.v.d. where **D** has dimensions $(r + 1) \times (r + 1)$, implying that **D** has a trailing zero on the diagonal, which is consequence of the singularity of the covariance matrices of the clr transformed data. If the s.v.d. in Eq. (22) is conceived that way, the corresponding normalization of the canonical coefficients is affected, and one has that $\mathbf{A}'\mathbf{S}_{xx}\mathbf{A} = \widetilde{\mathbf{I}}$ and $\mathbf{B}'\mathbf{S}_{yy}\mathbf{B} = \widetilde{\mathbf{I}}$, where $\widetilde{\mathbf{I}}$ is a diagonal matrix with r ones and one trailing zero on its diagonal. In the remainder, we conceive **D** of dimension $r \times r$, without trailing zero, such that the canonical coefficient matrices have no trailing column of zeros and can be considered to be full column rank, and satisfy the usual normalizations $\mathbf{A}'\mathbf{S}_{xx}\mathbf{A} = \mathbf{I}_r$ and $\mathbf{B}'\mathbf{S}_{yy}\mathbf{B} = \mathbf{I}_r$. Note that the columns of the matrices of canonical coefficients sum to zero. A justification for this is given in Appendix A. We complete this section enumerating some properties of the compositional canonical biplots obtained. For a treatment of compositional biplots, see also Section 5.4 of Pawlowsky-Glahn et al. (2015).

1. Biplot origin. The origin of the biplot represents the vector of geometric means of the n compositions. In $\mathbf{F}_s\mathbf{G}_p'$ scaling the origin corresponds to the geometric mean vector of the X compositions, whereas in $\mathbf{F}_p\mathbf{G}_s'$ scaling, the origin corresponds to the geometric mean vector of the Y compositions. This can be seen from equations $\mathbf{U} = \mathbf{X}_{c\text{clr}}\mathbf{A}$ and $\mathbf{V} = \mathbf{Y}_{c\text{clr}}\mathbf{B}$. If the double centring operation is applied to the vector of geometric means, a zero vector is obtained, and consequently the values of the canonical variables are zero. At the same time, the origin of the biplot is also the point from which the biplot vectors representing the clr components emanate.

2. Biplot vector (ray) length. Due to the symmetric nature of CCO, we can assume $D_x \leq D_y$ without loss of generality. The length of variable vectors plotted in standard coordinates is, for the smallest composition (the one with fewer parts), in the full space of the solution, equal to the standard deviation of the corresponding clr transformed part. This follows from

$$\mathbf{F}_s\mathbf{S}_s' = \mathbf{S}_{xx}\mathbf{A}\mathbf{A}'\mathbf{S}_{xx} = \mathbf{S}_{xx}, \quad (24)$$

where the last equality follows from the fact that $\mathbf{AA}'$ is the Moore-Penrose inverse of $\mathbf{S}_{xx}$. If a two-dimensional biplot is used as an approximation of the data set, the ray length will underestimate the observed sample standard deviation. It also follows that the length of a biplot vector can never exceed the sample standard deviation of the corresponding clr component. For the larger composition (the one with more parts), we have, in the full r dimensional space

$$\mathbf{G}_s\mathbf{G}_s^{'} = \mathbf{S}_{yy}\mathbf{BB}'\mathbf{S}_{yy} \approx \mathbf{S}_{yy}, \tag{25}$$

where the left hand side has rank r, but $\mathbf{S}_{yy}$ has rank $D_y - 1 \geq r$. Thus, for the larger composition, the length of the rays will be smaller than the standard deviation of the corresponding clr transformed part. Finally, biplot rays of parts that are plotted in principal coordinates are shrunk with respect to the standard coordinates due to the postmultiplication by the canonical correlations (see Eqs. (3) and (4)) and will always fall short of the observed sample standard deviation, and give a worse approximation to it compared with the standard coordinates. This is consistent with previous work (Graffelman, 2005), where it was shown that the within-set covariance matrices are better approximated with biplot vectors in standard coordinates.

3. Inner products between biplot vectors. It follows from Eq. (24) that the inner product between two biplot vectors of the *same* set (again in the full space, using standard coordinates, and correspondingly the set with the smaller composition) equals the covariance of the corresponding clr components. Inner products of biplot vectors *between* subsets (one set in standard and the other set in principal coordinates) approximate the between-set covariance matrix of clr transformed parts. This is justified by

$$\mathbf{F}_s\mathbf{G}_p^{'} = \mathbf{S}_{xx}\mathbf{ADB}'\mathbf{S}_{yy} = \mathbf{S}_{xy}\,. \tag{26}$$

This approximation is optimal in the generalized least squares sense as guaranteed by the s.v.d in Eq. (22), and it is the same in both biplot scalings, and in fact the focus of the analysis.

4. Cosine of angle between two biplot vectors. The cosine of the angle between the two vectors *within sets* (again referring to the standard coordinates of the smaller composition) equals the sample correlation of the clr components in the full space. In a two-dimensional subspace this will be “approximately so”, being it unknown if the approximation is optimal in some sense. Cosines of angles of biplot vectors *between* subsets will exaggerate the correlations between transformed clr components of the two subsets, even in the full space of the solution. This is because the ray lengths of the larger composition underestimate the standard deviation of the corresponding part (see the previous point 2). Importantly, the approximation to the correlations offered by using cosines depends on the biplot scaling. It is not the same in the $\mathbf{F}_s\mathbf{G}_p{}'$ and the $\mathbf{F}_p\mathbf{G}_s{}'$

scaling. This is because the length of the biplot vectors in the rows of $\mathbf{G}_p$ and $\mathbf{F}_p$ fall short of the corresponding standard deviation to a different extent.

5. Link length. A biplot link is the difference vector of two biplot rays. In CoDA biplot interpretation, the links are very important because they represent the log-ratio of the connected parts. For the composition that is represented in standard coordinates, the length of a link in the full space of the solution equals the standard deviation of the corresponding log-ratio. Let $\mathbf{f}_i$ and $\mathbf{f}_j$ represent the rays of parts i and j respectively (rows of $\mathbf{F}_s$). The squared length of their link is given by

$$\begin{aligned}\|\mathbf{f}_i - \mathbf{f}_j\|^2 &= \mathbf{f}_i'\mathbf{f}_i + \mathbf{f}_j'\mathbf{f}_j - 2\mathbf{f}_i'\mathbf{f}_j \\ &= \mathrm{Var}(\mathrm{clr}(x_i)) + \mathrm{Var}(\mathrm{clr}(x_j)) - 2\mathrm{Cov}(\mathrm{clr}(x_i),\ \mathrm{clr}(x_j)) \\ &= \mathrm{Var}\left(\ln\left(\frac{x_i}{\mathrm{g_m}(\mathbf{x})}\right) - \ln\left(\frac{x_j}{\mathrm{g_m}(\mathbf{x})}\right)\right) = \mathrm{Var}\left(\ln\left(\frac{x_i}{x_j}\right)\right).\end{aligned} \tag{27}$$

Under the considered scaling, the links of the larger composition will necessarily be represented in principal coordinates. Let $\mathbf{g}_i$ and $\mathbf{g}_j$ represent the rays of parts i and j respectively (rows of $\mathbf{G}_p$). The squared length of their link is given by

$$\begin{aligned}\|\mathbf{g}_i - \mathbf{g}_j\|^2 &= \mathbf{g}_i'\mathbf{g}_i + \mathbf{g}_j'\mathbf{g}_j - 2\mathbf{g}_i'\mathbf{g}_j \\ &\approx \mathrm{Var}(\mathrm{clr}(y_i)) + \mathrm{Var}(\mathrm{clr}(y_j)) - 2\mathrm{Cov}(\mathrm{clr}(y_i),\ \mathrm{clr}(y_j)) \\ &= \mathrm{Var}\left(\ln\left(\frac{y_i}{\mathrm{g_m}(\mathbf{y})}\right) - \ln\left(\frac{y_j}{\mathrm{g_m}(\mathbf{y})}\right)\right) = \mathrm{Var}\left(\ln\left(\frac{y_i}{y_j}\right)\right).\end{aligned} \tag{28}$$

This shows there is no corresponding full space result for the length of the links in principal coordinates (note the use of ≈ in the last equation). As argued above, the terms $\mathbf{g}_i'\mathbf{g}_i$ and $\mathbf{g}_j'\mathbf{g}_j$ underestimate the corresponding standard deviation, even in the full space. The principal links will equal the corresponding standard deviations in the full space only in the case of equally sized compositions ($p = q$) and all canonical correlations equal to 1.

6. Inner products between links. Since the focus of the analysis is on relationships between the log-ratios of the two sets, inner products and angles between X and Y links are of interest. Links are vectors of differences, and the inner product between two links corresponding to the log-ratios $\ln(x_i/x_j)$ and $\ln(y_r/y_s)$ is, in full space, the covariance between the two corresponding log-ratios because

$$\begin{aligned}(\mathbf{f}_i - \mathbf{f}_j)'(\mathbf{g}_r - \mathbf{g}_s) &= \mathbf{f}_i'\mathbf{g}_r - \mathbf{f}_i'\mathbf{g}_s - \mathbf{f}_j'\mathbf{g}_r + \mathbf{f}_j'\mathbf{g}_s \\ &= \mathrm{Cov}(\mathrm{clr}(x_i),\ \mathrm{clr}(y_r)) - \mathrm{Cov}(\mathrm{clr}(x_i),\ \mathrm{clr}(y_s)) \\ &\quad - \mathrm{Cov}(\mathrm{clr}(x_j),\ \mathrm{clr}(y_r)) + \mathrm{Cov}(\mathrm{clr}(x_j),\ \mathrm{clr}(y_s)) \\ &= \mathrm{Cov}\left(\ln\left(\frac{y_r}{y_s}\right), \ln\left(\frac{x_i}{x_j}\right)\right).\end{aligned} \tag{29}$$

This equation is exact in the full space and has interesting implications. Since all four clr covariances are optimally approximated in the analysis, the implication is that the covariances between log-ratios of the X set and the Y set are also

optimally approximated. An alternative way to construct a CoDA-CCO biplot is then to depict only links as arrows emanating from the origin and leave the clr components out of the biplot (e.g. see Figs. 2A and 3A in the Examples section, where the links in Fig. 2A are identified as the rays in Fig. 3A), this gives precisely the CoDA-CCO biplot obtained in the pairwise log-ratio approach (see Subsection 2.2.2).

7. Cosines of angles between links. Eqs. (27), (28) and (29) show that, in the full space, cosines of angles between links are "close to" the correlations of the corresponding log-ratios. However, because of the aforementioned inexact nature of Eq. (28), cosines of angles will not equal sample correlations between log-ratios exactly.

Up to this point, CoDA-CCO has been developed using a covariance-based approach, mainly because all clr transformed parts have the same log-ratio scale. This implies that inner products in the CoDA-CCO biplots (Eqs. (24) through (26)) represent covariances between clr transformed parts as well. Covariances are only indicative of the nature of the relationship (direct or indirect) but not about the strength of the observed relationship. For the latter purpose, correlations are far more useful. From the foregoing it is clear that in CoDA-CCO the approximation of the correlations by cosines is problematic for two reasons: first, for being inexact in the full space (when the larger composition is considered, or when principal coordinates are involved), and second, for having no justification that approximations in low-dimensional biplots are optimal. In order to avoid these problems, one might therefore consider to standardize the clr transformed data, such that the inner products in Eqs. (24) through (26) will approximate the correlations. This however, yields a biplot that approximates correlations between clr transformed parts, which do not seem particularly interesting. Note that the covariance on the right hand side of Eq. (29) is *not* converted into a correlation by standardizing the clr data. Potentially more interesting biplots, tightly related to the clr approach exposed here, are obtained in the pairwise log-ratio approach in the next section.

2.2.2. The pairwise log-ratio (plr) approach—An alternative approach to CoDA-CCO is to use the pairwise log-ratios (plr for short) of the X-set and the Y-set, and to submit these to a canonical analysis. First, we define two matrices X_{plr} and Y_{plr} with all possible log-ratios for the X and Y set respectively, having dimensions $n \times \frac{1}{2} D_x(D_x - 1)$ and $n \times \frac{1}{2} D_y(D_y - 1)$ respectively. We column-centre these matrices to obtain

$$\mathbf{X}_{cplr} = \mathbf{H}_c \mathbf{X}_{plr}, \qquad \mathbf{Y}_{cplr} = \mathbf{H}_c \mathbf{Y}_{plr}. \tag{30}$$

CoDA-CCO is now performed by the s.v.d. of the transformed between-set covariance matrices of $\mathbf{X}_{cplr}$ and $\mathbf{Y}_{cplr}$, that is, by applying Eq. (22) to the covariance matrices of the newly defined data matrices. Because of the structural singularity of $\mathbf{S}_{xx}$ and $\mathbf{S}_{yy}$, again the Moore-Penrose inverse of the latter two is used. It is immediately clear that the clr-approach and plr-approach are "equivalent" to a large extent. Any pairwise log-ratio is a linear combination of the clr transformed parts because

$$\ln\left(\frac{x_i}{x_j}\right) = \ln\left(\frac{x_i}{g_m(\mathbf{x})}\right) - \ln\left(\frac{x_j}{g_m(\mathbf{x})}\right). \quad (31)$$

It therefore follows that $\mathbf{X}_{plr}$ and $\mathbf{Y}_{plr}$ have the same rank as $\mathbf{X}_{\text{clr}}$ and $\mathbf{Y}_{\text{clr}}$ respectively, and the number of dimensions with non-zero singular values is the same in both analysis. Moreover, Eq. (29) already showed that the covariances of the plr data are linear combinations of the covariances of the clr data. Canonical correlation analysis is known to be invariant under linear transformations of the data. It is thus clear that the canonical correlations and the canonical variables obtained are the same in both types of analysis. The canonical coefficients are however, not invariant, and this implies, by virtue of Eqs. (3) and (4), that the biplot is affected. In the plr approach, biplots will generally be crowded with more rays, $n \times \frac{1}{2} D_x(D_x - 1)$ and $n \times \frac{1}{2} D_y(D_y - 1)$, respectively, for each set. These biplot vectors now *directly* represent the pairwise log-ratios. In the plr approach, biplot properties are straightforward to infer using the results in Subsection 2.2.1. We express these therefore more concisely, but emphasize some novelties.

1. Biplot origin. The origin of the biplot now represents the mean of each pairwise log-ratio, both for the pairwise log-ratios of the X set and the Y set.
2. Biplot vector (ray) length. The length of a variable vector plotted in standard coordinates is, for the smallest composition, in the full space of the solution, according to Eq. (24) now equal to the standard deviation of the corresponding log-ratio. Correspondingly, ray lengths in standard coordinates for the larger composition will underestimate the standard deviation of the corresponding log-ratio. Also correspondingly, biplot rays of parts plotted in principal coordinates give poorer approximations of the corresponding standard deviations of the log-ratios.
3. Inner products between biplot vectors. Eq. (24) now shows, with again the same conditions (full space, standard coordinates, the smaller composition), that the inner product between two biplot vectors of the *same* set equals the covariance of the corresponding log-ratios. Inner products of biplot vectors *between* subsets approximate the between-set covariance matrix of log-ratios, the latter being optimal in the generalized least squares sense.
4. Cosine of angle between two biplot vectors. The cosine of the angle between the two vectors *within sets* (standard coordinates, the smaller composition, full space) equals the sample correlation between two log-ratios. Cosines of angles of biplot vectors *between* subsets exaggerate the correlations between the log-ratios of the two subsets and depend on the biplots scaling for reasons previously described.
5. Links. A biplot link now becomes the difference vector of two log-ratios. If the two log-ratios share a part, having it both in the numerator, or both in the denominator, the link is another log-ratio because

$$\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{x_i}{x_k}\right) = \ln\left(\frac{x_k}{x_j}\right). \tag{32}$$

Representing this link is superfluous, as the biplot already shows all pairwise log-ratios as vectors emanating from the origin. If the two log-ratios don't share parts, we have

$$\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{x_k}{x_l}\right) = \ln\left(\frac{x_i}{x_k}\right) - \ln\left(\frac{x_l}{x_j}\right) = \ln\left(\frac{x_i x_j}{x_k x_l}\right), \tag{33}$$

showing that the biplot will have identical, duplicated links, to be interpreted as "differences in log-ratios". If the two log-ratios share a part, one having it in the numerator and one having it in the denominator, we have

$$\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{x_k}{x_i}\right) = \ln\left(\frac{x_i^2}{x_j x_k}\right). \tag{34}$$

Eq. (32) is a simple log-ratio, whereas Eqs. (33) and (34) are examples of *balances* (Egozcue and Pawlowsky-Glahn, 2005; Pawlowsky-Glahn et al., 2015). Balances can be very useful and can have substantive interpretation depending on the context of the data being analysed. At this point we refrain from developing inner products and cosines for links in the pairwise approach, and will focus mainly on the rays (pairwise log-ratios) for interpretation.

We argued above that in the clr approach standardization of the data did not seem very useful. In the plr approach, standardization can be highly useful, and it is probably often to be recommended. The reason is that standardization of the pairwise log-ratios now converts Eqs. (24), (25) and (26) into correlation matrices. In particular, Eq. (26) implies the biplot can now efficiently visualize the correlation structure of the pairwise log-ratios, and that optimal low-dimensional approximations to this correlation structure can be obtained. This was not possible in the clr approach given in Section 2.2.1.

3. Examples

In this section we present two examples of a compositional canonical correlation analysis. The first example concerns two synthetic 3-part compositions registered for the same set of subjects. The advantage of this example is that the between-set covariance matrix is of rank two, and that everything can be represented without error in two-dimensional space. The second example is geological and concerns the chemical composition (major oxides and trace elements) of European floodplain sediments.

3.1. Two sets of compositions of three parts

We show 100 observations on two 3-part compositions, **x** and **y**, in the ternary diagrams in Fig. 1. The ternary diagram of the X-set reveals a clear pattern, having an approximately constant x_1/x_2 ratio, whereas the Y-set shows, at first sight, no clear structure. These ternary

plots only reveal marginal information on the X and Y compositions, and are not informative about the relationships between the X-set and the Y-set.

We compute the centred log-ratio transformation of compositions **x** and **y** separately, and perform the clr-based compositional canonical correlation analysis developed in the previous section. Table 1 shows the classical numerical output of a CCO analysis. Initially, we use a covariance-based analysis, because all variables are in a commensurable log-ratio scale.

Table 1 shows that the first canonical correlation is very high, 0.94, implying that the two variable sets share a large part of their variation. All of the variance of clr(**x**) and clr(**y**), also known as total variance of the X and Y compositions, respectively, is accounted for by the two canonical variables, as expected. The goodness-of-fit of the between-set covariance matrix $\mathbf{S}_{xy}$ is also 100 percent, as predicted. Considering only one dimension, it is $0.944^2/(0.994^2 + 0.129^2) = 0.982$. This suggests there is only one important dimension. The cumulative adequacy coefficients ($R^2_{y\,|\,u}$) show that a two-dimensional $\mathbf{F}_s\mathbf{G}_p'$ biplot explains 100% of the total variance of the X composition, and 36.3% of the total variance of the Y composition. Most of the variance of the clr transformed parts is accounted for by the first dimension of the analysis. This dimension accounts for 91.3% of the variance of the X composition and for 35.3% of the variance of the Y composition. The first canonical variate U_1 correlates strongly with all X parts, and V_1 correlates strongly with y_1 and y_2. The second canonical correlation is small, and nonsignificant in a permutation test (see below). Log-ratio CoDA-CCO biplots are shown in various scalings in Fig. 2. Biplots have been overplotted with the canonical variables (multiplied by a single convenient scaling factor, using the rows of matrix **U** in Fig. 2A and C, and the rows of matrix **V** in Fig. 2B and D) in order to represent the original compositions in the biplot. The variable labels X_i, Y_j in the plot actually represent the clr transformed parts. A link between rays i and j *within* a subset represents the corresponding log-ratio $\ln(x_i/x_j)$. The key point of these biplots is to look for *parallel links of each subset that run parallel to a canonical variable* with a high correlation. The canonical variables "channel" the correlation structure of the variables and represent the most correlated feature of the data. Fig. 2A shows parallel links between (clr(x_1), clr(x_2)) and (clr(y_2), clr(y_3)), implying that the log-ratios $\ln(x_1/x_2)$ and $\ln(y_2/y_3)$ are correlated. However, the corresponding link is not parallel to the first canonical variable, and these log-ratios have only weak correlation. Moreover, Fig. 2B, C and D does not show this parallelism, suggesting that it is accidental. More interestingly, Fig. 2A also shows long parallel links through (clr(x_2), clr(x_3)) and through (clr (y_1), clr(y_2)) that run parallel to the first canonical variate, suggesting that the log-ratios $\ln(x_2/x_3)$ and $\ln(y_1/y_2)$ are highly correlated. These interpretations are confirmed by the sample correlations between these log-ratios; $r(\ln(x_1/x_2),\ln(y_2/y_3)) = -0.14$ and $r(\ln(x_2/x_3),\ln(y_1/y_2)) = -0.94$. Correlations inferred from the biplot can be corroborated by making a scatterplot matrix of all possible log-ratios, as is shown in Supplementary Fig. S1. An additional approximately parallel pair of links with some inclination is observed in Fig. 2A between (clr (x_1), clr(x_3)) and (clr(y_1), clr(y_3)). The corresponding log-ratios have a correlation of −0.56. The presence of the samples in the biplot aids interpretation and illustrates the observed correlations: the compositions projecting high onto the link through (clr(x_2), clr(x_3)) also project high onto

the link through (clr(y_1), clr(y_2)) and so confirm the correlated nature of the corresponding log-ratios.

An alternative biplot for the same data, using the $\mathbf{F}_p\mathbf{G}_s'$ scaling from Eq. (3), is shown in Fig. 2B. This biplot explains 81.5% of the variance of the clr transformed X parts, and 100% of the variance of the clr transformed Y parts. The goodness-of-fit of the between-set covariance matrix is the same as in Fig. 2A (100%). However, the biplot in Fig. 2B seems to be the more interesting option if the original compositions are added to the biplot, because overall it accounts for more variability of the clr transformed data. Note that the links corresponding to the log-ratios ln(x_1/x_2) and ln(y_2/y_3) are now not far from orthogonal, whereas in Fig. 2A they were virtually parallel. This shows that one needs to be cautious when interpreting the biplot, and that parallelism of links does not necessarily imply strong correlation of the corresponding log-ratios. Also note that the links corresponding to the log-ratios ln(x_2/x_3) and ln(y_1/y_2) are close to parallel in the direction of the first canonical variate, and that this is observed in *both* biplots A and B in Fig. 2. This is the most salient relationship between the two compositions. Fig. 2B also shows almost horizontal parallel links through (clr(x_1), clr(x_3)) and (clr(y_1), clr(y_2)), and more clearly reveals the correlation between the corresponding log-ratios. Fig. 2C and D shows CoDA-CCO biplots of the same data, but with the clr-data standardized prior to the canonical analysis. In these plots, inner products between the biplot vectors of both sets correspond to correlations between the clr components of each set. These plots resemble Fig. 2A and B, but with rescaled rays. This is precisely what is expected as a consequence of the invariance of CCO under linear transformations. Note that the goodness-of-fit of the between-set covariance matrix and the between-set correlation matrix is the same as expected. However, plots C and D in Fig. 2 do add value to the previous graphs in two ways: firstly, due to the presence of the unit circle it is possible to infer that the clr transformed X components are perfectly represented in Fig. 2C and the Y-parts in Fig. 2D. Secondly, Fig. 2C and D provides optimal approximations of the between-set correlation structure of the clr transformed components, whereas Fig. 2A and B do not. Because of the small size of the miniature example, and because $D_x = D_y = 3$, cosines of angles in Fig. 2A and B do coincide with the between-set sample correlations, but for larger compositions with $D_x \neq D_y$ this will generally not be the case.

CoDA-CCO biplots that are based on the analysis of pairwise log-ratios are shown in Fig. 3. Now, each biplot vector represents a log-ratio. Due to the aforementioned invariance, goodness-of-fit of the covariance and correlation matrices of the log-ratios is the same as in the previous clr-based approach. Note that in Fig. 3A and B, each biplot vector equals the sum or difference of the other two vectors of its set, which is a consequence of Eq. (32).

In all biplots in Fig. 3 the log-ratios ln(x_2/x_3) and ln(y_1/y_2) virtually coincide with the first canonical variate. Indeed, the first canonical variate can be interpreted as the difference between these two log-ratios, and confirms this is the most correlated aspect of the data. Results obtained with standardization of log-ratios shown in Fig. 3C and D do leave angles between vectors unaltered, but this is only because Fig. 3 represents full space results with $D_x = D_y$. Between-set inner products in panels C and D are now correlations and identify ln(y_2/y_3) as uncorrelated with all log-ratios except ln(y_1/y_3).

3.2. The composition of European floodplain sediments: major and minor components versus trace elements

The analysed data base is given by the chemical composition of floodplain sediments and is drawn from the FOREGS Geochemical Baseline Mapping Program initiated in 1998 to provide high quality environmental geochemical baseline data in Europe (http://weppi.gtk.fi/publ/foregsatlas/). The data set consists of 747 samples, stratified by European country and represents an interesting example to test the management of parallel sets of compositions obtained by using different experimental conditions and/or different portions of the same whole sample. A range of elements were determined by wavelength dispersive X-ray fluorescence spectrometry (WD-XRFS) and energy dispersive polarised X-ray fluorescence spectrometry (ED(P)XRFS). The instruments used were Philips PW1480 and PW2400 WD-XRFs, with W and Rh anode X-ray tubes respectively, and a Spectro X-LAB 2000 ED-XRF with a Pd anode X-ray tube. In practice, data for MgO, P_2O_5, K_2O, CaO, TiO_2, V, Cr, MnO, Cs, and Ba was taken from the ED technique; data from the WD technique was used for all other elements. Further details of the full range of elements are given on the FOREGS website. The concentrations, expressed as weight % for 10 major oxides (SiO_2, Al_2O_3, Na_2O, MgO, P_2O_5, K_2O, CaO, TiO_2, MnO, Fe_2O_3) and in ppm for 18 trace elements (V, Cr, Co, Ni, Cu, Zn, Ga, As, Rb, Sr, Zr, Nb, Sn, Cs, Ba, Pb, Th, U), were analysed as two parallel compositions with the aim to point out coherent geochemical behaviours for components characterized by different abundance. XRF spectrometry is one of the most widely used and versatile of all instrumental analytical techniques for bulk chemical analysis of materials in several fields (Fitton, 1997). An XRF spectrometer uses primary radiation from an X-ray tube to excite secondary X-ray emissions from the sample. The radiation emerging from the sample includes the characteristic X-ray peaks of major and trace elements present in the sample. Samples were prepared by mixing with a binder, then pressing into pressed powder pellets. Usually the technical apparatus and standards used for major oxides are not the same for trace elements, so that XRF analysis produces different compositional data sets for the same powdered sample. We applied CoDA-CCO to the XRF data set in order to investigate the relationships between the major and minor oxide compositions (%) and the trace element compositions (ppm). The relationship between major oxides and trace elements has been studied in various contexts, as well as using different techniques, thus remarking its interest (e.g. Tolosana-Delgado and McKinley, 2016). The use of different units in the major oxides (%) (X-composition) and the trace elements (ppm) (Y-composition) can draw the attention of a geologist, since traditional (non-compositional) analyses studying relationships between major oxides and trace elements require to have them in the same units. In the log-ratio approach to CoDA, a multiplicative change of units, like translating % of oxides into ppm, is a perturbation in the simplex, representing a shift or translation of the composition (e.g. Pawlowsky-Glahn et al., 2015). A shift does, as is typical for most statistical procedures, not influence variability measures. The same holds for compositional data (Pawlowsky-Glahn and Egozcue, 2001, Proposition 6); as CoDA-CCO deals with the variability of compositions, the change of units does not influence the results of the analysis. A mathematical demonstration of this statement is given in Appendix B. Our choice about the units of measurement follows the structure of the data of the FOREGS repository and the technical sheets associated with each oxide or element to interpret geochemical behaviour in solid materials. Before doing any biplot interpretation, we first comment the numerical

output of the analysis given in Table 2. This validation is important, because patterns detected in a biplot are unreliable when the overall goodness-of-fit is low, or if the involved variables are poorly represented. The full space of the solution of this data set has 9 dimensions, and Table 2 provides the numerical output for the first three dimensions of the analysis. The first three canonical correlations are high. This means that the two measurement domains, centred log-ratios of oxides and of trace elements, share variation to a large extent. The statistical significance of the canonical correlations was assessed by means of a permutation test. Such a test is performed by keeping one matrix fixed, say **X**, and randomly permuting the rows of **Y**. The permuted data set is analysed by CoDA-CCO, and the canonical correlations are registered. This procedure is repeated 10,000 times and in this way the distribution of the canonical correlations under the null hypothesis of no association between **X** and **Y** is generated. The observed canonical correlations of the original data set are compared against the generated distribution, and a p-value is calculated as the percentage of times the generated values exceed the observed canonical correlations. We found all nine canonical correlations to be highly significant with vanishingly small p-values. Results of the permutation test are given for all nine dimensions in Supplementary Fig. S2. Test results suggest all nine dimensions potentially could have a geological interpretation, though for reasons of space we limit ourselves to interpreting the first two dimensions.

Inspection of Table 2 shows that the oxides SiO_2, K_2O, Fe_2O_3 and TiO_2 and the trace elements V, Co, Rb, Ga, Zr, Nb and Th are important contributors to the first dimension of the analysis. For the second dimension these are the oxides CaO and TiO_2, trace Sr, and to a lesser extent trace elements Zr, Ni, Nb, Ba, Co and Ga. We focus on parts with canonical coefficient above 0.25 in absolute value. Often, though not always, these parts also have large canonical loadings. When interpreting the biplot, we will mainly focus on links involving these components. The biplot of the analysis is shown in Fig. 4A (major oxides in standard coordinates) and B (trace elements in principal coordinates). Plots A and B in Fig. 4 can be overlaid, but are presented separately to avoid an overcrowded display. We again look for links that run parallel to the canonical variables, which are represented by the perpendicular coordinate axes. Such links, representing approximately the standard deviation of the logarithm of ratios, can be expected to have particular strong correlations. The first three canonical correlations are 0.94, 0.89 and 0.81. Numerical output of the CoDA-CCO indicates that the two dimensional biplot accounts for 44% of the variation in the between-set covariance matrix of the clr transformed compositions, accounting, in the scaling used, for 46% of the total variance of the oxide composition, and 23% of the total variance of the trace composition. Log-ratios of the oxides that have large correlation with the first canonical variate are $\ln(K_2O/Fe_2O_3)$, $\ln(K_2O/TiO_2)$, and $\ln(SiO_2/Fe_2O_3)$, while for the trace elements these are ln(Rb/V), ln(Rb/Co), and ln(Ba/Co). The association among K_2O, SiO_2, Rb and Ba for negative values of the first canonical variate has a well defined geochemical meaning. They trace the behaviour of lithofile elements that follow Potassium geochemistry and the relative increase in Silica content (mainly presence of K-Feldspars) in the bedrock nature across Europe. On the other hand, positive values associated with Fe_2O_3, TiO_2, V and Co point out the presence of mafic and ultramafic lithologies (relative decreasing Silica content) as well as mineralizations and presence of clay-rich soil with

relatively high Al_2O_3 contents. The second canonical variate points out the association among Ca and Sr versus that of TiO_2, Fe_2O_3, MnO, Nb, Ga and Ba. This shows the presence of carbonatic lithologies versus the presence of mafic and ultramafic rocks, felsic crystalline rocks or clay-rich soils with high Al_2O_3 contents, as well as the presence of mineralization or pollution (i.e. Ba). We confirmed the relationships between log-ratios inferred from the biplot by making a scatterplot matrix, where the most prominent log-ratios are shown (see Supplementary Fig. S3 for log-ratios related to the first canonical variate). The canonical variates show clusters when represented in a geographical map (see Fig. S5). This shows, for instance, that the second canonical variate is large in the Mediterranean region, which has high values for log-ratios that carry CaO or Sr in the numerator. This result is comparable with the maps of CaO and Sr reported in the FOREGS repository (See weppi.gtk.fi/publ/foregsatlas/text/Ca.pdf or Sr.pdf), and describes the outcropping of calcareous lithologies and potential contributions from anthropogenic activities (addition of phosphate fertilisers or lime). When the canonical correlation analysis is stratified on a country-wise basis (biplots and tables not reported), using countries where a sufficient numbers of samples is available, interesting features emerge that can be related to general geochemical laws. In fact, the major oxides K_2O and Rb are closely related to the first canonical variate across all countries (ignoring Austria and Poland because of small sample size), while Al_2O_3 and Sr are quite related to the second one. The result describes well the geochemical affinity between K_2O and Rb. Both the elements pertaining to group 1 of the periodic table and the Rb^+ ion (ionic radius 152 pm) substitute for K^+ (138 pm) in several minerals, thus tracing its geochemical distribution in outcrops across all Europe. The association between Al_2O_3 and Sr appears to point out sedimentary processes, where the distribution of Sr may be affected by strong adsorption on clay minerals containing Al_2O_3. The 747 samples can be projected onto the biplot in Fig. 4 and this can aid interpretation (Graffelman, 2005). Supplementary Fig. S4 shows this more dense biplot. This plot shows some French samples (top-right) which are relatively high in Sr, MgO and CaO, a set of Spanish samples relatively high in Sr (in the Murcia region in south-east Spain, where there is widespread strontianite mineralization), and a set of Polish samples which are relatively low on the first canonical variate (low K_2O and Rb occurs over the glacial drift covered region extending from north Germany to Poland). However, in general, the samples of the different countries overlap to a large extent.

CoDA-CCO biplots based on a pairwise log-ratio approach are shown in Fig. 5, where oxides (panels A and C, in standard coordinates) and traces (panels B and D, in principal coordinates) are presented separately. Fig. 5A and B shows the covariance-based analysis, whereas Fig. 5C and D shows the correlation-based analysis. Panels A and B should be overlaid for interpretation, and panels C and D too. We emphasize that between-set inner products of the biplot vectors in Fig. 5A and B approximate between-set covariances, whereas between-set inner products of biplot vectors in Fig. 5C and D approximate correlations. Because of the large canonical correlations, there is little difference between the use of standard and principal coordinates. Because there are so many pairwise log-ratios, the pairwise log-ratios in Fig. 5 were filtered by goodness-of-fit, and only those log-ratios that have 75% or more of their variance accounted for are shown. For log-ratios in principal coordinates, this threshold was lowered to 60%, as the goodness-of-fit in this scaling is

typically worse. We summarize the main relationships uncovered by these biplots: $\ln(K_2O/TiO_2)$ and $\ln(K_2O/Fe_2O_3)$ are positively correlated, and have strong negative correlation with two log-ratios involving Rb, ln(Co/Rb) and ln(V/Rb). Samples with high values on the latter two log-ratios have low values on the log-ratios $\ln(K_2O/TiO_2)$ and $\ln(K_2O/Fe_2O_3)$. This is the most salient feature of the dataset uncovered by the first canonical variate. These relationships were also uncovered in the previous clr-based analysis. The second canonical variable is associated with at least six log-ratios that all involve CaO, and at the same time with at least eight traces that all involve Sr. The biplots in Fig. 5 represent in fact two approximately orthogonal sets, if CaO and Sr are consistently placed in the numerators (or denominators) of all involved log-ratios. Most of the high Sr values in floodplain sediments are due to its release from crystalline rocks, due to the weathering of feldspar, and from calcareous rocks, thus showing a strong relationship with Ca.

Note that the results are consistent with the clr-based analysis, where CaO and Sr had the longest biplot vectors and therefore many long links involving these components. The goodness-of-fit of the between-set covariance and correlation matrices in Fig. 5 is 44.4%, and coincides with the goodness-of-fit of the between-set clr covariances in Fig. 4. We note that the analysis in this section might have been performed by using the cation composition instead of the major oxide composition. Conversion of oxides to cations corresponds to a perturbation of the oxide composition. In Appendix B we show that the analysis is invariant under such perturbation, and therefore a cation-based analysis would have given the same results as the ones presented here.

4. Conclusions and discussion

Compositional canonical correlation analysis (CoDA-CCO) has been presented as a technique for analysing two compositional datasets, an X set and a Y set, potentially measuring different kinds of compositions, e.g. X may refer to a microbial composition and Y to the biochemical composition of the same set of subjects. We note that the method presented here has a wider scope of application, because it can also be applied to two subcompositions (that do not share parts) built from the *same* sample. The method will then act as a magnifying glass focusing on the relationships between the two subcompositions, allowing the investigation of geochemical paths on different concentration scales.

Canonical correlation analysis provides the best approximation, in the generalized least squares sense, to the between-set covariance matrix. In the context of compositional data, as treated in this paper, the CoDA-CCO biplot gives the optimal approximation for the between-set covariance matrix of the clr transformed parts in the clr-based approach, and the optimal approximation for the between-set covariance matrix of the pairwise log-ratios in the plr based approach. The same goodness-of-fit is obtained for both covariance matrices, as justified by Eq. (24), and as could be observed in both examples. Biplots in both approaches are fully equivalent if data are not standardized. If all within-set links in the clr-based biplot are “extracted” by calculation of all possible difference vectors, and plotted as vectors emanating from the origin, then the biplot of the pairwise log-ratio approach will be obtained. This property holds for any chosen dimensionality. Because the representation of

the links is explicitly optimized in the pairwise approach, it follows that the links are also optimally displayed in the clr-based biplots. This equivalence is clearly visible by comparing the full space solutions in Fig. 2A and B with Fig. 3A and B, but it also holds for approximate solutions like the ones given in Figs. 4A, B and 5A, B respectively.

The aforementioned equivalence between the clr and plr biplot may suggest the latter to be superfluous, but in our opinion this is not the case for several reasons. First, the clr-based approach is limited in the sense that links always represent pairwise log-ratios. In the plr approach, links between pairwise log-ratios correspond to balances, and the pairwise approach may uncover the existence of balances, or even correlations between balances, so allowing for a richer and more refined analysis. If the analysis is limited to the clr biplot, potentially interesting balances that invoke more than two parts may go unnoticed. Second, in the clr-based approach correlations between log-ratios are not optimally displayed, whereas they can be optimally represented in the pairwise approach. By standardization of the pairwise log-ratios by division by their standard deviation prior to canonical analysis, biplots with an optimal approximation to the correlation structure of the pairwise log-ratios are obtained. In the latter plots, unit circles are illustrative, as the goodness-of-fit of the pairwise log-ratios can be inferred from the ray's length. We note that these considerations carry over to compositional biplots made by principal component analysis as well.

Biplots are not unique, and in practical data analysis it may be daunting to choose the most appropriate plot for representing a given data set. This is particularly true for the rich family of compositional canonical biplots proposed in this paper. The analyst is confronted with at least three decisions in the analysis: a) whether to use a clr or plr based approach, b) to standardize the data prior to analysis or not and c) whether to use standard coordinates for rows and principal coordinates for columns or the other way around. We present some considerations on these issues, hoping this will help analysts to make a sensible choice.

The clr-based approach has the advantage of producing less dense plots having fewer rays. In principle, all pairwise log-ratios are present in this biplot by means of the links. The analyst will have to make the mental effort to search for interesting links, in particular by looking at links that run parallel to canonical variables with a high correlation. Standardization by division by the standard deviation of the clr transformed data, prior to the canonical analysis, will differentially scale the columns of the clr transformed data. This complicates the interpretation of the links, and seems therefore generally not indicated. Regarding the biplot scaling, if there is particular interest in representing the within-set correlation of one set, that set should be represented in standard coordinates to enhance the representation of its correlation structure. If both sets are equally important, then a pragmatic rule is to choose that scaling that explains most of the total variance of the data, as expressed by the adequacy and redundancy coefficients. For instance, biplot A in Fig. 2 of the artificial data in Section 3.1 explains 100% and 36.3% of the variance of the X and Y data in $\mathbf{F}_s\mathbf{G}_p'$ scaling, but respectively 81.5% and 100% in $\mathbf{F}_p\mathbf{G}_s'$ scaling. The latter may be preferred for giving, overall, a better approximation to the transformed data matrices. To safeguard against erroneous interpretations, we recommend always to explore the data using both biplot scalings used in this paper (Eqs. (3) and (4)). Patterns like parallel links that show up in *both* biplots are more likely to be real, and not an artifact due to the projection.

The pairwise log-ratio biplot has the advantage that pairwise log-ratios are directly displayed as rays in the biplot. With large compositions the number of links can be prohibitive, and produce very dense biplots. However, as shown in the geological example in Subsection 3.2, by removing all links with a low goodness-of-fit, these plots can be improved, and salient features of the data can be made visible.

In compositional data analysis, several log-ratio transformations are in use. In particular, the isometric log-ratio (ilr) transformation is increasing in popularity, as it provides Cartesian coordinates to represent the compositions. We show in Appendix A that a CoDA-CCO of the ilr transformed compositions will yield the same canonical correlations and the same canonical variables. An advantage of using ilr coordinates is that the singularity of the covariance matrices is avoided, which frees one from the need to calculate generalized inverses. For biplot construction however, a clr based or plr based analysis seems to be the most straightforward approach to CoDA-CCO. Given that the clr, plr and ilr transformations are linearly related to each other, the proposed compositional biplots could also be derived from a canonical analysis of ilr transformed compositions.

5. Software

Functions that perform compositional canonical analysis have been developed for the R environment (R Core Team, 2017), and are included in the R-package ToolsForCoDa.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants MTM2015-65016-C2-1-R and MTM2015-65016-C2-2-R (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund, by grant 2030_M1488-BUCBASI09 of PROG-GEOBASI TOSCANA, by the University of Florence funds, 2015 and 2016, and by Grant GM075091 from the United States National Institutes of Health.

Appendix

Appendix A

In this appendix we show the invariance of the canonical correlations and the canonical variables when the ilr transformation is used instead of the clr transformation.

The singular value decomposition in Eq. (22) can be rewritten as an equivalent eigenvalue decomposition:

$$(\mathbf{S}_{yy}^{\text{clr}})^{+}\mathbf{S}_{yx}^{\text{clr}}(\mathbf{S}_{xx}^{\text{clr}})^{+}\mathbf{S}_{xy}^{\text{clr}}\mathbf{B} = \mathbf{B}\mathbf{D}^{2} = \mathbf{B}\mathbf{D}_{\lambda}, \tag{35}$$

where $\mathbf{D}_{\lambda}$ contains the eigenvalues (squares of the singular values) of the spectral decomposition. The clr and ilr coordinates are linearly related by the following expressions (Egozcue et al., 2003)

$$\mathbf{X}_{\text{clr}} = \mathbf{X}_{\text{ilr}}\overline{\mathbf{U}}_x \text{ and } \mathbf{X}_{\text{ilr}} = \mathbf{X}_{\text{clr}}\overline{\mathbf{U}}_x', \tag{36}$$

where $\overline{\mathbf{U}}_x$ is a $D_x - 1 \times D_x$ matrix with orthonormal rows, satisfying $\overline{\mathbf{U}}_x\overline{\mathbf{U}}_x' = \mathbf{I}$ and $\overline{\mathbf{U}}_x'\overline{\mathbf{U}}_x = \mathbf{I} - \frac{1}{D_x}\mathbf{1}\mathbf{1}'$. We use the $\overline{\mathbf{U}}_x$ notation in order to follow the usual notation in CoDA (Egozcue et al., 2003), but put a bar in order not to create confusion with the previously defined canonical X variables ($\mathbf{U}$), and use a subindex x to show that it applies to the X composition. Note that $\overline{\mathbf{U}}_x'\overline{\mathbf{U}}_x$ is an idempotent centring matrix. The analogous transformation for the Y variables is given by a $D_y - 1 \times D_y$ matrix $\overline{\mathbf{U}}_y$, as the X and Y set may not have the same number of parts. By substitution we obtain straightforward expressions for the relationships between ilr and clr within-set and between-set covariance matrices of the X and Y compositions, given by:

$$\mathbf{S}_{xx}^{\text{clr}} = \overline{\mathbf{U}}_x'\mathbf{S}_{xx}^{\text{ilr}}\overline{\mathbf{U}}_x,\ \mathbf{S}_{xy}^{\text{clr}} = \overline{\mathbf{U}}_x'\mathbf{S}_{xy}^{\text{ilr}}\overline{\mathbf{U}}_y,\ \mathbf{S}_{yy}^{\text{clr}} = \overline{\mathbf{U}}_y'\mathbf{S}_{yy}^{\text{ilr}}\overline{\mathbf{U}}_y,\ \mathbf{S}_{yx}^{\text{clr}} = \overline{\mathbf{U}}_y'\mathbf{S}_{yx}^{\text{ilr}}\overline{\mathbf{U}}_x. \tag{37}$$

Premultiplication by $\overline{\mathbf{U}}_x$ and postmultiplication by $\overline{\mathbf{U}}_x'$ (or $\overline{\mathbf{U}}_y$ and $\overline{\mathbf{U}}_y'$, as corresponds) allows us to obtain the ilr covariance matrices from the clr covariance matrices:

$$\mathbf{S}_{xx}^{\text{ilr}} = \overline{\mathbf{U}}_x\mathbf{S}_{xx}^{\text{clr}}\overline{\mathbf{U}}_x',\ \mathbf{S}_{xy}^{\text{ilr}} = \overline{\mathbf{U}}_x\mathbf{S}_{xy}^{\text{clr}}\overline{\mathbf{U}}_y',\ \mathbf{S}_{yy}^{\text{ilr}} = \overline{\mathbf{U}}_y\mathbf{S}_{yy}^{\text{clr}}\overline{\mathbf{U}}_y',\ \mathbf{S}_{yx}^{\text{ilr}} = \overline{\mathbf{U}}_y\mathbf{S}_{yx}^{\text{clr}}\overline{\mathbf{U}}_x'. \tag{38}$$

Substituting Eq. (37) in Eq. (35) gives

$$(\overline{\mathbf{U}}_y'\mathbf{S}_{yy}^{\text{ilr}}\overline{\mathbf{U}}_y)^{+}\overline{\mathbf{U}}_y'\mathbf{S}_{yx}^{\text{ilr}}\overline{\mathbf{U}}_x(\overline{\mathbf{U}}_x'\mathbf{S}_{xx}^{\text{ilr}}\overline{\mathbf{U}}_x)^{+}\overline{\mathbf{U}}_x'\mathbf{S}_{xy}^{\text{ilr}}\overline{\mathbf{U}}_y\mathbf{B} = \mathbf{B}\mathbf{D}_\lambda. \tag{39}$$

At this point we note that $\overline{\mathbf{U}}_x$ has rank $D - 1$, and that the rows of $\overline{\mathbf{U}}_x$ are linearly independent. In that case, the Moore-Penrose inverse of $\overline{\mathbf{U}}_x$ is given by

$$\overline{\mathbf{U}}_x^{+} = \overline{\mathbf{U}}_x'(\overline{\mathbf{U}}_x\overline{\mathbf{U}}_x')^{-1} = \overline{\mathbf{U}}_x'.$$

Similarly, we also have $\overline{\mathbf{U}}_y^{+} = \overline{\mathbf{U}}_y'$. Eq. (39) can now be simplified to

$$\overline{\mathbf{U}}_y'(\mathbf{S}_{yy}^{\text{ilr}})^{+}\mathbf{S}_{yx}^{\text{ilr}}(\mathbf{S}_{xx}^{\text{ilr}})^{+}\mathbf{S}_{xy}^{\text{ilr}}\overline{\mathbf{U}}_y\mathbf{B} = \mathbf{B}\mathbf{D}_\lambda. \tag{40}$$

Since the covariance matrices of the ilr coordinates are invertible, premultiplying by $\overline{\mathbf{U}}_y$ this can be rewritten as

$$\mathbf{S}_{yx}^{\text{ilr}}(\mathbf{S}_{xx}^{\text{ilr}})^{-1}\mathbf{S}_{xy}^{\text{ilr}}(\mathbf{S}_{yy}^{\text{ilr}})^{-1}\mathbf{B}^{\text{ilr}} = \mathbf{B}^{\text{ilr}}\mathbf{D}_\lambda. \tag{41}$$

with $\mathbf{B}^{\text{ilr}} = \overline{\mathbf{U}}_y\mathbf{B}$, satisfying $\mathbf{B}^{\text{ilr}'}\mathbf{S}_{yy}^{\text{ilr}}\mathbf{B}^{\text{ilr}} = \mathbf{I}$. Eq. (40) is the eigenvalue-eigenvector decomposition corresponding to a canonical correlation analysis of X and Y compositions in

ilr coordinates. Finally, canonical variables obtained in the clr based and in the ilr based approach will be identical because $\mathbf{V} = \mathbf{H}\mathbf{Y}_{\text{clr}}\mathbf{B} = \mathbf{H}\mathbf{Y}_{\text{ilr}}\overline{\mathbf{U}}_y\mathbf{B} = \mathbf{H}\mathbf{Y}_{\text{ilr}}\mathbf{B}_{\text{ilr}}$. Eq. (40) shows that a clr-based and ilr-based CoDA-CCO yield the same canonical correlations, yield canonical coefficients that are related by a linear transformation, and yield the same canonical variables. Eqs. (40) and (41) show that in the clr-based approach the canonical coefficients of one canonical variate (columns of matrices **A** and **B**) sum to zero. Since $\overline{\mathbf{U}}$ satisfies $\overline{\mathbf{U}}\overline{\mathbf{U}}' = \mathbf{I}$ and $\overline{\mathbf{U}}'\overline{\mathbf{U}} = \mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}'$ we have that $\mathbf{1}'\overline{\mathbf{U}}' = \mathbf{0}'$. From Eq. (40) it follows that the columns of **B** sum to zero. By using a spectral decomposition analogous to Eq. (35) with **A** as eigenvectors, the same property can also be shown for **A**.

Appendix B

In this appendix we show that the results of the analysis are invariant with respect to the multiplication of the parts of the compositions by a scalar, using a different scalar for each part. This operation is, when followed by closure, known as a perturbation in CoDa (Aitchison, 1986b).

Let $\mathbf{x} = (x_1, x_2, \ldots, x_D)'$ be the original composition, written as a column vector. If, for instance, oxides are to be translated into cations, each element of the composition is multiplied by a corresponding constant a_i. We so obtain a new data vector $\widetilde{\mathbf{x}}$, given by $\widetilde{\mathbf{x}} = (a_1x_1,\ a_2x_2, \ldots, a_Dx_D)'$ and we define the coefficient vector $\mathbf{a} = (a_1, a_2, \ldots, a_D)'$. Prior to canonical analysis we log-transform the data followed by a double-centring operation (see Eq. (20)), such that first

$$\begin{aligned}\ln(\widetilde{\mathbf{x}}) &= (\ln(a_1x_1),\ \ln(a_2x_2), \ldots, \ln(a_Dx_D))' \\ &= (\ln(a_1) + \ln(x_1),\ \ln(a_2) + \ln(x_2), \ldots, \ln(a_D) + \ln(x_D))' \\ &= \ln(\mathbf{a}) + \ln(\mathbf{x}) \\ &= \mathbf{a}_\ell + \ln(\mathbf{x}),\end{aligned}$$

where $\mathbf{a}_\ell$ is the log-transformed coefficient vector. In matrix terms, with **X** a matrix having compositions in its rows, and $\mathbf{X}_\ell$ its log-transform, this amounts to

$$\widetilde{\mathbf{X}}_\ell = \mathbf{X}_\ell + \mathbf{A}_\ell,$$

where $\mathbf{A}_\ell = \mathbf{1}\mathbf{a}'_\ell$. Next, we double-centre this matrix, as in Eq. (20), obtaining

$$\widetilde{\mathbf{X}}_{cclr} = \mathbf{H}_c\widetilde{\mathbf{X}}_\ell\mathbf{H}_r = \mathbf{H}_c\mathbf{X}_\ell\mathbf{H}_r + \mathbf{H}_c\mathbf{A}_\ell\mathbf{H}_r = \mathbf{H}_c\mathbf{X}_\ell\mathbf{H}_r = \mathbf{X}_{cclr},$$

with $\mathbf{H}_c = \mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$ and $\mathbf{H}_r = \mathbf{I} - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D'$, as defined previously. The second term on the right hand side vanishes because

$$\mathbf{H}_c\mathbf{A}_\ell\mathbf{H}_r = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\right)\mathbf{1}_n\mathbf{a}'_\ell\left(\mathbf{I} - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D'\right) = (\mathbf{1}_n - \mathbf{1}_n)\mathbf{a}'_\ell\left(\mathbf{I} - \frac{1}{D}\mathbf{1}_D\mathbf{1}_D'\right) = \mathbf{O}.$$

Consequently, the matrix that enters into CCO is the same, whether an elementwise rescaling has been applied or not.

References

Aitchison J, 1982 The statistical analysis of compositional data (with discussion). J. R. Stat. Soc. Ser. B Methodol 44 (2), 139–177.

Aitchison J, 1983 Principal component analysis of compositional data. Biometrika 70 (1), 57–65.

Aitchison J, 1986a The Statistical Analysis of Compositional Data. The Blackburn press, Caldwell, NJ 2003 printing.

Aitchison J, 1986b The Statistical Analysis of Compositional Data. Chapman & Hall.

Aitchison J, 1990 Relative variation diagrams for describing patterns of compositional variability. Math. Geol 22 (4), 487–511.

Aitchison J, Greenacre M, 2002 Biplots of compositional data. J. R. Stat. Soc. Ser. C. Appl. Stat 51 (4), 375–392.

Anderson TW, 1984 An Introduction to Multivariate Statistical Analysis, second edition. John Wiley, New York.

Dillon WR, Goldstein M, 1984 Multivariate Analysis Methods and Applications. John Wiley & Sons, New York.

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

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

Fitton G, 1997 X-ray fluorescence spectrometry In: Gill r. (Ed.), Modern Analytical Geochemistry. Longman, Singapore, pp. 41–66.

Gabriel KR, 1971 The biplot graphic display of matrices with application to principal component analysis. Biometrika 58 (3), 453–467.

Gittins R, 1985 Canonical Analysis. Springer Verlag.

Graffelman J, 2005 Enriched biplots for canonical correlation analysis. J. Appl. Stat 32 (2), 173–188.

Graffelman J, Aluja-Banet T, 2003 Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis. Biom. J 45 (4), 491–509.

Greenacre MJ, 1984 Theory and Applications of Correspondence Analysis. Academic press.

Haber M, Gabriel KR, 1976 Weighted Least Squares Approximation of Matrices and Its Application to Canonical Correlations and Biplot Display Technical report. University of Rochester, Department of statistics.

Hotelling H, 1935 The most predictable criterion. J. Educ. Psychol 26, 139–142.

Hotelling H, 1936 Relations between two sets of variates. Biometrika 28, 321–377.

Johnson RA, Wichern DW, 2002 Applied Multivariate Statistical Analysis, fifth edition. Prentice Hall, New Jersey.

Manly BFJ, 1989 Multivariate Statistical Methods: A Primer. Chapman and Hall, London.

Mardia KV, Kent JT, Bibby JM, 1979 Multivariate Analysis. Academic press London.

Mateu-Figueras G, Daunis-i Estadella J, Coenders G, Ferrer-Rosell B, Serlavós R, Batista-Foguet JM, 2016 Exploring the relationship between two compositions using canonical correlation analysis. Metodoloski zvezki 13 (2), 131–150.

Otero N, Tolosana-Delgado R, Soler A, Pawlowsky-Glahn V, Canals A, 2005 Relative vs. absolute statistical analysis of compositions: a comparative study of surface waters of a Mediterranean River. Water Res. 39 (7), 1404–1414. [PubMed: 15862341]

Pawlowsky-Glahn V, Buccianti A, 2011 Compositional data analysis: theory and applications. In: Pawlowsky-Glahn V, Buccianti A (Eds.), John Wiley & Sons., Chichester, United Kingdom.

Pawlowsky-Glahn V, Egozcue JJ, 2001 Geometric approach to statistical analysis on the simplex. Stoch. Env. Res. Risk A. (SERRA) 15 (5), 384–398.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R, 2015 Modeling and Analysis of Compositional Data. John Wiley & Sons, Chichester, United Kingdom.

R Core Team, 2017 R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

Reyment RA, Savazzi E, 1999 Aspects of Multivariate Statistical Analysis in Geology. Elsevier Science B.V., Amsterdam, The Netherlands.

Searle SR, 1982 Matrix Algebra Useful for Statistics. John Wiley and Sons.

Stewart DK, Love WA, 1968 A general canonical correlation index. Psychol. Bull 70, 160–163. [PubMed: 5681306]

Ter Braak CJF, 1986 Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67 (5), 1167–1179.

Ter Braak CJF, 1990 Interpreting canonical correlation analysis through biplots of structure correlations and weights. Psychometrika 55 (3), 519–531.

Ter Braak CJF, Looman CWN, 1994 Biplots in reduced-rank regression. Biom. J 36 (8), 983–1003.

Ter Braak CJF, Smilauer P, 2002 CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (Version 4.5) Authors.

Thompson B, 1984 Canonical Correlation Analysis: Uses and Interpretation Sage Publications, Beverly Hills and London Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-047.

Tolosana-Delgado R, McKinley J, 2016 Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). Appl. Geochem 75, 263–276.

Tolosana-Delgado R, Otero N, Pawlowsky-Glahn V, Soler A, 2005 Latent compositional factors in the Llobregat River Basin (Spain) hydrogeoeochemistry. Math. Geol 37 (7), 681–702.

Van den Boogaart KG, Tolosana-Delgado R, 2013 Analysing Compositional Data With R Use R!. Springer, Berlin.

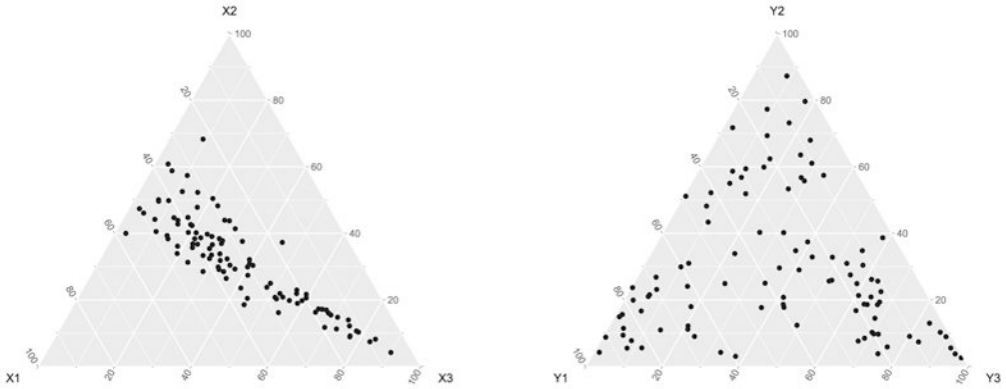

Fig. 1.
Ternary diagrams of two compositions, **x** and **y**, of three parts.

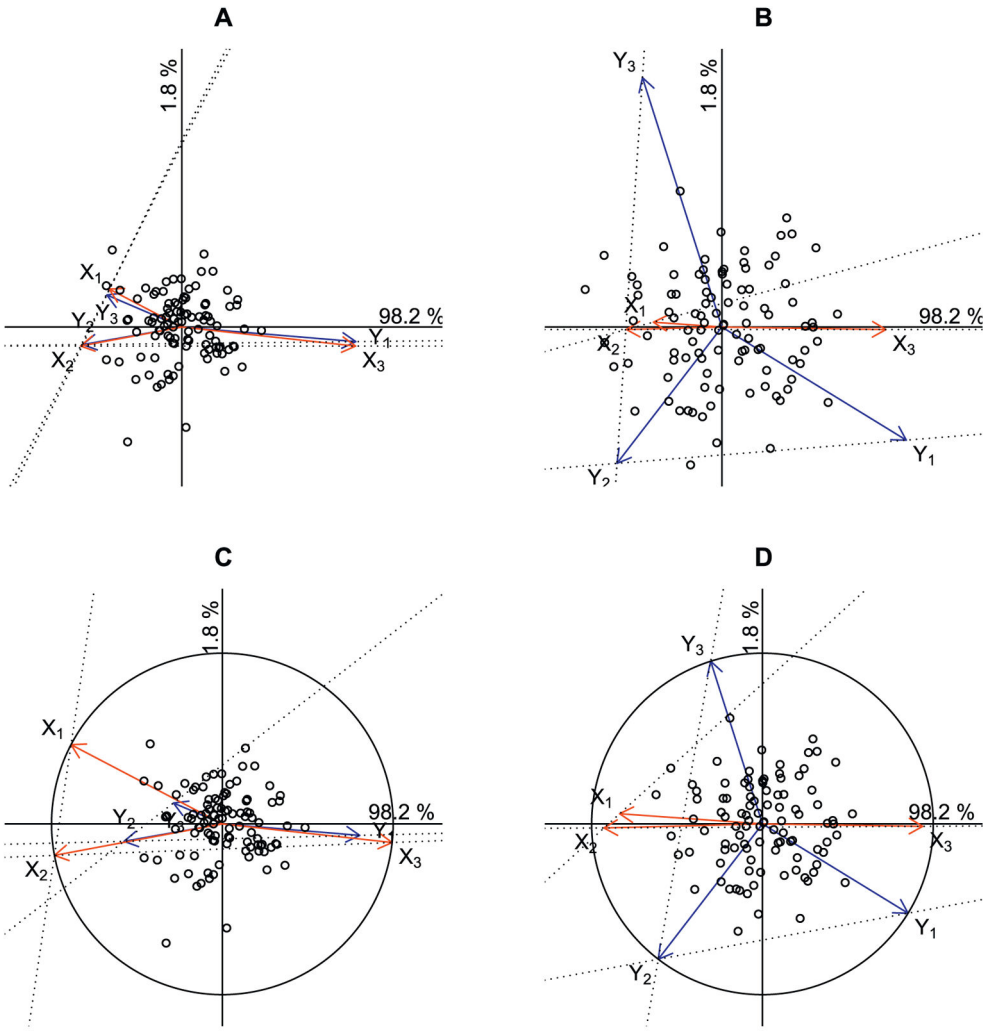

Fig. 2.

CoDA-CCO clr biplots of two three-part compositions using different scalings. Rays represent clr-transformed parts. Links (clr(x_1), clr(x_2)), (clr(x_2), clr(x_3)), (clr(y_1), clr(y_2)), (clr(y_2), clr(y_3)) are indicated by dotted lines. Panels A ($\mathbf{F}_s\mathbf{G}_p'$ scaling) and B ($\mathbf{F}_p\mathbf{G}_s'$ scaling) are biplots made with a covariance-based analysis. Panels C ($\mathbf{F}_s\mathbf{G}_p'$ scaling) and D ($\mathbf{F}_p\mathbf{G}_s'$ scaling), with unit circle, are biplots made with a correlation-based analysis.

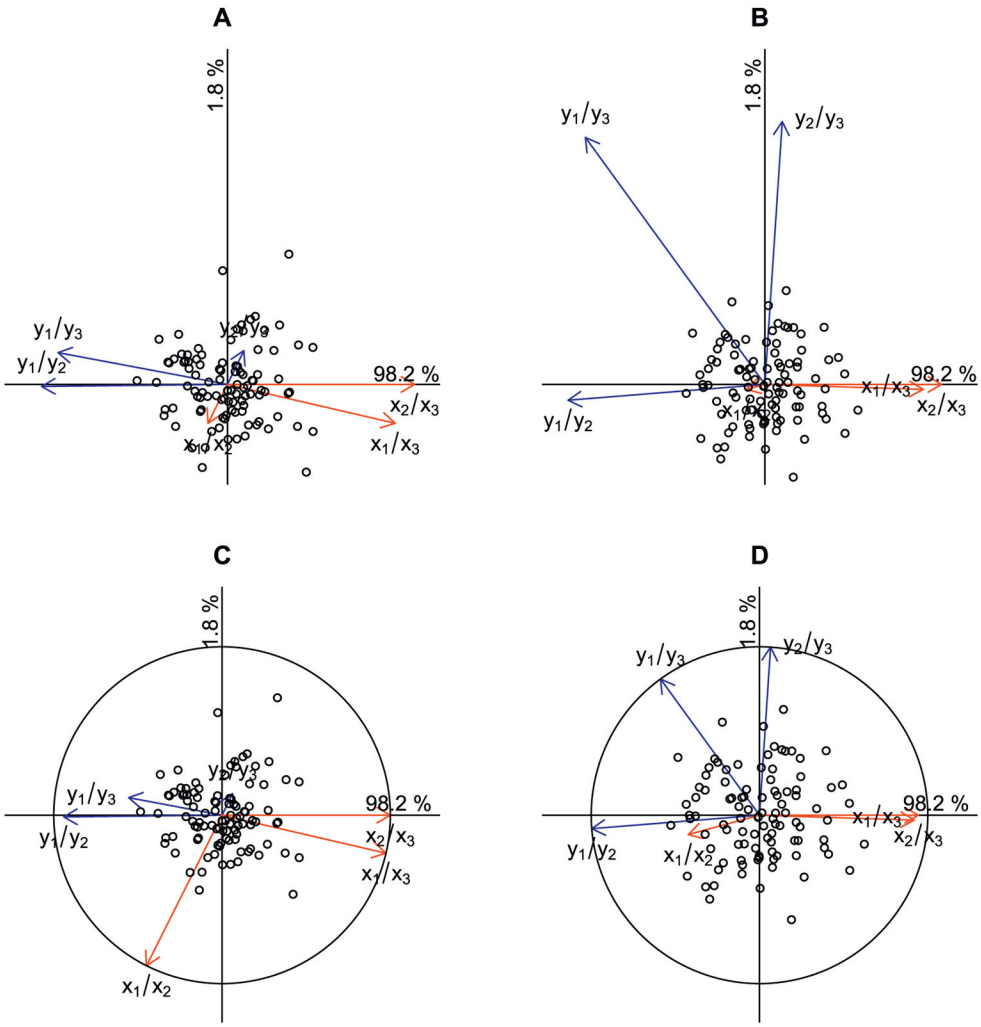

Fig. 3.

CoDA-CCO biplots using pairwise log-ratios of two three-part compositions using different scalings. Rays represent log-ratios. Panels A ($\mathbf{F}_s\mathbf{G}_p'$ scaling) and B ($\mathbf{F}_p\mathbf{G}_s'$ scaling) are biplots made with a covariance-based analysis. Panels C ($\mathbf{F}_s\mathbf{G}_p'$ scaling) and D ($\mathbf{F}_p\mathbf{G}_s'$ scaling), with unit circle, are biplots made with a correlation-based analysis.

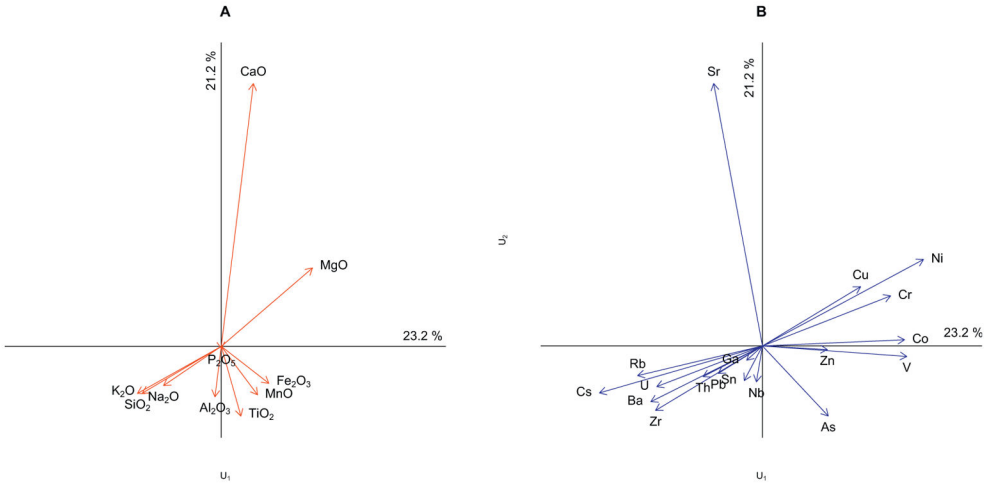

Fig. 4.

CoDA-CCO biplot of major oxides (in standard coordinates) and trace elements (in principal coordinates). Rays represent clr-transformed parts.

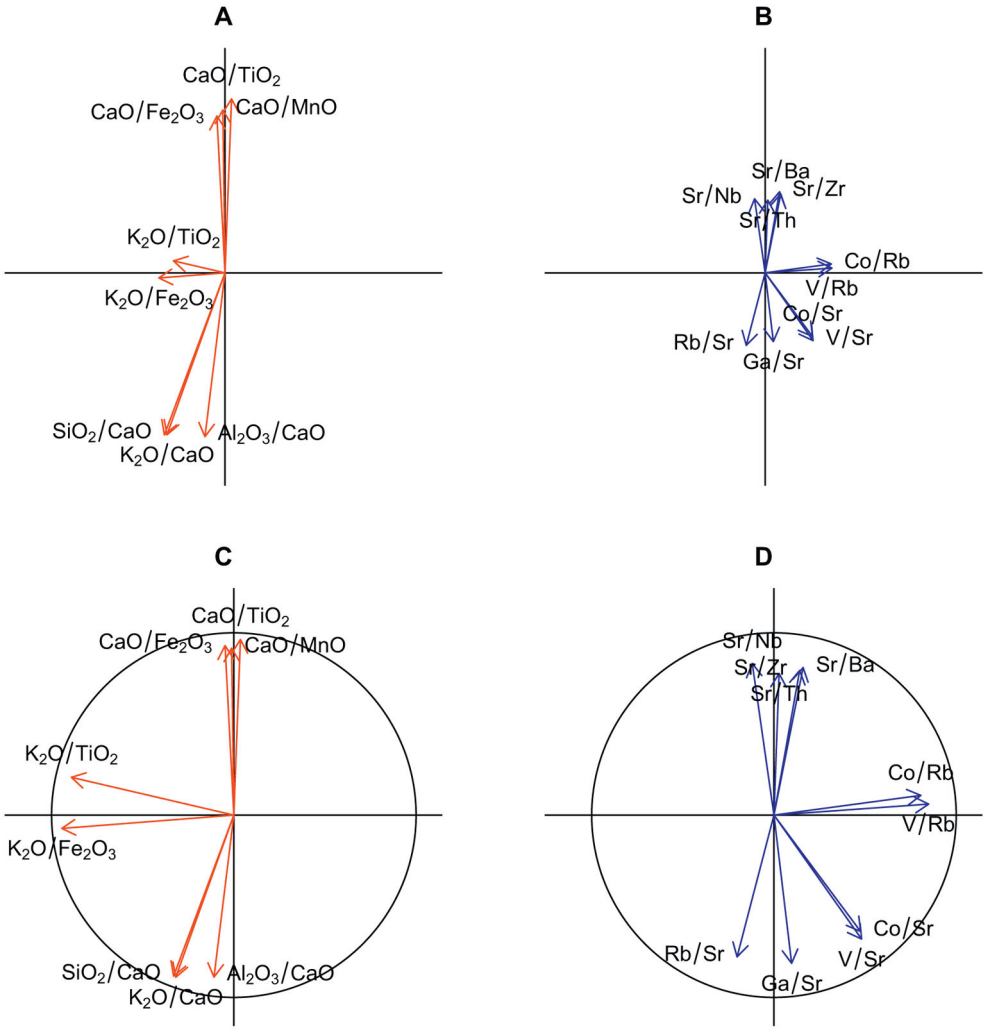

Fig. 5.

CoDA-CCO biplot of major oxides (in standard coordinates) and trace elements (in principal coordinates). Rays represent pairwise log-ratios. Panels A and B show a covariance-based analysis, whereas C and D show a correlation-based analysis.

Table 1

Canonical correlations (r_1,r_2), canonical weights, canonical loadings (between parentheses), adequacy coefficients ($R^2_{x\mid u}$, and cumulative $R^2_{x\mid u}$) and redundancy coefficients ($R^2_{x\mid v}$, and cumulative $R^2_{x\mid v}$) obtained in a CoDA-CCO of two sets of clr transformed compositions of three parts.

	$r_1 = 0.944$		$r_2 = 0.129$	
	U_1		U_2	
clr(x_1)	0.001	(−0.886)	3.847	(0.464)
clr(x_2)	−0.799	(−0.983)	−3.447	(−0.184)
clr(x_3)	0.798	(0.994)	−0.401	(−0.109)
$R^2_{x\mid u}$	0.913		0.087	
$R^2_{x\mid u}$	0.913		1.000	
$R^2_{x\mid v}$	0.813		0.001	
$R^2_{x\mid v}$	0.813		0.815	
	V_1		V_2	
clr(y_1)	0.762	(0.852)	−0.050	(−0.523)
clr(y_2)	−0.717	(−0.610)	−0.521	(−0.793)
clr(y_3)	−0.046	(−0.303)	0.572	(0.953)
$R^2_{y\mid v}$	0.397		0.603	
$R^2_{y\mid v}$	0.397		1.000	
$R^2_{y\mid u}$	0.353		0.010	
$R^2_{y\mid u}$	0.353		0.363	

Table 2

Canonical correlations (r_1,r_2,r_3), canonical weights, canonical loadings (between parentheses), adequacy coefficients ($R^2_{x\,|\,u}$, and cumulative $R^2_{x\,|\,u}$) and redundancy coefficients ($R^2_{x\,|\,v}$, and cumulative $R^2_{x\,|\,v}$) obtained in a CoDA-CCO of major oxides and trace elements of European floodplain sediments. Major oxides and trace elements with a canonical coefficient larger than 0.25 in absolute value are marked in bold for the first two dimensions.

	r_1 = 0.936		r_2 = 0.895		r_3 = 0.814	
	U_1		U_2		U_3	
SiO_2	−0.253	(−0.653)	0.092	(−0.366)	−1.501	(−0.556)
Al_2O_3	0.110	(−0.056)	0.092	(−0.446)	0.401	(0.372)
Na_2O	−0.057	(−0.344)	0.263	(−0.234)	0.438	(0.439)
MgO	0.145	(0.415)	0.083	(0.357)	0.071	(0.426)
P_2O_5	0.017	(0.001)	0.028	(−0.016)	−0.140	(−0.328)
K_2O	−1.524	(−0.812)	−0.145	(−0.481)	1.314	(0.190)
CaO	−0.033	(0.112)	0.551	(0.918)	−0.096	(−0.227)
TiO_2	0.738	(0.236)	−1.069	(−0.838)	−0.840	(−0.094)
MnO	−0.054	(0.302)	−0.244	(−0.404)	−0.180	(−0.263)
Fe_2O_3	0.910	(0.628)	0.350	(−0.488)	0.532	(−0.019)
$R^2_{x\,\|\,u}$	0.193		0.269		0.110	
$R^2_{x\,\|\,u}$	0.193		0.463		0.572	
$R^2_{x\,\|\,v}$	0.169		0.216		0.073	
$R^2_{x\,\|\,v}$	0.169		0.385		0.458	
V	0.579	(0.693)	−0.114	(−0.053)	0.426	(0.334)
Cr	−0.067	(0.467)	0.106	(0.191)	−0.144	(−0.182)
Co	0.372	(0.700)	−0.399	(0.032)	0.180	(0.154)
Ni	0.012	(0.577)	0.266	(0.324)	−0.215	(−0.042)
Cu	−0.046	(0.450)	0.074	(0.284)	−0.125	(−0.158)
Zn	0.144	(0.281)	−0.006	(−0.016)	−0.239	(−0.289)
Ga	0.541	(−0.110)	−0.214	(−0.104)	0.858	(0.670)
As	0.091	(0.181)	0.038	(−0.201)	−0.133	(−0.231)
Rb	−1.848	(−0.762)	0.138	(−0.189)	1.309	(0.517)
Sr	−0.137	(−0.153)	1.200	(0.859)	−0.087	(0.108)
Zr	−0.443	(−0.463)	−0.218	(−0.292)	−1.236	(−0.441)
Nb	0.622	(−0.046)	−0.544	(−0.284)	−0.220	(0.102)
Sn	0.089	(−0.046)	−0.067	(−0.091)	−0.011	(−0.012)
Cs	−0.057	(−0.502)	−0.107	(−0.151)	−0.303	(−0.195)
Ba	−0.140	(−0.507)	−0.450	(−0.265)	0.105	(0.162)
Pb	−0.097	(−0.168)	0.114	(−0.109)	−0.017	(−0.307)
Th	0.401	(−0.410)	0.191	(−0.222)	−0.275	(0.208)

	$r_1 = 0.936$		$r_2 = 0.895$		$r_3 = 0.814$	
	U_1		U_2		U_3	
U	−0.016	(−0.351)	−0.007	(−0.141)	0.126	(0.342)
$R^2_{y \mid v}$	0.194		0.078		0.088	
$R^2_{y \mid v}$	0.194		0.272		0.360	
$R^2_{y \mid u}$	0.170		0.062		0.058	
$R^2_{y \mid u}$	0.170		0.233		0.291	