Metodološki zvezki, Vol. 13, No. 2, 2016, 131-150

Exploring the Relationship between two Compositions using Canonical Correlation Analysis

Glòria Mateu-Figueras[1], Josep Daunis-i-Estadella[2], Germà Coenders[3], Berta Ferrer-Rosell[4], Ricard Serlavós[5], Joan Manuel Batista-Foguet[6]

Abstract

The aim of this article is to describe a method for relating two compositions which combines compositional data analysis and canonical correlation analysis (CCA), and to examine its main statistical properties. We use additive log-ratio (alr) transformation on both compositions and apply standard CCA to the transformed data. We show that canonical variates are themselves log-ratios and log-contrasts. The first pair of canonical variates can be interpreted as the log-contrast of a composition that has the maximum correlation with a log-contrast of the other composition. The second pair can be interpreted as the log-contrast of a composition that has the maximum correlation with a log-contrast of the other composition, under the restriction that they are uncorrelated with the first pair, and so on.

Using properties from changes of basis, we prove that both canonical correlations and canonical variates are invariant to the choice of divisors in alr transformation. We show how to implement the analysis and interpret the results by means of an illustration from the social sciences field using data from Kolb's Learning Style Inventory and Boyatzis' Philosophical Orientation Questionnaire, which distribute a fixed total score among several learning modes and philosophical orientations.

[1] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi, 17003 Girona, Spain; gloria.mateu@udg.edu

[2] Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi, 17003 Girona, Spain; pepus.daunis@udg.edu

[3] Department of Economics, University of Girona, Campus Montilivi, 17003 Girona, Spain; germa.coenders@udg.edu

[4] Department of Economics, University of Girona, Edifici Sant Domènec, Plaça Ferrater Mora 1, 17004 Girona, Spain; berta.ferrer@udg.edu

[5] Department of People Management and Organization, ESADE, University Ramon Llull, Av. Pedralbes, 60-62, 08034 Barcelona, Spain; ricard.serlavos@esade.edu

[6] Department of People Management and Organization, ESADE, University Ramon Llull, Av. Pedralbes, 60-62, 08034 Barcelona, Spain; joanm.batista@esade.edu

1 Introduction

Compositional data lie in a constrained positive space with a fixed sum and convey information on the relative importance of components. Typical examples are chemical and geological compositions (adding to 100% in weight or volume), genotype frequencies (adding to 1), time use (adding to 24 hours), voting (adding to 100% of votes), or household budget allocation (adding to 100% of the budget). The fixed sum is typically normalized to one, and a D-term composition $(x_1, x_2,..., x_D)$ is thus constrained as follows:

$$0<x_d<1 \text{ and } \sum_{d=1}^{D} x_d = 1 \tag{1.1}$$

Serious problems arise when using standard statistical analysis tools on compositional data (Aitchison, 1986, 2001; Pawlowsky-Glahn & Buccianti, 2011):

1. Compositional data have a bounded distribution. This implies at least non-normality and heteroscedasticity (lower variance close to the boundary).

2. One component can only increase if some others decrease. This results in negative spurious correlations among the components (Pearson, 1897) and prevents interpreting effects of linear models in the usual way "keeping everything else constant".

3. The true dimensionality of a set of compositional variables is $D-1$. Analysis of all D dimensions leads to perfect collinearity.

4. Compositional data lie in a $(D-1)$-dimensional Euclidean space called the simplex, with different operations and distance from real space (Billheimer et al., 2001; Pawlowsky-Glahn & Egozcue, 2001).

The compositional data analysis (CoDa) tradition started with Aitchison's seminal work (1986) on treating chemical and biological compositions. Nowadays, however, it spans almost all of the hard sciences and has started to be used in the social sciences, which often face similar problems (Batista-Foguet et al., 2015; Coenders et al., 2011; van Eijnatten et al., 2015; Ferrer-Rosell & Coenders, 2016; Ferrer-Rosell et al., 2015, 2016a, 2016b; Fry, 2011; Hlebec et al., 2012; Kogovšek et al., 2013; Vives-Mestres et at., 2016).

The literature on CoDa has extensively dealt with relating one composition to non-compositional data (Egozcue et al., 2012; Hron et al., 2012; Martín-Fernández et al., 2015) and with analyzing one single composition. As far as the exploratory data analysis of one

single composition is concerned (Egozcue & Pawlowsky-Glahn, 2011), available methods include the variation array (Aitchison, 1986), principal component analysis (Aitchison, 1983; Aitchison & Greenacre, 2012), the CoDa-dendrogram (Pawlowsky-Glahn & Egozcue, 2011), and the CoDa-biplot (Aitchison & Greenacre, 2012). As regards exploratory tools to relate two compositions, the natural choice is canonical correlation analysis – CCA (Aitchison, 1986). Typical problems relating two compositions include the relationship between the composition of species and the chemical composition of the environment (ter Braak, 1996); between the composition of foods and the composition of their energy and nutrients; or between the composition of materials and the composition of spectral curves in image processing. The use of CCA for compositional data was foreshadowed in Aitchison (1986), without much mention of its properties or interpretation. At a later date, van den Boogaart and Tolosana-Delgado (2013) devised an advanced procedure for compositional CCA requiring software designed for this purpose.

Drawing from Aitchison (1986), in this article we develop and illustrate a simple procedure for carrying out CCA of two compositional vectors and examine its interpretation and main statistical properties. Even if specialized techniques for compositional data have appeared (van den Boogaart & Tolosana-Delgado, 2013; Pawlowsky-Glahn & Buccianti, 2011; Pawlowsky Glahn et al., 2015; Thió-Henestrosa & Martín-Fernández, 2005), compositional data can also be transformed so that they can be subject to standard and well-understood statistical techniques carried out using standard software. This is the approach we take in this article.

Given the fact that only information on the relative size of components is available in a compositional data context, logarithms of ratios between component values are a meaningful way of expressing the data and guaranteeing the principles of CoDa (Aitchison, 2001). A logarithm of a ratio is scale invariant, meaning that it does not change if the values involved are multiplied by an arbitrary constant. Adding or dropping components from a composition does not modify the log-ratios computed from the remaining components. This is related to the principles of scale invariance and subcompositional coherence. For full details on CoDa principles, see Pawlowsky Glahn et al. (2015).

Several log-ratio transformations have been suggested in the literature (Egozcue et al., 2003). Additive log-ratio transformation (alr) is the easiest to compute since it is simply the log-ratio between each component and the last:

$$y_d=\ln(x_d/x_D)=\ln(x_d)-\ln(x_D) \text{ with } d=1,2,...,D-1 \quad (1.2)$$

Alr-transformed y_d variables recover the full unconstrained real space. It must be noted that one dimension is lost. Although alr transformation is used in this article due to its simplicity, there are alternatives (see Egozcue et al., 2003 for a general background on the transformations and Section 3.3. for a discussion of their applicability to CCA).

Since the decision on which component to leave in last place and serve as a reference in the alr transformation is often arbitrary, there is concern regarding whether the results of a statistical analysis are invariant to this arbitrary choice. Of course, different log-ratios constitute different variables and the raw results will never be invariant. However, it is considered desirable that overall goodness of fit measures be invariant to this choice. Once results are reexpressed as a function of the log components $\ln(x_d)$, they should ideally also be invariant.

The structure of the article is as follows. First, we review the basics of CCA. We then come to the particular case in which CCA is applied to compositions that have been subjected to alr transformation, showing how to interpret the key results, proving that they are invariant to the choice of reference component, and discussing alternative transformations. Following this, we present an illustration from the field of education using data from Kolb's Learning Style Inventory (Batista-Foguet et al., 2015; Kolb, 1984, 1999) and Boyatzis' Philosophical Orientation Questionnaire (Boyatzis et al., 2000). The final section discusses the strengths and weaknesses of the method.

2 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a multivariate analysis technique which studies the relationships between two sets of variables $\mathbf{Y}_a=(y_{a1}, y_{a2}, ..., y_{ap})$ and $\mathbf{Y}_b=(y_{b1}, y_{b2}, ..., y_{bq})$ usually defined in the real space. The method was first introduced in Hotelling (1936) and a non-technical description can be found in Hair et al. (2009).

CCA builds pairs of linear combinations of each set of variables called canonical variates. The first canonical variate cv_{a1} for set $\mathbf{Y}_a$ is derived so that it is maximally correlated with the first canonical variate cv_{b1} for set $\mathbf{Y}_b$. The second canonical variate cv_{a2} for set $\mathbf{Y}_a$ is derived so that it is maximally correlated with the second canonical variate cv_{b2} for set $\mathbf{Y}_b$ under the restriction that both new canonical variates are uncorrelated with cv_{a1} and cv_{b1}. The following

pairs are extracted in a similar manner and have the maximum mutual correlation, while being uncorrelated with the previous pairs. The process may be continued up to $\min\{p,q\}$ times.

The raw canonical coefficients w_{aij} and w_{bij} are the weights used to compute the i-th pair of canonical variates from the j-th original variables:

$$\begin{aligned} cv_{a1} &= w_{a11}y_{a1} + w_{a12}\,y_{a2} + \cdots + w_{a1p}\,y_{ap} \\ cv_{b1} &= w_{b11}y_{b1} + w_{b12}\,y_{b2} + \cdots + w_{b1q}\,y_{bq} \\ cv_{a2} &= w_{a21}y_{a1} + w_{a22}\,y_{a2} + \cdots + w_{a2p}\,y_{ap} \\ cv_{b2} &= w_{b21}y_{b1} + w_{b22}\,y_{b2} + \cdots + w_{b2q}\,y_{bq} \\ &\cdots\cdots \end{aligned} \tag{2.1}$$

In practice, the canonical coefficients are computed from three covariance matrices: the square matrix $\mathbf{S}_{aa}$ contains covariances in the first variable set, the square matrix $\mathbf{S}_{bb}$ covariances in the second set, and the rectangular matrix $\mathbf{S}_{ab}$ covariances between variables of one set and the other. Canonical variates are obtained from an eigenvalue analysis of matrix:

$$\mathbf{S}_{aa}^{-1}\mathbf{S}_{ab}\mathbf{S}_{bb}^{-1}\mathbf{S}_{ba} \tag{2.2}$$

The correlation between cv_{a1} and cv_{b1} is the first canonical correlation $\hat{\rho}_1$, the correlation between cv_{a2} and cv_{b2} is the second canonical correlation $\hat{\rho}_2$, and so on. These canonical correlations are obtained as the square root of the eigenvalues of the matrix in Equation (2.2).

The maximum number of canonical variates that can be extracted is the smallest dimension of the two sets of variables. For instance, if p=5 and q=8, then a maximum of 5 pairs of variates can be obtained. As with many other multivariate analysis techniques, the researcher is interested in a parsimonious solution and in interpreting only the most relevant variates. The relevance of a pair of canonical variates can be assessed by the sheer size of the canonical correlation, the interpretability of the canonical variates from the canonical weights, or the statistical significance of the canonical correlations according to Wilks' Λ tests, which are also a function of the eigenvalues. Since, $\hat{\rho}_1 > \hat{\rho}_2 > \ldots > \hat{\rho}_{\min\{p,q\}}$, a common strategy is to sequentially test the following hypotheses:

$$H_{01}: \rho_1=\rho_2=\rho_3=\cdots=\rho_{\min\{p,q\}}=0$$

$$H_{02}: \rho_2=\rho_3=\cdots=\rho_{\min\{p,q\}}=0$$

$$\ldots\ldots \qquad (2.3)$$

$$H_{0\min\{p,q\}-1}: \rho_{\min\{p,q\}-1}=\rho_{\min\{p,q\}}=0$$

$$H_{0\min\{p,q\}}: \rho_{\min\{p,q\}}=0$$

The rejection of H_{01} to H_{0i} and the failure to reject H_{0i+1} to $H_{0\min\{p,q\}}$ shows the first i canonical correlations to be statistically significant.

Other common results of a CCA which provide a useful aid to interpreting the canonical variates require standardization in some form or other (Hair et al., 2009) and are:

1. Standardized canonical coefficients (coefficients used to compute canonical variates from standardized *y* variables).

2. Canonical loadings (correlations between the canonical variates and the *y* variables they are computed from).

3. Canonical cross-loadings (correlations between canonical variates and the other set of *y* variables).

4. Redundancy analysis (percentages of variance for the *y* variables explained by their own canonical variates and from the canonical variates computed from the other set of *y* variables).

3 Canonical Correlation Analysis of Compositional Data Transformed by Means of alr

3.1 Interpretation

Given two compositions with D_a and D_b components, $X_a=(x_{a1},x_{a2},\ldots,x_{aDa})$ and $X_b=(x_{b1},x_{b2},\ldots,x_{bDb})$, following Aitchison (1986) we first apply alr transformation with the last component in the denominator. The results are the following two real vectors with $p=D_a-1$ and $q=D_b-1$ elements:

$$\mathbf{Y}_a = \left(\ln\left(\frac{x_{a1}}{x_{aDa}}\right), \ln\left(\frac{x_{a2}}{x_{aDa}}\right), \ldots, \ln\left(\frac{x_{ap}}{x_{aDa}}\right) \right)$$

$$\mathbf{Y}_b = \left(\ln\left(\frac{x_{b1}}{x_{bDb}}\right), \ln\left(\frac{x_{b2}}{x_{bDb}}\right), \ldots, \ln\left(\frac{x_{bq}}{x_{bDb}}\right) \right) \quad (3.1)$$

We can rewrite Equation (3.1) as:

$$\mathbf{Y}_a = \left(\ln(x_{a1}) - \ln(x_{aDa}), \ln(x_{a2}) - \ln(x_{aDa}), \ldots, \ln(x_{ap}) - \ln(x_{aDa})\right)$$

$$\mathbf{Y}_b = \left(\ln(x_{b1}) - \ln(x_{bDb}), \ln(x_{b2}) - \ln(x_{bDb}), \ldots, \ln(x_{bq}) - \ln(x_{bDb})\right) \quad (3.2)$$

$\mathbf{Y}_a$ and $\mathbf{Y}_b$ are two sets of real variables to which we can apply the standard CCA procedure from the covariance matrices of each set of transformed variables and the covariance matrix between the transformed variables of one set and the other in Equation (2.2).

The first pair of canonical variates in Equation (2.1), when expressed in terms of logarithms of components, becomes:

$$cv_{a1} = w_{a11}\ln(x_{a1}) + w_{a12}\ln(x_{a2}) + \cdots + w_{a1p}\ln(x_{ap}) - (w_{a11} + w_{a12} + \cdots + w_{a1p})\ln(x_{aDa})$$

$$cv_{b1} = w_{b11}\ln(x_{b1}) + w_{b12}\ln(x_{b2}) + \cdots + w_{b1q}\ln(x_{bq}) - (w_{b11} + w_{b12} + \cdots + w_{b1q})\ln(x_{bDb}) \quad (3.3)$$

Since the raw canonical coefficients are applied from $\ln(x_{a1})$ to $\ln(x_{ap})$ and again to $\ln(x_{aDa})$ with reversed signs, the weights of all D_a logarithms add up to zero, and the same occurs with the weights of the D_b logarithms of the x_b variables. This would also hold for the remaining canonical variates.

This is the same as saying that the canonical variates are log ratios of the product of components with a positive weight raised to a power equal to that weight, over the product of components with a negative weight raised to a power equal to the absolute weight. Let us show an example of the former for a canonical variate of a 5-dimensional composition with:

$$cv_{a1} = 1y_{a1} + 1.5y_{a2} + 0.5y_{a3} - 0.5\,y_{a4} \quad (3.4)$$

The reexpression of this canonical variate as a log-ratio is:

$$cv_{a1} = 1\ln(x_{a1}) + 1.5\ln(x_{a2}) + 0.5\ln(x_{a3}) - 0.5\ln(x_{a4}) - 2.5\ln(x_{a5}) = \ln\left(\frac{x_{a1}x_{a2}^{1.5}x_{a3}^{0.5}}{x_{a4}^{0.5}x_{a5}^{2.5}}\right) \quad (3.5)$$

The cv_{a1} log-ratio in this example is high mainly when x_{a1} and x_{a2} are high and x_{a5} is low. Since the sum of positive exponents equals the sum of negative exponents, the log-ratio is also a log-contrast, that is, a log-linear combination where the sum of the coefficients is 0 (Aitchison, 1986: 84).

The first pair of canonical variates can thus be interpreted as the log-contrast of one of the compositions that has the maximum correlation with a log-contrast of the other composition. The second pair can be interpreted as the log-contrast of one of the compositions that has the maximum correlation with a log-contrast of the other composition, under the restriction that they are uncorrelated with the first pair of canonical variates. A similar interpretation would hold for the third pair, subject to zero correlation with the first two pairs, and so on.

3.2 Invariance of the Results to the Choice of Reference Component in alr

Although the last component in each composition was chosen as the common divisor in our alr transformation, this could equally have been any other component. Consequently, for any analysis involving alr vectors, it is important to check the invariance of the key results to component permutations, or in other words, their invariance with respect to the choice of divisor in alr transformation. In this section we show specifically that Wilks' Λ tests, canonical correlations, and canonical variates as functions of log components –Equation (3.3)– are invariant to this choice.

It is easy to see how two alr-transformed vectors using different components as a divisor are related using a change-of-basis matrix. Following Mateu-Figueras et al. (2011), the elements of an alr vector are the coefficients of the original composition with respect to a particular non-orthonormal basis on the simplex, the sample space of compositional data. The effect of changing the common divisor is to obtain the coefficients with respect to another particular basis, which is analogous to performing an oblique rotation of the data.

Let $\mathbf{Y}_a$ and $\mathbf{Y}_b$ be the alr transformed vectors using the last components as common divisors and let $\mathbf{Y}_a^*$ and $\mathbf{Y}_b^*$ be the alr-transformed vectors using other components as denominators. Then, $\mathbf{Y}_a^*=\mathbf{Q}\mathbf{Y}_a$ and $\mathbf{Y}_b^*=\mathbf{P}\mathbf{Y}_b$. We can obtain the exact expression of matrices **Q** and **P** (see Aitchison, 1986: 94), but the important point here is that matrices **Q** and **P** are change-of-basis matrices. From the usual properties of covariance matrices we know that:

$$\mathbf{S}^*_{aa} = \mathbf{Q}\mathbf{S}_{aa}\mathbf{Q}' \tag{3.6}$$

$$\mathbf{S}^*_{bb} = \mathbf{P}\mathbf{S}_{bb}\mathbf{P}' \tag{3.7}$$

$$\mathbf{S}^*_{ab} = \mathbf{Q}\mathbf{S}_{ab}\mathbf{P}' \text{ and } \mathbf{S}^*_{ba} = \mathbf{P}\mathbf{S}_{ba}\mathbf{Q}' \tag{3.8}$$

When using different common divisors in alr transformation, the analyzed matrix in Equation (2.2) becomes:

$$\left(\mathbf{S}^*_{aa}\right)^{-1}\mathbf{S}^*_{ab}\left(\mathbf{S}^*_{bb}\right)^{-1}\mathbf{S}^*_{ba} \tag{3.9}$$

By using the relationships in Equations (3.6)–(3.8), Equation (3.9) becomes:

$$\begin{aligned}&\left(\mathbf{S}^*_{aa}\right)^{-1}\mathbf{S}^*_{ab}\left(\mathbf{S}^*_{bb}\right)^{-1}\mathbf{S}^*_{ba} = \left(\mathbf{Q}\mathbf{S}_{aa}\mathbf{Q}'\right)^{-1}\left(\mathbf{Q}\mathbf{S}_{ab}\mathbf{P}'\right)\left(\mathbf{P}\mathbf{S}_{bb}\mathbf{P}'\right)^{-1}\left(\mathbf{P}\mathbf{S}_{ba}\mathbf{Q}'\right) = \\ &\left(\mathbf{Q}'\right)^{-1}\mathbf{S}^{-1}_{aa}\mathbf{S}_{ab}\mathbf{S}^{-1}_{bb}\mathbf{S}_{ba}\mathbf{Q}'\end{aligned} \tag{3.10}$$

From linear algebra properties, we know that the eigenvalues of a matrix are invariant under changes of basis. Consequently, both the canonical correlations and Wilks' Λ tests are invariant under change of common divisor in alr transformation.

It is easy to see how the normalized eigenvectors of matrices in Equations (3.9) and (2.2), denoted as $\mathbf{w}^*_{ai}$ and $\mathbf{w}_{ai}$ respectively, must be related by $\mathbf{Q}'\mathbf{w}^*_{ai} = \mathbf{w}_{ai}$ or $\mathbf{w}^*_{ai} = \left(\mathbf{Q}'\right)^{-1}\mathbf{w}_{ai}$. Then we obtain the invariance of the corresponding canonical variates as:

$$cv^*_{ai} = \left(\mathbf{w}^*_{ai}\right)'\mathbf{Y}^*_a = \left(\left(\mathbf{Q}'\right)^{-1}\mathbf{w}_{ai}\right)'\mathbf{Q}\mathbf{Y}_a = \mathbf{w}'_{ai}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Y}_a = \mathbf{w}'_{ai}\mathbf{Y}_a = cv_{ai} \tag{3.11}$$

Conversely, all results that imply standardization, like standardized canonical coefficients, canonical loadings/cross-loadings and redundancy analysis, are not invariant to the choice of reference component in alr transformation. In the case of CoDa, however, given the facts that canonical variates can be readily interpreted as log-ratios and log-contrasts on their own, and that standardization is extremely uncommon for log-contrasts, standardized information is not needed to enhance interpretation and is not considered in this article.

3.3 Appropriateness of Alternative Log-ratio Transformations for Canonical Correlation Analysis

One key issue when working with CoDa is the choice of the log-ratio transformation, since different possibilities are available. Additive log-ratio (alr) and centered log-ratio (clr) transformations were introduced in Aitchison (1986), while isometric log-ratio transformation (ilr) was introduced in Egozcue et al. (2003).

Aitchison's (1986) proposal for compositional CCA involved alr transformation. Although alr transformation is simple and easy to interpret, it is asymmetric in its parts. By changing the part in the denominator, a different alr-transformed vector is obtained. For this reason, when alr transformation is used, it is important to check the invariance of the results with respect to the choice of common denominator, as we have done in Section 3.2. However, as Egozcue et al. (2003) noted, the main drawback of alr transformation is that it is not an isometric transformation from the simplex to the real space. It was later shown that an alr vector can be viewed as the coefficients of a composition with respect to a non-orthonormal basis on the simplex (Mateu-Figueras et al., 2011). Consequently, it is not suitable for statistical techniques that use distances or angles between alr vectors, such as cluster analysis. Note that these problems do not occur when using CCA because eigenvalues and eigenvectors of a product of covariance matrices are involved. Due to the non-orthonormality of the basis, the equality $\mathbf{Q}'\mathbf{w}_{ai}^{*} = \mathbf{w}_{ai}$ is only true if the vector product $\mathbf{Q}'\mathbf{w}_{ai}^{*}$ is normalized, although this does not affect the analyses considered in this article.

Clr transformation is defined as the logarithm of the ratio of each part over the geometric mean. It is a symmetric transformation with respect to the compositional parts and also an isometric transformation. Nevertheless, clr transformation has the disadvantage that the clr covariance matrix is singular. In our case, clr transformation would not be a good choice because CCA uses covariance matrices and their inverses. Conversely, it would be a good choice for cluster analysis or other statistical techniques in which distances are crucial and covariances do not need to be inverted.

Ilr transformation is isometric and consequently makes it possible to associate distances in the simplex with distances in the transformed space. Additionally, an ilr vector can be viewed as the coordinates of a composition with respect to an orthonormal basis on the simplex. Finally, covariance matrices can be inverted. It can thus be used in virtually all statistical analyses. The expression of the ilr using a particular orthonormal basis is given in Egozcue et

al. (2003). Nevertheless, in inner product spaces, an orthonormal basis is not uniquely determined and in some cases it is not straightforward to determine which basis is the most appropriate to solve a specific problem and how it can be interpreted. Faced with the problem of interpreting CCA on ilr transformation, van den Boogaart and Tolosana-Delgado (2013) devise a graphical back-transformation of the canonical coefficients. In any case, the invariance of the ilr results with respect to the choice of the orthonormal basis also holds.

In this article, although alr transformation was used due to its simplicity, ilr transformation could also have been used, and we actually did rerun the illustration analysis with ilr transformation. The final canonical variates expressed in terms of the log components and as log-contrasts are invariant, because alr and ilr vectors are also related through a change-of-basis matrix.

4 Illustration

4.1 Background

In this illustration of compositional CCA, our aim is to relate students' learning styles to their philosophical orientations. Philosophical orientation is a good means of understanding the relationship between people's values and beliefs, and their behavior and approach to learning (Boyatzis et al., 2000). Since a person's behavior is related to his or her values and beliefs, philosophy is important for comprehending and predicting behavior, with the added advantage that a person's philosophy goes beyond social context. Philosophical orientation is useful for answering questions such as how individuals 'act across various social settings' or 'think about establishing the value of things, activities and others' (Boyatzis et al., 2000: 50). Three major clusters of philosophical systems have traditionally been proposed. These clusters define the extent to which a person is pragmatic (PR), intellectual (IN) or humanistic (HU).

A person with a predominantly PR philosophical orientation will make decisions based on the benefits of the action, measured in terms of utility or comparing input and output. If the objectives to be achieved are not clear or measuring utility is difficult, then an activity will be less valuable to a person with this orientation.

Someone with a predominantly IN philosophical orientation will be rational, logical and focus on comprehending everything. The central concern underlying this philosophical orientation is analytical.

Someone with a predominantly HU orientation is thought to be committed to human values. This kind of person will tend to determine whether an activity is worthy in terms of its impact on other people and the quality of the relationship with these people. The central issue underlying HU orientation is a concern for close and personal relationships.

According to Experiential Learning Theory, learning is a process whereby knowledge is created through the transformation of experience (Kolb, 1984). Learning requires abilities to grasp and transform knowledge that are polar opposites. In grasping knowledge, some people perceive new information through experiencing the concrete, tangible, and felt qualities of the world, which is referred to as concrete experience (CE), while others tend to take hold of new information through symbolic representation or abstract conceptualization (AC). In transforming knowledge, some people tend to carefully watch others who are involved in the experience and reflect on what happens (reflective observation – RO), while others choose to start doing things (active experimentation - AE). Learning can also be conceived as a four-stage cycle, where each stage is represented by a learning mode.

At the CE stage, one tends to rely more on intuition than on a systematic focus. Moreover, in this stage, a learner relies on the ability to be open, receptive and adaptive to changes. At the RO stage, one comprehends situations by taking into account different perspectives. In this stage, a learner relies on patience and objectivity, as well as thoughts and feelings. At the AC stage, logic and ideas are needed to understand a problem, rather than feelings. A learner in this stage relies on systematic planning and the theoretical development of ideas. Finally, at the AE stage, one learns by experimenting with changing situations. In this stage, a learner will find it more useful to put ideas into practice and see what really works than to simply observe.

4.2 Data and Measures

Multidimensional forced-choice questionnaires to measure philosophical orientations and learning modes were designed in Boyatzis et al. (2000) and Kolb (1999). In these questionnaires, each question consists of a set of *D* statements, and each statement is an indicator of a different dimension, in our case, of a philosophical orientation (*D*=3) or a

learning mode (D=4). Respondents are instructed to rank these statements. In this article, we assume that ranks are coded as $D-1$ for the most preferred statement, $D-2$ to the second most preferred, down to 0 for the least preferred. The Philosophical Orientation Questionnaire consists of k=20 questions designed as in this example:

"I think of my value, or worth, in terms of:

(a) My relationships (e.g. family, friends).

(b) My ideas or ability to invent new concepts or ability to analyse things.

(c) My financial net worth or income."

Statement (a) reflects the HU orientation, (b) the IN orientation, and (c) the PR orientation.

The Learning Style Inventory includes k=12 questions designed as in this example:

"When I am learning:

(a) I like to experience sensations.

(b) I like to observe and listen.

(c) I like to think about ideas.

(d) I like to do things."

Statement (a) reflects the CE mode, (b) the RO mode, (c) the AC mode, and (d) the AE mode.

The ranks of each dimension are summed across the k questions to produce D global scores, one for each dimension. These D scores have a fixed sum for all respondents, equal to $kD(D-1)/2$. Once the global scores have been computed, forced-choice instruments can be understood as compositions, in which the $kD(D-1)/2$ total is allocated to the D dimensions (components), so that data only convey information about the relative importance of dimensions (learning modes and philosophical orientations) for a given individual. Under this coding scheme, the dimension score is the number of times the dimension has been preferred over other dimensions in all possible pair-wise comparisons over the k questions. For instance, if a component is always ranked as the lowest, it has never been preferred to any other mode and receives a 0 score. If a component is always ranked as the highest, it is preferred k times to the other $D-1$ modes and receives a $k(D-1)$ score. Scores can thus be understood as having ratio scale properties: a component with a score of 6 has been preferred

to other components twice as many times across the k items than a mode with score of 3 (Batista-Foguet et al., 2015). Alternative ways of coding these questionnaires are discussed in de Vries and van der Ark (2008).

In this illustration, we use the same data as those used by Batista-Foguet et al. (2015), which cover 7 consecutive years (2006-2013) of candidates on an international MBA program at a leading European business school. The sample size was 1,194 full time participants from 86 countries, of which the most common were Spain (15.9%), the US (13.7%), India (9.6%), and Germany (5.6%). 69.7% were male and 30.3% female. Average age was 31.4 years (SD 2.8 years). Previous student background was heterogeneous, including not only economics (11%) and management studies (32%), but also engineering (36.4%), social sciences (9.3%), arts (5.7%) and hard sciences (5.5%).

The philosophical orientation components were labeled x_{p1}=pragmatic (PR), x_{p2}=intellectual (IN), and x_{p3}=humanistic (HU); while the learning mode components were labeled x_{l1}=abstract conceptualization (AC), x_{l2}=concrete experience (CE), x_{l3}=active experimentation (AE), and x_{l4}=reflective observation (RO). The final two components, HU and RO, were used as a reference for the alr transformation:

$$\begin{aligned}
\text{log - ratio of PR over HU } y_{p1} &= \ln\left(\frac{x_{p1}}{x_{p3}}\right) = \ln\left(x_{p1}\right) - \ln\left(x_{p3}\right) \\
\text{log - ratio of IN over HU } y_{p2} &= \ln\left(\frac{x_{p2}}{x_{p3}}\right) = \ln\left(x_{p2}\right) - \ln\left(x_{p3}\right) \\
\text{log - ratio of AC over RO } y_{l1} &= \ln\left(\frac{x_{l1}}{x_{l4}}\right) = \ln\left(x_{l1}\right) - \ln\left(x_{l4}\right) \\
\text{log - ratio of CE over RO } y_{l2} &= \ln\left(\frac{x_{l2}}{x_{l4}}\right) = \ln\left(x_{l2}\right) - \ln\left(x_{l4}\right) \\
\text{log - ratio of AE over RO } y_{l3} &= \ln\left(\frac{x_{l3}}{x_{l4}}\right) = \ln\left(x_{l3}\right) - \ln\left(x_{l4}\right)
\end{aligned} \tag{4.1}$$

4.3 Results

After submitting the sets (y_{p1}, y_{p2}) and (y_{l1}, y_{l2}, y_{l3}) to a CCA using SPSS v.23, the resulting canonical correlations are $\hat{\rho}_1$=0.246 and $\hat{\rho}_2$=0.163. Their significance tests are in Table 1. The raw (unstandardized) canonical coefficients are in Table 2.

Table 1: Significance Tests for the Canonical Correlations

H_0	Wilk's Λ	χ^2	DF	*p*-value
$\rho_1=\rho_2=0$	0.914	93.854	6	0.000
$\rho_2=0$	0.973	28.295	2	0.000

Table 2: Raw Canonical Coefficients as a Function of the Log-ratios

	Variate 1	Variate 2
Philosophical orientations		
y_{p1} (log-ratio of PR over HU)	-0.524	1.730
y_{p2} (log-ratio of IN over HU)	2.085	-0.274
Learning modes		
y_{l1} (log-ratio of AC over RO)	1.720	-0.177
y_{l2} (log-ratio of CE over RO)	-0.447	-1.347
y_{l3} (log-ratio of AE over RO)	-1.032	1.311

The original canonical variates are functions of the log ratios and are easily re-expressed by hand as a function of the log-components as in Equation (3.3). For instance, in the philosophical orientation composition the first canonical variate is:

$$\begin{aligned} cv_{p1} &= -0.524\,y_{p1} + 2.085\,y_{p2} = \\ &-0.524\ln(x_{p1}) + 0.524\ln(x_{p3}) + 2.085\ln(x_{p2}) - 2.085\ln(x_{p3}) = \\ &-0.524\ln(x_{p1}) + 2.085\ln(x_{p2}) - 1.561\ln(x_{p3}) \end{aligned} \tag{4.2}$$

Table 3: Raw Canonical Coefficients as a Function of the Log-components

		Variate 1	Variate 2
Philosophical orientations			
	$ln(x_{p1})$ (PR)	-0.524	1.730
	$ln(x_{p2})$ (IN)	2.085	-0.274
	$ln(x_{p3})$ (HU)	-1.561	-1.456
Learning modes			
	$ln(x_{l1})$ (AC)	1.720	-0.177
	$ln(x_{l2})$ (CE)	-0.447	-1.347
	$ln(x_{l3})$ (AE)	-1.032	1.311
	$ln(x_{l4})$ (RO)	-0.241	0.213

Canonical variates as a function of log components are shown in Table 3. As in Equation (4.2), the coefficients in Table 2 apply to all rows in Table 3 but the last one of each composition, which receives their sum with reversed sign.

The canonical variates in Table 3 correspond to the following log-contrasts:

$$cv_{p1} = \ln\left(\frac{x_{p2}^{2.085}}{x_{p1}^{0.524} x_{p3}^{1.561}}\right) \qquad cv_{p2} = \ln\left(\frac{x_{p1}^{1.730}}{x_{p2}^{0.274} x_{p3}^{1.456}}\right)$$

$$cv_{l1} = \ln\left(\frac{x_{l1}^{1.720}}{x_{l2}^{0.447} x_{l3}^{1.032} x_{l4}^{0.241}}\right) \qquad cv_{l2} = \ln\left(\frac{x_{l3}^{1.311} x_{l4}^{0.213}}{x_{l1}^{0.177} x_{l2}^{1.347}}\right) \tag{4.3}$$

The first pair of canonical variates can therefore be interpreted as follows: when the IN (x_{p2}) orientation is high and the HU (x_{p3}) orientation is low, then the AC (x_{l1}) mode is high and the AE (x_{l3}) mode is low. The second pair of canonical variates can be interpreted as follows: when the PR (x_{p1}) orientation is high and the HU (x_{p3}) orientation is low, then the AE (x_{l3}) mode is high and the CE (x_{l2}) is low. Our results are similar to those of Boyatzis et al. (2000), who reported the PR orientation as correlating positively with AE and negatively with CE; and the IN orientation as correlating positively with AC and negatively with AE.

5 Discussion

The increasing awareness of CoDa leads to an increasing interest in problems involving more than one composition. Standard statistical analysis includes many tools for relating two sets of variables, and one of the most popular in multivariate exploratory analysis is CCA. Within CoDa, tools for relating several compositions are still underdeveloped. In this article we have shown how to adapt CCA to compositional data in order to explore the relationship between two compositions. In our illustration we have found learning styles to be related to philosophical orientations in an interpretable manner in accordance with the literature, which supports the practical usefulness of the method as an exploratory tool.

The appeal of the CoDa log-ratio approach for applied researchers lies in the fact that once the data have been transformed using appropriate log-ratios, standard and well-understood statistical techniques such as CCA can be used. Once log-ratios have been computed, a compositional CCA is no more complicated than a standard CCA and standard statistical software dealing with CCA can be used. In order to be used with compositional

data, software must be able to derive the canonical variates from the covariance product in Equation (2.2) and include raw canonical coefficients as a part of the output. We recommend either SPSS, the *cca* function in the *yacca* R library (setting *xscale=FALSE*, *yscale=FALSE*), or the *cc* function in the *CCA* R library. It must be taken into account that some software for CCA either analyzes correlation matrices rather than covariance matrices (like the *canocor* function in the R library of the same name) or reports only standardized coefficients (like the *CCorA* function in the *vegan* R library). For the computation of canonical correlations and their significance tests, standardization or the use of correlations are irrelevant.

In some cases, the interpretation of the results of a statistical method on compositional data differs to some extent from its interpretation on unconstrained data. In the case of CCA, standardized results are neither usable nor needed, because unstandardized canonical variates can be interpreted as log-contrasts in a straightforward manner. This way of interpreting the results as log-contrasts fits well with the CoDa way of thinking and increases the attractiveness of the approach within an exploratory CoDa. CCA can also be applied to relate one composition to a set of numeric variables defined in the real space. In this case, the canonical variates are log-contrasts in the composition and linear combinations of the set of numeric variables with maximum mutual correlation.

The CoDa approach focuses on relative rather than absolute differences in the data. Treating compositional data directly without the log-ratio transformation implies assuming that the difference between scores 1 and 2 is the same as the difference between scores 10 and 11, while in the former case they differ by 100% and in the second by only 10%. A commonly mentioned limitation of the CoDa approach is the presence of zeros in the x_d variables, which prevents the analyst from computing log-ratios. Details on methods available for treating zeros prior to analysis, which perform well if the percentage of cases with zeros is not large, can be found in Martín-Fernández et al. (2011).

Further research could include adapting other multivariate techniques that relate sets of variables to compositional data, such as redundancy analysis, in order to derive a specified number of new latent variables from a composition that explains as much variance as possible from the other compositions. Related methods in the statistical modeling arena include simultaneous regression systems in which both explanatory and dependent variables are compositional (Tolosana-Delgado and van den Boogaart, 2013) and compositional partial least squares (Kalivodová et al., 2015).

Acknowledgements

The authors would like to acknowledge the support provided by Spanish Health Ministry Grant CB06/02/1002 funding the research group "Epidemiology and Public Health (CIBERESP)"; Catalan Autonomous Government Consolidated Research Group Grants 2014SGR551 and 2014SGR582 funding the research groups "Compositional and Spatial Data Analysis (COSDA)" and "Leadership Development Research Centre (GLEAD)"; Spanish Economy and Competitiveness Ministry grants MINECO/FEDER-EU MTM2015-65016-C2-1-R and EDU2015-68610-R funding the projects "COmpositional Data Analysis and RElated meThOdS (CODA-RETOS)" and "Assessing Individual and Team Entrepreneurial Potential"; and University of Girona grants MPCUdG2016/069 and MPCUdG2016/098.

References

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

[2] Aitchison, J. (1986): *The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability.* London: Chapman and Hall.

[3] Aitchison, J. (2001): Simplicial inference. In M.A. Viana and D.S. Richards (Eds): *Algebraic Methods in Statistics and Probability*. *Contemporary Mathematics Series of the American Mathematical Society, vol. 287,* 1-22. Providence, RI: American Mathematical Society.

[4] Aitchison, J. and Greenacre, M. (2002): Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 375-392.

[5] Batista-Foguet, J.M., Ferrer-Rosell, B., Serlavós, R., Coenders, G. and Boyatzis, R.E. (2015): An alternative approach to analyze ipsative data. Revisiting Experiential Learning Theory. *Frontiers in Psychology*, **6**, 1742.

[6] Billheimer, D., Guttorp, P. and Fagan, W. (2001): Statistical interpretation of species composition. *Journal of the American Statistical Association,* **96**, 1205-1214.

[7] Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013): *Analyzing Compositional Data with R*. Berlin: Springer.

[8] Boyatzis, R.E., Murphy, A.J. and Wheeler, J.V. (2000): Philosophy as a missing link between values and behaviour. *Psychological Reports,* **86**, 47-64.

[9] ter Braak, C.J.F. (1996). *Unimodal Models to Relate Species to Environment*. Wageningen, NL: DLO-Agricultural Mathematics Group.

[10] Coenders, G., Hlebec, V. and Kogovšek, T. (2011): Measurement quality in indicators of compositions. A compositional multitrait-multimethod approach. *Survey Research Methods,* **5**, 63-74.

[11] Egozcue, J.J., Daunis-i-Estadella, J., Pawlowsky-Glahn, V., Hron, K. and Filzmoser, P. (2012): Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, **6**, 87–108.

[12] Egozcue, J.J., and Pawlowsky-Glahn, V. (2011): Basic concepts and procedures. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 12-28. New York: Wiley.

[13] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003): Isometric logratio transformations for compositional data analysis. *Mathematical Geology,* **35**, 279-300.

[14] van Eijnatten, F.M., van der Ark, L.A. and Holloway, S.S. (2015): Ipsative measurement and the analysis of organizational values: an alternative approach for data analysis. *Quality & Quantity*, **49**, 559-579.

[15] Ferrer-Rosell, B. and Coenders, G. (2016): Destinations and crisis. Profiling tourists' budget share from 2006 to 2012. *Journal of Destination Marketing & Management*. doi: 10.1016/j.jdmm.2016.07.002

[16] Ferrer-Rosell, B., Coenders, G. and Martínez-Garcia, E. (2015): Determinants in tourist expenditure composition- the role of airline types. *Tourism Economics*, **21**, 9-32.

[17] Ferrer-Rosell, B., Coenders, G. and Martínez-Garcia, E. (2016a): Segmentation by tourist expenditure composition. An approach with compositional data analysis and latent classes. *Tourism Analysis,* **21**, 589-602.

[18] Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2016b): Understanding low cost airline users' expenditure pattern and volume. *Tourism Economics*, **22**, 269–291.

[19] Fry, T. (2011): Applications in economics. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 318-326. New York: Wiley.

[20] Hair, J.F, Black, W.C., Babin, B.J. and Anderson, R.E. (2009): *Multivariate Data Analysis. A Global Perspective (7th ed.).* Upper Saddle River, NJ: Pearson/Prentice Hall.

[21] Hlebec, V., Kogovšek, T. and Coenders, G. (2012): The measurement quality of social support survey measurement instruments. *Metodološki Zvezki*, **9**, 1-24.

[22] Hotelling, H. (1936): Relations between two sets of variates. *Biometrika,* **28**, 321-377.

[23] Hron, K., Filzmoser, P. and Thompson, K. (2012): Linear regression with compositional explanatory variables. *Journal of Applied Statistics,* **39**,. 1115-1128.

[24] Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H. and Adam, T. (2015): PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics*, **29**, 21-28.

[25] Kogovšek, T., Coenders, G. and Hlebec, V. (2013): Predictors and outcomes of social network compositions. A compositional structural equation modeling approach. *Social Networks,* **35**, 1-10.

[26] Kolb, D.A. (1984): *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice Hall.

[27] Kolb, D.A. (1999): *Learning Style Inventory, Version 3.* Boston, MA: Hay Resources Direct.

[28] Martín-Fernández, J.A., Daunis-i-Estadella, J. and Mateu-Figueras, G. (2015): On the interpretation of differences between groups for compositional data. *SORT- Statistics and Operations Research Transactions*, **39**, 231–252.

[29] Martín-Fernández, J.A., Palarea-Albaladejo, J. and Olea, R.A. (2011): Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 47-62. New York: Wiley.

[30] Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J.J. (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn and A. Buccianti (Eds): *Compositional Data Analysis. Theory and Applications,* 31-42. New York: Wiley.

[31] Pawlowsky-Glahn, V. and Buccianti, A. (2011): *Compositional Data Analysis. Theory and Applications*. New York: Wiley.

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

[33] Pawlowsky-Glahn, V. and Egozcue, J.J. (2011): Exploring compositional data with the CoDa-dendrogram. *Austrian Journal of Statistics,* **40**, 103–113.

[34] Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015): *Modeling and Analysis of Compositional Data*. Chichester: Wiley.

[35] Pearson, K. (1897): Mathematical contributions to the theory of evolution. On a form of spurious correlations which may arise when indices are used in the measurements of organs. *Proceedings of the Royal Society*, **60**, 489-498.

[36] Thió-Henestrosa, S. and Martín-Fernández, J.A. (2005): Dealing with compositional data: The freeware CoDaPack. *Mathematical Geology,* **37**, 773-793.

[37] Tolosana-Delgado, R. and van den Boogaart, K.G. (2013): Regression between compositional data sets. In K. Hron, P. Filzmoser and M. Templ (Eds): *Proceedings of the 5th International Workshop on Compositional Data Analysis (CoDaWork 2013)*, 163-176. Vienna: Vienna University of Technology.

[38] Vives-Mestres, M., Martín-Fernández, J.A. and Kenett, R. (2016): Compositional data methods in customer survey analysis. *Quality and Reliability Engineering International,* **32**, 2115-2125.

[39] de Vries, A.L.M. and van der Ark, L.A. (2008): Scoring methods for ordinal multidimensional forced-choice items. In J. Daunis-i-Estadella and J.A. Martín-Fernández (Eds): *Proceedings of the 3rd International Workshop on Compositional Data Analysis (CoDaWork 2008)*, 1-18. Girona: University of Girona.