Comput Stat (2015) 30:1079–1096
DOI 10.1007/s00180-015-0570-1

ORIGINAL PAPER

Principal component analysis for compositional data vectors

Huiwen Wang · Liying Shangguan · Rong Guan · Lynne Billard

Received: 30 November 2012 / Accepted: 20 February 2015 / Published online: 4 March 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Since Aitchison's founding research work, compositional data analysis has attracted growing attention in recent decades. As a powerful technique for exploratory analysis, principal component analysis (PCA) has been extended to compositional data. Despite extensive efforts in PCA on compositional data parts as variables, this paper contributes to modeling PCA for compositional data vectors. Based on algebraic operators in Simplex space, the PCA process is deduced and transformed into calculating some inner products. Properties of principal components are also investigated. Two real-data examples illustrate the merits of the proposed PCA for compositional data vectors.

Keywords Compositional data · Principal component analysis (PCA) · Simplex space · Logratio transformation

H. Wang
School of Economics and Management, Beihang University, Beijing 100191, China
e-mail: wanghw@vip.sina.com

L. Shangguan
Postdoctoral Programme, China Galaxy Securities, Beijing 100033, China
e-mail: shgliying@outlook.com

R. Guan (✉)
School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China
e-mail: rongguan77@gmail.com; guanrong721@qq.com

L. Billard
Department of Statistics, University of Georgia, Athens, GA 30602, USA
e-mail: lynne@stat.uga.edu

1 Introduction

Principal component analysis (PCA), a mathematical procedure for screening multivariate data, aims to reduce the dimensionality of a data set (Jolliffe 2002). By using an orthogonal transformation, PCA converts a set of correlated variables into a small set of uncorrelated variables that are called principal components (PCs), which retain most of the variation presented in all the original variables. By successfully reducing the dimensionality, PCA can thus help to identify new meaningful underlying variables and to understand better the correlation structure among the original variables (Jolliffe 2002). The classical PCA method has been confronted with difficulties in dealing with complex types of data, such as interval-valued data (Cazes et al. 1997; Gioia and Lauro 2006; Wang et al. 2012), compositional data (Aitchison 1983, 1984), functional data (Ramsay and Silverman 2005; Valderrama 2007; Bali et al. 2011; Sawant et al. 2012) and the like. In this paper, our concern is compositional data.

Compositional data are usually described as proportions or percentages that only carry relative information (Aitchison 1986; Bacon-Shone 2011). In mathematical notation, D-part **compositional data** are expressed as $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, subject to the so-called unit-sum constraint, i.e.,

$$x_j > 0, \quad (j = 1, 2, \ldots, D) \tag{1}$$

and

$$\sum_{j=1}^{D} x_j = 1. \tag{2}$$

In this paper, we call x_j a **compositional data part** (or part for short) of $\mathbf{x}$. A vector $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]'$ constituted by p D-part compositional data will be referred to as a **p-dimensional compositional data vector** (or compositional data vector for short). A set of compositional data vectors are a **multiple compositional data set** or a **multiple compositional data matrix**.

Compositional data occur in many fields (Filzmoser and Hron 2011). For instance, in the economic sphere, economists often discuss *gross domestic products* (*GDP*), *employment* and *investment* across *primary*, *secondary*, and *tertiary* sectors in a region or a country (Fisher 1939; Wang et al. 2007; Orlik 2011). That is, three compositional data are concerned, each of which is organized by three industrial-sector parts. To understand better the relationship among *GDP*, *employment* and *investment* across three industry sectors for n different regions in a country, where n is the sample size, a PCA technique for compositional data vectors can be considered. Another example concerns industrial production of China. Here, we want to reduce the dimensionality of the output proportions of major industrial products (including *coal*, *crude oil*, *electricity*, *crude steel*, and *cement*) across four regions in the country, i.e., *Eastern China*, *Northeastern China*, *Central China* and *Western China*. PCA shall be used to quantify the internal structure of the 5 compositional data variables, rather than 5×4 parts. This case will be investigated later in our application analysis.

However, neither classical PCA or PCA method for compositional data analysis existed in current literature could cope with the abovementioned situations. Due to the

closure property, the correlation structure of compositional data will be spoiled and the analytical results will become doubtful, if the classical PCA is directly applied on raw data or components (Aitchison 1986). Additionally, the existing PCA methods for compositional data analysis concerns only component variables. In this way, the classical PCA is applied on transformed data by centered logratio (*clr*) transformation for compositional data (Aitchison 1982, 1983, 1984; Aitchison and Greenacre 2002). Later, researchers discussed robust PCA on components as variables by using isometric logratio (*ilr*) transformation (Filzmoser 1999; Filzmoser et al. 2009). Indeed, previous methods have not yet addressed the proposed problem. In light of this, we are motivated to investigate a novel PCA modeling procedure for compositional data vectors in this paper.

The remaining parts of this paper proceed as follows: in Sect. 2, the algebraic system of compositional data vectors is briefly introduced. Section 3 investigates PCA for compositional data vectors, where both covariance and correlation coefficient are proposed, based on the definition of inner product. To illustrate the usefulness of the proposed methods, two real-life examples are presented in Sect. 4. Finally, there are some conclusions and remarks in Sect. 5.

2 Preliminaries

To start with, we will present some basic notations in this section. For convenience, in what follows, we use "[. . .]" to denote a compositional data in Simplex space, and "(. . .)" to represent a vector in real space.

2.1 Compositional data

In general, a Simplex space of D parts, or Simplex for short, is defined as

$$S^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \,\middle|\, x_j > 0, j = 1, 2, \ldots, D; \sum_{j=1}^{D} x_j = 1 \right\}. \quad (3)$$

Thus, Simplex elements $\mathbf{x} \in S^D$ are referred to as D-part compositional data.

Algebraic operations on the Simplex space have great importance for compositional data analysis. Given any two compositional data $\mathbf{x}, \mathbf{y} \in S^D$, two basic algebraic operations, i.e., perturbation operator and power transformation, have been introduced in the literature (Aitchison 1986). Based on the operations, Aitchison (Aitchison 1986) stated some principles including scale invariance and subcompositional coherence for the analysis of compositional data.

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D), \quad (4)$$

$$\beta \otimes \mathbf{x} = \mathcal{C}\left(x_1^{\beta}, x_2^{\beta}, \ldots, x_D^{\beta}\right), \quad \forall \beta \in \mathbb{R}, \quad (5)$$

where Ç denotes the closure operator,

$$\text{Ç}\,(z_1, z_2, \ldots, z_D) = \left[\frac{z_1}{\sum_{i=1}^{D} z_i}, \frac{z_2}{\sum_{i=1}^{D} z_i}, \ldots, \frac{z_D}{\sum_{i=1}^{D} z_i}\right]. \tag{6}$$

Accordingly, $\mathbf{x} \ominus \mathbf{y}$ can be deduced with the perturbation operator and power transformation,

$$\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus ((-1) \otimes \mathbf{y}) = \text{Ç}\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \ldots, \frac{x_D}{y_D}\right). \tag{7}$$

Clr transformation (Aitchison 1986), a well-known transformation for compositional data, is the basic element in the definition of the inner product, norm and distance for compositional data (Pawlowsky-Glahn and Egozcue 2001). Given any compositional data $\mathbf{x}$, *clr* transformation is defined as the mapping $clr : S^D \to \mathbb{R}^D$,

$$clr(\mathbf{x}) = \mathbf{z} = (z_1, z_2, \ldots, z_D) = \left(\log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \ldots, \log \frac{x_D}{g(\mathbf{x})}\right), \tag{8}$$

where $g(\mathbf{x}) = \sqrt[D]{\prod_{i=1}^{D} x_i}$ is the geometric mean of $\mathbf{x}$. It is worth noting that *clr* transformation can not be used for compositional data containing zero part. If that is the case, compositional data shall be prepared in advance by a multiplicative log-normal replacements, or other methods (Palarea-Albaladejo and Martín-Fernández 2013). The corresponding inverse transformation $clr^{-1} : \mathbb{R}^D \to S^D$ is given by

$$\begin{aligned} clr^{-1}(\mathbf{z}) &= \mathbf{x} = [x_1, x_2, \ldots, x_D] \\ &= \left[\frac{\exp(z_1)}{\sum_{k=1}^{D} \exp(z_k)}, \frac{\exp(z_2)}{\sum_{k=1}^{D} \exp(z_k)}, \ldots, \frac{\exp(z_D)}{\sum_{k=1}^{D} \exp(z_k)}\right]. \end{aligned} \tag{9}$$

Equipped with the *clr* transformation, the Aitchison inner product (Pawlowsky-Glahn and Egozcue 2001) for any two compositional data $\mathbf{x}, \mathbf{y} \in S^D$ is defined as

$$\langle \mathbf{x}, \mathbf{y} \rangle_{S^D} = \langle clr(\mathbf{x}), clr(\mathbf{y}) \rangle = \sum_{i=1}^{D} \log \frac{x_i}{g(\mathbf{x})} \log \frac{y_i}{g(\mathbf{y})}. \tag{10}$$

The subscript "S^D" emphasizes that the inner product is used in the Simplex S^D. The inner product can further induce the Aitchison norm and the Aitchison distance in S^D,

$$\|\mathbf{x}\|_{S^D}^2 = \langle \mathbf{x}, \mathbf{x} \rangle_{S^D}, \tag{11}$$

$$d_{S^D}^2(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{S^D}^2 = \sum_{i=1}^{D} \left(\log \frac{x_i}{g(\mathbf{x})} - \log \frac{y_i}{g(\mathbf{y})}\right)^2. \tag{12}$$

Pawlowsky-Glahn and Egozcue (2001) proved that the vector space S^D is a Hilbert space of D-1 dimension.

2.2 Multiple compositional data geometry

Let $\mathbf{U}_{n\times(D\times p)}$ be a multiple compositional data matrix with n vectors constituted by p D-part, we have

$$\mathbf{U}_{n\times(D\times p)} = (\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p) = \begin{pmatrix} \mathbf{O}_1^{'} \\ \mathbf{O}_2^{'} \\ \vdots \\ \mathbf{O}_n^{'} \end{pmatrix} = \begin{pmatrix} \mathbf{u}_{11} & \mathbf{u}_{12} & \ldots & \mathbf{u}_{1p} \\ \mathbf{u}_{21} & \mathbf{u}_{22} & \ldots & \mathbf{u}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_{n1} & \mathbf{u}_{n2} & \ldots & \mathbf{u}_{np} \end{pmatrix}, \tag{13}$$

where $\mathbf{U}_j = (\mathbf{u}_{1j}, \mathbf{u}_{2j}, \ldots, \mathbf{u}_{nj})'$ represents the j-th variable with n D-part compositional data, and $\mathbf{O}_i = (\mathbf{u}_{i1}, \mathbf{u}_{i2}, \ldots, \mathbf{u}_{ip})'$ is the i-th compositional data vector containing p D-part compositional data $\mathbf{u}_{ij} = [u_{ij1}, u_{ij2}, \ldots, u_{ijD}]$. Noticeably, $\mathbf{u}_{ij}$ belongs to S^D, thus $\mathbf{O}_i^{'}$ is in the space $S^D \times S^D \times \ldots \times S^D$. We denote this space as S^{Dp} to remark that the space is the cartesian product of p D-part Simplex spaces.

Based on the operations of compositional data, all operations of p-dimensional compositional data vectors can be easily deduced by using a component-wise manner (Gallo 2012, 2013). That is,

$$\mathbf{O}_i^{'} \oplus \mathbf{O}_{i'}^{'} = \left(\mathbf{u}_{i1} \oplus \mathbf{u}_{i'1},\ \mathbf{u}_{i2} \oplus \mathbf{u}_{i'2},\ \ldots,\ \mathbf{u}_{ip} \oplus \mathbf{u}_{i'p}\right), \tag{14}$$

$$\beta \otimes \mathbf{O}_i^{'} = \left(\beta \otimes \mathbf{u}_{i1},\ \beta \otimes \mathbf{u}_{i2},\ \ldots,\ \beta \otimes \mathbf{u}_{ip}\right), \quad \forall \beta \in \mathbb{R}, \tag{15}$$

$$\mathbf{O}_i^{'} \ominus \mathbf{O}_{i'}^{'} = \left(\mathbf{u}_{i1} \ominus \mathbf{u}_{i'1},\ \mathbf{u}_{i2} \ominus \mathbf{u}_{i'2},\ \ldots,\ \mathbf{u}_{ip} \ominus \mathbf{u}_{i'p}\right), \tag{16}$$

$$\langle \mathbf{O}_i^{'}, \mathbf{O}_{i'}^{'} \rangle_{S^{Dp}} = \sum_{j=1}^{p} \langle \mathbf{u}_{ij}, \mathbf{u}_{i'j} \rangle_{S^D}, \tag{17}$$

$$\|\mathbf{O}_i^{'}\|_{S^{Dp}}^2 = \sum_{j=1}^{p} \|\mathbf{u}_{ij}\|_{S^D}^2, \tag{18}$$

$$d_{S^{Dp}}^2(\mathbf{O}_i^{'}, \mathbf{O}_{i'}^{'}) = \sum_{j=1}^{p} d_{S^D}^2(\mathbf{u}_{ij}, \mathbf{u}_{i'j}). \tag{19}$$

It is easily proved that the inner product for compositional data vectors, defined in Eq. (17), satisfies the usual axioms as listed below, which will account for the inner product space.

Proposition 1 *For any compositional data vectors* $\mathbf{O}_i^{'}, \mathbf{O}_{i'}^{'}$ *and* $\mathbf{O}_{i''}^{'} \in S^{Dp}$*, the inner product satisfies*

(1) *Positive definiteness, i.e.,* $\langle \mathbf{O}_i^{'}, \mathbf{O}_i^{'} \rangle_{S^{Dp}} \geq 0$*, and the equality holds if and only if*

$$\mathbf{O}_i^{'} = \left(\left[\tfrac{1}{D}, \tfrac{1}{D}, \ldots, \tfrac{1}{D}\right], \left[\tfrac{1}{D}, \tfrac{1}{D}, \ldots, \tfrac{1}{D}\right], \ldots, \left[\tfrac{1}{D}, \tfrac{1}{D}, \ldots, \tfrac{1}{D}\right]\right); \tag{20}$$

(2) *Symmetry, i.e.,* $\langle \mathbf{O}_i^{'}, \mathbf{O}_{i'}^{'} \rangle_{S^{Dp}} = \langle \mathbf{O}_{i'}^{'}, \mathbf{O}_i^{'} \rangle_{S^{Dp}}$*;*

(3) *Linearity, i.e.,* $\langle \mathbf{O}_i^{'} \oplus \mathbf{O}_{i''}^{'}, \mathbf{O}_{i''}^{'} \rangle_{S^{Dp}} = \langle \mathbf{O}_i^{'}, \mathbf{O}_{i''}^{'} \rangle_{S^{Dp}} + \langle \mathbf{O}_{i'}^{'}, \mathbf{O}_{i''}^{'} \rangle_{S^{Dp}}$, *and* $\langle \beta \otimes \mathbf{O}_i^{'}, \mathbf{O}_{i'}^{'} \rangle_{S^{Dp}} = \beta \langle \mathbf{O}_i^{'}, \mathbf{O}_{i'}^{'} \rangle_{S^{Dp}}, \forall \beta \in \mathbb{R}$.

3 The PCA method for compositional data vectors

To facilitate derivation, it is necessary to define the sample mean, sample variance, sample covariance and sample correlation coefficient for compositional data variables. Along with an algebraic viewpoint, we will prove that the modeling process of PCA will then be transformed into calculating the algebraic operations of compositional data vectors.

3.1 Sample statistics

For n observations of any random compositional data variable $\mathbf{U}_j$, the sample mean is given by Martín-Fernández et al. (1998), Pawlowsky-Glahn and Egozcue (2002)

$$E_{S^D}(\mathbf{U}_j) = \overline{\mathbf{u}}_j = \text{Ç}\left(g(\mathbf{L}_{j1}), g(\mathbf{L}_{j2}), \ldots, g(\mathbf{L}_{jD})\right), \tag{21}$$

where $\mathbf{L}_{jk} = (u_{1jk}, u_{2jk}, \ldots, u_{njk})'$, $k = 1, \ldots, D$, and $g(\cdot)$ is the geometric mean. For the sake of convenience, we denote $cen(\mathbf{U})_j = \left(\overline{\mathbf{u}}_j, \overline{\mathbf{u}}_j, \ldots, \overline{\mathbf{u}}_j\right)'$, and therefore express the centralization of $\mathbf{U}_j$ as

$$\mathbf{U}_j \ominus cen(\mathbf{U})_j. \tag{22}$$

Definition 1 For n observations of any random compositional data variable $\mathbf{U}_j$, the sample variance is defined by

$$Var_{S^D}(\mathbf{U}_j) = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{u}_{ij} \ominus \overline{\mathbf{u}}_j \right\|_{S^D}^2. \tag{23}$$

Remark According to Aitchison (1997), the concept of "total variance" is established as

$$totVar(\mathbf{U}_j) = \sum_{k=1}^{D} Var\left[clr_k(\mathbf{U}_j)\right], \tag{24}$$

where $clr_k(\mathbf{U}_j)$ denotes the k-th column in the transformed-data matrix $clr(\mathbf{U}_j)$ with

$$clr(\mathbf{U}_j) = \begin{pmatrix} clr(\mathbf{u}_{1j}) \\ clr(\mathbf{u}_{2j}) \\ \vdots \\ clr(\mathbf{u}_{nj}) \end{pmatrix}. \tag{25}$$

We could easily deduce that the variance defined in Eq. (23) is equal to the total variance, i.e.,

$$Var_{S^D}(\mathbf{U}_j) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{D}\left[clr_k(\mathbf{u}_{ij}) - clr_k(\overline{\mathbf{u}}_j)\right]^2 = \sum_{k=1}^{D} Var\left[clr_k(\mathbf{U}_j)\right]. \quad (26)$$

According Pawlowsky-Glahn and Egozcue (2001), the standardization of $\mathbf{U}_j$ can be given by

$$\frac{1}{\sqrt{Var_{S^D}(\mathbf{U}_j)}} \otimes (\mathbf{U}_j \ominus cen(\mathbf{U})_j). \quad (27)$$

Definition 2 For any two random compositional data variables $\mathbf{U}_j$ and $\mathbf{U}_{j'}$, the covariance and the correlation coefficient are, respectively, given by

$$Cov_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) = \frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{u}_{ij} \ominus \overline{\mathbf{u}}_j, \mathbf{u}_{ij'} \ominus \overline{\mathbf{u}}_{j'}\rangle_{S^D}, \quad (28)$$

$$r_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) = \frac{Cov_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'})}{\sqrt{Var_{S^D}(\mathbf{U}_j)}\sqrt{Var_{S^D}(\mathbf{U}_{j'})}}. \quad (29)$$

Remark The correlation coefficient defined in Eq. (29) follows the paradigm of Pearson's product–moment correlation coefficient. We could conclude that two compositional data variables $\mathbf{U}_j$ and $\mathbf{U}_{j'}$ are totally linearly correlated, namely, $\mathbf{U}_j = \beta \otimes \mathbf{U}_{j'} \oplus \mathbf{U}_{j''}$, where β is a real number and $\mathbf{U}_{j''}$ is a constant compositional data vector, if and only if

$$r_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) = \begin{cases} 1, & \text{if } \beta > 0 \\ -1, & \text{if } \beta < 0 \end{cases}. \quad (30)$$

For p-dimensional Euclidean space $\mathcal{R}^p$, the subset is $\mathcal{R}^q$, where $q \leq p$. Along this line of thought, the subset of S^{Dp}, namely, compositional data subvector is defined as follows.

Definition 3 Given a composition data vector $\mathbf{O}_i = (\mathbf{u}_{i1}, \ldots, \mathbf{u}_{ip})' \in S^{Dp}$, containing p D-part compositional data $\mathbf{u}_{ij} = [u_{ij1}, u_{ij2}, \ldots, u_{ijD}]$, a composition data subvector is $\mathbf{O}_i^{p^*} = (\mathbf{u}_{i1^*}, \ldots, \mathbf{u}_{ip^*})' \in S^{Dp^*}$ with p^* compositional data where the subindexes $(1^*, \ldots, p^*)$ indicate which compositional data are selected in the composition data vector, not have to be the first p^* ones.

According to Definition 3, the sample statistics obtained from compositional data subvector are not contradictory with those obtained from the full compositional data vector. Again, we emphasize that this definition focuses on reducing the dimension "p" rather than the part "D". Subcomposition data vector is obtained when the part "D" is reduced and the sample statistics could be different. Based on the sample statistics in our paper, the results are not subcompositional coherent.

Proposition 2 *For any two random compositional data variables* $\mathbf{U}_j, \mathbf{U}_{j'}$ *and any scalar* $\beta \in \mathbb{R}$*, we have*

$$E_{S^D}(\beta \otimes \mathbf{U}_j \oplus \mathbf{U}_{j'}) = \beta \otimes E_{S^D}(\mathbf{U}_j) \oplus E_{S^D}(\mathbf{U}_{j'}), \tag{31}$$

$$Var_{S^D}(\beta \otimes \mathbf{U}_j \oplus \mathbf{U}_{j'}) = \beta^2 Var_{S^D}(\mathbf{U}_j) + 2\beta Cov_{S^D}(\mathbf{U}_j, \mathbf{U}_{j'}) + Var_{S^D}(\mathbf{U}_{j'}). \tag{32}$$

The notations in Proposition 2 will be useful later on in the context of PCA.

3.2 The PCA algorithm

Suppose there is a sample of n compositional data vectors described by p variables $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$. Using the notations from previous paragraphs, we know that the k-th ($1 \leq k \leq p$) principal component $\mathbf{V}_k$ is a linear combination of $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$, i.e.,

$$\mathbf{V}_k = \bigoplus_{j=1}^{p} (e_{kj} \otimes \mathbf{U}_j) = e_{k1} \otimes \mathbf{U}_1 \oplus e_{k2} \otimes \mathbf{U}_2 \oplus \ldots \oplus e_{kp} \otimes \mathbf{U}_p, \tag{33}$$

where the vector $\mathbf{e}_k = (e_{k1}, e_{k2}, \ldots, e_{kp})' \in \mathbb{R}^p$ is chosen such that $Var_{S^D}(\mathbf{V}_k)$ is maximum, subject to $\|\mathbf{e}_k\| = 1$ and $\mathbf{e}_k' \mathbf{e}_l = 0, l = 1, 2, \ldots, p, l \neq k$.

For notational simplicity, we assume that the given compositional data variables $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$ have already been centralized. To find $\mathbf{e}_k$, the variance of the k-th principal component is expressed as

$$\begin{aligned} Var_{S^D}(\mathbf{V}_k) &= Var_{S^D}(e_{k1} \otimes \mathbf{U}_1 \oplus e_{k2} \otimes \mathbf{U}_2 \oplus \ldots \oplus e_{kp} \otimes \mathbf{U}_p) \\ &= \sum_{i=1}^{p} \sum_{j=1}^{p} e_{ki} e_{kj} Cov_{S^D}(\mathbf{U}_i, \mathbf{U}_j) \\ &= \mathbf{e}_k' \mathbf{W} \mathbf{e}_k, \end{aligned} \tag{34}$$

where $\mathbf{W}$ denotes the variance-covariance matrix of $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$.

The first m principal components ($m \leq p$) are determined by the m orthonormalized vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ that maximize $\sum_{k=1}^{m} Var_{S^D}(\mathbf{V}_k)$ subject to $Var_{S^D}(\mathbf{V}_1) \geq Var_{S^D}(\mathbf{V}_2) \geq \ldots \geq Var_{S^D}(\mathbf{V}_m)$ by the following optimization problem:

$$\max_{\mathbf{e}_k} \sum_{k=1}^{m} \mathbf{e}_k' \mathbf{W} \mathbf{e}_k$$
$$s.t. \begin{cases} \| \mathbf{e}_k \| = 1, \quad for\ k = 1, 2, \ldots, m; \\ \mathbf{e}_k' \mathbf{e}_l = 0, \quad for\ k, l = 1, 2, \ldots, m, l \neq k; \\ \mathbf{e}_1' \mathbf{W} \mathbf{e}_1 \geq \mathbf{e}_2' \mathbf{W} \mathbf{e}_2 \geq \ldots \geq \mathbf{e}_m' \mathbf{W} \mathbf{e}_m, \ for\ m \leq p. \end{cases} \tag{35}$$

The solution to this optimization problem gives the first m eigenvectors of $\mathbf{W}$, respectively denoted as $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, corresponding to the first m eigenvalues of $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0$.

Clearly, the derivation of PCA for multiple compositional data variables is thereby transformed into the eigen-decomposition of the variance-covariance matrix $\mathbf{W}$. According to the derivation of variance and covariance in Sect. 3.1, we calculate $\mathbf{W}$ for compositional data matrix by Eqs. (23) and (28).

Eventually, the modeling process of PCA for multiple compositional data variables can be summarized as the following steps.

- **Step 1:** Centralize each compositional data variable $\mathbf{U}_j$ $(j = 1, 2, \ldots, p)$ by using Eq. (22).
- **Step 2:** Calculate the variance-covariance matrix $\mathbf{W}$ by using Eqs. (23) and (28).
- **Step 3:** Eigen-decompose $\mathbf{W}$ to obtain the first eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ corresponding to the first eigenvalues in the decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > 0$.
- **Step 4:** Compute the k-th principal component by $\mathbf{V}_k = \bigoplus_{j=1}^{p}(e_{kj} \otimes \mathbf{U}_j)$ $(k = 1, \ldots, m)$.

As expected, the proposed method can reduce the dimensionality from the space S^{Dp} into the space S^{Dm}, $m \leq p$. What shall be stressed here is that, it depends on the real situation whether to use variance-covariance matrix or correlation-coefficient matrix in **Step 2**. Generally speaking, the range and scale of an compositional data variable can be measured by its variance. When the variables are similar in terms of range and scale or in the same units of measure, we shall use the variance-covariance matrix. Otherwise, the correlation-coefficient matrix shall be adopted to deal with too-large differences of within-variable range and scale.

3.3 Properties of PCA

Given that the original compositional data variables $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_p$ have been centralized, the obtained PCs $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m$ $(m \leq p)$ satisfy the following properties. Proofs of these are similar to those for classical PCA.

Proposition 3 $E_{S^D}(\mathbf{V}_k) = \left[\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right]$, $1 \leq k \leq p$.

Proposition 4 $Var_{S^D}(\mathbf{V}_k) = \lambda_k$, $1 \leq k \leq p$.

Proposition 5 $Cov_{S^D}(\mathbf{V}_k, \mathbf{V}_l) = 0$, $1 \leq k, l \leq p$, $l \neq k$.

Proposition 6 $\mathbf{U}_j = \bigoplus_{k=1}^{p}(e_{kj} \otimes \mathbf{V}_k)$, $1 \leq j \leq p$.

Proposition 7 $\sum_{j=1}^{p} Var_{S^D}(\mathbf{U}_j) = \sum_{k=1}^{p} Var_{S^D}(\mathbf{V}_k)$.

Similar to the process of traditional PCA, $Var_{S^D}(\mathbf{V}_k)$ indicates the variance carried by the k-th PC. Accordingly, the Cumulative Contribution Rate (CCR) of the first m PCs is

$$CCR_m = \frac{\sum_{k=1}^{m} Var_{S^D}(\mathbf{V}_k)}{\sum_{j=1}^{p} Var_{S^D}(\mathbf{V}_j)} = \frac{\sum_{k=1}^{m} \lambda_k}{\sum_{j=1}^{p} \lambda_j}. \tag{36}$$

If the PCA process is based on the correlation-coefficient matrix, the CCR_m is

$$CCR_m = \frac{1}{p}\sum_{k=1}^{m}\lambda_k. \tag{37}$$

Proposition 8 $r_{S^D}(\mathbf{V}_k, \mathbf{U}_j) = \frac{\sqrt{\lambda_k}}{\sqrt{Var_{S^D}(\mathbf{U}_j)}}e_{kj}$, $1 \le k \le p, 1 \le j \le p$.

In Proposition 8, $r_{S^D}(\mathbf{V}_k, \mathbf{U}_j)$ is the association between the k-th PC $\mathbf{V}_k$ and the j-th original variable $\mathbf{U}_j$. If all the variables $\mathbf{U}_j$ have been standardized, namely, $Var_{S^D}(U_j) = 1$, Proposition 8 can be simplified into $r_{S^D}(\mathbf{V}_k, \mathbf{U}_j) = \sqrt{\lambda_k}e_{kj}$.

4 Application

In this section, we aim to demonstrate the usefulness of the proposed PCA method for compositional data variables by two real-life applications. The first case is to investigate the industrial outputs across four regions of China. The second case mainly concerns a comprehensive evaluation of scientific journals from different disciplines.

4.1 Case study: industrial outputs

In developing countries such as China, industrial production has been considered as the significant drive of economic growth. Generally speaking, a great amount of industrial outputs is a good indicator of an active economy. In this case, we will mainly focus on the proportional distribution of industrial outputs across different regions in the country, and expect to investigate the innate structure and associations of industrial outputs.

According to National Bureau of Statistics of China, major industrial products include coal, crude oil, electricity, crude steel and cement. The percentages of each product in four regions, i.e., *Eastern China*, *Northeastern China*, *Central China* and *Western China*, construct a 4-part compositional data. We collect these proportional data from 2005 to 2011, from China Statistics Yearbook. The data set thus includes 7 observations featured by 5 compositional data variables, i.e., *coal* ($\mathbf{U}_1$), *crude oil* ($\mathbf{U}_2$), *electricity* ($\mathbf{U}_3$), *crude steel* ($\mathbf{U}_4$), and *cement* ($\mathbf{U}_5$), as listed in Table 1. It is worthy to note that the case study serves only as an illustrative application of our approach, but not a thorough investigation of the outputs of industrial products which would need results from more observations.

According to Table 1, the variances are, respectively, 0.0476, 0.0131, 0.0113, 0.0018, 0.0356 for the five variables $\mathbf{U}_1, \ldots, \mathbf{U}_5$. Considering the large differences in variable variances, the proposed analytical methods are performed on the correlation-coefficient matrix in this case. Table 2 has displayed the variance of each PC, denoted as $\mathbf{V}_k (1 \le k \le 5)$, the corresponding CCR, and the correlation coefficients between the PCs and the original variables.

Table 1 The multiple compositional data set of industrial outputs

Year	$\mathbf{U}_1(\%)$	$\mathbf{U}_2(\%)$	$\mathbf{U}_3(\%)$	$\mathbf{U}_4(\%)$	$\mathbf{U}_5(\%)$
2005	[12.9,8.4,41.8,36.9]	[37.1,34.9,3.2,24.8]	[45.1,7.7,22.7,24.4]	[55.2,10.7,21.2,12.9]	[51.6,5.3,22.5,20.6]
2006	[11.8,8.7,40.3,39.1]	[37.2,33.8,3.1,25.9]	[44.8,7.3,22.2,25.7]	[54.7,10.9,21.5,12.9]	[50.9,5.3,23.4,20.4]
2007	[11.2,7.8,40.3,40.6]	[36.8,32.2,3.1,27.9]	[43.2,7.0,22.9,26.9]	[54.2,10.6,22.2,13.0]	[47.6,5.5,25.0,21.9]
2008	[9.7,7.2,39.0,44.0]	[37.0,31.2,2.9,28.9]	[42.2,6.8,23.0,28.0]	[55.2,10.4,21.6,12.8]	[45.0,6.2,25.7,23.0]
2009	[9.6,6.7,35.9,47.8]	[38.4,29.8,2.9,28.9]	[41.1,6.5,23.2,29.1]	[55.5,10.8,20.7,13.0]	[42.0,6.7,26.0,25.3]
2010	[10.1,6.0,33.8,50.1]	[40.5,27.9,2.9,28.8]	[41.5,6.4,23.1,29.1]	[54.3,11.0,21.5,13.1]	[40.4,6.1,25.1,28.4]
2011	[9.1,5.5,33.4,51.9]	[38.7,28.3,2.8,30.2]	[41.2,6.2,22.8,29.8]	[54.4,10.2,21.3,14.1]	[38.4,6.7,25.4,29.6]

Table 2 PC variance, CCR and correlation coefficients

		$\mathbf{V}_1$	$\mathbf{V}_2$	$\mathbf{V}_3$	$\mathbf{V}_4$	$\mathbf{V}_5$
$r_{S_n^D}(\mathbf{V}_k, \mathbf{U}_j)$	$\mathbf{U}_1$	0.9751	0.0105	0.1273	−0.0671	−0.1686
	$\mathbf{U}_2$	0.7635	−0.6015	−0.0094	−0.2165	0.0911
	$\mathbf{U}_3$	0.9230	−0.1013	0.2102	0.3005	0.0580
	$\mathbf{U}_4$	0.5613	0.2634	−0.7840	0.0279	0.0118
	$\mathbf{U}_5$	0.4801	0.8217	0.2687	−0.1295	0.0722
PC variance		2.9312	1.1168	0.7474	0.1592	0.0454
$CCR(\%)$		58.62	80.96	95.91	99.09	100.00

Table 3 The *PC1* scores of each year

Year	*PC1* scores (%)
2005	[32.7,48.1,17.3,1.9]
2006	[28.5,50.6,16.7,4.2]
2007	[22.8,28.7,35.9,12.5]
2008	[18.6,23.6,32.2,25.6]
2009	[16.6,18.6,18.2,46.6]
2010	[17.0,8.8,15.7,58.5]
2011	[5.4,2.2,6.2,86.2]

Remarkably, all the original variables are positively correlated to $\mathbf{V}_1$, which accounts for 58.62 % of the total variance. $\mathbf{V}_1$ can be written as

$$\mathbf{V}_1 = 0.5695 \otimes \mathbf{U}_1 \oplus 0.4460 \otimes \mathbf{U}_2 \oplus 0.5391 \otimes \mathbf{U}_3 \oplus 0.3278 \otimes \mathbf{U}_4 \oplus 0.2804 \otimes \mathbf{U}_5. \quad (38)$$

Then, the *PC1* scores can be calculated and are shown in Table 3. A stacked bar chart is used to visualize the compositional data scores of $\mathbf{V}_1$ for each of the 7 observations (see Fig. 1a). As can be seen, there exists an increase in output percentages of Western China year by year, while the proportion of Eastern China or Northeastern China fall in general. The proportions of Central China increase first and then decrease. Most

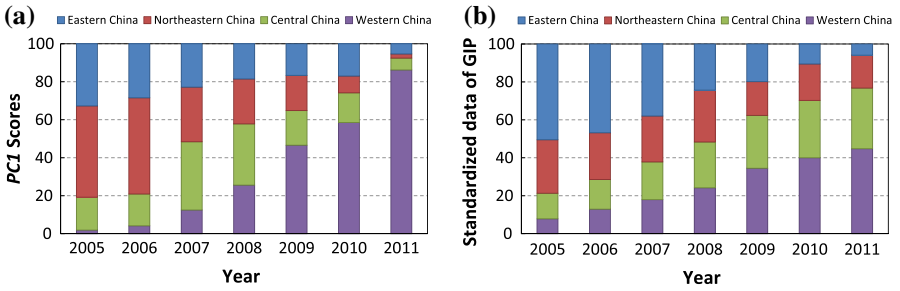

Fig. 1 Industrial outputs: **a** $PC1$ Scores, **b** standardized data of GIP (color figure online)

Fig. 2 Distance of $PC1$ Scores in adjacent years

probably, this is due to the advantages of having rich natural resources and low labor cost in the Western China. Moreover, since the launch of the well-known Great Western Development Strategy, both investments and supporting policies have offered a boost to local economic activities.

According to the meaning of $PC1$, we may expect to observe the changing consistency between $PC1$ scores and Gross Industrial Production (GIP) (also collected from National Bureau of Statistics of China), since industrial outputs generally reflect the prosperity of industrial economy. By using Eq. (29), the correlation coefficient between GIP variable and $PC1$ is 0.6658. Comparing Fig. 1a, b, we see that the moving tendency of the $PC1$ score roughly agrees with the standardization data of GIP. This has well demonstrated that the analytical result of the proposed PCA method is coherent to real life.

As aforementioned, compositional data work as a description of structure. Distances between compositional data in the neighboring years are proposed to reflect the similarity/dissimilarity of the industrial production activities. Figure 2 exhibits the dissimilarity *PC1* scores in adjacent years. This suggests that regional proportions of industrial outputs have experienced two big changes, i.e., from 2006 to 2007 and from 2010 to 2011. As shown in Fig. 1a, compared with the year of 2006, the output percentage in Central China doubled yet the proportion in Northeastern China halved in 2007. In 2011, the proportion score in Western China climbed over 80 %.

Table 4 The compositional data set of disciplines

Disciplines	$\mathbf{U}_1(\%)$	$\mathbf{U}_2(\%)$	$\mathbf{U}_3(\%)$	$\mathbf{U}_4(\%)$	$\mathbf{U}_5(\%)$
Earth Sci.	[9.70,20.8,69.5]	[58.3,30.5,11.2]	[16.6,20.9,62.5]	[12.5,36.1,51.4]	[26.3,26.4,47.3]
Compr. Sci.	[48.2,46.4,5.40]	[25.0,48.2,26.8]	[42.8,46.4,10.8]	[41.0,44.7,14.3]	[32.1,37.5,30.4]
Info. Sci.	[40.0,36.3,23.7]	[23.6,20.0,56.4]	[54.5,25.5,20.0]	[29.0,31.0,40.0]	[52.7,34.5,12.8]
Math. Phys. Sci.	[51.7,28.6,19.7]	[58.9,32.1,9.00]	[44.6,30.4,25.0]	[62.5,23.2,14.3]	[30.3,21.4,48.3]
Life Sci.	[38.2,32.0,29.8]	[27.0,37.8,35.2]	[30.6,36.0,33.4]	[37.8,32.0,30.2]	[31.0,34.3,34.7]
Chem. Sci.	[22.3,35.9,41.8]	[22.3,38.8,38.9]	[26.8,44.8,28.4]	[25.3,34.4,40.3]	[40.2,40.3,19.5]
Man. Sci.	[26.0,56.6,17.4]	[43.4,39.2,17.4]	[34.7,52.2,13.1]	[47.8,43.5,8.70]	[65.2,30.4,4.40]
Eng. Mat. Sci.	[29.4,41.1,29.5]	[33.6,28.5.37.9]	[35.7,31.6,32.7]	[25.2,42.1,32.7]	[26.3,45.2,28.5]

Taking a closer look at Table 2, we may notice that $\mathbf{V}_1$ has not been well explained by either $\mathbf{U}_4$ or $\mathbf{U}_5$, but is strongly associated with $\mathbf{U}_1$, $\mathbf{U}_2$ and $\mathbf{U}_3$. Since all of these three variables represent energy resources, we can refer to $PC1$ as an "energy factor". The $\mathbf{V}_2$ and $\mathbf{V}_3$ ($PC2$ and $PC3$) are, respectively, significantly correlated with $\mathbf{U}_5$ and $\mathbf{U}_4$. Consequently, we can refer to $PC2$ and $PC3$, respectively, as "cement factor" and "steel factor". The changing scores trend of *PC2* and *PC3* can also be discussed similarly to *PC1*. This real case has illustrated the merits of the proposed PCA for compositional data vectors. Noticeably, it may lead to spurious results since "time" has been adopted as observation in this case. To avoid this problem, some autocorrelation techniques should be used in further analysis.

4.2 Case study: journal evaluation

Another application is to consider the evaluation records of scientific journals for Chinese Science Citation Database (CSCD) in 2007. This case study includes 8 disciplines, specifically *Earth sciences (Earth Sci.)*, *Comprehensive sciences (Compr. Sci.)*, *Information sciences (Info. Sci.)*, *Mathematical & physical sciences (Math. Phys. Sci.)*, *Life sciences (Life Sci.)*, *Chemical sciences (Chem. Sci.)*, *Management sciences (Man. Sci.)*, and *Engineering & material sciences (Eng. Mat. Sci.)*. Each discipline is featured by five variables, i.e., *impact factor* ($\mathbf{U}_1$), *number of published paper* ($\mathbf{U}_2$), *immediacy index* ($\mathbf{U}_3$), *total citation* ($\mathbf{U}_4$), and *cited half-life* ($\mathbf{U}_5$).

For each discipline on each variable, we count the number of journals that fall into three disjoint categories and then calculate the proportions to construct a 3-part compositional data. Take $\mathbf{U}_1$ as an example. There are three disjoint categories: low (impact factor $\leq$ 0.2516), medium (0.2516 $<$ impact factor $\leq$ 0.4324), and high (impact factor $>$ 0.4324). Thus, the compositional data of Earth sciences means that, 9.7 % of journals in this discipline have low impact factors, while 20.8 and 69.5 % of journals, respectively, have medium and high impact factors. Similarly, compositional data of the other four variables can be obtained. As listed in Table 4, an 8×5 multiple compositional data set is formed.

Table 5 PC variances and CCR

PCs	PC variances	
	Value	CCR (%)
1	1.7016	56.72
2	0.7022	80.13
3	0.4695	95.78
4	0.1124	99.53
5	0.0142	100.00

According to Table 4, the variances are, respectively, 0.7687, 0.5492, 0.4605, 0.5835, 0.6378 for the five variables $\mathbf{U}_1, \ldots, \mathbf{U}_5$. Considering the small differences in variable variances, the proposed analytical method is performed on the covariance matrix in this case. By using the proposed analytical approach on the covariance matrix, the extracted PC variances and the CCR are achieved (see Table 5). The first and second components, respectively, account for 56.72 and 23.41 % of the variance information. As a consequence, the first two components summarize more than 80 % of the total information in the original variables. In other words, the original variables set can be sufficiently represented by these two components, expressed as follows

$$\mathbf{V}_1 = 0.6185 \otimes \mathbf{U}_1 \ominus 0.0504 \otimes \mathbf{U}_2 \oplus 0.4895 \otimes \mathbf{U}_3 \oplus 0.5099 \otimes \mathbf{U}_4 \oplus 0.3393 \otimes \mathbf{U}_5, \tag{39}$$

$$\mathbf{V}_2 = 0.1322 \otimes \mathbf{U}_1 \oplus 0.7717 \otimes \mathbf{U}_2 \ominus 0.0329 \otimes \mathbf{U}_3 \oplus 0.3069 \otimes \mathbf{U}_4 \ominus 0.5402 \otimes \mathbf{U}_5. \tag{40}$$

For interpretation, the relationship between the PCs and the original variables has been visualized in Fig. 3. As can be seen, the first PC, *PC1*, is positively associated with the variables $\mathbf{U}_1$, $\mathbf{U}_3$, $\mathbf{U}_4$ and $\mathbf{U}_5$, all of which are relevant indicators for "journal citation evaluation". The second PC, *PC2*, positively relates to $\mathbf{U}_2$, i.e., number of published papers, which denotes the volume of each journal. As a result, we refer to *PC2* as "journal volume".

According to Eqs. (39) and (40), scores of the extracted *PC1* and *PC2* can be calculated for each discipline in Table 6. Furthermore, the scores of the eight disciplines can be visualized in ternary diagrams (see Fig. 4).

Taking a close look at Fig. 4a, we remark on several aspects between journals from different disciplines. *Earth sciences* stand out from other journal groups in that an overwhelming majority of journals have shown a good performance in terms of citation. Probably, this is due to the fact that not only are these commonly seen within-discipline citations, but also other disciplines regularly cite research articles from *Earth sciences*. Compared with *Earth sciences*, a significant difference can be observed in both *Comprehensive sciences* and *Management sciences*. For each of these two disciplines, journal citations tend to fall into the middle or even low cited levels. Concerning the disciplines of *Chemical sciences*, *Life sciences*, and *Engineering* &

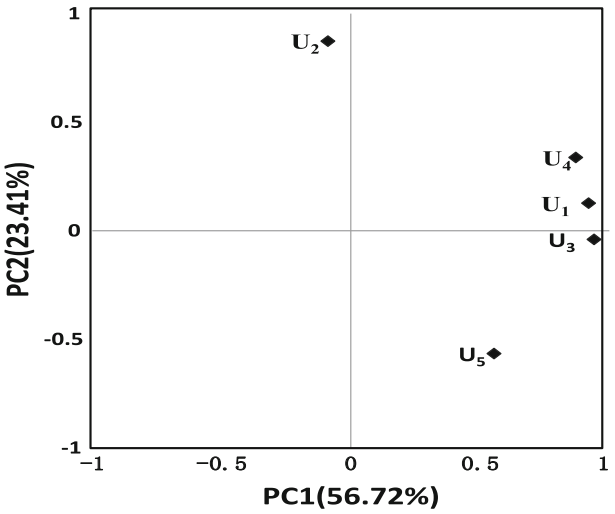

Fig. 3 Correlation of the first two PCs and the original variables

Table 6 The *PC1* and *PC2* scores of these 8 disciplines

Disciplines	*PC1* scores (%)	*PC2* scores (%)
Earth Sci.	[2.9,8.0,89.1]	[44.0,37.1,18.9]
Compr. Sci.	[48.2,45.4,6.4]	[32.2,45.9,21.9]
Info. Sci.	[48.8,24.9,26.3]	[14.4,14.9,70.7]
Math. Phys. Sci.	[58.3.9,16.1,25.6]	[62.8,29.9,7.3]
Life Sci.	[31.1,24.5,44.4]	[31.5,32.6,35.9]
Chem. Sci.	[19.0,32.5,48.5]	[17.4,28.1,54.5]
Man. Sci.	[41.3,52.5,6.2]	[26.9,36.6,36.5]
Eng. Mat. Sci.	[22.1,34.9,43.0]	[33.4,24.7,41.9]

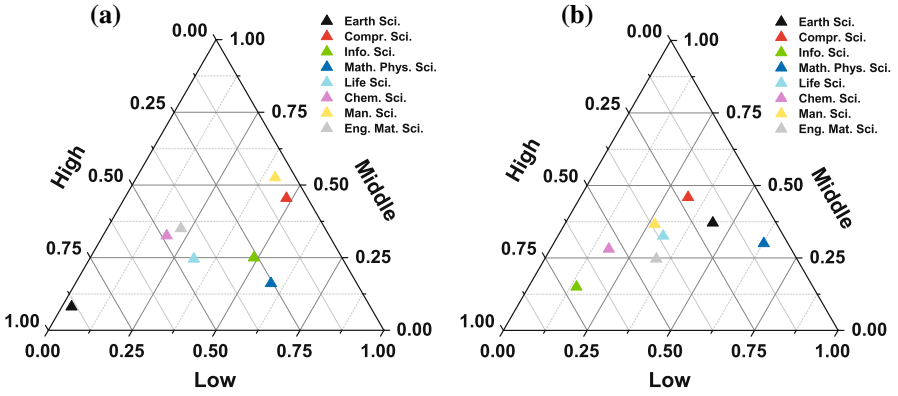

Fig. 4 $PC1 \times PC2$ scores of these 8 disciplines **a** $PC1$, **b** $PC2$ (color figure online)

Materials sciences, the proportions between three components are not much of a difference, with higher citations. Regarding the disciplines of *Information sciences*, and *Mathematical & physicals sciences*, a majority of journals have lower citations. Indeed, these observations have suggested that journal groups greatly distinguish from each other with respect to citation performance.

With regard to the *PC2* scores, interesting conclusions can also be achieved (see Fig. 4b). A remarkable observation exists in that most journals in *Information sciences* have published a large number of scientific papers. A similar phenomenon can also be found in the discipline of *Chemical sciences*. However, for the disciplines of *Earth sciences* and *Mathematical & physicals sciences*, only a small proportion ($<20\,\%$) of journals could be considered as high-volume periodicals. With respect to the other disciplines, i.e., *Comprehensive sciences*, *Management sciences*, *Life sciences*, and *Engineering & Materials sciences*, we observe that the low-volume journal percentages of these disciplines have relatively little difference between them.

In summary, the two case studies indicate that the proposed PCA modeling offers an effective approach on compositional data vectors.

5 Conclusions and remarks

This paper has proposed a new approach for modeling PCA on compositional data vectors. To facilitate deduction, an algebraic system of compositional data vectors has been proposed, including linear algorithms and algebraic operations such as inner product, norm and distance. Special emphasis is then placed on estimating the variance and covariance, since the derivation of PCA can be transformed into eigen-decomposition of the variance-covariance matrix. To facilitate the interpretation of PCA results, a definition of correlation coefficient is introduced.

It should be noticed that the compositional data discussed in this paper, see Eq. (13), is a three-way data. In the literature, there have been well-established methods for three-way data. Perhaps the most common approach is Tucker analysis, which is also known as N-mode PCA (Gallo and Lucadamo 2008; Gallo 2013). Concerning three-way compositional data, Tucker analysis, including Tucker-1, Tucker-2, Tucer-3 methods, are performed on the log-ratio transformed data (Gallo 2013). Undoubtedly, both Tucker analysis and our proposed method are capable of exploring the interrelations in three-way compositional data. However, Tucker-2 and Tucker-3 analysis, respectively, focus on reducing 2-dimensions and 3-dimensions. Our method concentrates on reducing only 1-dimension. It is necessary to make a comparison between our proposed method and Tucker-1 method. We argue that differences exist in at least three aspects. To start with, the two methods follow different analytical paradigms. Our proposed method is essentially an extension of classical PCA on an "individuals × variables" matrix, but to consider its each entry as a compositional data. The modeling process then uses Aitchison's geometry to accomplish traditional procedures. Tucker-1 method, however, considers multiple compositional data vectors in an "individuals × variables × compositions" form. That is, composition is deemed as a dimension in Tucker-1 method. On the basis of the first point, we notice that a second difference is the definition of the sample mean. Because of using Aitchison's geome-

try, our proposed method yields the sample mean for each compositional variable in the form of a compositional data (see Eq. (21)), instead of a numerical sample mean that Tucker-1 method obtains. This difference will lead to different results on the calculation of variance-covariance matrix, and eventually different PCA results. Third, the proposed method deserves further applications when an additional dimension, say time dimension, is considered in multiple compositional vectors. Due to its analytical paradigms as mentioned above, the proposed method can be extended to a four-way data in an "individuals × variables × compositions × times" form. The corresponding method deserves further discussion.

Acknowledgments The authors would like to express their gratitude to the editor and two anonymous reviewers for their valuable comments and suggestions that lead to an improved version of the work. This work was financially supported by the National Natural Science Foundation of China (Grant Nos. 71031001, 71420107025, 71401192), the National High Technology Research and Development Program of China (863 Program) (SS2014AA012303) and the Innovation Foundation of BUAA for Ph.D. Graduates (YWF-14-YJSY-027).

References

Aitchison J (1982) The statistical analysis of compositional data. J R Stat Soc Ser B (Methodol) 44:139–177
Aitchison J (1983) Principal component analysis of compositional data. Biometrika 70(1):57–65
Aitchison J (1984) Reducing the dimensionality of compositional data sets. Math Geol 16(6):617–635
Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman and Hal, London, New York Reprinted in 2003 with additional material by The Blackburn Press. Caldwell, NJ
Aitchison J (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn V (ed) Proceedings of IAMG'97-The third annual conference of the International Association for Mathematical Geology: International Center for Numerical Methods in Engineering (CIMNE), Barcelona pp 3–35
Aitchison J, Greenacre M (2002) Biplots of compositional data. J R Stat Soc Ser C Appl Stat 51(4):375–392
Bacon-Shone J (2011) A short history of compositional data analysis. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications. Wiley, Chichester, pp 3–11
Bali JL, Boente G, Tyler DE, Wang JL (2011) Robust functional principal components: a projection-pursuit approach. Ann Stat 39(6):2852–2882
Cazes P, Chouakria A, Diday E, Schektman Y (1997) Extension de l'analyse en composantes principales à des donnés de type intervalle. Revue de Statistique Appliqueé XLV(3):5–24
Filzmoser P (1999) Robust principal component and factor analysis in the geostatistical treatment of environmental data. Environmetrics 10:363–375
Filzmoser P, Hron K (2011) Robust statistical analysis. In: Pawlowsky-Glahn V, Buccianti A (eds) Compositional data analysis: theory and applications. Wiley, Chichester, pp 59–72
Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics 20(6):621–632
Fisher AGB (1939) Production, primary, secondary and tertiary. Econ Rec 15(1):24–38
Gallo M (2012) Coda in three-way arrays and relative sample space. Electron J Appl Stat Anal 5(3):400–405
Gallo M (2013) Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data. Adv Dyn Model Econ Soc Syst Stud Comput Intell 448:209–221
Gallo M, Lucadamo A (2008) Parafac/candecom analysis for compositional data. In: 10th European symposium on statistical methods for the food industry, Louvain-La-Neuve, pp 91–99
Gioia F, Lauro CN (2006) Principal component analysis on interval data. Comput Stat 21:343–363
Jolliffe IT (2002) Principal component analysis, 2nd edn. Springer, New York
Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (1998) A critical approach to non-parametric classification of compositional data. In: Rizzi A, Vichi M, Bock HH (eds) Advances in data science and classification. Springer, Berlin, pp 49–56
Orlik T (2011) Getting to grips with China's GDP data. FT Press, New Jersey

Palarea-Albaladejo J, Martín-Fernández J (2013) Values below detection limit in compositional chemical data. Anal Chim Acta 764:32–43
Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. Stoch Environ Res Risk Assess 15(5):384–398
Pawlowsky-Glahn V, Egozcue JJ (2002) Blu estimators and compositional data. Math Geol 34(3):259–274
Ramsay J, Silverman BW (2005) Functional data analysis. Springer, New York
Sawant P, Billor N, Shin H (2012) Functional outlier detection with robust functional principal component analysis. Comput Stat 27:83–102
Valderrama MJ (2007) An overview to modelling functional data. Comput Stat 22:331–334
Wang H, Liu Q, Mok HM, Fu L, Tse WM (2007) A hyperspherical transformation forecasting model for compositional data. Eur J Oper Res 179:459–468
Wang H, Guan R, Wu J (2012) CIPCA: complete-information-based principal component analysis for interval-valued data. Neurocomputing 86:158–169