**Rong GUAN, PhD**
rguan@cufe.edu.cn
Central University of Finance and Economics, Beijing, China

**Huijuan WANG, PhD**
huijuan-wang@163.com
Central University of Finance and Economics, Beijing, China

**Lijuan GUO, PhD (corresponding author)**
guolijuan@th.btbu.edu.cn
Beijing Technology and Business University, Beijing, China

# Factor Analysis for Multivariate Compositional Data and an Economic Application

**Abstract.** *This paper aims to address the lack of appropriate methods for analysing multivariate compositional data by introducing FAMCoDa, a factor analysis model tailored for such data, along with a two-step estimation method. First, we outline the mathematical framework of FAMCoDa and the procedures for estimating factor loadings and scores, validating its effectiveness through simulation experiments. We then apply FAMCoDa to analyse the industrial consumption structure data of 41 countries. The results reveal that FAMCoDa efficiently manages multivariate compositional data, identifying correlations between variables. Our contributions are threefold: (1) presenting a novel factor analysis model for multivariate compositional data, focusing on inter-variable correlations, unlike existing models; (2) devising a two-step estimation process, starting with multivariate CoDa PCA for initial loadings and refining them through variance rotation, followed by ilr transformation and OLS regression for factor scores; (3) ensuring the compositional structure of the extracted factors remains consistent with the original variables. This work has significant implications for economic applications.*

**Keywords**: *multivariate compositional data, factor analysis, isometric logratio transformation, inter-variable correlations，industrial consumption structure.*

**JEL Classification:** C38.

## 1. Introduction

There are now various data types in many fields with complex forms, such as compositional data (Fiksel et al., 2022; Greenacre et al., 2023), interval-valued data (Alcacer et al., 2024; Guan et al., 2020), distribution-valued data (Vo-Van & PhamToan, 2024; Zhao et al., 2022), and functional data (Li et al., 2022). Compositional data (CoDa) describe the relative information of components or parts in the whole, usually measured as a ratio or percentage, and are often found in practice, such as budget shares in microeconomics (Fry et al., 2000), gross domestic

DOI: 10.24818/18423264/58.4.24.05

product (GDP) at the industry level (Wang et al., 2013), sentiment structure in social media (Lu et al., 2021), environmental sciences (Buccianti & Pawlowsky-Glahn, 2005; Engle et al., 2014; Wei et al., 2018), and geochemistry (Tolosana-Delgado & von Eynatten, 2009; Rieser and Filzmoser 2023).

The use of CoDa reveals information on relative quantities often hidden behind absolute values, making it superior to numerical data. However, the unit-sum constraint introduces challenges for statistical modelling of CoDa. Classical statistical models and algorithms cannot be directly applied to CoDa due to this constraint. A common approach to addressing this issue is to eliminate the unit-sum constraint through some transformation methods before statistical modelling. This idea has led to the development of a family of log-ratio transformations, including additive log-ratio (alr, Aitchison, 1982), centred log-ratio (clr, Aitchison, 1983), isometric log-ratio (ilr, Egozcue et al., 2003) and $\alpha$-transformation (Tsagris et al., 2016). Detailed descriptions and comparisons of these transformation methods can be found in Alenazi (2023).

Multivariate problems tend to receive more attention as people's access to data grows. When multiple variables are measured through compositional data (we refer to this situation as multivariate compositional data), it is the correlation structure of these multivariate compositional data that we often wish to analyse. For example, the proportion of a country's industrial products produced in each region constitutes compositional data, and when we consider multiple industrial products (e.g., coal, crude oil, cement, etc.), we obtain multivariate compositional data (Wang et al., 2015). This kind of multivariate compositional data is also prevalent in other economic problems, such as the analysis of the industrial consumption structure and the consumption structure (Lu et al., 2024). Therefore, in recent years, many scholars have also begun to pay attention to the theory of statistical analysis of multivariate compositional data. Wang et al. (2015) proposed a principal component analysis method for multivariate compositional data, which is used to detect the correlation structure of a number of compositional data variables and to perform dimensionality reduction. Gu et al. (2021) established a new technological framework for the classification of multivariate compositional data, which both improves the quality of the data and solves the problem of the multivariate compositional data being recognised by machine learning methods.

Factor analysis is one of the commonly used statistical tools to show properties inherent in multivariate data by summarising the multivariate information in a compact form (Basilevsky, 1995; Woods & Edwards, 2007). Approaches for factor analysis with compositional data have been suggested (e.g., Filzmoser et al., 2009; Tolosana-Delgado et al., 2005). However, these methods are only applicable to the case of single compositional data variable. Regarding the factor model for multivariate compositional data, to the best of our knowledge, only the PARAFAC model (Di Palma et al., 2018) is relevant. This model solves the problem of dimensionality reduction of observations, variables, and components simultaneously. However, the PARAFAC approach reduces the dimensional *D*-part to a (*D-m*) part, where 0<*m*<*D*, making it difficult to understand the analysis results. We argue that the compositional structure within the same variable should remain unchanged before and after factor analysis. That is, the dimension reduction should be only for the variables, which is similar to the classical factor analysis. To this end, we innovatively propose factor analysis for multivariate compositional data (FAMCoDa), aimed at exploring the correlation structure among the observed multiple compositional variables. The proposed factor analysis model assumes the existence of a small set of uncorrelated CoDa variables or factors that explain the relationships between the observed CoDa variables. With these extracted factors, we can detect potential correlation structures between observed compositional variables, visualise these variables in a lower dimensional space spanned by fewer factors, and thus obtain an overall picture of the entire data set.

Our main contributions are reflected in three aspects. First, we propose a novel factor analysis model for multivariate compositional data, which had not yet been studied to the best of our knowledge. This model focuses only on the correlation between multiple CoDa variables, rather than on the correlation between multiple components within the same variable, which is very different from the existing works (Filzmoser et al., 2009; Scealy et al., 2015). Therefore, our proposed model has a higher practical value in economic application scenarios with multiple compositional data variables. Second, we provide a two-step approach to estimate the factor loadings and factor scores of the model. Specifically, we first obtain an initial solution for the factor loadings using the multivariate CoDa variable PCA method, and then use variance rotation to obtain the loadings, which explains the extracted factors well; then we build a series of regression models to estimate the factor scores by ilr transformation and OLS, so that each factor is still a CoDa variable. Third, we assumed that the compositional structure of the factors extracted by the model remained consistent with that of the original observed variables, which are all D-parts. We believe that this assumption is more in line with realistic application scenarios because the observed composition structure of CoDa usually has a specific meaning, e.g., industrial consumption structure and consumption structure. This structure should thereby remain unchanged after the factor analyses.

The rest of the paper is organised as follows. Section 2 presents some preliminaries concerning CoDa. Section 3 provides a detailed introduction to our proposed model and a two-step estimation method. Sections 4 and 5 present a simulation study and an economic application, respectively. The conclusion is presented in Section 6.

## 2. Preliminaries

In this section, we will briefly review some algebraic operations of CoDa, the isometric logratio transformation, and the sample statistics of CoDa variable(s).

### *2.1 Compositional data and algebraic operations*

A simplex space spanned by $D$-part CoDa is denoted as $S^D = \left\{\mathbf{x} = [x_1, x_2, \dots, x_D]'; x_i \geq 0, i = 1,2,\dots,D; \sum_{i=1}^{D} x_i = 1\right\}$. For any two CoDa $\mathbf{x}, \mathbf{y} \in S^D$

and any real number $\alpha \in \mathbb{R}$, the operations of addition and scalar-multiplication are defined, respectively, as (Aitchison, 1982; Greenacre et al., 2023):

$$\mathbf{x} \oplus \mathbf{y} = \boldsymbol{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D),$$
$$\alpha \odot \mathbf{x} = \boldsymbol{C}(x_1{}^{\alpha}, x_2{}^{\alpha}, \dots, x_D{}^{\alpha}),$$

with $\boldsymbol{C}(\cdot)$ being a closure operation that converts any $D$-dimensional positive real number vector $\boldsymbol{z} = (z_1, z_2, \cdots, z_D)' \in \mathbb{R}_+^D$ into a $D$-part CoDa by

$$\boldsymbol{C}(\mathbf{z}) = \left[\frac{z_1}{\sum_{i=1}^{D} z_i}, \frac{z_2}{\sum_{i=1}^{D} z_i}, \dots, \frac{z_D}{\sum_{i=1}^{D} z_i}\right]'.$$

The inner product in simplex space is defined by:

$$\langle \mathbf{x}, \mathbf{y} \rangle_{S^D} = \sum_{i=1}^{D} ln \frac{x_i}{g(\mathbf{x})} ln \frac{y_i}{g(\mathbf{y})},$$

where $g(\mathbf{x}) = (\prod_{i=1}^{D} x_i)^{1/D}$, i.e., the geometric mean of a $D$-part CoDa. Accordingly, norm and distance in simplex space can be defined, respectively, as follows:

$$\|\mathbf{x}\|_{S^D} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_{S^D}},$$
$$d_{S^D}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_{S^D}.$$

## *2.2 Isometric logratio transformation*

To perform statistical analysis on CoDa, it is common to first apply a transformation to the data. Typically, CoDa in simplex space is transformed into numerical data in Euclidean space. This transformation addresses the unit-sum constraint of CoDa and allows the application of classical statistical models and algorithms. The isometric log-ratio (ilr) transformation (Egozcue et al., 2003) is a commonly used method. Based on a set of orthogonal bases, ilr converts CoDa to real coordinates without changing its metric properties. Taking $\mathbf{e}_i$ $(i = 1, 2, \cdots, D - 1)$ to be a set of orthogonal bases in simplex space, the ilr transformation and its inverse transformation are respectively defined as:

$$\mathbf{x}^* = \mathrm{ilr}(\mathbf{x}) = \left(\langle \mathbf{x}, \mathbf{e}_1 \rangle_{S^D}, \langle \mathbf{x}, \boldsymbol{e}_2 \rangle_{S^D}, \cdots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_{S^D}\right),$$
$$\mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{x}^*) = (x_1^* \odot \mathbf{e}_1) \oplus \cdots \oplus (x_{D-1}^* \odot \mathbf{e}_{D-1}),$$

where $\mathbf{x} = [x_1, x_2, \dots, x_D]' \in S^D$ is a $D$-part CoDa, and $\mathbf{x}^* = (x_1^*, x_2^*, \cdots, x_{D-1}^*) \in R^{D-1}$ represents the numerical data vector after transformation.

Obviously, the above two equations depend on the specific form of orthogonal basis $\mathbf{e}_i$. One commonly used orthogonal basis is:

$$\mathbf{e}_i = \boldsymbol{C}\left(\exp\left(\underbrace{\sqrt{\frac{1}{i(i+1)}},\cdots,\sqrt{\frac{1}{i(i+1)}}}_{i\text{ elements}},-\sqrt{\frac{i}{i+1}},0,\cdots,0\right)\right).$$

Thus, the corresponding ilr transformation and its inversed transformation can be written as:

$$x_i^* = \sqrt{\frac{i}{i+1}}ln\left(\frac{g(x_1,\cdots,x_i)}{x_{i+1}}\right),\ \text{for } i = 1,2,\cdots,D-1,$$
$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \boldsymbol{C}(\exp(\,w_1),\exp(\,w_2),\cdots,\exp(\,w_D)),$$

where $w_i = \sum_{j=i}^{D} x_j^* \sqrt{1/j(j+1)} - x_{i-1}^*\sqrt{(i-1)/i}$ with $x_0^* = x_D^* = 0$.

For any two CoDa $\mathbf{x},\mathbf{y} \in S^D$ and two real numbers $\alpha,\beta \in \mathbb{R}$, it can be proved that the ilr transformation satisfies two properties as shown in the following two equations:

$$\text{ilr}(\alpha \odot \mathbf{x} \oplus \beta \odot \mathbf{y}) = \alpha\text{ilr}(\boldsymbol{x}) + \beta\text{ilr}(\mathbf{y}),$$
$$\langle \mathbf{x},\mathbf{y}\rangle_{S^D} = \langle \text{ilr}(\mathbf{x}),\text{ilr}(\mathbf{y})\rangle.$$

In other words, the ilr transformation preserves the metric properties of CoDa **x** and $\mathbf{y}$. As a consequence, the following two analytical strategies for CoDa are equivalent: (S1) perform a statistical analysis on CoDa in simplex space using algebraic operations, such as those defined in Section 2.1; (S2) first transform CoDa to real coordinates using ilr transformation, then perform statistical analysis in Euclidean space, and back-transform the analytical results to simplex space using the inverse transformation of ilr. In this paper, a hybrid strategy is adopted; S1 is used to solve factor loading and S2 for calculating factor scores. More details will be provided in the next section.

### *2.3 Compositional data variables and sample statistics*

We use **X** to denote a multivariate CoDa dataset with *n* observations described by *p* *D*-part CoDa variables, i.e.,

$$\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_p) = \begin{pmatrix} \mathbf{O}_1' \\ \mathbf{O}_2' \\ \vdots \\ \mathbf{O}_n' \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1p} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \cdots & \mathbf{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{n1} & \mathbf{x}_{\boldsymbol{n}2} & \cdots & \mathbf{x}_{np} \end{pmatrix},$$

where $\mathbf{x}_{ij} = [x_{ij1},x_{ij2},\ldots,x_{ijD}]' \in S^D(1 \le i \le n, 1 \le j \le p)$ is a *D*-part CoDa, $\mathbf{X}_j = (\mathbf{x}_{1j},\mathbf{x}_{2j},\cdots,\mathbf{x}_{nj})'\ (j = 1,2,\ldots,p)$ is a CoDa variable, and $\mathbf{O}_i = (\mathbf{x}_{i1},\mathbf{x}_{i2},\cdots,\mathbf{x}_{ip})'\ (i = 1,2,\ldots n)$ is a CoDa observation.

For any two CoDa variables $\mathbf{X}_j$ and $\mathbf{X}_{j'}(1 \leq j, j' \leq p)$ and any real number $\alpha \in \mathbb{R}$, addition and scalar-multiplication of CoDa variables are defined, respectively, as follows:

$$\mathbf{X}_j \oplus \mathbf{X}_{j'} = (\mathbf{x}_{1j} \oplus \mathbf{x}_{1j'}, \mathbf{x}_{2j} \oplus \mathbf{x}_{1j'}, \cdots, \mathbf{x}_{nj} \oplus \mathbf{x}_{1j'})'$$
$$\alpha \odot \mathbf{X}_j = (\alpha \odot \mathbf{x}_{1j}, \alpha \odot \mathbf{x}_{2j}, \cdots, \alpha \odot \mathbf{x}_{nj})'$$

The sample mean, sample variance, and sample covariance are respectively defined by(Wang et al., 2015):

$$E_{S^D}(\mathbf{X}_j) = \overline{\boldsymbol{X}}_j = \boldsymbol{C}(g(\tilde{\mathbf{x}}_{j1}), g(\tilde{\mathbf{x}}_{j2}), \cdots, g(\tilde{\mathbf{x}}_{jD})),$$
$$Var_{S^D}(\mathbf{X}_j) = \frac{1}{n-1}\sum_{i=1}^{n}\left\|\mathbf{x}_{ij} \ominus \overline{\boldsymbol{X}}_j\right\|_{S^D}^2,$$
$$Cov_{S^D}(\mathbf{X}_j, \mathbf{X}_{j'}) = \frac{1}{n-1}\sum_{i=1}^{n}\left\langle \mathbf{x}_{ij} \ominus \overline{\boldsymbol{X}}_j, \mathbf{x}_{ij'} \ominus \overline{\boldsymbol{X}}_{j'}\right\rangle_{S^D},$$

where $\tilde{\mathbf{x}}_{jk} = (x_{1jk}, x_{2jk}, \ldots, x_{njk})'$ for $k = 1, 2, \ldots, D$.

## 3. Model

### *3.1 Factor model for multivariate compositional data*

For a random *D*-part CoDa variable $\mathbf{X}_j(1 \leq j \leq p)$, the factor analysis model for is defined as

$$\mathbf{X}_j = \left(a_{j1} \odot \mathbf{F}_1\right) \oplus \cdots \oplus \left(a_{jm} \odot \mathbf{F}_m\right) \oplus \boldsymbol{\varepsilon}_j, \tag{1}$$

with $m(m < p)$ *D*-part CoDa factors $\mathbf{F}_r(1 \leq r \leq m)$, the error term $\boldsymbol{\varepsilon}_j$, and the loading coefficient $a_{jr} \in \mathbb{R}$. Usually, we use $\mathbf{A} = (a_{jr})_{p\times m}$ $(j = 1,2,\cdots,p;\ r = 1,2,\cdots,m)$ to denote the loading matrix.

The model in Eq. (1) follows the usual factor model assumptions (Basilevsky, 1995). Thus, we need to estimate the factor loading matrix $\mathbf{A} \in \mathbb{R}^{p\times m}$ and the factor score $\mathbf{F}_r = (\boldsymbol{f}_{1r}, \boldsymbol{f}_{2r}, \cdots, \boldsymbol{f}_{nr})'$ for $r = 1,2,\cdots,m$, where $\boldsymbol{f}_{ir} \in S^D(1 \leq i \leq n)$.

### *3.2 Estimation*

There are many estimation methods for factor analysis (Woods & Edwards, 2007). In this paper, we mainly follow the principal factor analysis (PFA) procedure for factor loading estimation. Similar to the PCA method proposed by Wang et al. (2015), we propose the following algorithm.

*Step 1*: estimate the covariance matrix of $\mathbf{X}_j$ by the sample covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jj'})_{p\times p}$ with $\sigma_{jj'} = Cov_{S^D}\left(\mathbf{X}_j, \mathbf{X}_{j'}\right)$ for $j \neq j'$ and $\sigma_{jj} = Var_{S^D}(\mathbf{X}_j)$.

*Step 2*: estimate $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \cdots, \lambda_m)$ and $\boldsymbol{Q} = (\mathbf{e_1}, \mathbf{e_2}, \cdots, \mathbf{e_m})$ by decomposing $\boldsymbol{\Sigma} = \boldsymbol{Q\Lambda Q'}$, where $\lambda_j$ and $\mathbf{e_j}$ are respectively the $j$-th eigenvalue and eigenvector.

*Step 3*: compute the loading matrix by $\widehat{\mathbf{A}} = \left(\sqrt{\lambda_1}\mathbf{e_1}, \sqrt{\lambda_2}\mathbf{e_2}, \cdots, \sqrt{\lambda_m}\mathbf{e_m}\right)$.

***Remark 1.*** The number of factors $m$ can be determined by experience. Usually, a threshold value, such as 70% or 80%, should be set according to the application scenario. The cumulative contribution rate (CCR) of the first $m$ eigenvalues, i.e., $CCR(m) = \sum_{r=1}^{m} \lambda_r / \sum_{j=1}^{p} \lambda_j$, should not be less than the pre-defined threshold value. Then, $m$ can be determined.

***Remark 2.*** It is very often that the original solution of $\widehat{\mathbf{A}}$ has a relatively low power of interpretation. The common practice is to use varimax rotation technique, i.e., multiplying $\widehat{\mathbf{A}}$ by an orthogonal matrix so that sum of variance of the squared loadings in the rotated matrix is maximised. For convenience, we still denote the rotated loading matrix as $\widehat{\mathbf{A}}$.

In the case of compositional data, the estimation of factor score $\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_m$ cannot be traditionally done. The problem can be solved by considering Eq. (1) as a regression model, where $\mathbf{X}_j$ is the dependent variable, the rotated factor loadings $a_{j1}, a_{j2}, \cdots, a_{jm}$ are independent variables, factors $\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_m$ are the to-be-estimated coefficients, and $\boldsymbol{\varepsilon}_j$ is the error term. Such a model is called simplicial-simplicial regression, where both sides of the equation contain compositional data.

Most published papers regarding simplicial-simplicial regression involve transformations of both simplicial sides (Chen et al., 2017; Han & Yu, 2022; Hron et al., 2012; Wang et al., 2013). A log-ratio transformation is used for both the response and predictor variables and performed a multivariate linear regression model. Inspired by this, we propose the following regression method for factor score estimation.

*Step 1*: perform ilr transformation on both sides of Eq. (1) to obtain the new regression models as:

$$\begin{cases} x_{ij1}^* = a_{j1} f_{i11}^* + \cdots + a_{jm} f_{im1}^* + \epsilon_{ij1}^* \\ x_{ij2}^* = a_{j1} f_{i12}^* + \cdots + a_{jm} f_{im2}^* + \epsilon_{ij2}^* \\ \cdots \\ x_{ij,D-1}^* = a_{j1} f_{i1,D-1}^* + \cdots + a_{jm} f_{im,D-1}^* + \epsilon_{ij,D-1}^* \end{cases}, \tag{2}$$

where $ilr(\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}^* = \left(x_{ij1}^*, x_{ij2}^*, \cdots, x_{ij,D-1}^*\right)'$ , $ilr(\boldsymbol{f}_{ir}) = \boldsymbol{f}_{ir}^* = \left(f_{ir1}^*, f_{ir2}^*, \cdots, f_{ir,D-1}^*\right)'$ and $ilr(\boldsymbol{\varepsilon}_{ij}) = \boldsymbol{\varepsilon}_{ij}^* = \left(\epsilon_{ij1}^*, \epsilon_{ij2}^*, \cdots, \epsilon_{ij,D-1}^*\right)'$.

*Step 2*: for $l = 1,2, \cdots, D-1$, estimate the $l$-th model in Eq. (2) via OLS method and obtain $\left(\hat{f}_{i1l}^*, \hat{f}_{i2l}^*, \cdots, \hat{f}_{iml}^*\right)' = \left(\widehat{\mathbf{A}}'\widehat{\mathbf{A}}\right)^{-1} \widehat{\mathbf{A}}^T \left(x_{i1l}^*, x_{i2l}^*, \cdots, x_{ipl}^*\right)^T$.

*Step 3*: perform the inversed ilr transformation on $\hat{\boldsymbol{f}}^*_{ir} = (\hat{f}^*_{ir1}, \hat{f}^*_{ir2}, \cdots, \hat{f}^*_{ir,D-1})'$ to obtain the estimated factor score $\hat{\boldsymbol{f}}_{ir} = [\hat{f}_{ir1}, \hat{f}_{ir2}, \cdots, \hat{f}_{irD}]$.

So far, we have introduced the factor analysis model for multivariate compositional data and proposed a two-step estimation methos, referred to as FAMCoDa method hereinafter. The figure presents the framework of the FAMCoDa method.

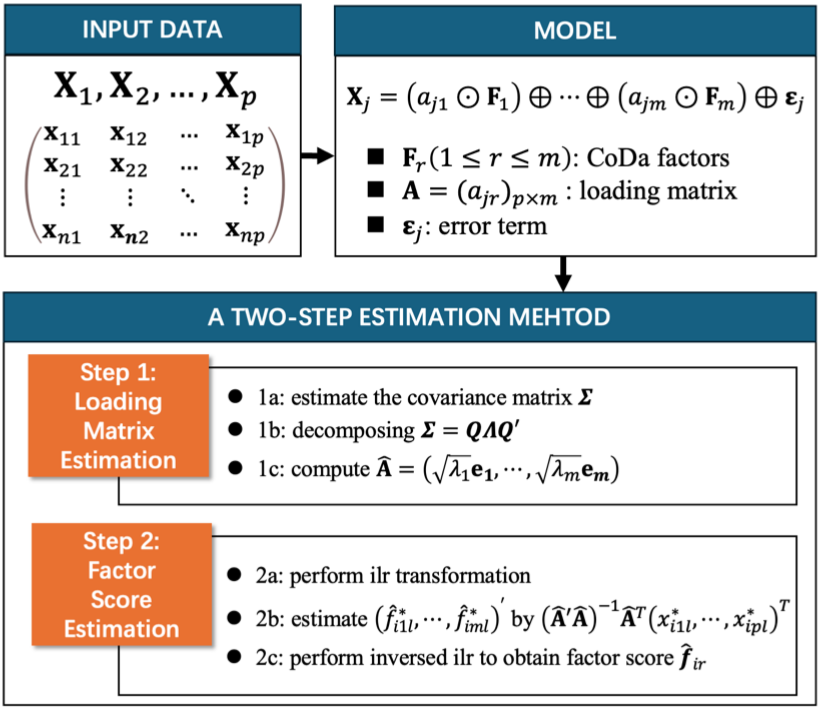
INPUT DATA
$\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p$
$\begin{pmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1p} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \cdots & \mathbf{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{n1} & \mathbf{x}_{n2} & \cdots & \mathbf{x}_{np} \end{pmatrix}$
MODEL
$\mathbf{X}_j = (a_{j1} \odot \mathbf{F}_1) \oplus \cdots \oplus (a_{jm} \odot \mathbf{F}_m) \oplus \boldsymbol{\varepsilon}_j$
■ $\mathbf{F}_r (1 \le r \le m)$: CoDa factors
■ $\mathbf{A} = (a_{jr})_{p \times m}$ : loading matrix
■ $\boldsymbol{\varepsilon}_j$: error term
A TWO-STEP ESTIMATION MEHTOD
Step 1: Loading Matrix Estimation
● 1a: estimate the covariance matrix $\boldsymbol{\Sigma}$
● 1b: decomposing $\boldsymbol{\Sigma} = \boldsymbol{Q\Lambda Q'}$
● 1c: compute $\hat{\mathbf{A}} = (\sqrt{\lambda_1}\mathbf{e}_1, \cdots, \sqrt{\lambda_m}\mathbf{e}_m)$
Step 2: Factor Score Estimation
● 2a: perform ilr transformation
● 2b: estimate $(\hat{f}^*_{i1l}, \cdots, \hat{f}^*_{iml})'$ by $(\hat{\mathbf{A}}'\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^T(x^*_{i1l}, \cdots, x^*_{ipl})^T$
● 2c: perform inversed ilr to obtain factor score $\hat{\boldsymbol{f}}_{ir}$

**Figure 1. Flowchat of the proposed FAMCoDa method**
*Source:* Illustration by authors.

## 4. Simulation studies

To demonstrate the performance of the proposed method, we present here a number of simulation experiments using synthetic data with predefined correlation structures. In each experiment, we generate a matrix $\mathbf{X}_{n\times(D\times p)}$, which contains $n$ observations described $p$-dimensional $D$-part CoDa, by the following three steps.

*Step 1:* generate $D-1$ $(p \times p)$-dimensional correlation matrices $R_l(l = 1,2,\dots,D-1)$ satisfying a given correlation structure, whose details will be described later. For each $R_l$, conduct an eigen-decomposition, i.e.,$R_l = Q_l \Lambda_l Q_l^T$, where $Q_l$ is a matrix composing by the orthogonal eigenvectors of $R_l$, and $\Lambda_l$ is a diagonal matrix whose non-zero entries are the eigenvalues of $R_l$.

*Step 2:* Generate an $(n \times p)$-dimensional matrix $\mathbf{X}_l^*$, whose $i$-th $(i = 1,2,\cdots,n)$ row $\left(x_{i1l}^*, x_{i2l}^*, \dots x_{ipl}^*\right)$ equals to $Q_l \Lambda_l^{1/2} Q_l^T \mathbf{U}_i$, where $\mathbf{U}_i$ is a vector randomly selected from $p$-dimensional normal distribution $N(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma$ as an identity matrix.

*Step 3:* Let $\mathbf{x}_{ij}^*$=$\left(x_{ij1}^*, x_{ij2}^*, \dots x_{ij(D-1)}^*\right)$. Conduct an inversed ilr transformation to $\mathbf{x}_{ij}^*$ and obtain a $D$-part compositional data $\mathbf{x}_{ij} = [x_{ij1}, x_{ij2}, \dots, x_{ijD}]'$.

Concerning the correlation structure in $R_l$, we design eight cases as shown in Table 1. In Case 1-2, there is only one factor. Case 3-5 and Case 6 respectively consider two and three factors. Within each group, it is regarded as highly correlated if correlation coefficient values are 0.8~1, and moderately correlated between 0.5~0.8. For variables from different factors, correlation coefficient values are set to 0~0.3. A total of 100 experiments are conducted. The parameters are set as $p = 8$, $n = 100$, $D = 3$.

**Table 1. Experimental design**

| Case | Factor number | Correlation structure |
|---|---|---|
| 1 | 1 | $X_1 - X_8$ are highly correlated. |
| 2 | 1 | $X_1 - X_8$ are moderately correlated. |
| 3 | 2 | $X_1 - X_3$ are highly correlated.<br>$X_4 - X_8$ are highly correlated. |
| 4 | 2 | $X_1 - X_3$ are highly correlated.<br>$X_4 - X_8$ are moderately correlated. |
| 5 | 2 | $X_1 - X_3$ are moderately correlated.<br>$X_4 - X_8$ are moderately correlated. |
| 6 | 3 | $X_1 - X_3$ are highly correlated.<br>$X_4 - X_6$ are highly correlated.<br>$X_7 - X_8$ are highly correlated. |

*Source:* Values used by authors.

In each case shown in Table 1, we build factor models based on the number of factors and observe whether the estimates of the factor models correctly reflect the correlation structure of the original variables designed in Table 1. For statistical analysis, we define a metric called the *Correct Detection Ratio* (*CDR*), which is the ratio of times in 100 experiments that the factor model correctly identifies no less than *q* of the original variables. We set *q* = 8, 7, 6 and 5. Next, we need to define what we mean by "correctly identified". Take Case 1 for example, at this point we set the number of factors to 1 and observe the loading coefficients of that factor on $X_1 - X_8$. If the loading coefficients are all greater than 0.5 (in factor analysis, when the loading coefficients are greater than 0.5, the correlation between the factor and the original variable is generally considered to be high), then we can assume that $X_1 - X_8$ are all correctly identified by the model. That is, the factor analysis model correctly captures the correlation structure in the original data.

Table 2 reports the results of the simulated studies. We can see that, in Case 1, 2, 3 and 4, FAMCoDa method shows a nearly perfect performance in recovering the pre-defined correlation structure in 100 experiments. Nevertheless, the method behaves slightly different in Case 5 and 6. In Case 5, when two variable groups are designed as moderately correlated, our method well captures the structure in a frequency of only 34%. In Case 6, although the three variable groups are all highly correlated within each group, our method shows a relatively low ability of identifying such a structure.

**Table 2. Correct Detect Ratio in six simulation cases**

| Case | q=8 | q=7 | q=6 | q=5 |
|---|---|---|---|---|
| 1 | 100 | 100 | 100 | 100 |
| 2 | 100 | 100 | 100 | 100 |
| 3 | 100 | 100 | 100 | 100 |
| 4 | 98 | 100 | 100 | 100 |
| 5 | 34 | 43 | 66 | 89 |
| 6 | 52 | 69 | 80 | 80 |

*Source*: Calculation by authors.

## 5. Case study

Industrial consumption structure reflects the technological and economic links among different industrial sectors, referred to simply as ‘sectors’ hereinafter. On the one hand, it reflects the degree of dependence among these sectors, while, on the other hand, a difference in the industrial consumption structures of different countries can reflect a difference in the technological level of different sectors. Therefore, it is certainly of economic significance to show the differences among countries through industrial consumption structure data. The main purposes of this case is to use the factor model proposed in this paper to identify countries with a high correlation in industrial consumption structure and to classify these countries according to the compositional data that describes their industrial consumption structures.

### *5.1 Data and explanatory analysis*

We use the World Input-Output Table (WIOD) published by the European Union, where country and sectoral classifications are harmonized over long time series, and which provides consistent single-country input-output tables for individual countries and sectors, which is more suitable for the analytical scenario of this case. The latest updated version of this data is up to the year 2014. According to the input-output table, the proportion of products or services consumed by each sector in each country and sector in the total global consumption of the sector is calculated and the industrial consumption structure data is obtained as compositional data. For example, consumption of products or services in 34 industry sectors of Australia's *agriculture, forestry, animal husbandry, and fishery industry* in 2014 can be expressed as a 34-part CoDa, i.e., (0.335, 0.004, …, 0.001, 0.015). For the remaining 33 sectors in Australia, we used a similar expression. Therefore, a country's consumption of products or services is considered as a 34-part CoDa variable with 34 observations. We collected such data for 41 countries and regions and in this way obtained a dataset with 41 CoDa variables and 34 observations, as shown in Table 3.

**Table 3. Industrial consumption data expressed by multivariate compositional data**

| **Industries** | **Australia** | **…** | **USA** |
|---|---|---|---|
| A01-A03 | (0.335, 0.004, ... , 0.001, 0.015) | … | (0.372, 0.011, ... , 0.002, 0.004) |
| B4 | (0.008, 0.094, ... , 0.003, 0.056) | … | (0.001, 0.265, ... , 0.001, 0.003) |
| … | … | … | … |
| P85 | (0.004, 0.004, ... , 0.026, 0.050) | … | (0.003, 0.002, ... , 0.019, 0.023) |
| O84&R_S&T&U | (0.017, 0.004, ... , 0.007, 0.077) | … | (0.002, 0.005, ... , 0.011, 0.063) |

*Source*: Raw data from WIOD and processed by authors.

## *5.2 Factor analysis*

The eigenvalues and the cumulative contribution rates are listed in Table 4. With regard to the cumulative contribution rate, we know that the first five factors are able to represent more than 75% of the original variables. Therefore, five factors are selected for further analysis.

**Table 4. Eigenvalues and cumulative contribution rate**

| ***r*** | $\lambda_r$ | $CCR(m)$ | ***r*** | $\lambda_r$ | $CCR(m)$ | ***r*** | $\lambda_r$ | $CCR(m)$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 28.103 | 68.545 | 15 | 0.358 | 86.411 | 29 | 0.214 | 95.606 |
| 2 | 0.833 | 70.576 | 16 | 0.341 | 87.244 | 30 | 0.202 | 96.099 |
| 3 | 0.680 | 72.235 | 17 | 0.330 | 88.047 | 31 | 0.193 | 96.570 |
| 4 | 0.627 | 73.764 | 18 | 0.320 | 88.828 | 32 | 0.188 | 97.027 |
| 5 | 0.624 | 75.287 | 19 | 0.305 | 89.572 | 33 | 0.180 | 97.467 |
| 6 | 0.586 | 76.716 | 20 | 0.297 | 90.297 | 34 | 0.172 | 97.886 |
| 7 | 0.522 | 77.988 | 21 | 0.289 | 91.003 | 35 | 0.164 | 98.285 |
| 8 | 0.515 | 79.244 | 22 | 0.272 | 91.667 | 36 | 0.162 | 98.681 |
| 9 | 0.480 | 80.415 | 23 | 0.264 | 92.311 | 37 | 0.149 | 99.045 |
| 10 | 0.476 | 81.576 | 24 | 0.237 | 92.890 | 38 | 0.141 | 99.388 |
| 11 | 0.435 | 82.637 | 25 | 0.236 | 93.466 | 39 | 0.135 | 99.717 |
| 12 | 0.424 | 83.671 | 26 | 0.224 | 94.012 | 40 | 0.088 | 99.932 |
| 13 | 0.392 | 84.628 | 27 | 0.222 | 94.553 | 41 | 0.028 | 100.000 |
| 14 | 0.373 | 85.538 | 28 | 0.218 | 95.085 | | | |

*Source*: Calculation by authors.

The rotated factor loading matrix, as shown in Table 5, was solved using the maximum variance method. The fact that all the communalities were above 0.6 indicates that the extracted factors are a better explanation for the variance of the original CoDa variables. To understand the meaning of the extracted factors, we took a close look at the loadings. The maximum loading in each row is highlighted in bold for the sake of clarity. Accordingly, factors 1–5 correspond to country groups 1–5, respectively. The countries represented by Factor 1 are mainly concentrated in Central and Northern Europe. Their technological development level and socioeconomic status both rank first among European countries. The countries represented by Factor 2 include most non-European countries and four European countries closer to North Africa and West Asia. Factor 3 consists mainly of three countries that border the Mediterranean Sea: Cyprus, Greece and Malta. Services, mainly including tourism, transportation and finance, is the main industry for each of these three countries, accounting for more than 70% of each of their GDP in 2014. The countries represented by Factor 4 are concentrated in the Central and Eastern Europe. Most of these countries have an advantage in a particular aspect of industry. For instance, The Czech Republic, Slovakia, Slovenia, and Italy are prominent in the automobile industry, and Croatia and Lithuania are superior in the light industry. Factor 5 mainly represents India, a country that has developed rapidly in recent years. The outstanding feature of its industrial structure is that the proportion of its industry dedicated to agriculture is much higher than most other countries.

**Table 5. Factor loadings after rotation**

| Country | $F_1$ | $F_2$ | $F_3$ | $F_4$ | $F_5$ |
|---|---|---|---|---|---|
| AUT | **0.542** | 0.403 | 0.274 | 0.453 | 0.310 |
| BEL | **0.571** | 0.427 | 0.266 | 0.424 | 0.235 |
| CHE | **0.539** | 0.387 | 0.290 | 0.382 | 0.343 |
| DEU | **0.526** | 0.458 | 0.266 | 0.412 | 0.261 |
| DNK | **0.512** | 0.445 | 0.362 | 0.359 | 0.203 |
| EST | **0.484** | 0.440 | 0.246 | 0.517 | 0.217 |
| FIN | **0.471** | 0.383 | 0.208 | 0.430 | 0.318 |
| GBR | **0.552** | 0.490 | 0.277 | 0.317 | 0.199 |
| HUN | **0.542** | 0.417 | 0.289 | 0.478 | 0.189 |
| IRL | **0.744** | 0.278 | 0.182 | 0.189 | 0.016 |
| LVA | **0.504** | 0.377 | 0.125 | 0.435 | 0.227 |
| NLD | **0.577** | 0.408 | 0.303 | 0.407 | 0.162 |
| NOR | **0.469** | 0.416 | 0.256 | 0.350 | 0.248 |
| POL | **0.498** | 0.489 | 0.209 | 0.488 | 0.241 |
| PRT | **0.502** | 0.479 | 0.298 | 0.436 | 0.246 |
| SWE | **0.510** | 0.453 | 0.254 | 0.363 | 0.305 |
| FRA | 0.376 | **0.642** | 0.259 | 0.288 | 0.205 |
| AUS | 0.330 | **0.558** | 0.173 | 0.440 | 0.180 |

| **Country** | $\boldsymbol{F_1}$ | $\boldsymbol{F_2}$ | $\boldsymbol{F_3}$ | $\boldsymbol{F_4}$ | $\boldsymbol{F_5}$ |
|---|---|---|---|---|---|
| BGR | 0.460 | **0.476** | 0.285 | 0.294 | 0.352 |
| BRA | 0.406 | **0.523** | 0.169 | 0.436 | 0.102 |
| CAN | 0.301 | **0.746** | 0.197 | 0.315 | 0.166 |
| CHN | 0.437 | **0.510** | 0.294 | 0.438 | 0.270 |
| ESP | 0.460 | **0.573** | 0.276 | 0.352 | 0.237 |
| IDN | 0.289 | **0.600** | 0.279 | 0.197 | 0.250 |
| JPN | 0.354 | **0.641** | 0.238 | 0.272 | 0.260 |
| KOR | 0.404 | **0.656** | 0.248 | 0.289 | 0.237 |
| ROU | 0.120 | **0.528** | 0.348 | 0.487 | 0.165 |
| TUR | 0.349 | **0.566** | 0.265 | 0.342 | 0.249 |
| TWN | 0.379 | **0.650** | 0.286 | 0.307 | 0.190 |
| USA | 0.383 | **0.556** | 0.354 | 0.269 | 0.197 |
| CYP | 0.245 | 0.378 | **0.663** | 0.266 | 0.054 |
| GRC | 0.399 | 0.242 | **0.556** | 0.029 | 0.453 |
| MLT | 0.210 | 0.242 | **0.616** | **0.474** | 0.112 |
| ITA | 0.488 | 0.465 | 0.229 | **0.516** | 0.188 |
| CZE | 0.309 | 0.215 | 0.288 | **0.654** | 0.223 |
| HRV | 0.396 | 0.398 | 0.416 | **0.454** | 0.236 |
| LTU | 0.249 | 0.432 | 0.107 | **0.619** | 0.232 |
| SVK | 0.429 | 0.400 | 0.289 | **0.598** | 0.136 |
| SVN | 0.475 | 0.381 | 0.265 | **0.556** | 0.092 |
| IND | 0.124 | 0.333 | 0.113 | 0.275 | **0.794** |

*Source*: Calculation by authors.

Using FAMCoDa, we also obtained the factor scores of the five country groups (see Table 6). The score of each factor, represented in CoDa, indicates the products or services consumed from a particular sector in this group as a proportion of their total consumption.

**Table 6. Factor scores**

|  | $\boldsymbol{F_1}$ | … | $\boldsymbol{F_5}$ |
|---|---|---|---|
| A01-A03 | (0.026, 0.023, …, 0.031, 0.030) | … | (0.017, 0.035, …, 0.035, 0.035) |
| B4 | (0.027, 0.020, …, 0.037, 0.031) | … | (0.043, 0.023, …, 0.014, 0.029) |
| … | … | … | … |
| P85 | (0.035, 0.025, …, 0.016, 0.025) | … | (0.025, 0.041, …, 0.020, 0.035) |
| O84&R_S&T&U | (0.029, 0.032, …, 0.029, 0.021) | … | (0.028, 0.037, …, 0.028, 0.025) |

*Source*: Calculation by authors.

## 6. Conclusions

In this paper, we propose factor analysis for multivariate compositional data and its two-step estimation method, i.e., FAMCoDa. We firstly present the mathematical model of FAMCoDa and provide the estimation methods of factor loadings and factor scores, and then validate the efficacy of FAMCoDa through comprehensive simulation experiments. Furthermore, we apply FAMCoDa to analyse the industrial consumption structure data of 41 countries. The results demonstrate that our proposed FAMCoDa can effectively handle multivariate compositional data, identifying complex inter-variable correlations and uncovering latent compositional factors.

This study not only enriches the methodological toolkit for compositional data analysis but also highlights the potential of FAMCoDa in addressing practical economic challenges, particularly in fields dealing with large-scale compositional datasets. Future research can extend this work by exploring robust estimation methods for factor analysis of multivariate compositional data, developing confirmatory factor analysis techniques, and investigating potential applications in other domains such as environmental sciences and public health.

***Acknowledgements:*** *The research of Rong Guan is supported by General Project of MOE (Ministry of Education) Foundation on Humanities and Social Sciences (Grant No. 23YJC910002) National Natural Science Foundation of China (Grant No. 11971504), and the Disciplinary Funding of Central University of Finance and Economics. The research of Lijuan Guo is supported by National Social Science Foundation of China (Grant No. 20CTJ006) and Capital Circulation Industry Research Base.*

## References

[1] Aitchison, J. (1982), *The statistical analysis of compositional data. Journal of the Royal Statistical Society, Series B*, 44, 139-177.

[2] Aitchison, J. (1983), *Principal component analysis of compositional data. Biometrika*, 70(1), 57–65, https://doi.org/10.1093/biomet/70.1.57.

[3] Alcacer, A., Martinez-Garcia, M., Epifanio, I. (2024), *Ordinal classification for interval-valued data and interval-valued functional data. Expert Systems with Applications*, 238, 122277, https://doi.org/10.1016/j.eswa.2023.122277.

[4] Alenazi, A. (2023), *A review of compositional data analysis and recent advances. Communications in Statistics - Theory and Methods*, 52(16), 5535-5567, https://doi.org/10.1080/03610926.2021.2014890.

[5] Basilevsky, A. (1994), *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley, New York, USA.

[6] Buccianti, A., Pawlowsky-Glahn, V. (2005), *New Perspectives on Water Chemistry and Compositional Data Analysis. Mathematical Geology*, 37(7), 703-727, https://doi.org/10.1007/s11004-005-7376-6.

[7] Chen, J., Zhang, X., Li, S. (2017), *Multiple linear regression with compositional response and covariates. Journal of Applied Statistics*, 44(12), 2270-2285, https://doi.org/10.1080/02664763.2016.1157145.

[8] Di Palma, M.A., Filzmoser, P., Gallo, M., Hron, K. (2018), *A robust Parafac model for compositional data. Journal of Applied Statistics*, 45(8), 1347-1369, https://doi.org/10.1080/02664763.2017.1381669.

[9] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C. (2003), *Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology*, 35(3), 279-300.

[10] Engle, M.A., Gallo, M., Schroeder, K.T., Geboy, N.J., Zupancic, J.W. (2014), T*hree-way compositional analysis of water quality monitoring data. Environmental and Ecological Statistics*, 21(3), 565-581, https://doi.org/10.1007/s10651-013-0268-x.

[11] Fiksel, J., Zeger, S., Datta, A. (2022), *A Transformation-Free Linear Regression for Compositional Outcomes and Predictors. Biometrics*, 78(3), 974-987, https://doi.org/10.1111/biom.13465.

[12] Filzmoser, P., Hron, K., Reimann, C., Garrett, R. (2009), *Robust factor analysis for compositional data. Computers & Geosciences*, 35(9), 1854-1861, https://doi.org/10.1016/j.cageo.2008.12.005.

[13] Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., Quinn, T. (2023), *Aitchison's Compositional Data Analysis 40 Years on: A Reappraisal. Statistical Science*, 38(3), https://doi.org/10.1214/22-STS880.

[14] Gu, J., Cui, B., Lu, S. (2021), *A classification framework for multivariate compositional data with Dirichlet feature embedding. Knowledge-Based Systems*, 212, 106614, https://doi.org/10.1016/j.knosys.2020.106614.

[15] Guan, R., Wang, H., Zheng, H. (2020), *Improving accuracy of financial distress prediction by considering volatility: An interval-data-based discriminant model. Computational Statistics*, 35(2), 491-514, https://doi.org/10.1007/s00180-019-00916-9.

[16] Han, H., Yu, K. (2022), *Partial linear regression of compositional data. Journal of the Korean Statistical Society*, 51(4), 1090–1116, https://doi.org/10.1007/s42952-022-00177-1.

[17] Hron, K., Filzmoser, P., Thompson, K. (2012), *Linear regression with compositional explanatory variables. Journal of Applied Statistics*, 39(5), 1115-1128, https://doi.org/10.1080/02664763.2011.644268.

[18] Li, Y., Qiu, Y., Xu, Y. (2022), *From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas. 50th Anniversary Jubilee Edition*, 188, 104806, https://doi.org/10.1016/j.jmva.2021.104806.

[19] Lu, S., Wang, W., Guan, R. (2024), *Kent feature embedding for classification of compositional data with zeros. Statistics and Computing*, 34(2), 69, https://doi.org/10.1007/s11222-024-10382-z.

[20] Lu, S., Zhao, J., Wang, H. (2021), *MD-MBPLS: A novel explanatory model in computational social science. Knowledge-Based Systems*, 223, 107023, https://doi.org/10.1016/j.knosys.2021.107023.

[21] Scealy, J.L., De Caritat, P., Grunsky, E.C., Tsagris, M.T., Welsh, A.H. (2015), *Robust Principal Component Analysis for Power Transformed Compositional Data. Journal of the American Statistical Association*, 110(509), 136-148, https://doi.org/10.1080/01621459.2014.990563.

[22] Tolosana-Delgado, R., Otero, N., Pawlowsky-Glahn, V., Soler, A. (2005), *Latent Compositional Factors in The Llobregat River Basin (Spain) Hydrogeochemistry. Mathematical Geology*, 37(7), 681-702, https://doi.org/10.1007/s11004-005-7375-7.

[23] Tsagris, M., Preston, S., Wood, A.T.A. (2016), I*mproved Classification for Compositional Data Using the α-transformation. Journal of Classification*, 33(2), 243-261, https://doi.org/10.1007/s00357-016-9207-5.

[24] Vo-Van, T., PhamToan, D. (2024), *A supervised learning algorithm based on the quasi-Bayesian method for the probability density functions and application for medical data. Knowledge-Based Systems*, 299, 112003, https://doi.org/10.1016/j.knosys.2024. 112003.

[25] Wang, H., Shangguan, L., Guan, R., Billard, L. (2015), *Principal component analysis for compositional data vectors. Computational Statistics*, 30(4), 1079-1096, https://doi.org/10.1007/s00180-015-0570-1.

[26] Wang, H., Shangguan, L., Wu, J., Guan, R. (2013), *Multiple linear regression modeling for compositional data. Neurocomputing*, 122, 490-500, https://doi.org/10.1016/ j.neucom.2013.05.025.

[27] Wei, Y., Wang, Z., Wang, H., Yao, T., Li, Y. (2018), *Promoting inclusive water governance and forecasting the structure of water consumption based on compositional data: A case study of Beijing. Science of The Total Environment*, 634, 407-416, https://doi.org/10.1016/j.scitotenv.2018.03.325.

[28] Woods, C.M., Edwards, M.C. (2007), *12 Factor Analysis and Related Methods*. In Rao, C.R., Miller, J.P., Rao, D.C. (Eds.), *Handbook of Statistics*, 27, 367-394, Elsevier. https://doi.org/10.1016/S0169-7161(07)27012-9.

[29] Zhao, Q., Wang, H., Lu, S. (2022), *M-LDQ feature embedding and regression modeling for distribution-valued data. Information Sciences*, 609, 121-152, https://doi.org/10.1016/j.ins.2022.07.064.