IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment

BOOK OF SHORT PAPERS

Editors: Rosaria Lombardo, Ida Camminatiello and Violetta Simonacci

Book of Short papers
10th International Conference **IES 2022**
Innovation and Society 5.0: Statistical and Economic
Methodologies for Quality Assessment

Department of Economics, University of Campania "L. Vanvitelli",
January 27th - 28th 2022

Scientific Committee of the group of the Italian Statistical Society on Statistics for the Evaluation and Quality of Services – SVQS

Pietro Amenta -University of Sannio
Matilde Bini -European University of Roma
Luigi D'Ambra -University of Naples "Federico II"
Maurizio Carpita -University of Brescia
Paolo Mariani -University of Milan "Bicocca"
Marica Manisera -University of Brescia
Monica Palma -University of Salento
Pasquale Sarnacchiaro -University of Rome Unitelma Sapienza

Program Committee of the conference IES 2022

Chair: Rosaria Lombardo -University of Campania "L. Vanvitelli"
Fabio Bacchini -ISTAT
Laura Baraldi -University of Campania "L. Vanvitelli"
Eric Beh -University of Newcastle, Australia
Wicher Bergsma -The London School of Economic and Political Science, UK
Enrico Bonetti -University of Campania "L. Vanvitelli"
Eugenio Brentari -University of Brescia
Clelia Buccico -University of Campania "L. Vanvitelli"
Rosalia Castellano -University of Naples "Parthenope"
Ida Camminatiello -University of Campania "L. Vanvitelli"
Carlo Cavicchia -University of Rotterdam, The Netherlands
Enrico Ciavolino -University of Salento
Corrado Crocetta -University of Foggia
Claudio Conversano -University of Cagliari
Antonello D'Ambra -University of Campania "L. Vanvitelli"
Antonio D'Ambrosio -University of Naples "Federico II"
Alfonso Iodice D'Enza -University of Naples "Federico II"
Tonio Di Battista -University of Chieti "G. D'Annunzio"
Michele Gallo -University of Naples "L'Orientale"
Francesco Gangi -University of Campania "L. Vanvitelli"
Michele La Rocca -University of Salerno
Amedeo Lepore -University of Campania "L. Vanvitelli"
Riccardo Macchioni -University of Campania "L. Vanvitelli"
Filomena Maggino -University of Rome "La Sapienza"
Angelos Markos -Demokritus University of Thrace, Greece
Lucio Masserini -University of Pisa
Stefania Mignani -University of Bologna
Nicola Moscariello -University of Campania "L. Vanvitelli"
Francesco Palumbo -University of Naples "Federico II"
Alessandra Petrucci -University of Florence
Alessio Pollice -University of Bari

Donato Posa -University of Salento
Luca Secondi -University of Tuscia
Amalia Vanacore -University of Naples “Federico II”
Michel van de Velden -University of Rotterdam, The Netherlands
Rosanna Verde -University of Campania “L. Vanvitelli”
Donatella Vicari -University of Rome “La Sapienza”
Grazia Vicario -Polytechnic University of Turin
Maurizio Vichi -University of Rome “La Sapienza”

Organizing Committee

Ida Camminatiello -University of Campania “L. Vanvitelli”
Antonello D’Ambra -University of Campania “L. Vanvitelli”
Rosaria Lombardo -University of Campania “L. Vanvitelli”
Elvira Romano -University of Campania “L. Vanvitelli”
Luca Rossi -University Cusano
Violetta Simonacci -University of Naples “L’Orientale”

Editors

Rosaria Lombardo - University of Campania “L. Vanvitelli”, Italy
Ida Camminatiello - University of Campania “L. Vanvitelli”, Italy
Violetta Simonacci - University of Naples “L’Orientale”, Italy

pke
PROFESSIONAL
KNOWLEDGE
EMPOWERMENT

PKE - Professional Knowledge Empowerment s.r.l.
Sede legale: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Sede operativa: Villa Marelli - Viale Thomas Alva Edison, 45 - 20099 Sesto San Giovanni (MI)
Ufficio Di Rappresentanza: Via Giacomo Peroni, 400 - 00131 Roma (RM)
CF / P.I. 03167830920 — www.pke.it; e-mail info@pke.it — Privacy

February 2022 PKE s.r.l.
ISBN 978-88-94593-35-8 on print
ISBN 978-88-94593-36-5 online

CP decomposition of 4th-order tensors of compositions

Decomposizione CP di tensori composizionali di ordine 4

Violetta Simonacci, Tullio Menini and Michele Gallo

Abstract Multifold data structures are generally stored in high-dimensional objects defined as *n*th-order tensors. Generalization of trilinear decompositions such as the CANDECOMP/PARAFAC model can be used for modelling 4th order tensors. The application of these techniques is, however, quite limited due to procedural complexity and interpretational issues. These concerns increase when tensors contain data with a compositional structure. This work aims at addressing these difficulties through an application on Italian university staff.

Abstract *Strutture di dati complesse sono generalmente memorizzate in oggetti multidimensionali definiti come tensori di ordine n. Per modellare tensori di ordine 4, è possibile utilizzare generalizzazioni delle decomposizioni trilineari come il modello CANDECOMP/PARAFAC. L'applicazione di queste tecniche è, tuttavia, piuttosto limitata a causa della loro complessità procedurale e interpretativa. Tali difficoltà aumentano poi nel caso in cui i tensori contengano dati con una struttura composizionale. Questo lavoro mira ad affrontare tali problemi attraverso un'applicazione sul personale universitario italiano.*

Key words: CoDa, CANDECOMP/PARAFAC, logratio, higher order decomposition, parameter estimation

Violetta Simonacci
Dept. of Social Science, University of Napels Federico II, Naples, Italy
e-mail: violetta.simonacci@unina.it

Tullio Menini
Dept. of Human and Social Sciences, University of Naples - "L'Orientale", Naples, Italy
e-mail: menini@unior.it

Michele Gallo
Dept. of Human and Social Sciences, University of Naples - "L'Orientale", Naples, Italy
e-mail: mgallo@unior.it

1 Introduction

Complex social phenomena are the results of different layers of information continuously interacting at repeated occasions. As data-storing capabilities become virtually unbounded, finding effective ways of modeling together multiple entities has become an ongoing challenge.

Tensors are the preferred algebraic architecture for storing complex data and describing multilinear relationships between entities in a compact form. A generic nth-order tensor stores data along n indices and can be described as a generalization of simple structures such as scalars, vectors and matrices which are special cases of 0-order (no index), 1st-order (1 index) and 2nd-order (2 indices) tensors.

Tensor data structure may presents additional challenges besides a multidimensional variability structure. Let us think of tensor with proportion values (e.g. percentages, shares, parts of a total), defined in statistical literature data as Compositional Data (CoDa). Such data are characterized by a biased covariance structure which can be modeled only in relative terms [1] and requires special tools.

Tensor decompositions techniques can come quite handy when dealing with multilinear data. These tools allow capturing the multidimensional information in a tensor by breaking it down in sets of simpler objects, generally lower order tensors. The two most commonly used techniques for the decomposition of nth-order tensors are the Higher-Order TUCKER and CANDECOMP/PARAFAC (CP) models [9, 5]. The TUCKER model is more suitable for summarizing large information into condensed sets of variables, thus, it is the preferred method for tensor compression and variability structure descriptions. The CP method is more appealing when trying to retrieve a meaningful underlying structure. This is because this model provides a unique solution under mild conditions [8].

The higher order CP model can be easily adapted to compositional data by use of log-ratio transformations which, applied prior to the decomposition, do not alter its procedural steps but call for an additional interpretability effort.

Multilinear decomposition for tensors of order higher than 3 are occasionally used in Chemistry related fields, however, their applications in social sciences is uncommon. This is mainly due to model complexity which makes these tools unfriendly for non-experts. For tensors of compositions the degree of complexity increases even more, thus, compositional adaptations of n-th order decompositions are completely absent in social sciences.

Given these considerations the aim of this work is to address two issues which cause the infrequent use of these tools, namely, parameter estimation ambiguities and interpretability concerns. The focus will be only on the CP procedure because its desirable uniqueness makes it more vulnerable to efficiency and algorithmic problems.

In order to reach this goal, an application on University teaching staff in Italy recorded by macro-region, disciplinary field, role and year will be presented. Specifically, a 4th-order tensor is considered in which disciplinary field shares are treated as compositional data. After following CoDa methodology by extending the strategy proposed for tridimensional arrays to a 4-way tensor, the CP model will be computed. Results will be analyzed by paying careful attention to the advantages of using such procedure and to the estimating problems of current algorithms in a compositional setting.

In Section 2 tensor notation is explained and the dataset is briefly introduced; in Section 3 the methodology is outlined for the four-way CoDa-CP procedure and in Section 4 some initial consideration are conveyed.

2 Tensor notation a data

Let us consider a 4th-order tensor $\mathscr{T}$ with data arranged over the four indices $[1,\dots,i,\dots,I]$, $[1,\dots,j,\dots,J]$, $[1,\dots,k,\dots,K]$ and $[1,\dots,l,\dots,L]$. Its generic element is denoted by t_{ijkl}. The information contained in such tensor can be rearranged in many ways to focus on index relationships. The simplest way is to consider its composing vectors, generally referred to fibers. There are four types of fibers, one for each index so that I-,J-, K- and L-dimensional vectors can be identified as a generalization of rows and columns of a matrix. It is clear that there are as many fibers of a type as the product of the remaining indices, e.g. there are IKL fibers or rows $\mathbf{t}_{i:kl}$ with dimension J.

The tensor $\mathscr{T}$ can also be rearranged in 3rd-order blocks obtained by combining two of the four modes together into pseudo-fully stretched arrays $\underline{\mathbf{T}}_I(I \times JK \times L)$, $\underline{\mathbf{T}}_J(J \times KL \times I)$, $\underline{\mathbf{T}}_K(K \times LI \times J)$ and $\underline{\mathbf{T}}_L(L \times IJ \times K)$ [6].

Each of these tridimensional blocks can be seen as a set of slices, namely 2nd-order sections obtained by fixing one the three indices of the pseudo-fully stretched arrays and varying the remaining two. Specifically, it is possible to identify four sets of frontal slices $\mathbf{T}_{::l}(I \times JK)$, $\mathbf{T}_{::i}(J \times KL)$, $\mathbf{T}_{::j}(K \times LI)$ and $\mathbf{T}_{::k}(L \times IJ)$.

These alternative notations are only some of the many ways tensor information can be rearranged presented here to aid methodological explanations.

A 4th-order tensor presents a compositional structure if the elements of at least one of the fiber types describe the parts of a whole. Following conventions, let us assume that the J-dimensional fibers or rows are CoDa. Formally we have that the generic row $\mathbf{t}_{i:kl}$ is a compositional vector if it describes a point bounded in a subspace of $\Re^J_+$ known as simplex and defined as:

$$S^J = \left\{ \left(t_{i1kl_{(1)}}, \dots, t_{iJkl} \right) : t_{i1kl} \geq 0, \dots, t_{iJkl} \geq 0; t_{i1kl} + \dots + t_{iJkl} = \kappa \right\} \tag{1}$$

where κ is a positive constant. To operate within this subspace special operations and rules known as Aitchison geometry must be followed. Alternatively CoDa vectors can be conveyed in real space coordinates by transforming them into log-ratios. Several transformations have been proposed in the literature, however, for brevity purposes only centered log-ratio (clr) coordinates are introduced. This function generates an isometric mapping between S^J and a hyperplane of $\Re^J$ in this fashion:

$$\mathbf{z}_{i:kl} = \mathrm{clr}(\mathbf{t}_{i:kl}) = \left[\log\frac{t_{i1kl}}{\mathrm{g}(\mathbf{t}_{i:kl})},\ldots,\log\frac{t_{ijkl}}{\mathrm{g}(\mathbf{t}_{i:kl})},\ldots,\log\frac{t_{iJkl}}{\mathrm{g}(\mathbf{t}_{i:kl})}\right] \text{ with } \mathrm{g}(\mathbf{t}_{i:kl}) = \sqrt[J]{\prod_{j=1}^{J} t_{ijkl}} \quad (2)$$

These coordinates have the limit of yielding a pure multicollinear structure, which may cause estimating issues. As demonstrated in [2, 3] for 3rd-order tensors, *clr*-coordinates can be directly modeled with standard statistical tools. For the 4th-order tensor $\mathscr{T}$ a four-way CP model can be implemented, than results are translated back into compositional terms.

After clarifying tensor notation, the application of interest can be described in these terms. The dataset contains information on University teaching staff in Italy arranged over 4 directions with the following dimensions: 5 macro-region, 14 disciplinary fields, 3 role and 5 year, yielding a small tensor $\mathscr{T}$ with dimensions $(I = 5 \times J = 14 \times K = 3 \times L = 5)$.

For each macro-region, the partitioning among different disciplinary fields of the total number employee can be described as a compositional problem. Each row vector can thus be transformed as shown in eq.2 obtaining a new 4th-order tensor $\mathscr{Z} \in \mathbb{R}^{5\times14\times3\times5}$. This tensor can be decomposed with the CoDa-CP model as showed in the following section.

3 Four-way CoDa-CP model

Four-way CoDa-CP is an estimating model based on the polyadic decomposition which aims at providing the best low rank approximation of the tensor $\mathscr{Z} = \hat{\mathscr{Z}} + \mathscr{E}$, where $\mathscr{E}$ is the tensor of residuals. Here, the tensor is decomposed into the sum of a finite $f = 1,\ldots,F$ number of 1st-order factors $\mathbf{a}_f$, $\mathbf{b}_f$, $\mathbf{c}_f$ and $\mathbf{d}_f$:

$$\hat{\mathscr{Z}} = \sum_{f=1}^{F} \mathbf{a}_f \circ \mathbf{b}_f \circ \mathbf{c}_f \circ \mathbf{d}_f \quad (3)$$

The F terms of this decomposition can be arranged in four factor matrices $\mathbf{A} = [\mathbf{a}_1,\ldots,\mathbf{a}_f,\ldots,\mathbf{a}_F]$, $\mathbf{B} = [\mathbf{b}_1,\ldots,\mathbf{b}_f,\ldots,\mathbf{b}_F]$, $\mathbf{C} = [\mathbf{c}_1,\ldots,\mathbf{c}_f,\ldots,\mathbf{c}_F]$ and $\mathbf{D} = [\mathbf{d}_1,\ldots,\mathbf{d}_f,\ldots,\mathbf{d}_F]$.

The model can also be rewritten using the pseudo fully starched array slice notation as follows:

$$\mathbf{Z}_{::l} = \mathbf{A}\mathrm{diag}(\mathbf{d}_{(l)})(\mathbf{C} \odot \mathbf{B})^t + \mathbf{E}_{::l} \quad l = 1,\cdots,L \quad (4)$$

$$\mathbf{Z}_{::i} = \mathbf{B}\mathrm{diag}(\mathbf{a}_{(i)})(\mathbf{D} \odot \mathbf{C})^t + \mathbf{E}_{::i} \quad i = 1,\cdots,I \quad (5)$$

$$\mathbf{Z}_{::j} = \mathbf{C}\mathrm{diag}(\mathbf{b}_{(j)})(\mathbf{A} \odot \mathbf{D})^t + \mathbf{E}_{::j} \quad j = 1,\cdots,J \quad (6)$$

$$\mathbf{Z}_{::k} = \mathbf{D}\mathrm{diag}(\mathbf{c}_{(k)})(\mathbf{B} \odot \mathbf{A})^t + \mathbf{E}_{::k} \quad k = 1,\cdots,K \quad (7)$$

Here $\odot$ is the Khatri-Rao product and diag$(\mathbf{d}_{(l)})$, diag$(\mathbf{a}_{(i)})$, diag$(\mathbf{b}_{(j)})$ and diag$(\mathbf{c}_{(k)})$ denote the diagonal matrices extracting the lth, ith, jth and kth rows of the factor matrices respectively.

The four-way CoDa-CP model is unique under mild conditions and is generally estimated through a least-squares loss function. Estimation problems may, however, occur, such as solution degeneracies [10] and slow convergence, especially for collinear data [7].

4 Preliminary considerations

One of the best ways to unveil the latent structure of 4th-order tensor is to carry out a CP decomposition. The uniqueness of the CP model makes this procedure both appealing and harder to estimate with respect to other techniques for the decomposion of 4th-order tensors as the TUCKER model [9]. Many difficulties may arise when estimating CP parameters connected to both efficiency and accuracy of the solution. Multicollinearity, typical of *clr*-coordinates, makes this issues even more pressing.

Several procedure have been proposed over the years to cope with these difficulties, all with different points of strenght and fallacies. The problem, however, is generally dealt with for the simpler case of 3rd-order tensors.

In this work, by considering the 4th-order tensor of University teaching staff data we are going to tackle two challenges: 1) show the potential of the four-way CoDa CP methodology with respect to other, more common, modeling tools; 2) explore the estimation problem of the CP model in the generalized framework of 4-way compositional data by extending the work of [4].

References

1. Aitchison, J.: The statistical analysis of compositional data. Chapman & Hall (1986)
2. Gallo, M.: Log-ratio and parallel factor analysis: an approach to analyze three-way compositional data. In: Advanced dynamic modeling of economic and social systems, pp. 209-221, Springer (2013)
3. Gallo, M., Simonacci, V.: A procedure for the three-mode analysis of compositions. Electronic Journal of Applied Statistical Analysis **6**(2), 202–210 (2013)
4. Gallo, M., Simonacci, V., Di Palma, M.A.: An integrated algorithm for three-way compositional data. Quality & Quantity **53**(5), 2353–2370 (2019)
5. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an "explantory" multi-modal factor analysis. UCLA working papers in phonetics **16**, 1–84 (1970)
6. Kang, C., Wu, H.L., Yu, Y.J., Liu, Y.J., Zhang, S.R., Zhang, X.H., Yu, R.Q.: An alternative quadrilinear decomposition algorithm for four-way calibration with application to analysis of four-way fluorescence excitation–emission–ph data array. Analytica chimica acta **758**, 45–57 (2013)
7. Kiers, H.A.: A three-step algorithm for CANDECOMP/PARAFAC analysis of large data sets with multicollinearity. Journal of Chemometrics **12**(3), 155–171 (1998)
8. Sidiropoulos, N.D., Bro, R.: On the uniqueness of multilinear decomposition of n-way arrays. Journal of chemometrics **14**(3), 229–239 (2000)

9. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
10. Zijlstra, B.J., Kiers, H.A.: Degenerate solutions obtained from several variants of factor analysis. Journal of Chemometrics: A Journal of the Chemometrics Society **16**(11), 596–605 (2002)