Three-way compositional data: a multi-stage trilinear decomposition algorithm

Dati composizionali a tre vie: un algoritmo multi-stadio di decomposizione trilineare

Gallo M., Simonacci V., and Di Palma M.A.

Abstract The CANDECOMP/PARAFAC model is an extension of bilinear PCA and has been designed to model three-way data by preserving their multidimensional configuration. The Alternating Least Squares (ALS) procedure is the preferred estimating algorithm for this model because it guarantees stable results. It can, however, be slow at converging and sensitive to collinearity and over-factoring. Dealing with these issues is even more pressing when data are compositional and thus collinear by definition. In this talk the solution proposed is based on a multi-stage approach. Here parameters are optimized with procedures that work better for collinearity and over-factoring, namely ATLD and SWATLD, and then results are refined with ALS.

Abstract *Il modello CANDECOMP/PARAFAC è una generalizzazione per matrici a tre indici dell'ACP. Per stimare i parametri di tale modello la procedura di stima più usata è l'Alternating Least Squares (ALS). Tale algoritmo è il più usato in quanto garantisce risultati stabili, tuttavia, presenta anche degli inconvenienti, quali essere lento e sensibile alla multicollinearità e alla sovra-fattorizzazione. Affrontare questi problemi diventa poi particolarmente impegnativo quando i dati sono multicollineari per costruzione, come nel caso dei dati composizionali. Come soluzione di tali problemi, nel presente lavoro si propone un approccio multi-stadio in cui i parametri sono prima ottimizzati con procedure che funzionano meglio quando vi llinearità e sovra-fattorizzazione, cioè ATLD e SWATLD, e successivamente i risultati finali sono individuati con l'ALS.*

Michele Gallo
DISUS University of Naples "L'Orientale", Largo S. Giovanni Maggiore 30, 80134 Naples, Italy.
e-mail: mgallo@unior.it

Violetta Simonacci
DISUS University of Naples "L'Orientale", Largo S. Giovanni Maggiore 30, 80134 Naples, Italy.
e-mail: vsimonacci@unior.it

Maria Anna Di Palma
DISUS University of Naples "L'Orientale", Largo S. Giovanni Maggiore 30, 80134 Naples, Italy.
e-mail: madipalma@unior.it

Key words: CP model, PARAFAC-ALS, ATLD, SWATLD, Compositional Data

1 Introduction

Observations over a set of variables can be recorded in different occasions, such as time or location. These data present a tridimensional structure and the only way to obtain a low rank approximation without confusing the variability of two dimensions together is using multi-linear techniques such as the CANDECOMP/PARAFAC (CP) model [2, 10]. This model estimates three separate sets of parameters, one for each mode of the analysis, thus is highly complex and the search for innovative ways to improve its efficiency without compromising accuracy of results is of great relevance.

The most widely used algorithm for the CP model is currently PARAFAC-ALS (ALS) thanks to the merit of granting stable results, a least square solution and an always monotonically decreasing fit. It does, however, present some problematic aspects such as slow convergence and sensitiveness to over-factoring, multicollinearity and factor collinearity. These issues are even more significant when dealing with data that present particular challenges such as Compositional Data (CoDa) [1, 11] . CoDa can be defines as positive vectors with a purely multicollinear structure as their elements describe the parts of a whole and thus only carry relative information. Given these considerations, in [9] an alternative way to overcome these difficulties in a compositional framework is presented. Specifically it is suggested that in order to mitigate ALS inefficiencies this procedure can be integrated by adding an initialization/recovery stage where parameters are optimized through the Self-Weighted TriLinear Decomposition (SWATLD). In this manner a novel two-stage procedure is implemented (INT-1).

SWATLD proposed by [3] was chosen amongst other alternative because it can be seen as complementary to ALS given than its strengths are fast convergence and robustness to over-factoring and collinearity while its fallacies are finding a solution in a non-least-square sense and unstable results [5, 12, 14, 16].

INT-1 appears to work quite well in the simulations presented in the cited article, however several ways to improve its performance and reliability were suggested in future developments but not yet verified. In this perspective the purpose of this contribution is to explore the possibility of improving the performance of INT-1 by trying to answer two unresolved queries.

The first question is the consequence of a methodological comparison with [15] where it is argued that the Alternating TriLinear Decomposition (ATLD) proposed in [13] works better than SWATLD for initializing random numbers, multicollinearity and speediness. We thus wondered if ATLD could be considered as an initialization step. To resolve this, a second multi-stage procedure (INT-2) was devised, this time with three stages, to see if adding an ATLD step to start off could improve performance.

The second problem concerns the identification of an optimal transition point from

one stage to the next for the integrated procedure, i.e. is there an optimal convergence criteria or number of iterations capable of making INT-1 and INT-2 perform at their best? This question is addressed in a simulation study on stage transition parameters. Once these two aspects are dealt with, a new comparative study can be carried out to verify three points of interest: whether both INT-1 and INT-2 perform better than ALS for compositional data and in what terms; which between INT-1 and INT-2 is a better alternative; and how do data characteristics such as noise level and factor collinearity influence results.

2 Multidimensional data with a compositional structure

Let us consider a three-way array $\underline{\mathbf{V}}$ $(I \times J \times K)$ with generic positive element v_{ijk} where $i = 1 \ldots I$, $j = 1 \ldots J$, and $k = 1 \ldots K$. If its row vectors $\mathbf{v}_{ik} = [v_{i1k}, ..., v_{iJk}]$ present a biased covariance structure due to an implicit or explicit sum constraint $v_{i1k} + \ldots + v_{iJk} = \kappa$, where κ is a positive constant, the array has a compositional structure and should be processed with compositional methodology.
This bounded covariance imposes a purely multicollinear structure to the data since the elements of a compositional vector are not linearly independent and thus the covariance matrix for each of the K frontal slabs $\mathbf{V}_k(I \times J)$ of the array $\underline{\mathbf{V}}$ will be singular.
From a geometric stand point these row vectors are forced in a subspace of $\Re_+^J$ known as simplex and defined as:

$$S^J = \{(v_{i1k}, \ldots, v_{iJk}) : v_{i1k} \geq 0, \ldots, v_{iJk} \geq 0;\ v_{i1k} + \ldots + v_{iJk} = \kappa\} \tag{1}$$

To operate within this subspace a non-Euclidean set of rules, known as Aitchison geometry, is used to identify a linear vector space [11]. Compositional vectors can, however, be converted into Euclidean space coordinates by using log-ratio transformations: pairwise, centered, additive [1] or isometric [4].
For the purpose of this contribution we will only be referring to centered log-ratio (*clr*) coordinates which can be expressed as:

$$\mathbf{z}_{ik} = clr(\mathbf{v}_{ik}) = \left[ln\frac{v_{i1k}}{g(\mathbf{v}_{ik})}, \ldots, ln\frac{v_{iJk}}{g(\mathbf{v}_{ik})} \right] \quad \text{with } g(\mathbf{v}_{ik}) = \sqrt[J]{\prod_{j=1}^{J} v_{ijk}} \tag{2}$$

By applying this transformation the tridimensional array of compositions $\underline{\mathbf{V}}$ can easily be changed into an array of *clr*-coordinates $\underline{\mathbf{Z}}$ so that standard algorithms can be applied as long as results are interpreted in compositional terms [6, 8]. It is important to note that *clr*-coordinates by providing an S^J to $\Re^J$ projection, do not remove the collinearity problem.

3 CP model and estimating procedures

An array of *clr*-coordinates $\underline{\mathbf{Z}}$ can be decomposed with the CP model in three sets of parameters, one for each mode of the analysis. Let F be the number of considered factors, using a slab-wise notation we can write:

$$\mathbf{Z}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^t + \mathbf{E}_k \quad k = 1, ..., K \tag{3}$$

where $\mathbf{A}$ $(I \times F)$ and $\mathbf{B}$ $(J \times F)$ are the loading matrices for the first and second mode, respectively; $\mathbf{D}_k$ is a diagonal matrix containing the kth row of $\mathbf{C}$ $(K \times F)$, loading matrix of third mode; $\mathbf{Z}_k$ $(I \times J)$ is the kth frontal slab of $\underline{\mathbf{Z}}$; and $\mathbf{E}_k$ $(I \times J)$ is the corresponding frontal slab of the error array $\underline{\mathbf{E}}$.
Different algorithms can be used to fit the data to the model. The most common one is ALS. This is an iterative procedure where sets of parameters are estimated in three successive least-square steps. On the other hand, ATLD and SWATLD are also three-step iterative procedures but do not follow a least-square approach and are characterized by the use of three distinct objective function, one for each mode, which focus on prioritizing the trilinear structure of the data.

4 Multi-stage procedures

The described algorithms all present some qualities and weaknesses directly derived from the properties of their loss functions. ATLD is the fastest at converging and it is robust to over-factoring, collinearity and initial values. It does not, however, find a least-square solution, it may not monotonically decrease, it is sensitive to noise and often does not converge properly. On the opposite end there is ALS, the slowest at converging, stable in its results, capable of finding a solution in the least square sense but sensitive to collinearity and over-factoring. SWATLD occupies a middle ground: it is more stable than ATLD but not quite as reliable as ALS, it is pretty fast at converging but slower than ATLD while still robust to over-factoring and collinearity. In addition it may still not have a monotonically decreasing fit and not converge to a least square solution.
Given these considerations two multi-stage procedures were devised to try and maximize the advantages and counter-balance the inefficiencies of these algorithms. INT-1 is structured in the following manner: in a first stage (recovery stage) parameters are estimated by SWATLD with the purpose of identifying the correct underlying components in case of over-factoring, to deal better with multicollinearity and to speed up the procedure; successively in a second stage the solution is adjusted through ALS steps (refinement stage) to obtain a least square solution and avoid SWATLD instabilities.
INT-2 presents a similar outline but also includes an additional initialization ATLD stage, which could help when dealing with multicollinearity and bad initial values.
It is important to note that for both algorithms at least one iteration has to be per-

formed at each stage. A schematic overview of the procedures is displayed in Fig.1. In both cases step transition can be user defined in terms of relative fit and number of iterations. However these transition parameters can hugely hinder or improve performance of both INT-1 and INT-2, thus ideal values will be identified through a threshold simulation study. Once optimal parameters are found, they will be included as defining elements of the procedures.

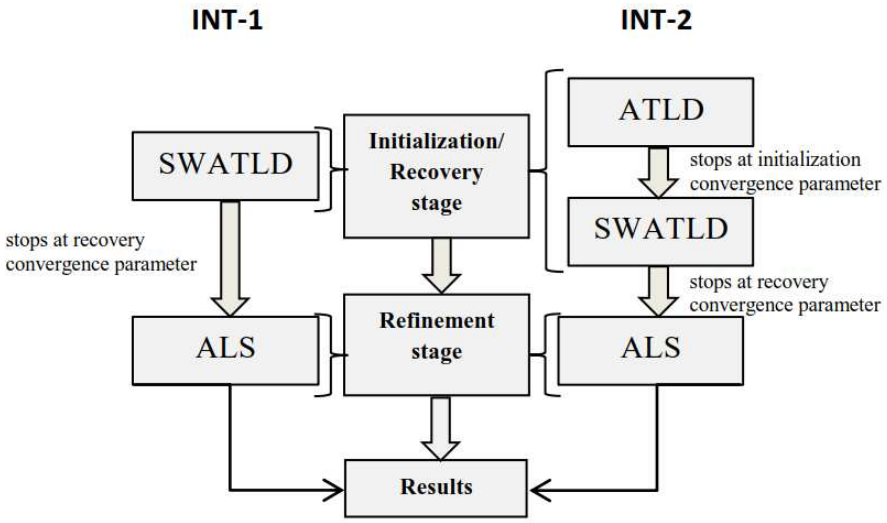

Fig. 1 Multi-stage procedures outline

5 Discussion

With the purpose of further developing the findings presented in [9] where a two stage SWATLD-ALS is introduced, this contribution proposes two important advancements: 1) devising a three-step INT-2 procedure to see if initializing with ATLD grants additional benefits; and 2) setting up a study to identify ideal stage transition parameters for both INT-1 and INT-2. A comparative simulation study is then carried out in a compositional setting to compare INT-1 and INT-2 to ALS, once ideal parameters are set. Multiple scenarios will be considered with different levels of noise and factor collinearity

Given that only partial results are available, at this stage we can only make the following considerations. In terms of ideal transition parameters, there is a trade-off between accuracy and efficiency: stricter relative fit convergence criteria (10^{-3} or 10^{-4}) generally render the algorithms more efficient but more unstable. On the other hand looser criteria are less fast but more reliable (10^{-1} or 10^{-2}) and for this rea-

son most likely preferable. In terms of comparative results we thus expect to see the INT-1 and INT-2 with ideal parameters performing similarly to ALS in terms of reliability (better in case of over-factoring) while still being far more efficient, with INT-1 slightly more reliable but a little slower than INT-2. Complete and in-depth results will be discussed during presentation.

References

1. Aitchison, J.: The Statistical Analysis of Compositional Data. Chapman & Hall (1986)
2. Carroll, J. D.,Chang, J.J.: Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrika 35(3), 283–319 (1970)
3. Chen, Z.P., Wu, H.L., Jiang, J.H., Li, Y., Yu, R.Q.: A novel trilinear decomposition algorithm for second-order linear calibration. Chemometrics and Intelligent Laboratory Systems 52(1):75–86 (2000)
4. Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras G., Barcelo-Vidal C.: Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300 (2003)
5. Faber, N.K.M., Bro, R., Hopke, P.K.: Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. Chemometrics and Intelligent Laboratory Systems 65(1):119–137 (2003)
6. 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, Springer, pp 209–221(2013)
7. Gallo, M., Buccianti, A.: Weighted principal component analysis for compositional data: application example for the water chemistry of the arno river (Tuscany, central Italy). Environmetrics 24(4):269–277 (2013)
8. Gallo, M., Simonacci V.: A procedure for the three-mode analysis of compositions. Electronic Journal of Applied Statistical Analysis 6(2):202–210 (2013)
9. Gallo, M., Di Palma, M.A., Simonacci V.: Integrated SWATLD-ALS algorithm for Compositional Data (2016). Submitted
10. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis (1970)
11. Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado R.: Modeling and analysis of compositional data. John Wiley & Sons (2015).
12. Tomasi, G., Bro, R.: A comparison of algorithms for fitting the PARAFAC model. Computational Statistics & Data Analysis 50(7):1700–1734 (2006)
13. Wu, H.L., Shibukawa, M., Oguma, K.: An alternating trilinear decomposition algorithm with application to calibration of hplc–dad for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. Journal of Che-mometrics 12(1), 1–26 (1998)
14. Yu, Y.J., Wu, H.L., Nie, J.F., Zhang, S.R., Li, S.F., Li, Y.N., Zhu, S.H.,Yu, R.Q.: A comparison of several trilinear second-order calibration algorithms. Chemometrics and Intelligent Laboratory Systems 106(1):93–107 (2011)
15. Yu, Y. J., Wu, H. L., Kang, C., Wang, Y., Zhao, J., Li, Y. N., Liu, Y.J., Yu, R. Q. Algorithm combination strategy to obtain the secondorder advantage: simultaneous determination of target analytes in plasma using threedimensional fluorescence spectroscopy. Journal of Chemometrics, 26(5), 197-208 (2012)
16. Zhang, S.R., Wu, H.L., Yu, R.Q.: A study on the differential strategy of some iterative trilinear decomposition algorithms: PARAFAC-ALS, ATLD, SWATLD,and APTLD. Journal of Chemometrics 29(3):179—192 (2015)