Matilde Bini, Pietro Amenta, Antonello D'Ambra, Ida Camminatiello
Editors

Società Italiana di Statistica

Statistical Methods for Service Quality Evaluation

Book of short papers

9th International Conference **IES 2019** - Innovation & Society -

Statistical evaluation systems at 360°: techniques, technologies and new frontiers

organized by Statistics for the Evaluation and Quality in Services Group of the Italian Statistical Society and European University of Rome

SIS Società Italiana di Statistica

Cuzzolin

Matilde Bini - European University of Rome, Italy
Pietro Amenta - University of Sannio, Italy
Antonello D'Ambra - University of Campania "L. Vanvitelli", Italy
Ida Camminatiello - University of Campania "L. Vanvitelli", Italy
Editors

Prima Edizione: Luglio 2019

Traversa Michele Pietravalle, 8 - 80131 Napoli
Tel. 081 5451143 - Fax 081 7707340
cuzzolineditor@cuzzolin.it
www.cuzzolineditore.com

ISBN: 978-88-86638-65-4

CUZZOLIN EDITORE

Table of Contents

CP model estimation with incorrect rank of factorization on large data sets.

Stima del modello CP con selezione erronea del rango di fattorizzazione su dati di grandi dimensioni.

Violetta Simonacci and Michele Gallo and Enrico Ciavolino

Abstract The accuracy of the ALS procedure for fitting the CP decomposition is affected by incorrect model selection due to the properties of its objective function. A study on ALS performance is presented in order to show that for large data sets this occasional shortcoming becomes prevalent. This deficiency should warn off researchers from employing ALS unless the rank of the underlying trilinear structure can be established in advance. Multi-optimization provides a possible solution: ALS can be initialized with procedures insensitive to over-factoring such as SWATLD and ATLD. In this manner it is possible to overcome factorization issues and provide a gain in efficiency without relying on computationally expensive model selection procedures.
Abstract *La precisione della procedura ALS per la decomposizione PARAFAC è condizionata dalla selezione del modello a causa delle proprietà della sua funzione di perdita. Uno studio sulle prestazioni dell'ALS verrà presentato per dimostrare che per dataset di grandi dimensioni questo problema diventa diffuso e dovrebbe scoraggiare l'impiego dell'ALS nel caso in cui il rango della struttura trilineare non sia conosciuto in anticipo. La multi-ottimizzazione rappresenta una possibile soluzione: l'ALS può essere inizializzato con procedure insensibili alla sovra-fattorizzazione come lo SWATLD e l'ATLD. Questo metodo risolve i problemi di fattorizzazione e garantisce un aumento di efficienza senza ricorrere a procedure per la selezione del modello, dispendiose dal punto di vista computazionale.*

Key words: ALS, ATLD, model selection, over-factoring, SWATLD

Michele Gallo
Università degli Studi di Napoli - "L'Orientale", DISUS, Largo S. Giovanni Maggiore, 30, Napoli, e-mail: mgallo@unior.it

Violetta Simonacci
Università degli Studi di Napoli - "L'Orientale", DISUS, Largo S. Giovanni Maggiore, 30, Napoli, e-mail: vsimonacci@unior.it

Enrico Ciavolino
Università del Salento, DSSSU, Via di Valesio, 24, LECCE, e-mail: enrico.ciavolino@unisalento.it

1 Introduction

The CANDECOMP/PARAFAC (CP) model is a decomposition tool for three-way tensors [1, 4]. This method works under the hypothesis of trilinearity of the data and preserves the multi-modal structure of the variance by estimating three groups of parameters, one for each dimension. As a result its calculation on large data sets can be quite demanding in terms of memory usage and computational time.

PARAFAC-ALS (ALS) is the estimating procedure proposed in the original formulation of the model to fit the data. It is by far the method of choice because it is based on the minimization of squared errors and provides stable, least squares solutions. Nevertheless, ALS presents well documented problems when the model is not correctly specified. The algorithm struggles to converge properly in case of over-factoring, namely, if the number of factors used for the approximation is larger than the number of components describing the real underlying solution.

Specifically, ALS is likely to encountered two possible issues: temporary and permanent degeneracies. The first problem, also referred to as two-factor degeneracy or swamp, is a typical difficulty of the least squares estimation method which consists in the algorithm slowing down in correspondence to local minima and employing an excessive amount of iterations to converge [5]. This issue is much more likely to arise in case of over-factoring but it can also be connected to bad initialization.

In addition, over-factoring may also cause more severe consequences: ALS occasionally suffers from permanent degeneracies, yielding incorrect solutions. The ALS capability of identifying the real latent structure in case of erroneous model specification is reduces because of the differential properties of its objective function [10].

This works aims to demonstrate that the problem of permanent degeneracies becomes more relevant as the dimensions of the data set increase. ALS approximation of large three-way tensors almost always extracts a wrong or highly inaccurate solution when over-factoring. This limitation makes the algorithm unsuitable for large data sets unless the latent structure is known a priori.

As a solution, two multi-optimization methods, INT and INT-2, are suggested. They work by initializing ALS with procedures known to be insensitive to over-factoring, i.e. the Self-Weighted TriLinear Decomposition (SWATLD) [2] and the Alternating TriLinear Decomposition (ATLD) [9] respectively. The performance of ALS compared the ones of the proposed alternatives is thus evaluated in an experimental design under different conditions.

2 CP estimation and model selection

A three-way array $\underline{\mathbf{X}}$ $(I \times \;J \times K\;)$ containing measurements of $i = 1,...,I$ observations on $j = 1,...,J$ variables at $k = 1,...,K$ occasions describing a trilinear phnomenon can be decomposed using the CP model. This multi-way technique yields a representation of $\underline{\mathbf{X}}$ by estimating the three loading matrices $\mathbf{A}$ $(I \times F)$,

$\mathbf{B}$ $(J \times F)$ and $\mathbf{C}$ $(K \times F)$ where F is the number of components, i.e. factors of the low-dimensional solution.

$\underline{\mathbf{X}}$ can be seen as a collection of two-entry matrices and subdivided in K matrices $\mathbf{X}_k$ of dimension $(I \times J)$ referred to as frontal slices. Similarly it can also be arranged in I horizontal slices $\mathbf{X}_i$ $(J \times K)$ or in J vertical slices $\mathbf{X}_j$ $(I \times K)$.

Using this notation the CP model can be written in one of the following ways:

$$\mathbf{X}_i = \mathbf{B}\mathbf{D}_i\mathbf{C}^t + \mathbf{E}_i \quad i = 1,...,I \tag{1}$$

$$\mathbf{X}_j = \mathbf{C}\mathbf{D}_j\mathbf{A}^t + \mathbf{E}_j \quad j = 1,...,J \tag{2}$$

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

where $\mathbf{D}_i$, $\mathbf{D}_j$ and $\mathbf{D}_k$ are diagonal operators extracting the ith, jth and k-th row of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ respectively and $\mathbf{E}_i$, $\mathbf{E}_j$ and $\mathbf{E}_k$ are the horizontal, vertical and frontal slices of the error array $\underline{\mathbf{E}}$.

The CP model presents the advantage of providing a unique solution under mild conditions [7]. However this determinacy is associated with the risk of a degenerate solution [11]. Choosing the incorrect dimensionality of the trilinear model is one of the main causes of degeneracy. This is particularly true if the ALS algorithm is used. Despite its stability and well-defined properties, ALS has the shortcoming of being highly dependent on the rank of factorization.

In the literature, issues with over-factoring are generally addressed by using additional procedures for model selection [6]. These methods, however, are not always reliable and are too computationally expensive for large data sets.

The two algorithms ATLD and SWATLD have been proposed as alternative to bypass over-factoring degeneracies. Nonetheless, these methods present other instabilities due to the use of three loss functions rather than just one: they do not ensure a monotonically decreasing fit and tend to return noisier solutions. ATLD and SWATLD can however be integrated with ALS into two variants, INT and INT-2 [3]. These procedures have two stages as shown in Tab.1 where the loss functions for each stage and procedure are illustrated. Using two optimization steps allows retaining the advantages of both ALS and SWATLD/ATLD. In this manner over-factoring inefficiencies can be addressed without having to employ expensive model selection procedures, which is of utmost importance for large data.

Table 1 INTegrated ALS procedures

	Loss func.	INT	INT-2
		SWATLD	ATLD
STAGE 1	$L(\mathbf{A})$	$\sum_{i=1}^{I} \left\|(\mathbf{B}^{+}\mathbf{X}_i - \mathbf{D}_i\mathbf{C}^t)^t\mathbf{D}_C^{-1}\right\|^2 + \sum_{i=1}^{I} \left\|(\mathbf{X}_i\mathbf{C}^{+t} - \mathbf{B}\mathbf{D}_i)\,\mathbf{D}_B^{-1}\right\|^2$	$\sum_{i=1}^{I} \|\mathbf{X}_i - \mathbf{B}\mathbf{D}_i\mathbf{C}^t\|^2$
	$L(\mathbf{B})$	$\sum_{j=1}^{J} \left\|(\mathbf{C}^{+}\mathbf{X}_j - \mathbf{D}_j\mathbf{A}^t)^t\mathbf{D}_A^{-1}\right\|^2 + \sum_{j=1}^{J} \left\|(\mathbf{X}_j\mathbf{A}^{+t} - \mathbf{C}\mathbf{D}_j)\,\mathbf{D}_C^{-1}\right\|^2$	$\sum_{j=1}^{J} \|\mathbf{X}_j - \mathbf{C}\mathbf{D}_j\mathbf{A}^t\|^2$
	$L(\mathbf{C})$	$\sum_{k=1}^{K} \left\|(\mathbf{A}^{+}\mathbf{X}_k - \mathbf{D}_k\mathbf{B}^t)^t\mathbf{D}_B^{-1}\right\|^2 + \sum_{k=1}^{K} \left\|(\mathbf{X}_k\mathbf{B}^{+t} - \mathbf{A}\mathbf{D}_k)\,\mathbf{D}_A^{-1}\right\|^2$	$\sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t\|^2$
		$\searrow$	$\swarrow$
		ALS	
STAGE 2	$L(\mathbf{A},\mathbf{B},\mathbf{C})$	$\sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^t\|^2$	

3 Experimental design

In order to investigate the degree of inadequacy of standard ALS in case of over-factoring a simulation study was implemented. ALS and its variants INT and INT-2 are tested under different conditions. The design includes multiple levels of noise contamination, both homescedastic and heterscedastic, and varied levels of congruence among factors. In addition, arrays of increasing dimensions are considered in order to evaluate the fall in performance of ALS as the dimensionality of the data increases.

It was found that for smaller data sets the disadvantage of ALS when over-factoring is mainly limited to a convergence slow down. This means that generally the correct solution is identified, however, it takes an enormous amount of iterations to converge.

Conversely, as the dimensions of the array increase, the algorithm becomes unable to identifying the correct solution rather than simply encountering temporary degeneracies. The multi-stage variants INT and INT-2, on the other hand, are capable of providing stable results in case of over-factoring together with an over-all increase if efficiency even in case of correct model selection.

References

1. 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)
2. 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)
3. Gallo, M., Simonacci, V., Di Palma, M.A.: An integrated algorithm for three-way compositional data. Quality & Quantity, pp.1-18 (2018)
4. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multi-modal factor analysis (1970)
5. Mitchell, B.C., Burdick, D.S.: Slowly converging parafac sequences: Swamps and two-factor degeneracies. Journal of Chemometrics 8(2), 155-168 (1994)
6. Nguyen, T. H., Şimşekli, U., Richard, G., Cemgil, A. T.: Efficient Bayesian Model Selection in PARAFAC via Stochastic Thermodynamic Integration. IEEE Signal Processing Letters, 25(5), 725-729 (2018)
7. ten Berge, J.M., Sidiropoulos, N.D.: On uniqueness in CANDECOMP/PARAFAC. Psychometrika, 67(3), 399-409. Chicago (2002)
8. Timmerman, M.E., Kiers, H.A.: Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima. British journal of mathematical and statistical psychology, 53(1), 1-16 (2000)
9. 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)
10. 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)
11. Zijlstra, B.J., Kiers, H.A.: Degenerate solutions obtained from several variants of factor analysis. Journal of Chemometrics, 16(11), 596-605 (2002)