A novel estimation procedure for robust CP model fitting

Perfezionamento della procedura per la stima robusta del modello CP

Valentin Todorov and Violetta Simonacci and Michele Gallo and Nikolay Trendafilov

Sommario The usual way of parameter estimation in CANDECOM/PARAFAC (CP) is an alternating least squares (ALS) procedure that yields least-squares solutions and provides consistent outcomes but at the same time has several deficiencies, like sensitivity to the presence of outliers in the data, slow convergence, and susceptibility to degeneracy conditions. A number of works have addressed these weaknesses, but to our knowledge, there is no outlier-robust procedure that is highly computationally efficient at the same time, especially for large data sets. We propose a robust procedure based on an integrated estimation algorithm, alternative to ALS, which guards against outliers and is computationally efficient at the same time.

Sommario *Il metodo comunemente usato per la stima dei parametri nel modello CP è una procedura dei minimi quadrati alternati (ALS). Questo algoritmo fornisce soluzioni ai minimi quadrati e produce risultati stabili ma, allo stesso tempo, registra diverse carenze come la sensibilità alla presenza di valori anomali nei dati, convergenza lenta, e suscettibilità a condizioni di degenerazione. Numerosi lavori hanno affrontato queste debolezze ma, per quanto ne sappiamo, non esiste una procedura robusta che sia allo stesso tempo altamente efficiente dal punto di vista computazionale, specialmente per grandi insiemi di dati. Proponiamo una procedura robusta basata su un algoritmo di stima integrato, alternativo all'ALS, che protegge dai valori anomali ed è computazionalmente efficiente allo stesso tempo.*

Valentin Todorov
United Nations Industrial Development Organization (UNIDO), VIC, Vienna, e-mail: valentin@todorov.at

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

Michele Gallo
University of Naples-L'Orientale, Naples, 80134, Italy e-mail: mgallo@unior.it

Nikolay Trendafilov
University of Naples-L'Orientale, Naples, 80134, Italy e-mail: ntrendafilov@unior.it

Key words: ALS, ATLD-ALS, robustness, outliers, computational efficiency

1 Introduction

The standard multivariate analysis addresses data sets represented as two-dimensional matrices. In recent years, an increasing number of application areas like chemometrics, computer vision, econometrics and social network analysis involve analysis of data sets that are represented as multidimensional arrays and multiway data analysis becomes popular as an exploratory analysis tool. Different techniques exist to analyze such multi-way data but CANDECOMP/PARAFAC (CP) is one of the most popular. The usual way of parameter estimation in CP is an alternating least squares (ALS) procedure which yields least-squares solutions and provides consistent outcomes. Together with these desirable features, the ALS procedure suffers several major flaws which might be particularly problematic for large-scale problems: slow convergence and sensitiveness to degeneracy conditions such as over-factoring, collinearity, bad initialization and local minima. Furthermore, it is well-known that algorithms which rely on least squares easily break down in the presence of outliers. The issue of non-robustness of the ALS procedure was addressed by [2] and software is available in the R package **rrcov3way**. The other issues were addressed in a number of works proposing algorithms more efficient than ALS. However, often these do not provide stable results because the increased speed might come at the expense of accuracy. An integrated algorithm was proposed in [4] which seems to combine improved speed and stability. The purpose of this work is to develop further this algorithm by adding capabilities for dealing with outliers present in the data.

2 The CP model, the ALS algorithm and its robust version

The CP model [1, 3] decomposes the 3-way data array $\underline{\mathbf{X}}(I \times J \times K)$ with a generic element x_{ijk} into three loading matrices $\mathbf{A}$ $(I \times R)$, $\mathbf{B}$ $(J \times R)$, $\mathbf{C}$ $(K \times R)$ with R components (using the same number for each mode). The CP model can be written formally as

$$\mathbf{X}_A = \mathbf{A}(\mathbf{C} \otimes \mathbf{B})^\top + \mathbf{E}_A, \tag{1}$$

where $\mathbf{X}_A$ and $\mathbf{E}_A$ are are the original array and the error array unfolded with respect to mode A and the symbol $\otimes$ represents the *Kronecker product* between two matrices. To estimate the optimal component matrices the residual sum of squares

$$||\mathbf{E}_A||^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2 = \sum_{i=1}^{I}||\mathbf{x}_i - \hat{\mathbf{x}}_i||^2 = \sum_{i=1}^{I} RD_i^2 \tag{2}$$

is minimized. The residual distance (RD) for observation i is thus given by

$$RD_i = ||\mathbf{x}_i - \hat{\mathbf{x}}_i|| = \sqrt{\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \hat{x}_{ijk})^2} \tag{3}$$

and the estimation is equivalent to the minimization of the sum of the squared distances. With ALS the component matrices are estimated one at a time, keeping the estimates of the other component matrices fixed, i.e. we start with initial estimates of **B** and **C** and find an estimate for **A** conditional on **B** and **C** by minimizing the objective function. Estimates for **B** and **C** are found analogously. The iteration continues until the relative change in the model fit is smaller than a predefined constant.

The idea of a robust version of CP is to identify enough ”good“ observations and to perform the classical ALS on these observations. This is repeated until no significant change is observed. Finally, a reweighting step is carried out to improve the efficiency of the estimators. In order to identify the "good" observations a robust version of principal component analysis on the unfolded array is used. We will call this procedure R-ALS in the rest of the paper. It is obvious that the robust procedure will be much more time consuming than the classical one, repeating many times the ALS optimization. Therefore, any improvement of the performance of the parameter estimation procedure will contribute to the improvement of the performance of the complete robust procedure.

3 The ATLD and INT2 algorithms

The alternating trilinear decomposition (ATLD) proposed by [6] seems to be the most efficient method among the proposed alternatives to ALS. It is based on the use of three loss functions with different response surfaces. However, these advantages are obtained at the cost of unstable results and non-least-squares solutions. To cope with these issues [4] proposed a multi-optimization procedure in which ATLD is followed by ALS estimation steps which is demonstrated to be quite effective. The robust procedure R-ALS described in Section 2 is entirely based on ALS and thus suffers the slow convergence and other disadvantages of this algorithm. We propose to replace ALS by INT2 thus obtaining a new robust estimation procedure which we will call R-INT2. As before it starts by robust principal components to identify any outlying points and then iterates using the INT2 algorithm until no significant change is observed. After convergence a reweighting step with INT2 is conducted which produces the final solution.

4 Simulation study

The performance of the newly proposed algorithm R-INT2 for robust estimation of trilinear CP models will be demonstrated in a brief simulation study comparing

classical CP, R-ASL and R-INT2. First of, all we want to verify that R-INT2 works well on data sets with and without contamination by identifying the outliers at least as good as R-ALS retrieving solutions with good statistical quality. At the same time we want to verify that the convergence is improved significantly and thus the computational time is reduced.

These two aspects will be illustrated on three-way data generated as in [4], [5] and [2]. The three-way arrays have $I = 50$ observations, $J = 100$ variables and $K = 10$ occasions and the number of factors is $R = 2$. The loadings matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are generated as randomly multivariate normal distributed $N_R(\mathbf{0}, \boldsymbol{\Sigma}_R)$ where $\boldsymbol{\Sigma}_R$ is a diagonal matrix with (10, 2) on the diagonal.

Such data does not contain any contamination to start. Different configurations of outliers can then be considered. In this work, because of the space constraints, only the so-called bad leverage points will be generated. An example of such outliers is shown in the outlier map illustrated in the left panel of Fig. 1 - these are the observations lying in the upper right quadrant. Regular observations lie in the lower left quadrant. Other types of outliers, not exemplified here, include good leverage points (lower right quadrant) and residual outliers (upper left quadrant). For more details about the generation of the different types of outliers see [2, page 158]. In the simulation study, apart from the case with no contamination, two other cases will be studied - with 10% and 20% of bad leverage points. For each setup, in order to account for minor statistical fluctuations 20 replicates were computed yielding the following three quantities which will represent how close the obtained estimates are to the original data: the mean square error (MSE), the angle between the estimated subspace and the true subspace spanned by the B-loadings and analogously, for the C-loadings. Also, the computation time and the number of iterations were recorded. The mean squared error (MSE) is given by

$$MSE = \frac{1}{w} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} w_i (x_{ijk} - \hat{x}_{ijk})^2 \tag{4}$$

with $w = \sum_{i=1}^{I} w_i$ and $w_i = 0$ if the i-th observation is outlier or $w_i = 1$ otherwise. Thus the MSE will be computed only for the regular observations. The angle between the estimated subspace and the original one is given by

$$maxsub = \max_{\mathbf{b}_1} \min_{\mathbf{b}_2} \arccos(\mathbf{b}_1^\top \mathbf{b}_2) \tag{5}$$

This subspace angle has to be as small as possible and is reported in radians. We use the function `subspace()` from the **R** package **pracma** to compute *maxsub*.

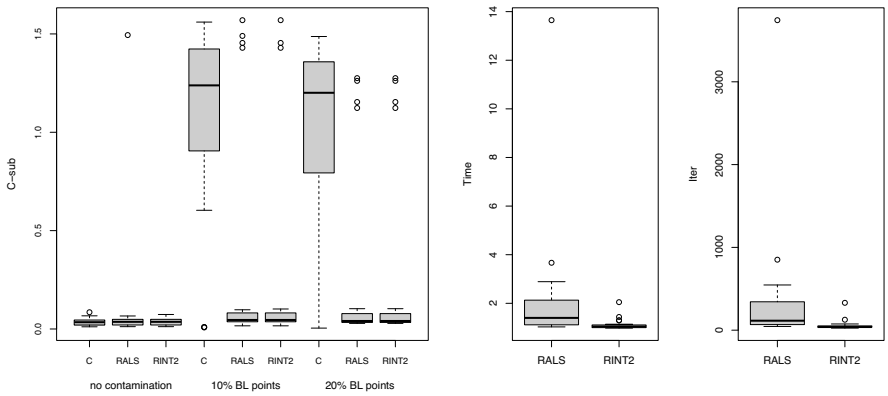

Figura 2 Angle of C-loadings (left) of classical CP (C), robust CP with ALS (R-ALS) and robust CP with INT2 (R-INT2) for simulation settings with no outliers, 10% and 20% bad leverage points and computational time and number of iterations (right) for 20% bad leverage points.

All three estimators perform equally well on clean data both in terms of MSE and maxsub (see Fig. 1, right panel and Fig. 2, left panel). However, when outliers are added to the data (10% and 20%) the classical CP is influenced - the MSE increases and the quality of the fit of the loadings decreases. These effects are even more pronounced when the outlier fraction is increased to 20%. There is no much difference in the performance of the two robust methods in terms of MSE and maxsub, howe-

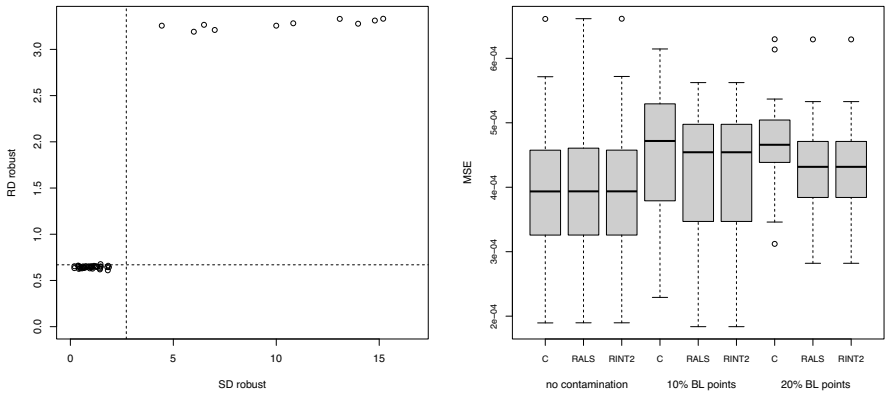

Figura 1 Example of an outlier map for simulated data with 20 percent bad leverage points (left) and MSE values of classical CP (C), robust CP with ALS (R-ALS) and robust CP with INT2 (R-INT2) for simulation settings without contamination and with bad leverage points (right).

ver, if we look at the right panel of Fig. 2 which presents their performance in terms of computational time and number of iterations the gain in performance is obvious. The median time of R-INT2 is more than 30% lower than that of R-ALS which also has much higher variance. This is due to the reduced number of iterations, as it is seen in the right part of the same Figure.

5 Summary and conclusions

We combine the robust procedure for CP proposed by [2] with the highly efficient estimation algorithm INT2 proposed by [4] in order to obtain a fast estimation and robust to outliers CP modeling technique. The conducted simulation study demonstrates the advantages of the new procedure in terms of computational time and at the same time shows that the robustness properties and the statistical efficiency have not been affected. Future work should bring a thorough investigation of the properties of the algorithm, comparison to the many existing fast alternatives and studying the possibilities for combination with other computational algorithms.

References

[1] Carroll J, Chang J (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. Psychometrica 35(3):283–319

[2] Engelen S, Hubert M (2011) Detecting outlying samples in a parallel factor analysis model. Analytica Chemica Acta 705:155–165

[3] Harshman RA (1970) Foundations of the PARAFAC procedure: Models and conditions for an ”explanatory“ multi-modal factor analysis. Tech. Rep. 10, UCLA

[4] Simonacci V, Gallo M (2020) An ATLD—ALS method for the trilinear decomposition of large third-order tensors. Soft Computing 18

[5] Tomasi G, Bro R (2006) A comparison of algorithms for fitting the PARAFAC model. Computational Statistics & Data Analysis 50(7):1700–1734

[6] Wu HL, Shibukawa M, Oguma K (1998) An alternating trilinear decomposition algorithm with application to calibration of HPLC-DAD for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. Journal of Chemometrics 12