SIS | 2023
SIS Società Italiana di Statistica
Statistical LEArning, Sustainability and Impact EvaluatioN
SEAS IN
UNIVERSITÀ POLITECNICA DELLE MARCHE
June 21-23, 2023
Ancona (Italy)

Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica

UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali DISES

LIUC Università Cattaneo

LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB

Pearson

Comparing three robust procedures for CANDECOMP/PARAFAC estimation

Valentin Todorov[a], Violetta Simonacci[b], Michele Gallo[c], and Nikolay Trendafilov[c]

[a]UNIDO, Vienna, Austria; valentin@todorov.at
[b]University of Naples Federico II, Naples, Italy; violetta.simonacci@unina.it
[c]University of Naples-L'Orientale, Naples, Italy; mgallo@unior.it, ntrendafilov@unior.it

Abstract

CANDECOMP/PARAFAC aims to identify the true components underlying data with a trilinear configuration. The search for a unique solution is not always an easy task, as degeneracies may occur. The presence of outlier contamination further complicates the matter by requiring the implementation of robust procedures. The most used robust approach R-ALS is based on the iterative repetition of the standard alternating least squares algorithm, which is known to be slow and vulnerable to over-factoring, collinearity, and bad initial values. Here the faster and stable robust alternative R-INT1, based on the SWATLD-ALS integrated scheme INT-1, is implemented. Its performance is tested against ALS, R-ALS, and R-INT2 (built on INT-2, an ATLD-ALS procedure already proposed in the literature). Performance is assessed in a simulation study with varied levels of outlier contamination.

Keywords: ALS, ATLD-ALS, SWATLD-ALS, outliers, computational efficiency

1. Introduction

Three-way data sets collect observations on a set of variables measured over several occasions (locations, times, conditions) and are represented as three-dimensional arrays rather than data matrices, as it is usual in the multivariate data analysis. Different techniques exist to analyze such three-way data but CANDECOMP/PARAFAC (CP), which can be seen as a generalization of principal component analysis (PCA) to higher-order tensors is one of the most popular. The idea of CP is to find a given number of components that jointly represent the data well.

The usual way of parameter estimation in CP is an alternating least squares (ALS) procedure which yields least-squares solutions and provides consistent outcomes. It is well-known that algorithms which rely on least squares easily break down in the presence of outliers. This is well recognized for PCA and a number of robust alternatives were proposed in the literature. In the multivariate case where two-way data are analyzed, the outliers are assumed to be rows (observations, objects, subjects, etc.) in the data set which lie significantly far from the other observations. Similarly, in the three-way case, we can assume that outliers are matrices (slices) that have a profile strongly deviating from the rest. (3) have demonstrated the influence of outlying samples on the classical ALS. They split the observations into four groups: regular observations, good leverage points, bad leverage points and residual outliers constructing a plot, the outlier map, similar to the one in robust regression or in robust PCA on two-way data. To cope with the presence of outlying samples they have also proposed a robust version of the ALS algorithm, R-ALS, which relies on the robust PCA method ROBPCA (5).

Apart from its proneness to outlying samples, the standard ALS-PARAFAC 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. A lot of research was invested to find a solution to these problems and a number of improved versions of ALS were created. Several alternatives to ALS were developed, like the alternating trilinear decomposition (ATLD) (10), self-weighted trilinear decomposition (SWATLD) (2) and their properties and comparative performances have been studied in several works, e.g. (9). These alternative procedures resist temporary degeneracies and over-factoring problems while ensuring a much speedier estimation process than ALS, thanks to a steeper convergence curve. These advantages are obtained at the cost of losing the stability of results and obtaining non-least-squares solutions. As a possible fix, an algorithm integration strategy was introduced in (6; 7) to combine the benefits of faster procedures with ALS stability.

The robust version of ALS proposed by (3) also shares these issues. In this approach, standard ALS is iteratively executed; thus, the effect of these disadvantages will be multiplied causing the whole procedure to become very slow, especially in the case of large data sets. The ATLD-based procedure INT2 (7) was extended to a robust version R-INT2 by (8) and its superior performance was demonstrated in an extensive simulation study and experimental data. Following these lines, in this paper we propose to robustify also the INT1 procedure based on SWATLD and call it R-INT1. We study its performance in the case of data without contamination or with increasing levels of contamination and compare to the other two known robust procedures.

2. The CP model, the ALS algorithm and its integrated versions

The CP model (1; 4) decomposes the 3-way data array $\underline{\boldsymbol{X}}(I \times J \times K)$ with a generic element x_{ijk} into the three loading matrices $\boldsymbol{A}$ $(I \times R)$, $\boldsymbol{B}$ $(J \times R)$, $\boldsymbol{C}$ $(K \times R)$ with R components (using the same number for each mode). The CP model can be written formally as

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

where $\boldsymbol{X}_A$ and $\boldsymbol{E}_A$ 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

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

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

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

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 $\boldsymbol{B}$ and $\boldsymbol{C}$ and find an estimate for $\boldsymbol{A}$ conditional on $\boldsymbol{B}$ and $\boldsymbol{C}$ by minimizing the objective function. Estimates for $\boldsymbol{B}$ and $\boldsymbol{C}$ are found analogously. The iteration continues until the relative change in the model fit is smaller than a predefined constant.

As already mentioned in the introduction, while ALS is still the algorithm of choice because of its many desirable characteristics, its use can be problematic especially in cases of large data sets due to slow convergence and sensitiveness to degeneracy conditions such as over-factoring, collinearity, bad initialization and local minima. The alternating trilinear decomposition (ATLD) proposed by (10) seems to be the most efficient method among the proposed alternatives to ALS and it is claimed to be less sensitive than ALS to over-factoring. It is based on the use of three loss functions with different response surfaces. SWATLD (2) does not attempt to find the minimum of (2), instead, it alternates between minimizing three different (non-least squares) loss functions, one per each of the loading matrices. It is important that the three matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ must have full column rank in order for the algorithm to resolve uniquely the components of interest. SWATLD seems to be the best algorithm in terms of recovery capability (factor congruence) (9).

However, these advantages of ATLD and SWATLD are obtained at the cost of unstable results and non-least squares solutions. To cope with these disadvantages a recent research development demonstrated that an integrated approach, combining algorithms with complementary points of strength, could provide a suitable solution. Two integrated algorithms INT-1 (6) and INT-2 (7) were proposed which combine SWATLD and ATLD steps with ALS, respectively, to ensure faster convergence, stability, and insensitivity to wrong model specification. For both integrated procedures the authors demonstrated the gain in performance in terms of computational efficiency and resistance to different undesirable effects. However, it is not known which of them is to be preferred in different situations since no comparison between them was conducted.

3. The robust alternatives to ALS

The idea of a robust version of CP proposed by (3) 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. To identify the “good” observations a robust version of PCA, e.g. ROBPCA (5), is used on the unfolded array. 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 parameter estimation procedure will contribute to the improvement of the performance of the complete robust procedure. R-ALS is entirely based on ALS and thus suffers the slow convergence and other disadvantages of this algorithm. (8) proposed to replace ALS by INT2 thus obtaining a new robust estimation procedure which they called R-INT2. As in R-ALS, it starts with 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.

Since the integrated procedure (6) based on SWATLD was demonstrated to have also very good performance as an alternative to ALS, we suggest extending it to a robust version in the same way as it was done for R-INT2 in (8). To verify that it can cope with outlying samples in the data and outperform R-ALS we conduct a simulation survey which is presented in Section 4. The purpose of this simulation study is also to find out which of the two integrated robust procedures performs better.

4. Simulation study

We will study the performance of the newly proposed procedure R-INT1 for robust estimation of trilinear CP models and will compare it to the classical ALS and the other two known robust procedures R-ALS and R-INT2 on a detailed simulation platform. Similarly as with the other integrated procedures the performance of the two-stage procedure R-INT1 will depend significantly on the transition parameters for switching from the initialization stage to the refinement stage. For this reason, the preliminary part of this simulation study is dedicated to the empirical estimation of these parameters, but due to the space limitation we will only state the final result - the values 10^{-2} and 10^{-3} seem to be most favorable and 10^{-2} will be used in all further computations. Successively we will compare the classical CP, the robust version based on ALS as proposed by Engelen and Hubert (3) R-ALS, the integrated robust version proposed by Todorov et al (8) R-INT2 and the newly proposed integrated procedure based on SWATLD R-INT1. First of all, we want to verify that R-INT1 and R-INT2 work well on data sets with and without contamination by identifying the outliers at least as well as R-ALS, retrieving solutions with good statistical quality. At the same time, we want to verify that the convergence of these two procedures is improved significantly and thus the computational time is reduced. And finally, we want to compare the computational performance of R-INT1 and R-INT2 in different scenarios.

These aspects will be illustrated on three-way data generated as in (7), (9), (3) and (8). The three-way arrays have $I = 50$ observations, $J = 100$ variables and $K = 10$ occasions and the number of factors is $R = 3$ or $R = 5$. For each data set random matrices $\boldsymbol{A} \in \mathbb{R}^{I\times R}$, $\boldsymbol{B} \in \mathbb{R}^{J\times R}$ and $\boldsymbol{C} \in \mathbb{R}^{K\times R}$ are generated from a uniform distribution. With the so defined loadings matrices a three-way data set can be constructed according to Equation 1 with different levels of homoscedastic (HO) and heteroscedastic (HE) noise $\boldsymbol{E}_A^{HO}$ and $\boldsymbol{E}_A^{HE}$ respectively defined as in (9).

Different types of outliers were then added following the scheme proposed by (3) and used by (8). First of all, we want to study the behavior of the four procedures on clean data and thus, the first setup is created by not including any outliers. In the other configurations 10% and 20% of the observations are modified to contain *bad leverage points*. For each setup, in order to account for minor statistical fluctuations 100 replicates were conducted. For evaluating efficiency performance CPU time, iterations, and incidence of temporary degeneracies (swamps) are considered. For accuracy, the computed diagnostics include the value of the objective function (FIT), the occurrence of fault recoveries (FR), the mean square error (MSE), and the angle between the estimated and original subspaces of the second and third mode. See (9) and (8) for a full description.

Overall, at 20% contamination with bad leverage points, the difference between R-ALS FIT and R-INT1 FIT is higher than $1e^{-4}$ in less than 1% of the cases. In more than half of the cases (55.9%) the fit of R-INT1 is better than that of R-ALS which demonstrates that R-INT1 is capable of identifying the best low rank approximation as well as R-ALS. The results for R-INT2 are similar, as shown in (8).

It is important to verify that the known instabilities of ATLD and SWATLD are not passed to the integrated procedures. We can check this by looking at the percentage of fault recoveries reported for all three algorithms for different levels of contamination and different ranks in Table 1. For $R = 3$, both in the correct rank estimation case and when over-factoring, the percentage of fault recoveries for all robust methods is not higher than 1%. Only for the classical estimates on data with 10% and 20% contamination the percentage increases drastically, coming close to 100%. For R = 5 all percentages are slightly higher (not shown here).

		$F = R = 3$				$F = R + 1 = 4$			
		C	RALS	RINT1	RINT2	C	RALS	RINT1	RINT2
	0%	0.0	0.0	0.0	0.1	1.2	1.1	0.0	0.0
FR	10%	98.9	0.0	0.0	0.1	0.0	1.0	0.0	0.0
	20%	99.4	0.0	0.2	0.1	0.0	0.9	0.0	0.0
	0%	0	0	0	0	22	18	0	0
SWAMPS	10%	0	0	0	0	1	25	0	1
	20%	0	0	0	0	2	15	0	0

Table 1: Total percentages of FR and number of swamps (out of 4500 repetitions) by rank and number of factors for different levels of contamination with bad leverage points.

The next measure of accuracy to look at is the MSE. The results for the four estimators and three types of data (no outliers, 10% bad leverage points and 20% bad leverage points) are presented in the box plots in the left panel of Figure 1. The right panel of the same figure presents the angles of the B-loadings for the four methods and the three contamination types.

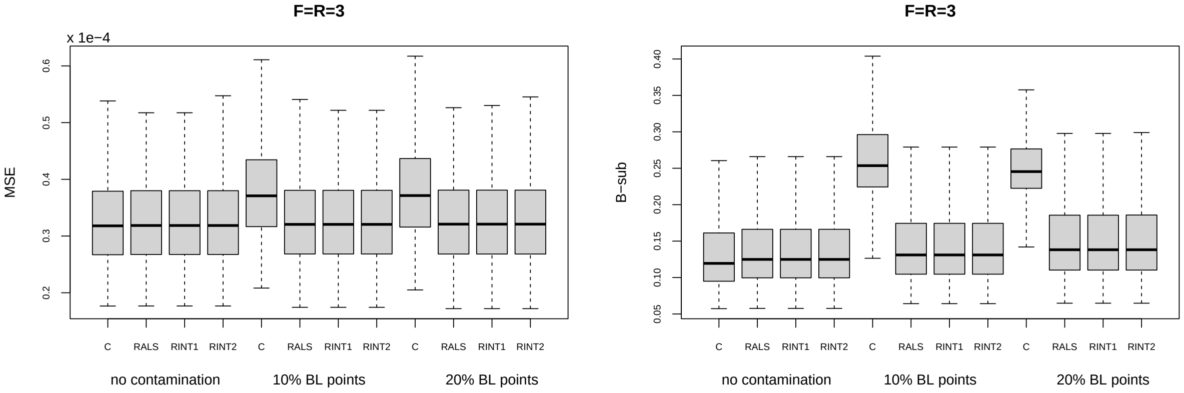

Figure 1: MSE values and angle of B-loadings of classical CP (C), robust CP with ALS (R-ALS), robust CP with INT1 (R-INT1) and robust CP with INT2 (R-INT2) on data sets without contamination, and with 10% and 20% bad leverage points respectively.

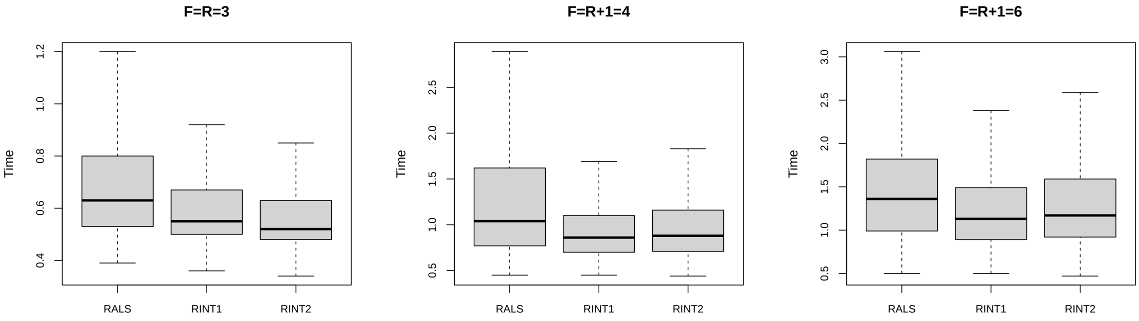

Figure 2: CPU time in seconds, robust CP with ALS (R-ALS), robust CP with INT1 (R-INT1) and robust CP with INT2 (R-INT2) on data sets 20% bad leverage points for different number of factors.

All four estimators perform equally well on clean data both in terms of MSE and maxsub, 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. There is non much difference in the performance of the three robust methods in terms of MSE and maxsub. However, if we look at Fig. 2 which presents their performance in terms of computational time the gain in performance in the two integrated procedures is obvious. The integrated procedures R-INT1 and R-INT2 perform better than R-ALS both in terms of median time and variance, with R-INT2 being slightly better in the case of correct factor decomposition ($F = R = 3$, left panel). In over-factoring ($F = R + 1 = 4$ and $F = R + 1 = 6$, middle and right panels) the roles change and R-INT1 becomes better.

The computational efficiency can also be judged by counting the number of swamps, i.e. the temporary degeneracies which continue for more than 10 iterations and thus slowdown the procedure. This problem was not significantly manifested in our simulation. As seen in the lower part of Table 1, no swamp cases are observed when estimating the correct rank and when $R = 3$, for none of the estimators and for none of the contamination levels. Only several cases were observed when over-factoring, for the classical ALS and the robust R-ALS, however the number of these cases is insignificant when compared to the total number of 4500 repetitions.

5. Summary and conclusions

The simulation study shows that R-INT1 is a viable alternative to R-ALS for dealing with outliers in a three-way setting. Throughout scenarios, it is stable in converging to the correct parameters and does not appear to model excessive noise, reaching least squares solutions. The integrated strategy ensures higher efficiency and also proves useful in terms of accuracy when over-factoring. As far as the differences between R-INT1 and R-INT2 go, they are both solid procedures. R-INT2 is slightly more efficient in terms of median values, however, in presence of over-factoring, R-INT1 is the top performer. A thorough presentation of the resulting diagnostics, as well as an analysis carried out on an application will be provided to further strengthen the comparison and complete this work. Future work should also study the possibilities for combination with other computational algorithms. The behavior of the algorithms if collinearity is present will be of great interest as well as their extension with additional constraints.

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] Chen ZP, Wu HL, Jiang JH, Li Y and Yu RQ (2000) A novel trilinear decomposition algorithm for second-order linear calibration. Chemometrics and Intelligent Laboratory Systems 52 75–86.

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

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

[5] Hubert M, Rousseeuw PJ and Vanden Branden K (2011) ROBPCA: A new approach to robust principal component analysis. Technometrics 47 64–79

[6] Simonacci V, Gallo M (2019) Improving PARAFAC-ALS estimates with a double optimization procedure. Chemometrics and Intelligent Laboratory Systems 192:103822.

[7] Simonacci V, Gallo M (2020) An ATLD—ALS method for the trilinear decomposition of large third-order tensors. Soft Computing 24:13535–13546

[8] Todorov V, Simonacci V, Gallo M and Trendafilov N (2023) A novel estimation procedure for robust CANDECOMP/PARAFAC model fitting. Submitted for publication.

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

[10] 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 1–26