Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

A novel estimation procedure for robust CANDECOMP/PARAFAC model fitting

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

[a] United Nations Industrial Development Organization (UNIDO), Vienna, Austria
[b] University of Naples Federico II, Naples, Italy
[c] University of Naples-L'Orientale, Naples, Italy

ARTICLE INFO

Article history:
Received 20 September 2022
Revised 7 July 2023
Accepted 8 July 2023
Available online xxx

Keywords:
CP
PARAFAC
ALS
ATLD-ALS
Robustness
Outliers
Computational efficiency

ABSTRACT

The parameter estimation in CANDECOMP/PARAFAC (CP) is carried out by alternating least squares (ALS) that yields least-squares solutions and provides consistent outcomes. At the same time it has several drawbacks, like sensitivity to the presence of outliers in the data, issues with the computational efficiency in terms of processing time and memory requirements, as well as susceptibility to degeneracy conditions. These weaknesses have been addressed, but there is no outlier-robust procedure that at the same time is highly computationally efficient, especially for large data sets. A novel procedure based on an integrated estimation algorithm is proposed. This is an alternative to ALS, which guards against outliers and is computationally efficient at the same time. The performance of the new method is demonstrated on an extensive simulation study and an empirical example.

1. Introduction

The CANDECOMP/PARAFAC (CP) model is an extension of PCA to higher-order data and decomposes the systematic part of a three-way array into the sum of trilinear factors (Carroll and Chang, 1970; Harshman, 1970). The model fitting aims to find a unique solution where the extracted components represent the same constructs throughout modes (Cattell, 1944).[1] By imposing component uniqueness, the CP model yields an easily interpretable latent structure, which can be also assessed with a confirmatory outlook. One of the main limitations of the CP model is the estimation process of trilinear parameters. Computations are generally carried out with the alternating least squares (ALS) procedure (Smilde et al., 2004). This algorithm is selected thanks to its in-built capability of minimizing the modeled noise and the well-defined properties of its objective function. The main issue with ALS estimation is its computational efficiency. Processing time and memory requirements may become prohibitive for large data. Additionally, degenerate solutions can occur, especially under challenging data conditions such as bad initialization, factor collinearity, and over-factoring (Yu et al., 2011b). A lot of research was invested to find a solution to this problem, extensive literature on this topic was published, and a number of improved versions of ALS were created. Several alternatives to ALS were developed, like the alternating trilinear decomposition (ATLD) (Wu et al., 1998), self-weighted trilinear decomposition (SWATLD) (Chen et al., 2000) and their properties and compara-

[*] Corresponding author.
E-mail address: valentin@todorov.at (V. Todorov).

[1] The analyses were carried out using R and the `rrcov3way` CRAN package, the code and replication material are available at https://www.github.com/valentint/robust-parafac-ecosta.

https://doi.org/10.1016/j.ecosta.2023.07.001
2452-3062/© 2023 EcoSta Econometrics and Statistics. Published by Elsevier B.V. All rights reserved.

tive performances have been studied in several works (Faber et al., 2003; Tomasi and Bro, 2006; Yu et al., 2011b; 2011a). These alternative procedures are resilient to temporary degeneracies and over-factoring problems. As a result they provide a steeper convergence curve and much faster estimation than ALS. However, the advantages are obtained at the cost of losing stability of results and obtaining non-least-squares solutions. As a possible remedy, an integrated algorithm strategy was introduced in (Simonacci and Gallo, 2019; 2020) to combine the benefits of faster procedures with ALS stability. In detail, the two integrated procedures INT and INT-2 were proposed, which initialize ALS estimation with SWATLD and ATLD, respectively.

Not less important is another disadvantage of the classical ALS algorithm: it is susceptible to the presence of outliers in the data and will break down in the same way as the classical least squares regression problem on which the ALS solution is based. The breakdown value (Donoho and Huber, 1983) of the linear least squares regression is 0, see for example Rousseeuw (1984, 1997) which means that a single outlier may be sufficient to destroy the estimates. The problem is well recognized in multivariate statistics and decades of research have yielded a number of robust techniques which can cope with outliers. There exist robust methods for computing the multivariate location and scatter which are the basis of many other multivariate techniques. The most popular such method is the Minimum Covariance Determinant (MCD) estimator proposed by Rousseeuw (1984) for which an efficient computational algorithm, FastMCD, was developed by Rousseeuw and Van Driessen (1999). Although the MCD estimator is limited to situations where the number of objects is much larger than the number of variables, recently it was extended to a regularized version (Boudt et al., 2020), MRCD, which demonstrates excellent performance even when applied to high-dimensional data sets. The C-step which is the core idea of the FastMCD algorithm was generalized to many other multivariate procedures, including the one which will be discussed in the present paper. An overview of existing estimators of multivariate location and scatter as well as the application of such estimators in other multivariate techniques is given in Hubert et al. (2017). Most of these methods are implemented in the **R** (R Core Team, 2022) package **rrcov** (Todorov, 2020).

Several attempts for dealing with outliers in the CP models are known in the literature, like Andersen and Bro (2003), Smilde et al. (2004), Riu and Bro (2003), see Kroonenberg (2008) for a review. However, the key breakthrough was done by Engelen and Hubert (2011) who proposed a sophisticated method based on initialization with robust principal components and then utilizing the standard ALS procedure. The method was shown to be effective in a simulation study, as well as on practical data examples. An extension for compositional data was developed by Di Palma et al. (2018) and different variants of these methods were implemented in the **R** (R Core Team, 2022) package **rrcov3way** Todorov et al. (2023). However, this procedure will suffer from all the computational disadvantages of the ALS approach mentioned above. Since ALS is executed many times iteratively this effect will be multiplied, and the whole procedure can become very slow, especially in case of large data sets. Here comes to rescue the novel proposal of an integrated estimating procedure which is both robust and computationally efficient. Based on INT-2, it will combine the good features and computational efficiency of the ATLD method with the stability of the ALS procedure, extended to handle data with outliers and provide robust solutions to the CP model.

The structure of the rest of the manuscript is as follows. In Section 2.1 the CP model and the standard ALS algorithm are revisited, then the computationally more efficient alternatives ATLD and INT-2 are discussed. Section 2.2 recalls the robust CP algorithm of Engelen and Hubert (2011) and introduces the new robust R-INT2 approach. In Section 3 the performance of the new procedure is demonstrated on a simulation study, while Section 4 illustrates the approach on an empirical example. Section 5 concludes with some remarks and outlines the further research on this topic.

2. Methodology

2.1. ALS and the faster alternatives

The CANDECOMP/PARAFAC model (see Carroll and Chang, 1970; Harshman, 1970) decomposes the 3-way data array $\underline{\boldsymbol{X}}(I \times J \times K)$ into three loading matrices $\boldsymbol{A}$ $(I \times F)$, $\boldsymbol{B}$ $(J \times F)$, $\boldsymbol{C}$ $(K \times F)$ with F components (using the same number of components for each mode). The CP model can be written formally as

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

where $\boldsymbol{X}_A$ and $\boldsymbol{E}_A$ are the original array and the error array matricized (rearranged into a two-way matrix by concatenating, for example, the horizontal slabs next to each other, (Kiers, 2000)) with respect to the mode A. The symbol $\odot$ represents the Khatri-Rao product of two matrices (Liu and Trenkler, 2008) that can be defined as the column-wise Kronecker product (see also Smilde et al., 2004, Section 2.3). The CP model can be considered a constraint version of the more general multilinear model Tucker3 (Tucker, 1966) with the same number of components for each mode $(P = Q = R = F)$, and no interaction between the components is allowed. It is suitable for data with a trilinear configuration which can be factorized into the sum of outer products of rank-1 tensors.

To estimate the optimal component matrices the residual sum of squares

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

Table 1
The standard ALS algorithm for the CP model.

	ALS algorithm
(1)	initialize $\boldsymbol{B}$ and $\boldsymbol{C}$
(2)	estimate $\hat{\boldsymbol{A}} = \boldsymbol{X}_A(\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}})((\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}})^\top(\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}}))^+$
(3)	estimate $\hat{\boldsymbol{B}} = \boldsymbol{X}_B(\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{A}})((\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{A}})^\top(\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{A}}))^+$
(4)	estimate $\hat{\boldsymbol{C}} = \boldsymbol{X}_C(\hat{\boldsymbol{A}} \odot \hat{\boldsymbol{B}})((\hat{\boldsymbol{A}} \odot \hat{\boldsymbol{B}})^\top(\hat{\boldsymbol{A}} \odot \hat{\boldsymbol{B}}))^+$
(5)	iterate until convergence

$\boldsymbol{Z}^+$ is the Moore-Penrose inverse of $\boldsymbol{Z}$
$\boldsymbol{X}_A^{I\times JK}$, $\boldsymbol{X}_B^{J\times IK}$ and $\boldsymbol{X}_C^{K\times IJ}$ are the unfoldings or matricizations (Kiers, 2000) of the three dimensional array $\underline{\boldsymbol{X}}$ in the A, B and C modes

is minimized. The residual distance 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. Three-way models are usually fitted by an iterative procedure based on alternating least squares (Smilde et al., 2004). The component matrices are estimated one at a time, keeping the estimates of the other component matrices fixed, i.e., 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. The standard ALS algorithm is summarized in Table 1.

Different types of degeneracy can occur when estimating the model with the ALS algorithm (Mitchell and Burdick, 1994). These can be due to bad initialization or the algorithm could be trapped in a local minimum. Another problem which is likely to cause degeneracies is the model misspecification, i.e. estimating more factors than the rank of the model. In some cases, a local minimum can cause a temporary degeneracy which slows down the algorithm—it might or might not get out of the local minimum but in any case too many iterations will be executed and even if the final solution is not affected the computational time will grow significantly. Mitchell and Burdick (1994) call these temporary degeneracies “swamps” - the objective function decreases very slowly and for large number of iterations little progress in fitting of a solution is achieved. While the degeneracies caused by bad initialization or issues with the local minima could be mitigated by repeated random runs (Harshman and Lundy, 1984; Simonacci and Gallo, 2019), no solution so far exists to cope with the occurrence of over-factoring degeneracies for ALS. One should either know in advance the correct number of factors or should use tools for estimating the number of factors. Such tools exist (Timmerman and Kiers, 2000; Ceulemans and Kiers, 2006), however they add complexity to the procedure and could increase significantly the necessary computational time.

The alternating trilinear decomposition (ATLD) is one of the fastest alternatives to ALS for fitting the CP model. It was proposed by Wu et al. (1998) specifically to overcome two of the main disadvantages of ALS: the slow convergence and the sensitiveness to over-factoring. The essential improvement in ATLD is that it uses three objective functions instead of one as in ALS which ensures different response surfaces and provides a steeper convergence rate. The resistance to over-factoring is explained by the differential properties of its objective functions which remain strictly convex even for $F > R$ where R is the number of factors of the true underlying trilinear solution. The algorithm performs exceptionally well for data that do not deviate significantly from perfect trilinearity as it maximizes the extraction of the diagonal data. However, this feature is also a disadvantage and causes instabilities because it will rarely yield a least squares solution. Perhaps this is why ATLD is infrequently utilized for data analysis. On the other hand, its convergence properties offer an ideal initialization tool. Simonacci and Gallo (2020) propose to use the ATLD algorithm as a first initialization step, not iterated to convergence, which is then followed by ALS refinement. This resulted in their INT-2 algorithm which showed very attractive performance in variety of simulated scenarios. In detail, ATLD steps are carried out until an interim convergence parameter set by the user is reached. Note that the choice of this interim convergence parameter may affect INT-2 final performance (see Simonacci and Gallo, 2020, Section 4.1). The INT-2 algorithm is summarized in Table 2.

2.2. Robust methods for CP and the robust INT-2 estimation procedure

It is well known that algorithms which rely on least squares break down in the presence of outliers (Rousseeuw and Leroy, 1987). This is the case of PCA, for which this problem was extensively studied (see Jolliffe, 2002, Chapter 10), (Devlin et al., 1981). A number of robust algorithms, resistant to outliers, were proposed (Huber and Ronchetti, 2009; Rousseeuw et al., 2006), see Ronchetti (Ronchetti, 2021) for a recent reference. For low dimensional data a robust covariance matrix can be utilized (see for example Croux and Haesbroeck, 2000) while for high dimensional data projection pursuit methods (Croux et al., 2007), combination of projection pursuit and robust covariance estimation (Hubert et al., 2005), spherical approach (Locantore et al., 1999) and estimators based on robust scale of residuals (Maronna, 2005) were proposed (see Hubert et al., 2008; Filzmoser and Todorov, 2013, for a review). Likewise, the ALS method is severely affected by anomalous

Table 2
The Integrated INT-2 algorithm for the CP model as proposed by (Simonacci and Gallo, 2020).

	Integrated INT-2 algorithm
(1)	initialize $\boldsymbol{B}$ and $\boldsymbol{C}$
(2)	estimate $\hat{\boldsymbol{a}}_i^\top = diag(\hat{\boldsymbol{B}}^+ \boldsymbol{X}_{i..} (\hat{\boldsymbol{C}}^\top)^+), i = 1, \ldots, I$
(3)	estimate $\hat{\boldsymbol{b}}_j^\top = diag(\hat{\boldsymbol{C}}^+ \boldsymbol{X}_{.j.} (\hat{\boldsymbol{A}}^\top)^+), j = 1, \ldots, J$
(4)	estimate $\hat{\boldsymbol{c}}_k^\top = diag(\hat{\boldsymbol{A}}^+ \boldsymbol{X}_{..k} (\hat{\boldsymbol{B}}^\top)^+), k = 1, \ldots, K$
(5)	iterate 2-4 until met interim convergence criteria
(6)	estimate $\hat{\hat{\boldsymbol{A}}}$, $\hat{\hat{\boldsymbol{B}}}$ and $\hat{\hat{\boldsymbol{C}}}$ using ALS (Table 1) starting from $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{C}}$
	$\boldsymbol{Z}^+$ is the Moore-Penrose inverse of $\boldsymbol{Z}$ $\hat{\boldsymbol{a}}_i$, $\hat{\boldsymbol{b}}_j$ and $\hat{\boldsymbol{c}}_k$ are the i-th, j-th and k-th row of the matrix $\hat{\boldsymbol{A}}$, $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{C}}$ respectively

Table 3
The Robust ALS algorithm for the CP model as proposed by (Engelen and Hubert, 2011).

	Robust ALS algorithm (R-ALS)
(1)	find initial $h-$subset by robust PCA (ROBPCA)
(2)	estimate $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{A}}_h$ with the standard ALS on the $h-$subset
(3)	estimate $\hat{\boldsymbol{A}} = \boldsymbol{X}_A((\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}})^+)^\top$
(4)	estimate $\hat{\boldsymbol{X}}_A = \hat{\boldsymbol{A}}(\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}})^\top$ and RD_i using Equation 3
(5)	create new $h-$subset from the samples with smallest RD_i
(6)	iterate (2)-(5) until convergence
(7)	perform a reweighting step, standard ALS on the samples with
	smallest RD_i $\boldsymbol{Z}^+$ is the Moore-Penrose inverse of $\boldsymbol{Z}$ $\boldsymbol{X}_A^{I \times JK}$ is the unfolding of $\underline{\boldsymbol{X}}$ in the A mode h is a number between $I/2$ and I, recommended is $0.75I$ Note that the robust initialization (1) is crucial for the selection of outlier-free h-subset in (5)

observations in the data and robust algorithms for three-way methods are also needed. A robust version of Tucker3 was proposed by Pravdova et al. (2001), later improved and adapted for CP by Engelen and Hubert (2011). This procedure is referred to as R-ALS henceforth. The idea of a robust version of CP is to identify enough “good” (say, h where $I/2 < h < I$) observations and then to perform the classical ALS procedure on those observations. Thereafter, results are used to compute the residual distances $RD_i, i = 1, \ldots, I$ for all observations according to Equation 3 and select as “good” observations those h observations that have the smallest residual distances. Note that the robust initialization at the beginning is crucial for obtaining an outlier-free sample on which to base the classical ALS and thus for the robustness of the procedure. Without this robust initialization the residual distances RD_i are no more a reliable measure for keeping possible outliers out of the selected sample which could affect the subsequent steps and lead to non-robust solutions. This procedure is repeated until no significant change is observed. It resembles and is inspired by the C-step in the Fast-MCD algorithm of Rousseeuw and Van Driessen (Rousseeuw and Van Driessen, 1999) for computing the MCD estimator. The convergence of the algorithm is guaranteed (Engelen and Hubert, 2011), but not to a global optimum. In order to initially identify h “good” observations, a robust version of principal component analysis, e.g., ROBPCA (Hubert et al., 2005; Todorov and Filzmoser, 2009) on the matricized array is used. The value of h is chosen to control the desired robustness of the solution. To obtain robustness close to 50% (i.e. the procedure can resist up to 50% outlying samples) $h \approx I/2$ can be chosen, but the recommended value of h is $0.75 \cdot I$ which is a good compromise between robustness and efficiency. This value of h will be used in the computations presented in Sections 3 and 4. Finally, a reweighting step is carried out to improve the efficiency of the estimates: samples are given weight zero or one if their residual distance is larger or smaller than a cutoff value and then the classical PARAFAC model is applied. The cutoff is computed as shown in Section 2.3, see also Engelen and Hubert (2011).

In order to label extreme points once the robust ALS-procedure is performed, two distances are computed: the robust residual distance calculated as in Equation 3 which indicates how well the fitted data correspond to the observations and the robust score distance, a Mahalanobis-type distance in the scores space given by Equation 4. These distances will be used for visualization of the fitted CP model in an outlier map, as described in Engelen and Hubert (2011). The robust ALS algorithm is presented schematically in Table 3.

Similarly, a robust procedure based on INT-2 can be constructed, see details in Table 4.

Table 4
The Robust integrated algorithm R-INT2 for the CP model.

	Robust INT-2 algorithm (R-INT2)
(1)	find initial $h-$subset by robust PCA (ROBPCA)
(2)	estimate $\hat{\boldsymbol{B}}$, $\hat{\boldsymbol{C}}$ and $\hat{\boldsymbol{A}}_h$ with INT-2 on the $h-$subset
(3)	estimate $\hat{\boldsymbol{A}} = \boldsymbol{X}_A((\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}})^+)^\top$
(4)	estimate $\hat{\boldsymbol{X}}_A = \hat{\boldsymbol{A}}(\hat{\boldsymbol{C}} \odot \hat{\boldsymbol{B}})^\top$ and RD_i using Equation 3
(5)	create new $h-$subset from the samples with smallest RD_i
(6)	iterate (2)-(5) until convergence
(7)	perform a reweighting step, INT-2 on the samples with smallest RD_i
	$\boldsymbol{Z}^+$ is the Moore-Penrose inverse of $\boldsymbol{Z}$ $\boldsymbol{X}_A^{I\times JK}$ is the unfolding of $\underline{\boldsymbol{X}}$ in the A mode h is a number between $I/2$ and I, recommended is $0.75I$ Note that the robust initialization (1) is crucial for the selection of outlier-free h-subset in (5)

2.3. Outliers in CP

Next, contamination and outliers are considered and the manner in which they appear in the context of the CP model is examined. Specifically, when analyzing multivariate data in a two-way scenario, it is assumed that the outliers consist of rows (observations, objects, subjects, etc.) within the data set that exhibit a significant deviation from the other observations. Similarly, in the three-way case, outliers can be regarded as matrices (slices) with profiles that exhibit a strong deviation from the remainder of the data. Following the literature (Pravdova et al., 2001; Riu and Bro, 2003; Engelen et al., 2007; Engelen and Hubert, 2011; Di Palma et al., 2018), outliers in the sample space (or in the first mode, or in mode A) are considered, and these are horizontal slices (Kroonenberg, 2008, Chapter 3.2). To investigate the presence of outliers across other dimensions, it is necessary to modify the configuration of the array in a manner that relocates the desired space to the sample space. This approach is akin to searching for outlying variables instead of outlying observations in a two-way problem. Recently a different type of outliers, called *cellwise* outliers, have received much attention in the literature on multivariate analysis (see (Hubert et al., 2019) for an extension of the robust PCA algorithm) which would be an important future development for the robustness approach in multi-way data analysis.

Following Engelen and Hubert (2011) the observations can be split into four groups: regular observations, good leverage points, bad leverage points and residual outliers. To do this a plot is constructed, the outlier map, similar to the one in robust regression or in robust PCA on two-way data (see Hubert et al., 2005). For this purpose two distances are used. The first one is the residual distance RD_i already defined in Equation 3. The second one, the score distance, a Mahalanobis-type distance in the scores space is defined as:

$$\mathrm{SD}_i = \sqrt{(\hat{\boldsymbol{a}}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{a}}_i - \hat{\boldsymbol{\mu}})}, \tag{4}$$

where $\hat{\boldsymbol{a}}_i$ is the i-th row (i-th score) of the matrix $\hat{\boldsymbol{A}}$, and $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are estimates of location and spread, respectively, of $\hat{\boldsymbol{A}}$. These estimates can be obtained by the Minimum Covariance Determinant (MCD) method (Rousseeuw, 1984). To flag observations as outliers the so defined distances are compared to the usual cutoff values from the two-way situation. The score distance is compared to the 97.5% quantile of the χ^2 distribution $\sqrt{\chi^2_{0.975,F}}$ with F degrees of freedom. The cutoff value for the residual distances is $(m + sz_{0.975})^{3/2}$ where m and s^2 are the mean and variance estimates of the residual distances and $z_{0.975} = \Phi^{-1}(0.975)$ is the 97.5% quantile of the standard normal distribution. The estimates m and s^2 are obtained using the univariate MCD. The four types of outliers are illustrated in the left panel of Figure 1. The regular observations have small residual and score distances while the residual outliers have deviating underlying structure (large residual distance) but small score distance. The good and bad leverage points have large score distance but the residual distance only for the bad leverage points exceeds the corresponding cutoff value. Section 3.1 describes how these different outlier types are generated in the simulated data. The right panel of Figure 1 shows the outlier map plot of the empirical data example presented in Section 4.

3. Simulation

The performance of the newly proposed procedure R-INT2 for robust estimation of trilinear CP models is investigated on a detailed simulation platform. As already mentioned in Section 2, the performance of the two-stage procedure R-INT2 will largely depend 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.

In this study, the performance of three methods is compared: the classical CP method, the robust version R-ALS based on ALS as proposed by Engelen and Hubert (2011) and the novel R-INT2. The initial objective is to validate that R-INT2 performs effectively on both contaminated and uncontaminated data sets by successfully identifying outliers, at least as well as R-ALS,

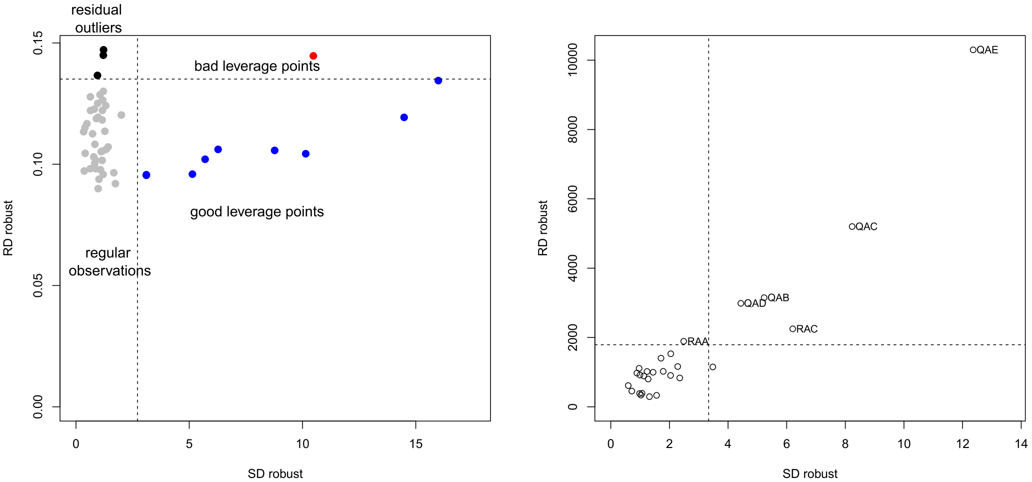

Fig. 1. The four types of outliers in CP (left pane) and the outlier map for the data of the empirical example discussed in Section 4

and producing solutions with good statistical quality. Simultaneously, we aim to confirm that R-INT2 exhibits significantly improved convergence, resulting in reduced computation time.

The general setup of the simulation is similar to the one proposed by Tomasi and Bro (2006), see also Simonacci and Gallo (2019, 2020); Engelen and Hubert (2011). To do a fair comparison all algorithms were initialized using SVD decomposition. This approach might have disadvantages since the ALS procedure is known to be affected by bad initial values, which could cause stopping prematurely at local minima. To mitigate this problem, one potential strategy is to utilize random starting values for the algorithms. By initiating the algorithms from multiple different starting points, the solution with the best fit can be identified or several runs can be found which yield the same solution. This approach was advocated by Harshman and Lundy (1984) and was used in a simulation study by Simonacci and Gallo (2019), however, in the case where the robust methods are by themselves extremely time consuming, multiplying the computational time by say, 10, might render the simulation not feasible.

3.1. Data generation

For the simulation study, artificial data sets of order $I\times J\times K$ were constructed, with known (three-dimensional) underlying CP structure. Here I, J and K are parameters representing the number of observations, variables and cases, respectively. In the determination of parameter values for I, J, K the decision was made to align with the existing literature on CP algorithm comparison (Tomasi and Bro, 2006; Simonacci and Gallo, 2020), where $J=K$ was used, in order to have general data without specific characteristics also in terms of dimensions. It is duly recognized that in the majority of existing applications, $J\gg K$. However, there is no reason to expect changes in the results, apart from the overall time, as was confirmed by the conducted limited simulations with $I=100, J=100, K=20$. Key results from these additional simulations are presented later in this section and the complete results are available in the Supplementary Material. The total number of real factors of the underlying trilinear model is denoted by R. 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 with uniformly distributed elements. Since collinearity in the data can affect the results, a predetermined level of factor collinearity is forced on these generated loadings matrices, controlled by the parameter *CONG*. To impose the desired level of factor collinearity the procedure starts by orthogonalizing them using QR decomposition. Then they are multiplied by an upper triangular matrix obtained by Cholesky decomposition of an $R\times R$ matrix with 1s on the diagonal and the desired value of collinearity between the loading vectors *CONG* elsewhere. In this way, all loading matrices generated with the same number of underlying factors R and equal factor collinearity *CONG* will have the same condition number (Kiers et al., 1999). Using these loadings matrices the pure three-way array $\tilde{\boldsymbol{X}}^{I\times JK}$ is generated as

$$\tilde{\boldsymbol{X}}^{I\times JK} = \boldsymbol{A}(\boldsymbol{C}\odot\boldsymbol{B})^\top \tag{5}$$

A given level of homoscedastic (HO) and heteroscedastic (HE) noise is added to it:

$$\boldsymbol{X}^{I\times JK} = \boldsymbol{A}(\boldsymbol{C}\odot\boldsymbol{B})^\top + \boldsymbol{E}_{HO}^{I\times JK} + \boldsymbol{E}_{HE}^{I\times JK} \tag{6}$$

The homoscedastic noise array is generated, as suggested by Tomasi and Bro (2006), from random normally distributed numbers $\tilde{\boldsymbol{E}}^{I\times JK}\sim N(\boldsymbol{0},\boldsymbol{I})$, scaled to a Frobenius norm of 1:

$$\boldsymbol{E}_{HO}^{I\times JK} = \sqrt{\frac{HO}{1-HO}}||\boldsymbol{X}^{I\times JK}||_F\tilde{\boldsymbol{E}}^{I\times JK} \tag{7}$$

where the term $||\boldsymbol{X}^{I\times JK}||_F$ normalizes the error to the Frobenius norm of the pure $\boldsymbol{X}^{I\times JK}$ and thus allows the noise level HO to reflect the percentage noise in the total variation of the data $\underline{\boldsymbol{X}}$. In a similar way the heteroscedastic noise $\boldsymbol{E}_{HE}^{I\times JK}$, with the desired level HE is generated, but in order to make it proportional each element of the noise array is multiplied by the corresponding element of $\underline{\boldsymbol{X}}$. Throughout the simulations three values of HO: 15%, 20% and 25% and three values of HE: 10%, 15% and 20% are used.

Different types of outliers are added to the generated data sets, following the scheme proposed by Engelen and Hubert (2011). The first objective of the study is to evaluate the performance of the three procedures on clean data. Therefore, the initial setup is devised without the inclusion of any outliers. Data contaminated with good leverage points are obtained by multiplying ε randomly selected slices of $\underline{\tilde{\boldsymbol{X}}}$ with a constant $c_1 = 10$. The noise components are added to the modified array as described above to obtain the final ε-contaminated array $\underline{\boldsymbol{X}}$. Bad leverage points are generated by adding a constant $c_2 = 0.1$ to the already generated good leverage points. In a last setting residual outliers were generated by adding a constant $c_2 = 0.1$ to randomly selected slices of $\underline{\boldsymbol{X}}$. In all contaminated settings $\varepsilon = 0.1$ or $\varepsilon = 0.2$ percentage of outliers is used. The primary focus is on observing the behavior of the procedures when applied to clean data as well as data that has been contaminated with bad leverage points. A brief comment is provided at the end of this Section regarding the remaining two scenarios: good leverage points and residual outliers.

3.2. Evaluation criteria

The performance of different estimation procedures for the CP model has been thoroughly compared in a number of comparative studies (Tomasi and Bro, 2006; Faber et al., 2003; Yu et al., 2011b; 2011a) using both simulations and empirical data and the criteria for comparison are relatively well studied. The main goal in this simulation study is to measure the computational performance of the new algorithm, to compare it to the already existing robust estimation procedure (Engelen and Hubert, 2011) and to confirm that the new procedure has better computational properties. At the same time the intention is to assess the accuracy of the novel procedure and validate that it delivers at least as good solutions as the existing methods. The performance in terms of computational efficiency is measured by the CPU time consumed in seconds (*time*), the number of iterations necessary to reach convergence (*iter*) and the occurrence of temporary degeneracies (swamps). These are revealed by the *triple cosine* (TC) (or congruence product) which can be calculated between all pairs of loadings. If TC becomes less than -0.8 for at least 10 iterations a temporary degeneracy is counted. The triple cosine for two factors is defined as

$$TC_{i,j} = \phi(\boldsymbol{a}_i\boldsymbol{a}_j)\phi(\boldsymbol{b}_i\boldsymbol{b}_j)\phi(\boldsymbol{c}_i\boldsymbol{c}_j) \tag{8}$$

where

$$\phi_{xy} = \frac{\sum \boldsymbol{xy}}{\sqrt{\sum \boldsymbol{x}^2 \sum \boldsymbol{y}^2}} \tag{9}$$

is the Tucker's congruence coefficient between two vectors (Lorenzo-Seva and ten Berge, 2006).

Regarding the quality of the solutions obtained, several different measures (Tomasi and Bro, 2006; Engelen and Hubert, 2011) can be considered: the value of the objective function (FIT), occurrence of full recoveries (FD) or rather its complementary value, the occurrence of fault recoveries (FR), the mean square error of the fit (MSE) and the angle between the estimated subspace and the true subspace spanned by the B- and C-loadings. The recovery of the correct solution is assessed by calculating the factor congruence, given by Equation 9 between the known underlying factors and the extracted factors (Tomasi and Bro, 2005; 2006; Mitchell and Burdick, 1994). This coefficient is calculated for each pair of true and estimated factors and is multiplied across the three modes. A solution is considered to be an FR if the obtained value for at least one factor is less than 0.95. Due to the permutation and sign indeterminacy of the CP solutions (Harshman, 1970) all possible permutations of the extracted factors and the underlying components need to be compared and the one yielding the highest sum of the coefficients will be considered the correct one (Mitchell and Burdick, 1994; 1993; Tomasi and Bro, 2006).

The mean squared error is given by

$$MSE = \frac{1}{w}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} w_i(x_{ijk} - \hat{x}_{ijk})^2 \tag{10}$$

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

$$maxsub = \max_{\vec{b}_1}\min_{\vec{b}_2}\arccos(\vec{b}_1^\top\vec{b}_2) \tag{11}$$

This subspace angle has to be as small as possible and is reported in radians. The function `subspace()` from the **R** (R Core Team, 2022) package **pracma** (Borchers, 2022) is used to compute *maxsub*.

Table 5
Parameters for generation of the data sets.

R	3, 5
I, J, K	100, 20, 20; 100, 100, 20[a]
HO	15%, 20%, 25%
HE	10%, 15%, 20%
CONG	0.2, 0.3, 0.4, 0.5, 0.6, 0.7
Outliers: ε	0%, 10%, 20%
Outliers: type	good L.P., bad L.P., residual outliers
Repetitions	100
Interim conv	10^{-2}
Final conv	10^{-8}

[a] Simulations with $I, J, K = 100, 100, 20$ were conducted only for $R = 3$ and Outlier type "bad L.P."

Table 6
Percentages of fault recovery (FR) cases by convergence threshold at CONG=0.7 for $F = R = 5$.

	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
FR%	0.44	1.33	3.33	6.33	8.89	9.44	9.67	10.11

3.3. Parameters of the optimization procedure

The simulation study begins with a specific task: to identify the best threshold parameters for switching from the initialization phase to the refinement phase for R-INT2. These thresholds can be specified in terms of an interim convergence parameter, i.e. instead of iterating the ATLD step to convergence the procedure stops earlier and moves to the second step. Different interim parameters seem to affect the computational efficiency of the procedure as well as the accuracy of the obtained results. Simonacci and Gallo (2020) used a similar simulation for the INT-2 algorithm and found out that one could successfully use the procedure with interim convergence parameter set to 10^{-2} or 10^{-3}. The aforementioned simulation is conducted for the robust version of the algorithm R-INT2. The data sets are generated as described above in Section 3.1 using six levels of factor congruence *CONG*: 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7, the three levels of homoscedastic and heteroscedastic noise respectively as given in Table 5 and two ranks of the trilinear array $R = 3, 5$. For each set of parameters $(2 \times 3 \times 3 \times 6)$, 100 data sets are generated (resulting in total of 10,800 data sets). For each data set the CP model is estimated eight times with the R-INT2 algorithm, each time using a different convergence threshold: $10^{-1}, \ldots, 10^{-8}$. The size of the arrays was chosen as $50 \times 50 \times 50$ and only correct factor estimation model fitting $(F = R)$ was applied. This is repeated for each type of outliers and levels of contamination. Since the effect of the contamination type and level on the choice of the interim convergence threshold was insignificant, only the results for the case with 20% bad leverage points are presented. The performance is evaluated by the criteria presented in Section 3.2.

The primary emphasis is placed on the key efficiency diagnostics, namely *time* and *iter*, as the primary focus of interest revolves around the computational efficiency of the new procedure. The aggregated values of computational time and number of iterations are shown as box plots in Figure 2. It is immediately seen that there are no drastic differences between the values of the intermediate convergence parameter. The median time as well as the variance increases with increasing the parameter and values 10^{-2} and 10^{-3} appear to be most favorable.

The number of swamps is insignificant and they occur only for extreme values of CONG larger than 0.7. The variances of FIT and MSE were close to 0 therefore these results are not shown (the box plots would have been identical). The percentage of fault recoveries is also monitored. For $F = R = 3$ all values are below 1% and for $F = R = 5$, where the values are already visibly different, all the incidences are concentrated in CONG=0.7. The values are presented in Table 6. It is evident that with the increase of the interim convergence criterion the procedure becomes less stable and the percentage of fault recoveries increases significantly. It is above 5% already for 10^{-4}. Taking this into account, but considering also the conducted simulations which are presented in the next Section 3.4 choosing the value of 10^{-2} is recommended.

3.4. Simulation results

In the presentation of the results, priority will be given to the case of bad leverage points, as it is expected these to inflict the most substantial impact on the procedures. The first objective of the simulation study is to assert that the solutions obtained by R-INT2 are not much different from those obtained by R-ALS in terms of stability and modeled noise. This can be achieved by comparing the FIT diagnostic and verifying that the variability explained by R-INT2 is not significantly different from that explained by R-ALS. Overall, at 20% contamination with bad leverage points, the difference of R-ALS FIT and R-INT2 FIT is higher than $1e^{-4}$ (Tomasi and Bro, 2006) in only 1% of the cases. In more than half of the cases (54%) the fit of R-INT2 is better than that of R-ALS which demonstrates that R-INT2 is capable of identifying the best low rank approximation as well as R-ALS does. This diagnostic is less relevant in the case of over factoring $(F = R + 1)$ because of

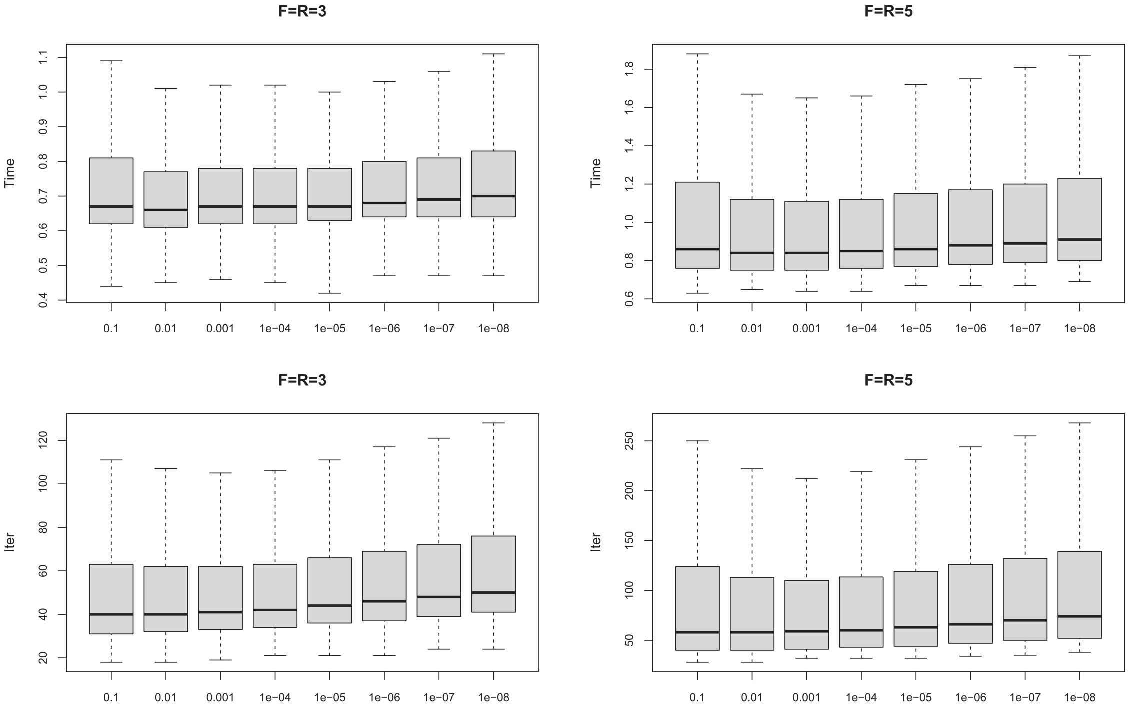

Fig. 2. CPU time in seconds and number of iterations for the robust CP with INT-2 estimation procedure (R-INT2) using various interim convergence parameters. The data sets are generated with 20% bad leverage points and the results are aggregated over all considered CONG and Noise levels.

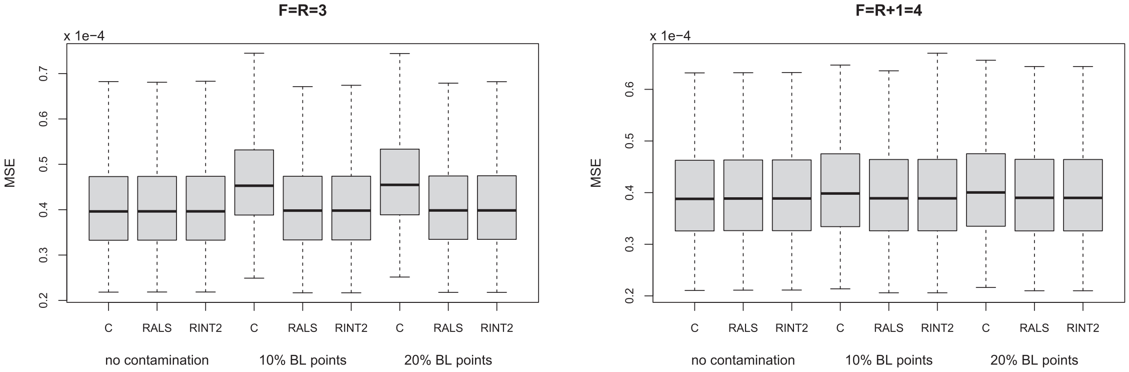

Fig. 3. MSE values for classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets without contamination, and with 10% and 20% bad leverage points respectively. The results are aggregated over all considered CONG and Noise levels.

the noise modeled by the additional component. Still, only in 27% of the cases the difference in fit is larger than $1e^{-4}$ and in half of the cases R-INT2 provides a better fit. The results for 10% bad leverage points and clean data are very similar and therefore are not reported here. These results as well as all the other results that did not find place in the article are provided in the Supplementary Material. The comparison of the FIT diagnostic in the case of larger data sets ($J \gg K$) is quite similar: in only 0.04% of the cases the difference is higher than $1e^{-4}$ and in 61% the R-INT2 fit is better than R-ALS. In the case of over factoring ($F = R + 1$) these percentages are 18% and 54% respectively.

The next measure of accuracy to look at is the mean squared error (MSE) given by Equation 10. The results for the three estimators and three types of data (no outliers, 10% bad leverage points and 20% bad leverage points) are presented in the box plots in Figure 3. The left panel shows the results for correct factor estimation ($F = R = 3$) and the right one presents the case of over-factoring ($F = R + 1 = 4$). The results are aggregated over all considered CONG and Noise levels. All three estimators give almost identical results in the case of no contamination. It is clear that the classical method should be preferred because of its much lower computational time. Further, it is easily seen that R-INT2 is not less accurate than R-ALS in all three contamination cases, however, due to different levels of MSE in the different simulation scenarios, the trend is quite flattened, especially in the case of over-factoring. Therefore, in Figure 4 the same data disaggregated by CONG with fixed noise level (left panel) and by Noise level (right panel) with fixed CONG are presented. The results are for data

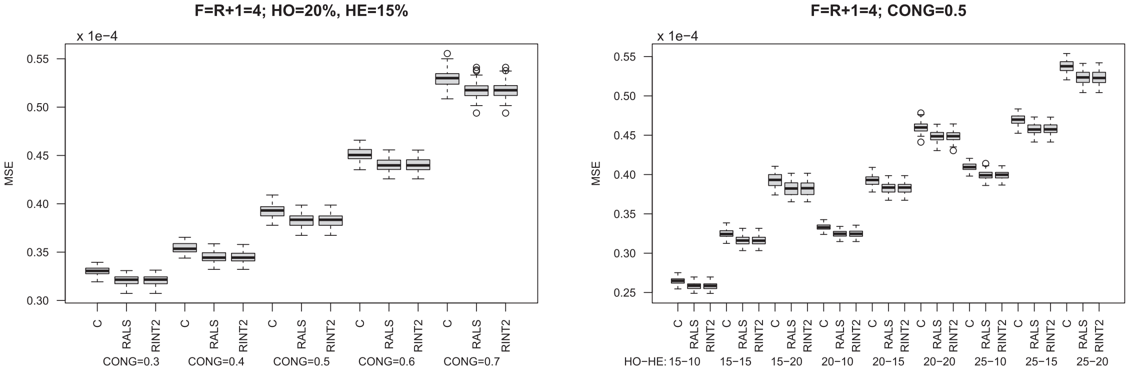

Fig. 4. MSE values for classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets with 20% bad leverage points aggregated by CONG (left) and by Noise (right).

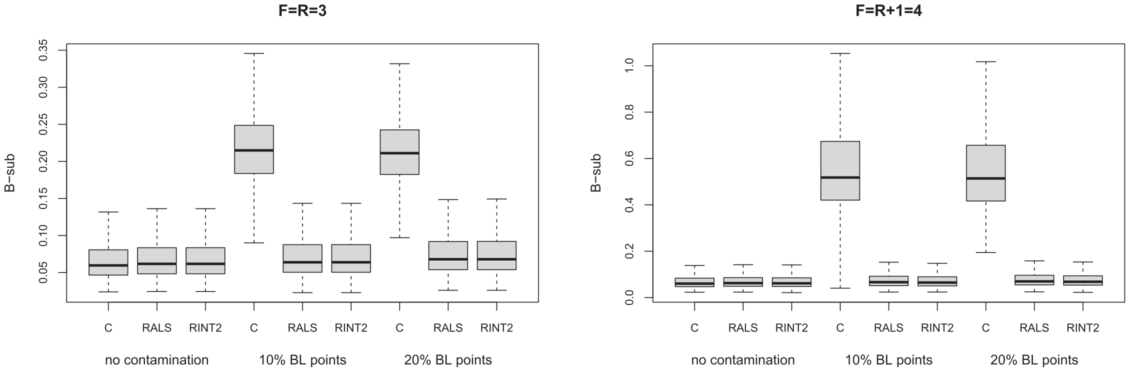

Fig. 5. Angle of B-loadings of classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets without contamination, and with 10% and 20% bad leverage points respectively. The results are aggregated over all considered CONG and Noise levels.

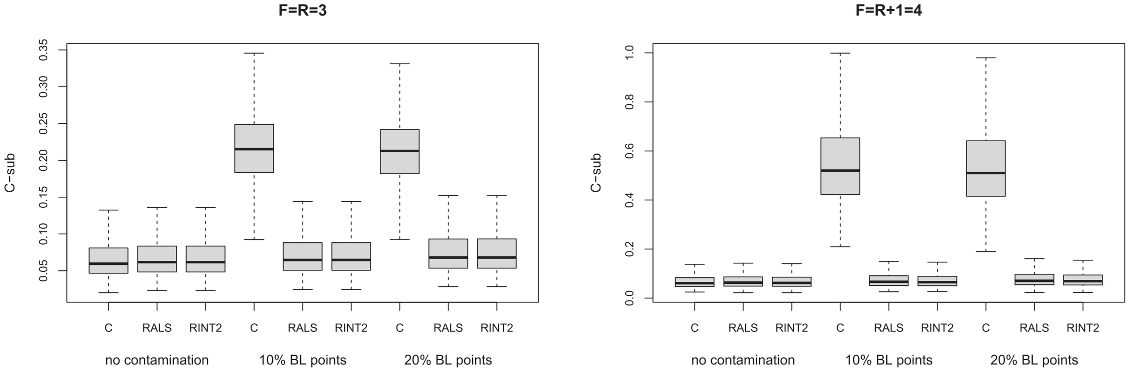

Fig. 6. Angle of C-loadings of classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets without contamination, and with 10% and 20% bad leverage points respectively. The results are aggregated over all considered CONG and Noise levels.

with 20% bad leverage points but for 10% BL points the pattern is similar and for no contamination as it could be expected the boxplots of the three methods look identical (not shown here).

The angles of the B-, respectively C-loadings for the three methods and the three contamination types are presented as box plots in Figures 5 and 6. Again, in case of no contamination the results of all three methods are similar, but in the presence of 10% or 20% contamination R-ALS and R-INT2 perform much better (much lower median value and lower variation). The performance of R-ALS and R-INT2 is almost identical.

It is important to verify that the known instabilities of ATLD are not transferred to the integrated procedure. This can be checked by looking at the percentage of fault recoveries reported for all three algorithms for different levels of contamination and different ranks in Table 7. For $R = 3$, both in the correct rank estimation case and when over-factoring the percentage of fault recoveries for the robust methods is below 1%, only for the classical estimates on data with 10% and 20% contamination

Table 7
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.

		$F = R = 3$			$F = R + 1 = 4$		
		C	RALS	RINT2	C	RALS	RINT2
FR	0%	0.4	0.2	0.6	0.6	1.0	0.1
	10%	98.2	0.2	0.6	0.0	0.7	0.1
	20%	98.8	0.2	0.8	0.0	0.4	0.1
SWAMPS	0%	0	0	0	13	7	0
	10%	0	0	0	1	14	1
	20%	0	0	0	0	11	5
		$F = R = 5$			$F = R + 1 = 6$		
		C	RALS	RINT2	C	RALS	RINT2
FR	0%	4.3	4.1	4.8	1.4	1.4	1.4
	10%	95.1	4.6	5.7	1.6	1.7	1.7
	20%	95.0	5.0	6.6	0.1	2.2	2.1
SWAMPS	0%	1	2	2	8	10	3
	10%	0	2	0	4	10	7
	20%	0	2	2	4	14	14

the percentage increases drastically, coming close to 100%. For $R = 5$ all percentages are slightly higher but still lower than 5% except for the classical ALS in case of correct factor estimation and contaminated data. The only difference in the results for the larger data sets ($J \gg K$) is that the instability of R-ALS increases in all contamination scenarios in the case of over factoring and the percentage of fault recoveries nears 10%. This of course increases the computational time, as it will be seen later.

Upon confirming that R-INT2 exhibits accuracy that is comparable to, or even surpasses, that of R-ALS across various scenarios, the focus shifts to the primary metrics of interest: computational efficiency, quantified by the CPU time in seconds (*time*), and the number of iterations (*iter*). The results aggregated over all levels of CONG and Noise for data with 20% bad leverage points are presented in Figure 7. The box plots for the number of iterations follow the pattern of the time box plots, as it is expected, and in all cases (correct factor estimation and over-factoring) R-INT2 outperforms R-ALS, both in terms of median values and variance. The computational time of both R-ALS and R-INT2 does not depend on the level of contamination as can be seen in Figure 8. In all three cases (no contamination, 10% BL and 20% BL) the median values of R-ALS and R-INT2 remain the same, but what is very interesting, the time consumed by the classical ALS increases drastically—three times in the case of correct factor estimation and even five times in the case of over-factoring, becoming almost equal to the time of the robust R-INT2. In Figure 8 the results for rank $R = 5$ are presented, but the pattern for $R = 3$ looks exactly the same. The fact that the classical estimation method becomes slower in case of contaminated data is one more advantage of the robust methods for fitting the CP model.

Figure 9 shows the computational time of the three procedures on data sets of higher dimensions ($J \gg K$) with different level of contamination. In the case of correct rank estimation the only difference is that the dependence of *time* on the contamination level for the classical estimates is even more pronounced. However, the right panel of Figure 9 shows completely different picture: the case of over factoring renders the robust CP estimation with ALS almost useless in all contamination scenarios by dramatically increasing the computational time both in terms of median and variance. This once again confirms the advantages of the robust CP estimation based on the integrated procedure R-INT2. It is reasonable to expect that the performance of the estimators will differ across various data setups characterized by different CONG and Noise levels. To gain a more detailed understanding of the performance measures, particularly in terms of computational time, the results are presented in a disaggregated manner based on CONG. Figure 10 illustrates the results for three specific CONG values (0.3, 0.5, and 0.7) and two ranks (R = 3 and R = 5). The cases of correct rank estimation and over-factoring are considered. In the left panel (with correct rank estimation, $F = R$), it becomes evident that the performance of both estimators improves when the rank is lower, specifically $R = 3$, compared to $R = 5$. In all cases R-INT2 is much better than R-ALS except for higher CONG (0.7) in $R = 5$. When over-factoring ($F = R + 1$) R-INT2 is better than R-ALS in terms of median value and variance in all cases but the difference in performance decreases with increasing the CONG level.

The computational efficiency can also be judged by counting the number of swamps, i.e. the temporary degeneracy which continue for more than 10 iterations and thus slows down the procedure. This problem was not significantly manifested in the simulation. As seen in the lower part of Table 7, 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, while only several cases were observed when over-factoring and when $R = 5$. However the number of such cases is insignificant when compared to the total number of 4500 repetitions.

All results presented so far were either for clean data or for bad leverage points with different percent of contamination. The complete set of simulations was conducted separately also for the setups with good leverage points and residual outliers.

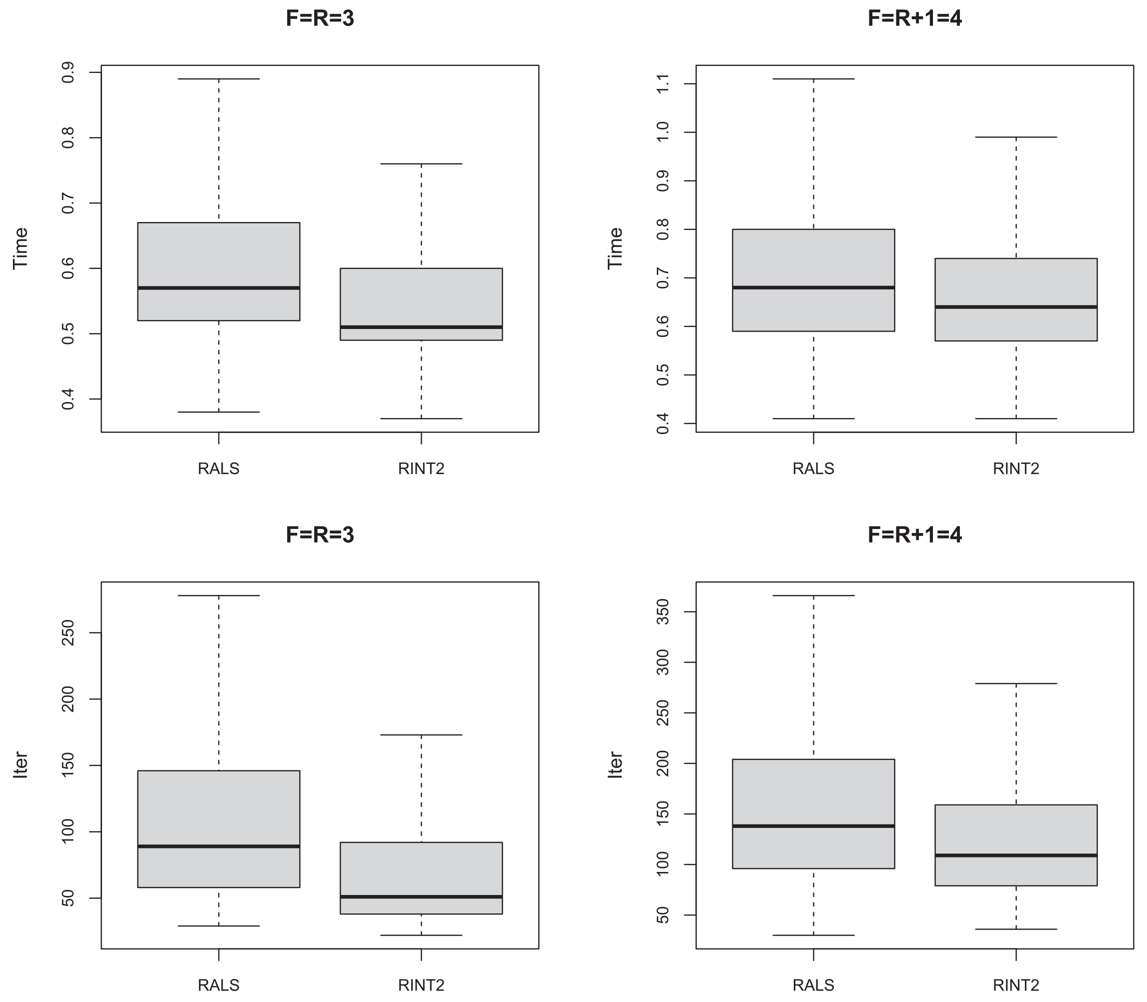

Fig. 7. CPU time in seconds and number of iterations, robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets with 20% bad leverage points. The results are aggregated over all considered CONG and Noise levels.

The obtained results in the latter case were quite similar to those for the setup with bad leverage points, both in terms of accuracy and computational efficiency, therefore are not shown here.

The case of good leverage points can be described in more detail. For correct factor estimation ($F = R$) all three methods are robust and the MSE stays almost the same for all levels of contamination. The estimates of the loadings measured by the angles of the B- and C-loadings are also not worse, even the higher the level of contamination the better estimates. This is not surprising since the good leverage points are apart of the bulk of the data but they follow the model and thus even help to increase the precision of the estimates. It can be concluded from these results that both the classical and robust methods cope with good leverage points and fit the data well, confirming the results published by Engelen and Hubert (2011). Additionally, it can be confirmed that both robust methods successfully identify all the good leverage points, whereas the classical ALS procedure fails to detect the majority of them.

The picture changes however in the over-factoring case $F = R + 1$. MSE and angles of B- and C-loadings remain similar to the correct rank estimation, but the overall difference of R-ALS FIT and R-INT2 FIT is higher than $1e^{-4}$ in more than 80% of the cases. In more than three quarters of the cases (79%) the fit of R-INT2 is better than that of R-ALS. But most importantly, the computational time changes drastically as shown in Figure 11. While for clean data all estimators are slightly slower on the over-factoring case (measured on the median CPU time in seconds), Table 8, with 10% contamination the classical ALS is 80 times slower and with 20% – 140 times slower. The robust ALS is 11 times slower on 10% good leverage points and 35 times slower on 20%. At the same time the timing of the integrated robust algorithm R-INT2 hardly changes showing fascinating stability against the over-factoring with good leverage points. This is one more remarkable property of the R-INT2 algorithm. This extends further to the fault recoveries. While in the correct rank estimation all FR values as percentage of the total cases, for all estimators are below 1%, in the over-factoring case the classical ALS jumps to 53% and 75% for 10%

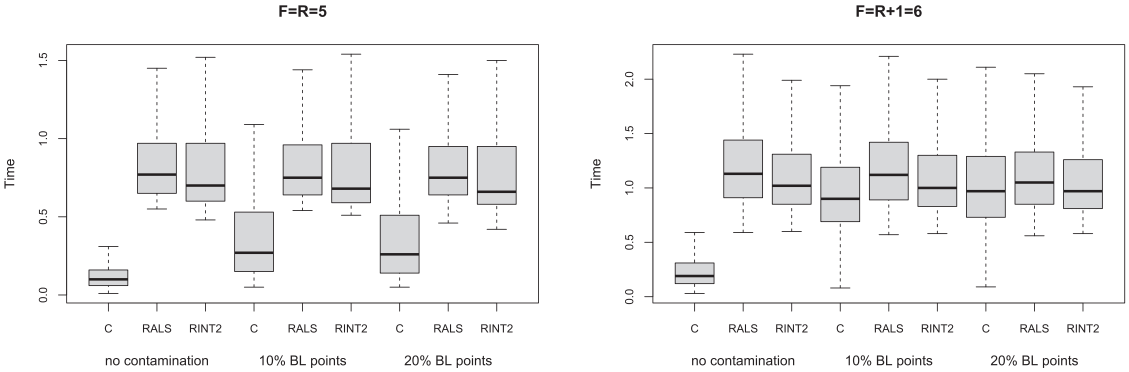

Fig. 8. CPU time in seconds, classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets with different level of contamination. The results are aggregated over all considered CONG and Noise levels.

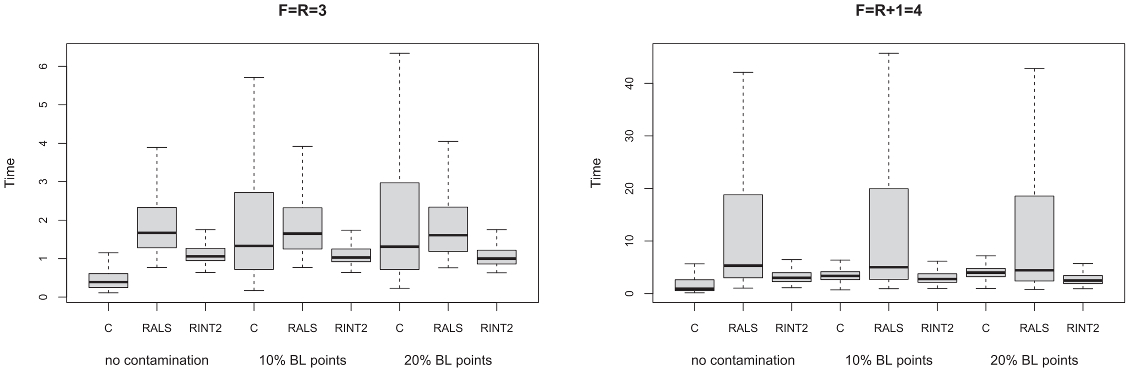

Fig. 9. CPU time in seconds, classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets of higher dimensions ($J \gg K$) with different level of contamination. The results are aggregated over all considered CONG and Noise levels.

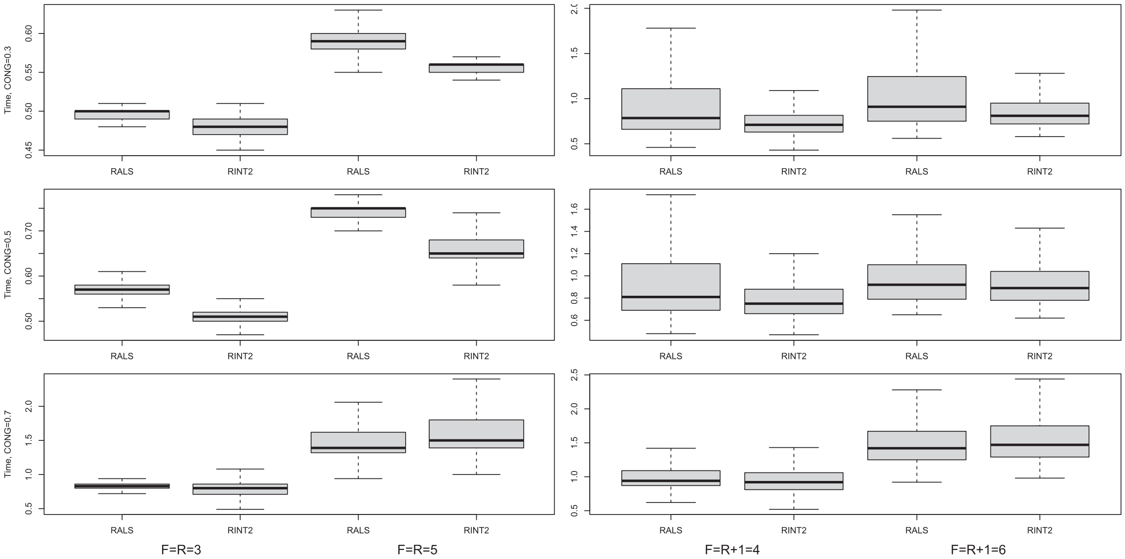

Fig. 10. CPU time in seconds, robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets with 20% BL. The results are aggregated over all considered Noise levels and presented by selected levels of CONG (0.3, 0.5 and 0.7) for different ranks, with correct rank estimation and over-factoring.

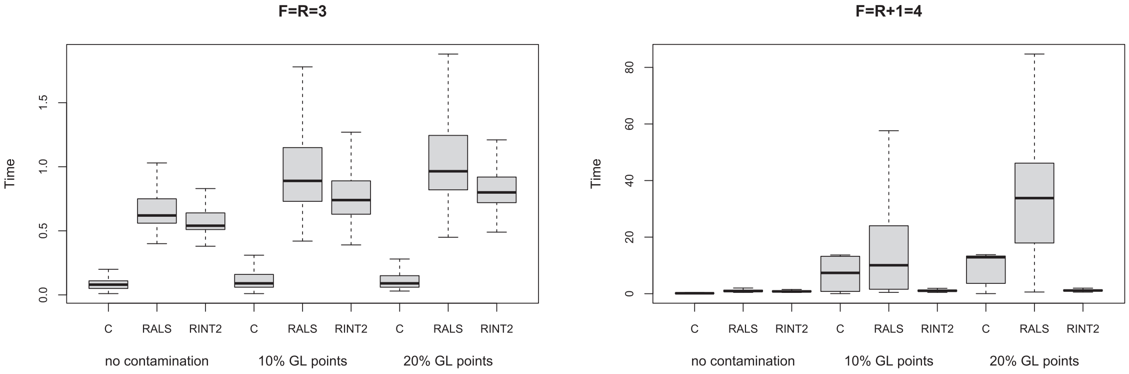

Fig. 11. CPU time in seconds, classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets with different level of good leverage points contamination: correct factor estimation (left) and over-factoring (right). The results are aggregated over all considered CONG and Noise levels.

Table 8
Median CPU time in seconds (TIME) and fault recoveries (FR), classical CP (C), robust CP with ALS estimation (R-ALS) and robust CP with INT-2 estimation (R-INT2) on data sets with different level of good leverage points contamination: correct factor estimation (left) and over-factoring (right).

		$F = R = 3$			$F = R + 1 = 4$		
		C	RALS	RINT2	C	RALS	RINT2
TIME	0%	0.08	0.62	0.54	0.15	0.94	0.81
	10%	0.09	0.89	0.74	7.35	10.05	1.03
	20%	0.09	0.96	0.80	12.88	33.79	1.13
FR	0%	0.42	0.36	0.73	0.38	0.62	0.04
	10%	0.00	0.09	0.13	52.80	36.40	0.07
	20%	0.00	0.11	0.22	75.04	70.31	0.16

and 20% good leverage points respectively. The robust ALS R-ALS is slightly better with 36% and 70%. The integrated robust algorithm R-INT2 remains also in this case below 1%.

4. Example: The Dorrit data set

In this section a data set from chemometrics will be used to demonstrate the computational advantages of the new estimation procedure. This data set was already used in the context of outlier detection (Riu and Bro, 2003) and robust estimation of the CP model (Engelen and Hubert, 2011) and shows the difference between classical and robust methods when severe outliers are present. The original Dorrit data set (Baunsgaard, 1999; Riu and Bro, 2003) represents 27 synthetic samples containing different concentrations of four analytes: hydroquinone, tryptophan, phenylalanine and dopa and thus consists of 27 fluorescence landscapes of 233 emission wavelengths (250–482 nm) and 24 excitation wavelengths (200–315 nm taken at 5 nm intervals). The data set is modified as described in Engelen and Hubert (2011): the emission wavelengths are taken at 2 nm, noisy parts situated at the excitation wavelengths from 200 to 230 nm and at emission wavelengths below 250 nm are excluded and the severe Rayleigh scattering areas present in all samples are replaced by interpolated values. The modified Dorrit data set is available in the **R** package **rrcov3way** (Todorov et al., 2023) as a three dimensional ($27 \times 116 \times 18$) array. For the purpose of this study the results obtained by Engelen and Hubert (2011) will be reproduced, i.e. the same outliers will be identified and at the same time this will be done significantly faster.

The estimated CP model for the Dorrit data with the R-INT2 procedure is presented graphically in the right panel of Figure 1. It will be shown that the novel method obtains the same results even when the model is misspecified (over-factoring) which can often happen in the practice and again, this is done much faster than with the ALS-based robust procedure. From Baunsgaard (1999) it is known that a CP model with four components should be fitted. Furthermore, it is known that the samples QAB, QAC and QAE (2, 3 and 5) are bad leverage points and sample QAD (4) is a border case. The sample QAA (10) which is identified by Riu and Bro (2003) as well as by the classical outlier plot as a residual outlier is actually also a border case.

Fitting the model with four components results in identical fit of 98.73% for both procedures. Also, the three loading matrices are identical if rounded to the third decimal position and reflected (adjusting the sign of the components as necessary). The comparison of the computational performance of the two procedures is presented in Table 9 which shows a noteworthy 38% improvement in speed in the case of correct factor estimation. In the case of over-factoring ($F = 5$) the fit value is the same (98.87%) but the gain in computational performance is even higher at 41%.

Table 9
Computational performance (CPU time in seconds and number of iterations) of R-ALS and R-INT2 on the Dorrit data set.

	$F = 4$			$F = 5$		
	R-ALS	R-INT2	(%)	R-ALS	R-INT2	(%)
iter	216	205		297	241	
time	5.97	4.08	38%	8.8	5.25	41%

5. Summary and conclusions

By combining the robust procedure for CP as proposed by Engelen and Hubert (2011) with the highly efficient estimation algorithm INT-2 as proposed by Simonacci and Gallo (2020), a fast and robust CP modeling technique is obtained. This integration leverages the speed advantage of the INT-2 algorithm and the ability of the procedure from Engelen and Hubert (2011) to handle outliers effectively. The 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. It is demonstrated in a separate simulation study that the interim convergence threshold of the robust integrated procedure should be set to 10^{-2} in order to maximize the computational advantage. The new procedure is then illustrated on an empirical data example already well known in the literature. The simulation study and the data example verify that R-INT2 is a fast converging procedure providing stable solutions. Thanks to these advantages computationally expensive tools for identifying the correct rank of the three-way data will not be needed.

All computations were performed in the **R** language and environment for statistical computing (R Core Team, 2022) using the package **rrcov3way** (Todorov et al., 2023), freely available from the Comprehensive R Archive Network (CRAN) at https://cran.r-project.org/. The package provides Tucker3 and CP functions with a robust option for standard and compositional three-way data as well as specific plotting functions for the resulting objects. All code scripts for replication of the simulations and the empirical example are available in the *GitHub* repository at https://github.com/valentint/robust-parafac-ecosta.

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, like these proposed in Simonacci and Gallo (2019). The behavior of the algorithm if collinearity is present (Di Palma et al., 2018) will be of great interest as well as its extension with additional constraints.

Acknowledgements

The present version of the paper has greatly benefited from detailed comments and constructive criticism by two anonymous referees and the editors. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the United Nations Industrial Development Organization.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ecosta.2023.07.001

References

Andersen, C., Bro, R., 2003. Practical aspects of PARAFAC modelling of fluorescence excitation-emission data. Journal of Chemometrics 17, 200–215.
Baunsgaard, D., 1999. Factors Affecting 3-way Modelling (PARAFAC) of Fluorescence Landscapes. Technical Report. Royal Veterinary and Agricultural University. Department of Dairy and Food Science
Borchers, H. W., 2022. pracma: Practical Numerical Math Functions. R package version 2.4.2. https://CRAN.R-project.org/package=pracma.
Boudt, K., Rousseeuw, P.J., Vanduffel, S., Verdonck, T., 2020. The minimum regularized covariance determinant estimator. Statistics and Computing 30, 113–218.
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.
Cattell, R.B., 1944. "Parallel proportional profiles" and other principles for determining the choice of factors by rotation. Psychometrika 9 (4), 267–283.
Ceulemans, E., Kiers, H.A.L., 2006. Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. British Journal of Mathematical and Statistical Psychology 59, 133–150.
Chen, Z.-P., Wu, H.-L., Jiang, J.-H., Li, Y., Yu, R.-Q., 2000. A novel trilinear decomposition algorithm for second-order linear calibration. Chemometrics and Intelligent Laboratory Systems 52 (1), 75–86.
Croux, C., Filzmoser, P., Oliveira, M., 2007. Algorithms for projection-pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems 87 (218), 218–225.
Croux, C., Haesbroeck, G., 2000. Principal components analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. Biometrika 87, 603–618.
Devlin, S.J., Gnanadesikan, R., Kettenring, J.R., 1981. Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association 76, 354–362.
Di Palma, M.A., Filzmoser, P., Gallo, M., Hron, K., 2018. A robust Parafac model for compositional data. Journal of Applied Statistics 45 (8), 1347–1369.
Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: Bikel, P., Doksum, K., Hodges, J.L. (Eds.), A Festschrift for Erich Lehmann. Wadsworth, Belmont, CA, pp. 157–184.

Engelen, S., Frosch-Møller, S., Hubert, M., 2007. Automatically identifying scatter in fluorescence data using robust techniques. Chemometrics and Intelligent Laboratory Systems 86, 35–51.

Engelen, S., Hubert, M., 2011. Detecting outlying samples in a parallel factor analysis model. Analytica Chemica Acta 705, 155–165.

Faber, N.M., Bro, R., Hopke, P.K., 2003. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. Chemometrics and Intelligent Laboratory Systems 65, 119–137.

Filzmoser, P., Todorov, V., 2013. Robust tools for the imperfect world. Information Sciences 245, 4–20.

Harshman, R.A., 1970. Foundations of the PARAFAC procedure: Models and conditions for an ”explanatory“ multi-modal factor analysis. Technical Report. UCLA.

Harshman, R.A., Lundy, M.E., 1984. The PARAFAC model for three-way factor analysis and multidimensional scaling. Research methods for multimode data analysis 122–215.

Huber, P.J., Ronchetti, E., 2009. Robust Statistics. John Wiley & Sons, New Jersey.

Hubert, M., Debruyne, M., Rousseeuw, P.J., 2017. Minimum covariance determinant and extensions. WIREs computational statistics 10, e1421.

Hubert, M., Rousseeuw, P., Vanden Branden, K., 2005. ROBPCA: A new approach to robust principal component analysis. Technometrics 47, 64–79.

Hubert, M., Rousseeuw, P.J., van Aelst, S., 2008. High-breakdown robust multivariate methods. Statistical Science 23, 92–119.

Hubert, M., Rousseeuw, P.J., den Bossche, W.V., 2019. Macropca: An all-in-one pca method allowing for missing values as well as cellwise and rowwise outliers. Technometrics 61 (4), 459–473.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer-Verlag, New York, NY, USA.

Kiers, H., 2000. Towards a standardized notation and terminology in multiway analysis. Journal of Chemometrics 14 (3), 105–122.

Kiers, H.A., ten Berge, J.M., Bro, R., 1999. PARAFAC2 – Part I. A direct fitting algorithm for the PARAFAC2 model. Journal of Chemometrics 13 (3–4), 275–294.

Kroonenberg, P.M., 2008. Applied multiway data analysis. John Wiley & Sons, Hoboken, NJ.

Liu, S., Trenkler, G., 2008. Hadamard, khatri-rao, kronecker and other matrix products. International Journal of Information and Systems Science 4 (1), 160–177.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., 1999. Robust principal components for functional data. Test 8, 1–28.

Lorenzo-Seva, U., ten Berge, J.M.F., 2006. Tucker's congruence coefficient as a meaningful index of factor similarity. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 2, 57–64.

Maronna, R.A., 2005. Principal components and orthogonal regression based on robust scales. Technometrics 47, 264–273.

Mitchell, B.C., Burdick, D.S., 1993. An empirical comparison of resolution methods for three-way arrays. Chemometrics and Intelligent Laboratory Systems 20 (2), 149–161.

Mitchell, B.C., Burdick, D.S., 1994. Slowly converging parafac sequences: Swamps and two-factor degeneracies. Journal of Chemometrics 8.

Pravdova, V., Estienne, F., Walczak, B., Massart, D., 2001. A robust version of the tucker3 model. Chemometrics and Intelligent Laboratory Systems 59 (1), 75–88.

R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Riu, J., Bro, R., 2003. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. Chemometrics and Intelligent Laboratory Systems 65 (1), 35–49.

Ronchetti, E., 2021. The main contributions of robust statistics to statistical science and a new challenge. METRON 79, 127–135.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 851–857.

Rousseeuw, P.J., 1997. Introduction to positive-breakdown methods. In: Maddala, G.S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 15. Elsevier, North-Holland, pp. 101–121.

Rousseeuw, P.J., Debruyne, M., Engelen, S., Hubert, M., 2006. Robustness and outlier detection in chemometrics. Critical Reviews in Analytical Chemistry 36, 221–242.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. John Wiley & Sons, New York.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

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

Simonacci, V., Gallo, M., 2020. An ATLD–ALS method for the trilinear decomposition of large third-order tensors. Soft Computing 24, 13535–13546.

Smilde, A., Bro, R., Geladi, P., 2004. Multi-way analysis with applications in the chemical sciences. John Wiley & Sons, Chichester, Hoboken (N.J.). http://opac.inria.fr/record=b1101787

Timmerman, M.E., Kiers, H.A., 2000. 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.

Todorov, V., 2020. **rrcov**: Scalable Robust Estimators with High Breakdown Point. R package version 1.5-3. https://CRAN.R-project.org/package=rrcov.

Todorov, V., Filzmoser, P., 2009. An object oriented framework for robust multivariate analysis. Journal of Statistical Software 32 (3), 1–47. http://www.jstatsoft.org/v32/i03/

Todorov, V., Simonacci, V., Di Palma, M. A., Gallo, M., 2023. **rrcov3way**: Robust Methods for Multiway Data Analysis, Applicable also for Compositional Data. R package version 1.0. http://CRAN.R-project.org/package=rrcov3way.

Tomasi, G., Bro, R., 2005. Parafac and missing values. Chemometrics and Intelligent Laboratory Systems 75 (2), 163–180.

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

Tucker, L., 1966. Some mathematical notes on three-mode factor analysis. Psychometrica 31 (3), 279–311.

Wu, H.-L., 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.

Yu, Y.-J., Wu, H.-L., Kang, C., Wang, Y., Zhao, Y., Li, Y.-N., Liu, Y.-J., Yu, R.-Q., 2011. Algorithm combination strategy to obtain the second-order advantage: simultaneous determination of target analytes in plasma using three-dimensional fluorescence spectroscopy. Journal of Chemometrics 26 (5), 197–208.

Yu, Y.-J., Wu, H.-L., Nie, J.-F., Zhang, S.-R., Li, S.-F., Li, Y.-N., Zhu, S.-H., Yu, R.-Q., 2011. A comparison of several trilinear second-order calibration algorithms. Chemometrics and Intelligent Laboratory Systems 106 (1), 93–107.