

Detecting Outlying Samples in a PARAFAC Model

Sanne Engelen and Mia Hubert

April 4, 2011

Summary

To explore multi-way data, different methods have been proposed. Here, we study the popular PARAFAC (Parallel factor analysis) model, which expresses multi-way data in a more compact way, without ignoring the underlying complex structure. To estimate the score and loading matrices, an alternating least squares procedure is typically used. It is however well known that least squares techniques suffer from outlying observations, making the models useless when outliers are present in the data. In this paper, we present a robust PARAFAC method. Essentially, it searches for an outlier-free subset of the data, on which we can then perform the classical PARAFAC algorithm. An outlier map is constructed to identify outliers. Simulations and examples show the robustness of our approach.

Keywords: Robustness, PARAFAC, multi-way data, outliers

1 Introduction

Many studies in chemistry often involve the simultaneous measurement of variables giving rise to a complex data structure. For example, the outcome of a fluorescence experiment can be ordered in a cube where for several samples the intensity of emitted light is measured at different wavelengths for many excitation wavelengths. It is shown in e.g. [1] that preserving the nature of the data set by arranging it in a higher dimensional tensor instead of forcing it into a matrix, leads to a better understanding and modeling of the data. Such complex data structures are called multi-way data sets. More specifically, the concepts one-way data and two-way data refer to vectors and matrices, respectively. Although generalizations to higher order data are possible, we will only consider three-way data throughout this paper, which can be organized in a cube. This implies that for each sample two sets of variables are measured, which can be organized in a matrix.

Different techniques exist to explore such multi-way data, among which PARAFAC and Tucker3 are the most famous ones. See e.g. [2, 3, 4] for PARAFAC and e.g. [5, 6, 7] for Tucker3. In both methods scores and loading matrices are constructed to express the data in a more comprehensible way. From this point of view, PARAFAC and Tucker3 can be seen as a generalization of principal components analysis (PCA) to higher order tensors [8, 9].

In this paper we focus on the robustness properties of PARAFAC. More precisely, we investigate how well the PARAFAC model can deal with outlying samples in three-way data. Therefore we elaborate first on what will be considered an outlier. For multivariate problems, in which two-way data sets are analyzed, outliers are usually described as observations that differ significantly from the majority of other samples. In case of three-way data, this definition can be copied, such that matrices that have a deviating profile compared towards the other ones, are called outliers.

It is well-known that least squares approaches break down in the presence of outliers, such that the estimated model fits the outliers better than the bulk of the data. As such, also the PCA and PARAFAC model, which both rely on a least squares algorithm for the computation of the scores and loadings, suffer from anomalous observations. For PCA, several robust alternatives have been developed over the last years. For low-dimensional data, methods can be used based on a robust covariance matrix, see e.g. [10]. To cope with high-dimensional data, recent proposals include projection pursuit techniques [11, 12], a combination of projec-

tion pursuit and robust covariance estimation [13], estimators based on minimizing a robust scale of the residuals [14] and spherical PCA for functional data [15]. However, much less research has been done to robustify the PARAFAC method.

For PARAFAC, two procedures for dealing with outliers in multi-way data, are proposed in [2, 8] and [16]. However, they do not come up to the mark. Both approaches work well in some particular situations, but in general they do not succeed in detecting all the anomalous samples. In [2, 8] a measurement of the influence of a sample on the estimates, and the residuals that reflect the distance between the observed data point and the estimated value, are used to mark outlying samples. If the residual and/or the influence is too large, this sample is excluded from further computations. However, the fit itself can be attracted by the outliers. As such, the outliers can have a small residual and/or a small influence and consequently the outliers can not be found anymore.

In [16] the outlier detection method is based on jack-knifing. For each observation the scores and loadings computed for the full data are compared with those estimated from the data minus that observation. Outliers are then indicated as samples for which both scores or loadings differ significantly. This technique works fine for single outliers, but suffers from masking when the outliers appear as groups.

Our proposal is inspired by [17], which presents a robust Tucker3 algorithm. This method starts by searching for an ideal subset of the samples, such that possible outliers can be excluded from the computations. Here, 'ideal' means the subset for which the sum of the squared residuals is minimal. Then, the classical Tucker3 method is performed on this ideal subset and residuals are computed for all samples using the obtained estimates. Finally, points are flagged as outlier if their residual is too large. This approach seems to work very well in most cases, although some disadvantages can be found. First, the choice of the initial subset seems not optimal, because it contains the observations for which the determinant of the covariance matrix is minimal, whereas the objective is to minimize the points for which the sum of the squared residuals is minimal. Secondly, the B - and C -loadings are not robustly initialized. Finally, the cutoff value for residuals to be flagged as outliers is based on the assumption that the residuals are normally distributed. A weighted χ^2 -distribution is probably a more accurate estimate of the distribution of the residuals, which is also suggested in [8, p. 170, 297]. Moreover it is shown in [13] that a cutoff based on this weighted χ^2 -distribution works well for two-way data. To summarize, the main ideas of this robust

Tucker3 method are fine, but some details have not been fully worked out.

In our robust alternative for PARAFAC we try to overcome the shortcomings of the robust Tucker3 algorithm. We start by explaining the PARAFAC model and the alternating least squares algorithm in Section 2. Our robust alternative is described in Section 3. In Section 4 a diagnostic plot to identify outliers is introduced. The results of a simulation study are presented in Section 5, whereas the new algorithm is illustrated on real-life data sets in Section 6.

Throughout the paper, we use the following notation (as in [8]). Arrays, matrices and vectors are always written in bold face. The number of samples is denoted by I . For each sample a $J \times K$ matrix \mathbf{X}_i is available, which corresponds to the measurements of a set of J and another set of K variables. By stacking these matrices we obtain the three-way data array $\underline{\mathbf{X}}$. When this three-way data cube $\underline{\mathbf{X}}$ is unfolded to a matrix, we write \mathbf{X} together with superscripts that indicate how the unfolding is performed, such as $\mathbf{X}^{(I \times JK)}$. The vec -operator $\text{vec}(\mathbf{X})$ is used to represent the vector obtained by unfolding a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \dots,]$ column-wise to one column.

2 The classical PARAFAC algorithm

To construct a PARAFAC model for three-way data an $(I \times F)$ score matrix \mathbf{A} and $(J \times F)$ - and $(K \times F)$ -loading matrices \mathbf{B} and \mathbf{C} are defined, such that the unfolded matrix $\mathbf{X}^{I \times JK}$ can be decomposed as:

$$\mathbf{X}^{I \times JK} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})' + \mathbf{E}^{I \times JK}. \quad (1)$$

In here, F is the number of factors to include in the model, $\mathbf{E}^{I \times JK}$ is the unfolded error term and $\mathbf{C} \odot \mathbf{B}$ is defined as $\mathbf{C} \odot \mathbf{B} = [\text{vec}(\mathbf{b}_1 \mathbf{c}'_1), \dots, \text{vec}(\mathbf{b}_F \mathbf{c}'_F)]$ (called the Kathri-Rao product).

An estimate of the i th observation for the j th variable and the k th occasion is given by:

$$\hat{x}_{ijk} = \sum_{f=1}^F \hat{a}_{if} \hat{b}_{jf} \hat{c}_{kf},$$

which is equivalent to the matrix notation $\hat{\mathbf{X}} = \hat{\mathbf{A}}(\hat{\mathbf{C}} \odot \hat{\mathbf{B}})'$. Using $\hat{\mathbf{X}}$, residuals can be defined as the difference between the estimated and the observed data. The residual for

observation i is thus given by $\mathbf{R}_i = \mathbf{X}_i - \hat{\mathbf{X}}_i$. The Frobenius norm of \mathbf{R}_i is called the residual distance (RD) and equals:

$$RD_i = \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|_F \quad (2)$$

$$= \sqrt{\sum_{j=1}^J \sum_{k=1}^K (x_{ijk} - \hat{x}_{ijk})^2}. \quad (3)$$

The scores and loadings of the PARAFAC model are estimated by minimizing the objective function:

$$\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2 = \|\mathbf{X} - \hat{\mathbf{A}}(\hat{\mathbf{C}} \odot \hat{\mathbf{B}})'\|_F^2, \quad (4)$$

$$= \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (x_{ijk} - \hat{x}_{ijk})^2, \quad (5)$$

which is equivalent with minimizing the sum of the squared residual distances. An algorithm based on alternating Least Squares (ALS) is frequently used for this purpose (see e.g. [8]). This means that given initial estimates for \mathbf{B} and \mathbf{C} , \mathbf{A} is estimated conditionally on \mathbf{B} and \mathbf{C} by minimizing (4). If we define $\mathbf{Z} = (\mathbf{C} \odot \mathbf{B})$ the optimization problem can be reduced to minimizing $\|\mathbf{X} - \mathbf{AZ}'\|_F^2$, which gives rise to the classical least squares regression problem. A least squares estimate for \mathbf{A} is therefore given by $\hat{\mathbf{A}} = \mathbf{XZ}(\mathbf{Z}'\mathbf{Z})^+$ with $(\mathbf{Z}'\mathbf{Z})^+$ the Moore-Penrose inverse of $(\mathbf{Z}'\mathbf{Z})$. Estimates for \mathbf{B} and \mathbf{C} are found analogously, which leads to the following algorithm:

1. Initialize \mathbf{B} and \mathbf{C} (see [8] for more details).
2. $\mathbf{Z} = (\hat{\mathbf{C}} \odot \hat{\mathbf{B}})$
 $\hat{\mathbf{A}} = \mathbf{X}^{(I \times JK)} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^+$
3. $\mathbf{Z} = (\hat{\mathbf{C}} \odot \hat{\mathbf{A}})$
 $\hat{\mathbf{B}} = \mathbf{X}^{(J \times IK)} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^+$
4. $\mathbf{Z} = (\hat{\mathbf{A}} \odot \hat{\mathbf{B}})$
 $\hat{\mathbf{C}} = \mathbf{X}^{(K \times IJ)} \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^+$
5. Go to step 2 until the relative change in the objective function (4) is smaller than a predefined constant (e.g. $1e^{-6}$).

It is a fairly easy-to-understand algorithm, but the least squares estimates in step 2, 3 and 4 will be heavily attracted by outliers. This is already shown for linear regression in e.g. [18] and it will be pointed out in the simulations of Section 5 and in the examples of Section 6. How this corruption can be prevented is explained in the next Section, where the robust algorithm is constructed.

3 The robust algorithm

The robustification of the PARAFAC method is broadly made up of three steps. The procedure starts with looking for $\frac{I}{2} < h < I$ points that minimize the objective function (4). On these h points the classical PARAFAC method is performed, followed by a reweighting step to increase the efficiency. Before we explain the algorithm in more detail, we first elaborate on the choice of h . Choosing the value h between half the number of the observations and the total number of observations, provides a tool to exclude possible outliers from the computations. This value of h actually plays the role of a trade-off between robustness and efficiency, because the larger h is, the more statistically efficient the outcome will be, but the less robust, and vice versa. The default value of h is put equal to 75% of the total number of observations, but can be adapted by the user. After fixing this h -value, the Robust PARAFAC algorithm proceeds as follows :

1. In the first step, an initial h -subset is constructed by performing ROBPCA [13], a robust PCA method, on the unfolded data matrix $\mathbf{X}^{I \times JK}$. The h observations with the smallest residuals towards the space spanned by the robustly estimated principal components, are taken as initial guess for the final h -subset, that should minimize (4). The number of components to retain in ROBPCA are chosen by looking at the screeplot and the robust PRESS [19].
2. The classical PARAFAC algorithm is then executed on these h points, resulting in loadings $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ and scores $\hat{\mathbf{A}}_h$.
3. In the next step, scores need to be computed for each observation \mathbf{X}_i . We use the approach proposed in [8] for determining scores of a new sample \mathbf{X}_i , which is a $(J \times K)$ -matrix. First we unfold \mathbf{X}_i to a column vector by putting all the columns of the matrix

below each other, denoted by $\text{vec}(\mathbf{X}_i)$. From (1) it follows that :

$$\text{vec}(\mathbf{X}_i) \sim (\hat{\mathbf{C}} \odot \hat{\mathbf{B}}) \mathbf{a}'_i,$$

from which we can estimate the scores $\hat{\mathbf{a}}_i$ by

$$\hat{\mathbf{a}}_i = (\hat{\mathbf{C}} \odot \hat{\mathbf{B}})^+ \text{vec}(\mathbf{X}_i).$$

Define $\hat{\mathbf{A}}$ as the matrix with $\hat{\mathbf{a}}'_i$ on the rows.

4. The unfolded data \mathbf{X} is estimated by :

$$\hat{\mathbf{X}} = \hat{\mathbf{A}}(\hat{\mathbf{C}} \odot \hat{\mathbf{B}})',$$

and residual distances RD_i are computed using (2).

5. A new h -subset is now created by storing the h samples with smallest residual distance.
6. The whole procedure starting from step 2 is iterated until the members of the h -subset do not change anymore. Remark that convergence to the global minimum must be attained in the classical PARAFAC procedure in step 2 of this algorithm. It sometimes requires to decrease the default stop-criteria of the PARAFAC algorithm. If this is not taken into account, similar h -subsets can give rise to different PARAFAC models. The relative change in fit then becomes large, which finally results in a non-convergent robust algorithm.
7. Finally, a reweighting is performed, where the classical PARAFAC model is applied on all the observations for which the residual is small enough. The cutoff is based on the assumption that the residual distances to the power $2/3$ are approximately normally distributed [13]. Remark that in [8] a cutoff-value is suggested using a χ^2 -distribution, but after applying the Wilson-Hilferty approximation of the χ^2 -distribution, the same cutoff is obtained (see [20, 21]).

In this procedure the drawbacks of the algorithm of [17] are circumvented. The choice of the initial h -subset here is more inspired by the objective function than the one used in [17]. Also the initialization is more robust, because the first time the classical PARAFAC method is applied, only h samples are taken into account. Moreover, by including a variable number

h , the user can define the level of robustness or efficiency instead of fixing $h = 51\%I$. Finally, the cutoff has been adapted to the one in correspondence with the underlying χ^2 -distribution.

Note that the convergence of the algorithm can be guaranteed. Suppose that we denote the objective function (4) for a h -subset H by $\sum_{i \in H} RD_i^2(\mathbf{A}_H, \mathbf{B}_H, \mathbf{C}_H)$, with \mathbf{A}_H , \mathbf{B}_H and \mathbf{C}_H the scores and loadings obtained by applying PARAFAC on the h -subset H . We also introduce H_m , which is the h -subset obtained in the m th iteration of the algorithm. Then, it holds that

$$\sum_{H_m} RD_i^2(\mathbf{A}_{H_{m-1}}, \mathbf{B}_{H_{m-1}}, \mathbf{C}_{H_{m-1}}) \leq \sum_{H_{m-1}} RD_i^2(\mathbf{A}_{H_{m-1}}, \mathbf{B}_{H_{m-1}}, \mathbf{C}_{H_{m-1}}), \quad (6)$$

because the newly created h -subset in the m th iteration step of the algorithm contains the h points with smallest residual distance. From the properties of a least squares estimator, it also follows that :

$$\sum_{H_m} RD_i^2(\mathbf{A}_{H_m}, \mathbf{B}_{H_m}, \mathbf{C}_{H_m}) \leq \sum_{H_m} RD_i^2(\mathbf{A}_{H_{m-1}}, \mathbf{B}_{H_{m-1}}, \mathbf{C}_{H_{m-1}}). \quad (7)$$

Combining (6) and (7) leads to

$$\sum_{H_m} RD_i^2(\mathbf{A}_{H_m}, \mathbf{B}_{H_m}, \mathbf{C}_{H_m}) \leq \sum_{H_{m-1}} RD_i^2(\mathbf{A}_{H_{m-1}}, \mathbf{B}_{H_{m-1}}, \mathbf{C}_{H_{m-1}}).$$

This means that updating the h -subset implies a decrease in the objective function. In addition, the objective function is bounded below by 0 and because only a finite number of h -subsets can be taken, the convergence of the algorithm is ensured. This approach is highly comparable to the C -step technique used in e.g. the FAST-MCD algorithm [22]. On the other hand, in general we can not be sure that the final h -subset is the optimal one minimizing (4), since only one h -subset is considered to start the iterations from. Another chosen initial h -subset could perhaps lead to a final h -subset for which the objective function is even smaller.

From the simulation study in Section 5 and the analysis of Section 6 we can however conclude that taking this single h -subset is enough to ensure the robustness of the results. In the next section, we first elaborate on the different types of outlying samples and we introduce a graphical tool for outlier identification.

4 The outlier map

It is possible to divide all the observations into four classes: the regular samples, the residual outliers, the good leverage points and the bad leverage points. For this classification, we have used a similar approach as in robust regression (see e.g. [23]) and in the robust principal components analysis on two-way data (see e.g. [13, 12]). Two distances will characterize the type of the observation: the residual distance (RD) and the score distance (SD). The residual distance has been introduced in (2) and can be considered as a measure of how well the fitted data and the observed data collapse. The second distance is the score distance (SD). It measures how much the estimated scores of a sample deviate from the center of the scores. The score distance of an observation \mathbf{X}_i is computed by taking the robust counterpart of the Mahalanobis distance of the score $\hat{\mathbf{a}}_i$ to the center of the scores matrix $\hat{\mathbf{A}}$:

$$SD_i = \sqrt{(\hat{\mathbf{a}}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\mathbf{a}}_i - \hat{\boldsymbol{\mu}})},$$

with $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ robust estimates of the center and covariance matrix of the scores respectively, which are provided by the Minimum Covariance Determinant (MCD) method [18].

Observations are considered as outliers if their score distance or residual distance is too large. For the score distance we have adopted the cutoff values from the two-way situation, i.e. $\sqrt{\chi_{0.975, F}^2}$, the 97.5% quantile of the χ^2 -distribution with F degrees of freedom. This stems from the assumption that the scores are approximately normally distributed and thus that the squared score distances are asymptotically χ^2 -distributed. As a consequence, for scores that do not follow a normal distribution, this cutoff must be taken with caution. For the residual distances on the other hand, the cutoff-value becomes $(m + s z_{0.975})^{3/2}$ with $z_{0.975} = \Phi^{-1}(0.975)$ the 97.5% quantile of the standard normal distribution and m and s^2 the estimated mean and variance of the residual distances respectively, robustly estimated by the univariate MCD procedure. This is in correspondence with the assumed underlying normal distribution of the residual distances to the power 2/3, like stated before.

With both cutoff values, we can classify the different types of outliers. A *regular point* is a sample for which the residual distance as well as the score distance are smaller than the corresponding cutoff value. Observations which have a highly deviating underlying structure, but a small score distance are called *residual outliers*. The leverage points have a large score distance, but the residual distance is small for *good leverage* points, whereas a large residual

distance typifies the *bad leverage* points. A summary can be found in Table 1.

Table 1 should be placed here.

To visualize this classification, an outlier map can be made, where the residual distance is plotted against the score distance, together with a vertical and horizontal line for the cutoff values. To allow comparison between the robust and classical PARAFAC method, a classical diagnostic plot, which is equivalent with the plot proposed in [2], is also introduced in the examples in Section 6. In here, both distances and their cutoff values are computed using classical mean and covariance estimators, instead of robust ones.

5 A simulation study

The aim of the simulations is two-fold. Firstly, the classical and robust PARAFAC methods are compared on data sets with and without contamination. Secondly, as the final h -subset on which the PARAFAC algorithm is applied, is crucial for the robustness of the method, we compare our proposal of using only one initial h -subset indicated by ROBPCA (method R1), with a similar algorithm, but where 100 random initial h -subsets are drawn instead of considering only one (method R2). Both aspects are assessed on three-way data generated with the following parameters : the number of observations $I = 50$, the number of variables $J = 100$, the number of occasions $K = 10$, and the number of factors $F = 2$.

The scores \mathbf{A} and loadings \mathbf{B} , and \mathbf{C} are drawn from a multivariate random normal distribution $N_F(\mathbf{0}, \mathbf{\Sigma}_F)$, where $\mathbf{\Sigma}_F$ is a diagonal matrix with $(10 \ 2)$ as diagonal elements. We have rescaled \mathbf{B} and \mathbf{C} such that their Frobenius norm is equal to one in order to make the comparison of the different PARAFAC models easier. This can be done without loss of generality, because of the scale indeterminacy of the PARAFAC model. The pure three-way data are constructed by :

$$\mathbf{X}_{pure}^{(I \times JK)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})',$$

and after adding a noise term $\mathbf{E}^{I \times JK}$, we obtain :

$$\mathbf{X}^{(I \times JK)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})' + \mathbf{E}^{I \times JK}.$$

The noise term is suggested in [24] and is equal to :

$$\mathbf{E}^{I \times JK} = \frac{\text{Noise}\%}{100 - \text{Noise}\%} \|\mathbf{X}_{\text{pure}}^{I \times JK}\|_F \tilde{\mathbf{E}}^{I \times JK}$$

with $\tilde{\mathbf{E}}^{I \times JK} \sim N(\mathbf{0}, \Sigma_{\tilde{\mathbf{E}}})$ and normalized to the Frobenius norm of 1. We have set $\Sigma_{\tilde{\mathbf{E}}}$ equal to the identity matrix. The factor $\|\mathbf{X}_{\text{pure}}^{I \times JK}\|_F$ is added to normalize the error terms $\tilde{\mathbf{E}}$ to the Frobenius norm of $\mathbf{X}_{\text{pure}}^{I \times JK}$. By using this construction, the Noise% reflects the percentage of noise in the total variance of the data \mathbf{X} . In all the settings, we used Noise% = 20%.

We have considered different settings of outliers. At first, we did not include any outliers and thus clean data are investigated.

In a second setting, we consider data with 10% and 20% good leverage points. This type of outliers is obtained by multiplying 10% or 20% of \mathbf{X}_{pure} with a constant c_1 , which leads to the final contaminated data \mathbf{X} after adding noise to \mathbf{X}_{pure} . We have chosen to take $c_1 = 10$. To verify that we have indeed generated good leverage points, we have made an outlier map in Figure 1 of the simulated data points with 20% contamination using the known scores and loadings. Clearly, the 10 outliers are marked as good leverage points. A good leverage point can thus be considered as a point that can be described with highly deviating scores, but for which the underlying structure can be described by the loadings, that also fit the regular data well.

Figure 1 should be placed here.

In a third situation 10% and 20% bad leverage points have been included in the simulated data by generating 10% or 20% good leverage points and afterwards adding a constant c_2 to these outlying observations. In our simulation setting the constant for the good leverage points $c_1 = 10$, and $c_2 = 0.1$. Again the outlier map for the setting with 20% outliers in Figure 2 shows that we have generated bad leverage points. Bad leverage points are thus characterized by anomalous underlying data structures and scores.

Figure 2 should be placed here.

Finally, we have also considered data with 10% and 20% residual outliers. These data are generated by adding a constant $c_2 = 0.1$ to 10% or 20% of \mathbf{X} . The outlier map in Figure 3

confirms the presence of residual outliers in this type of data set. The construction of such contamination involves samples that have scores in the range of the regular ones, but with a different underlying structure.

Figure 3 should be placed here.

In all the settings we have included at most 20% outliers. Therefore, we do not adapt the default value of $h = \lfloor 0.75I \rfloor$, which means that $h = 37$. Moreover, for all the different types of data sets, we have made a R-PRESS plot to decide on the number of components to retain in ROBPCA in the first step of the algorithm. It turned out that two or three components are optimal. The initial h -subsets for 2 and 3 components differ at most in two regular points. So, both choices do not yield important differences in the results. Therefore, we have fixed the number of components equal to $F = 2$.

We have generated $m = 100$ data sets and computed for each data set the following items:

- The mean squared error (MSE):

$$MSE = \frac{1}{wJK} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K w_i (x_{ijk} - \hat{x}_{ijk})^2 \quad (8)$$

with $w = \sum_{i=1}^I w_i$ and $w_i = 0$ if sample i is an outlier and else $w_i = 1$. Hence, the MSE is only computed on the regular points.

- The angle between the estimated subspace and the true subspace spanned by the B -loadings. The subspace angle between two linear subspaces B_1 and B_2 is defined as

$$\max_{\mathbf{b}_1 \in B_1} \min_{\mathbf{b}_2 \in B_2} \arccos(\mathbf{b}'_1 \mathbf{b}_2).$$

It represents the largest angle between a vector in B_1 and the vector most parallel to it in B_2 , and can be computed in MATLAB[®] with the 'subspace' function (see also the 'maxsub' criterion used in [13]). This subspace angle has to be as small as possible and is reported in radians.

- Analogously we computed the angle between the estimated subspace and the true subspace spanned by the C -loadings.

Remark that in order to compare the mean squared error for the scores and loadings, the permutation indeterminacy and scaling indeterminacy, which is inherent for the PARAFAC model, should be solved first. To remove the scale indeterminacy \mathbf{B} and \mathbf{C} are normalized as mentioned earlier. To ascertain the permutation freedom, the order of the factors is fixed by assuming decreasing variability in the A mode.

In Figures 4-9 we have plotted boxplots of all the simulation results. We also computed the average computation time for each simulation setting and each PARAFAC method. This computation time did not highly depend on the configuration of the outliers. Using our MATLAB implementation we roughly needed 0.5 seconds for the classical PARAFAC method, 5 seconds for our robust method with one initial h -subset, and around 250 seconds for the robust method with 100 random h subsets.

Figures 4-6 should be placed here.

The simulation results for the data without outliers can be found in the three most left boxplots of Figures 4, 5 and 6. We see that the classical and the two robust procedures fit the data almost equally well. Not surprisingly the classical method is in favor here because of the computation time, and its slightly better median performance.

Figures 4, 5 and 6 also show the outcomes for data with good leverage points. Both for the 10% and the 20% outliers, the results of all methods are still robust. We see a small increase in the MSE. This is essentially because in (8) we only consider the regular observations, and hence the MSE is computed on a smaller data set (similar results were obtained by setting $c_1 = 0.1$). The estimates of the loadings are not worse, as can be seen in Figures 5 and 6. Classical PARAFAC and the robust approach with many random subsets (R2) even yield better estimates for the B and the C loadings. This does not come as a surprise. Good leverage points are outlying from the bulk of the data, but they still follow the model. Hence including them in the estimation procedure increases the precision of the estimates. Our robust method with one initial subset (R1) excludes the good leverage points, and consequently yields estimates which are comparable with those of uncontaminated data. We can conclude that both the classical and robust methods can deal with good leverage points and fit the data well. Using many initial h -subsets leads to slight improvements, but at the cost of a huge computation time. By comparing the classical and robust outlier map (not

shown here), the classical procedure fails in marking the good leverage points as outliers. Therefore, the robust version is still highly recommended.

Figures 7-9 should be placed here.

The simulation results for the bad leverage points and the residual outliers are depicted in Figures 7, 8 and 9. We see that classical PARAFAC is influenced by any type and amount of contamination. Also the robust method with many initial random h -subsets (R2) faces many problems. With bad leverage points in the data, the R2 method is still often very robust. But in some cases it yields high MSE values, and badly estimated loadings. With residual outliers, R2 significantly decreases. This shows that considering 100 random h -subsets to determine the final h -subset, is not a good approach. The chance of including outliers in the random subsets h -subsets (and consequently in the final h -subset) is too large. Increasing the number of random subsamples could solve the problem, but would lead to a very time consuming algorithm. Instead, it is much more interesting to use the robust PARAFAC method with one, well chosen, robust initial subset (R1). The simulation results shows that this method is always able to fit the loadings well. Moreover, its computation time is still reasonable.

6 Examples

In this section we consider two data sets on which we perform both the classical and the robust PARAFAC algorithm. By means of the first data, we illustrate that the robust algorithm works well in case of no contamination. The second data set shows the difference between the classical and robust algorithm when severe outliers are involved.

6.1 The Aspirin data

The Aspirin data are kindly provided by Professor da Silva from the Faculty of Science at the University of Porto and is fully described in [25]. The three-way data set describes the synchronous fluorescence spectra of mixtures of the three major metabolites of acetylsalicylic acid. There are 41 synchronous fluorescence spectra measured for 24 pH values and 12

different concentrations of the three acids, giving a data array of size $(12 \times 24 \times 41)$. For a full chemical background of the experiment we refer to [25]. The data are examined without preprocessing and three factors were used because the mixtures consist of three analytes.

A classical and robust PARAFAC analysis were performed on the data and outlier maps are displayed in Figure 10.

Figure 10 should be placed here.

It can be seen that classical PARAFAC classifies sample 1 as a residual outlier, whereas our robust PARAFAC method flags the first three observations as residual outliers. However, from both outlier maps it is also evident that none of the samples is strongly outlying. It rather seems that the robust cutoff value for the residual distances is estimated too conservative (it is only based on $h = 9$ values). This immediately raises the question whether the robust results are precise enough as the final reweighting step in the robust PARAFAC method excludes these three observations. This can be validated by comparing the classical and the robust loadings. The classical pH-profiles and the differences between the robust and classical pH-profiles are depicted in Figure 11. They are clearly very similar (taking into account the different scales of the vertical axes). Also from the spectra, which can be seen in Figure 12, only minor differences between the classical and robust loadings can be detected. Only a small discrepancy is situated in the gentisic acid spectrum. With the classical estimates, the curve goes below zero around wavelength 300, whereas the robust estimates yield a curve that is smoothly increasing from wavelength 250 to 350. As the intensity of light is non-negative, the robust method is certainly not performing worse than the classical one. This example thus again illustrates that the robust PARAFAC works well for uncontaminated data.

Figure 11 should be placed here.

Figure 12 should be placed here.

6.2 The Dorrit data

The Dorrit data [26] is a laboratory-made fluorescence data set, where four fluorophores are mixed together for different sets of concentrations. The four chemical analytes are

phenylalanine, 3, 4-dihydroxyphenylalanine (DOPA), 1, 4- dihydroxybenzene and tryptopan. The data set contains 27 excitation-emission (EEM) landscapes measured with a Perkin-Elmer LS50 B fluorescence spectrometer with emission wavelengths ranging from 250 nm to 482 nm every 2 nm for excitation at wavelengths from 200 nm to 315 nm at 5 nm intervals. The noisy parts due to the condition of the Xenon lamp and the physical environment, situated at the excitation wavelengths from 200 - 230 nm and at emission wavelengths below 250 nm, are excluded before the PARAFAC models are built. This means that we end up with a $(27 \times 116 \times 18)$ data array.

In all the samples severe Rayleigh scattering is present. Consequently, the data set does not contain an outlier-free subset and our robust PARAFAC method will not work appropriately. For that reason, we have replaced the scattered areas by interpolated values using the automated scatter identification method proposed in [27].

Figure 13 should be placed here.

From previous investigations [26], it is known that four components should be used in the PARAFAC model, and that the loadings, that should be obtained, are as depicted in Figure 13. Moreover, the data has also been investigated on the presence of outliers in [16] and observations 2, 3, 5 and 10 were marked as anomalous points. This a priori knowledge makes this data highly suitable to assess the proposed robust algorithm versus the classical method.

Figure 14 should be placed here.

Our analysis starts by examining the presence of outliers in this data set. The outlier maps based on classical PARAFAC and robust PARAFAC are shown in Figure 14. On the classical outlier map, observation 10 is marked as a residual outlier and samples 2, 3, 4 and 5 are flagged as good leverage points. The robust method identifies different outliers: samples 2, 3 and 5 are indicated as bad leverage points and samples 1, 4, 12 and 27 as good leverage points. In order to find out which of both methods we should trust, the emission and excitation loadings are plotted in Figure 15. It is obvious that the classical method is corrupted by the outliers, as the emission loadings nor the excitation loadings are correct, compared to the reference loadings of Figure 13. Also the angle between the estimated and

the reference subspaces for the B -loadings 1.34 and the C -loadings 0.44 confirms the wrong classical estimates. On the other hand, based on visual inspection of Figure 15 and angles equal to 0.057 for the B -loadings and 0.17 for the C -loadings, we can conclude that the robust algorithm succeeded in estimating the underlying structure of the data very accurately.

Figure 15 should be placed here.

This discrepancy between the robust and the classical results can be understood by the corruption of the least squares PARAFAC algorithm due to the outliers. Samples 2, 3 and 5 have a large influence on the estimates of the classical model parameters, such that the final classical model approaches the underlying structure of these three outliers better than it fits the majority of the other points. This is reflected in a large score distance combined with a small residual distance. When the effect of these three outliers is removed in the robust algorithm, the three good leverage points become bad leverage points. This indicates that these three samples have a highly deviating underlying structure compared with the other samples, and should therefore indeed be considered as bad leverage points.

There is only one small problem with observation 10. On the classical outlier map and in [16], this sample is marked as a residual outlier, whereas on the robust outlier map in Figure 14 sample 10 is flagged as a border case. It exceeds the cutoff value for the residual distance, but is not really deviating from the other samples. Hence sample 10 is more like a border case than an unambiguous outlier. This is not in correspondence with the results of [16]. To clarify which analysis should be trusted, we have computed the classical PARAFAC angles between the subspaces spanned by the reference loadings, depicted in Figure 13 and by the B - and C -loadings for the data minus observations 2, 3 and 5, which results in 0.0524 for the B -loadings and 0.1589 for the C -loadings. The same is done for the data without observations 2, 3, 5 and 10, leading to a B -angle of 0.0559 and a C -angle of 0.1669. These angles are all similar and moreover comparable to the angles of the robust analysis (0.0569 and 0.1777 respectively). This confirms that sample 10 is flagged correctly by the robust outlier map as a point with only a minor influence on the PARAFAC estimates. Also on the Resample Influence Plot (RIP) of [16], sample 10 is misclassified. This can be explained as follows. In the RIP plot, for each observation the difference between the squared residual distance of that sample is plotted versus the estimated loadings on the full data and on the data minus that point. If this results in a large loading difference and/or a large squared

residual distance, the sample is marked as an outlier. Only deleting sample 10 from the data provides a wrong model, because the data is still contaminated by observations 2, 3 and 5. Therefore, the large B -loading difference for sample 10 in the RIP plot in [16] may not be interpreted as observation 10 being an outlier. For this reason, the RIP is not always the best choice as an outlier identification tool when more than one highly deviating point is present in the data. So, we can summarize that samples 2, 3 and 5 are the only outliers in the data. This is correctly detected by our robust PARAFAC algorithm.

7 Conclusion and Outlook

We have developed a robust PARAFAC method, which can deal with the pernicious effects of outlying samples. Our method searches for a subset of the data that minimizes the sum of the squared residual distances. An initial subset is found by unfolding the three-way data and applying a robust PCA on the unfolded data matrix. Next, an iterative algorithm is applied to improve the objective function. Furthermore, a robust diagnostic plot has been introduced as a graphical device for outlier detection. Simulations and analysis on real data have shown that the proposed method clearly outperforms the classical PARAFAC technique in case of contaminated data sets. Moreover, also for clean data sets, the robust PARAFAC method estimates the loadings and scores well and in a reasonable time.

This robust PARAFAC method opens opportunities for the development of other robust procedures for multiway data. It can be used to detect Rayleigh and Raman scatter in fluorescence data in an automated way [27]. When both scatter and outlying excitation-emission landscapes are present in the data, a fully robust PARAFAC method has been developed, which highly relies on the methodology presented in this paper [28]. It will also be interesting to study how the combination of our robust PARAFAC method with robust SIMCA [29] can lead to a robust classifier of multiway data, similar to the work presented in [30] for uncontaminated groups.

Matlab code to perform our robust PARAFAC method can be downloaded at

<http://wis.kuleuven.be/stat/robust>.

The program needs the LIBRA toolbox [31], the PLS toolbox 3.5 (or higher) [32] and some

functions of the N-way toolbox [33].

Acknowledgements: The authors would like to acknowledge Rasmus Bro, Randy Pell and Karlien Vanden Branden for providing constructive comments on this paper. Moreover, we thank Professor da Silva for providing the Aspirin data set.

References

- [1] R. Bro. *Multi-way Analysis in the Food Industry*. PhD thesis, Royal Veterinary and Agricultural university, Denmark, 1998.
- [2] C.M. Andersen and R. Bro. Practical aspects of PARAFAC modelling of fluorescence excitation-emission data. *Journal of Chemometrics*, 17:200–215, 2003.
- [3] R. Bro. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics and Intelligent Laboratory Systems*, 46:133–147, 1999.
- [4] R.D. Jiji, G.G. Andersson, and K.S. Booksh. Application of PARAFAC for calibration with excitation-emission matrix fluorescence spectra of three classes of environmental pollutants. *Journal of Chemometrics*, 14:171–185, 2000.
- [5] L.R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.
- [6] R. Leardi, C. Armanino, S. Lanteri, and L. Alberotanza. Three-mode principal component analysis of monitoring data from venice lagoon. *Journal of Chemometrics*, 14: 187–195, 2000.
- [7] G.R. Flaten, B. Grung, and Kvalheim O.M. Quantification of pollution levels by multiway modelling. *Journal of Chemometrics*, 18:173–182, 2004.
- [8] A. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis with Applications in the Chemical Sciences*. Wiley, England, 2004.

- [9] P.M. Kroonenberg. *Applied multiway data analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ, 2008.
- [10] C. Croux and G. Haesbroeck. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87:603–618, 2000.
- [11] C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95:206–226, 2005.
- [12] M. Hubert, P.J. Rousseeuw, and S. Verboven. A fast robust method for principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60:101–111, 2002.
- [13] M. Hubert, P.J. Rousseeuw, and K. Vanden Branden. ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47:64–79, 2005.
- [14] R.A. Maronna. Principal components and orthogonal regression based on robust scales. *Technometrics*, 47:264–273, 2005.
- [15] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, and K.L. Cohen. Robust principal component analysis for functional data. *Test*, 8:1–73, 1999.
- [16] J. Riu and R. Bro. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics and Intelligent Laboratory Systems*, 65:35–49, 2003.
- [17] V. Pravdova, B. Walczak, and D.L. Massart. A robust version of the Tucker3 model. *Chemometrics and Intelligent Laboratory Systems*, 59:75–88, 2001.
- [18] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.
- [19] M. Hubert and S. Engelen. Fast cross-validation for high-breakdown resampling algorithms for PCA. *Computational Statistics and Data Analysis*, 51:5013–5024, 2007.

- [20] G.E.P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in one-way classification. *The Annals of Mathematical Statistics*, 25:33–51, 1954.
- [21] P. Nomikos and J.F. MacGregor. Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37:41-59, 1995.
- [22] P.J. Rousseeuw and K. Van Driessen. A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, 41:212–223, 1999.
- [23] P.J. Rousseeuw and B.C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–651, 1990.
- [24] G. Tomasi and R. Bro. A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics and Data analysis*, 50:1700–1734, 2006.
- [25] J.C.G. Esteves da Silva and S.A.G. Novais. Trilinear PARAFAC decomposition of synchronous fluorescence spectra of mixtures of the major metabolites of acetylsalicylic acid. *The Analyst*, 123:2067–2070, 1998.
- [26] D. Baunsgaard. *Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes*. PhD thesis, Royal Veterinary and Agricultural University, Department of Dairy and Food technology, Frederiksberg, Denmark, 1999.
- [27] S. Engelen, S. Frosch Møller, and M. Hubert. Automatically identifying scatter in fluorescence data using robust techniques. *Chemometrics and Intelligent Laboratory Systems*, 86:35–51, 2007.
- [28] S. Engelen, S. Frosch Møller, and B.M. Jorgensen. A fully robust PARAFAC method for analyzing fluorescence data. *Journal of Chemometrics*, 23:124–131, 2009.
- [29] K. Vanden Branden and M. Hubert. Robust classification in high dimensions based on the SIMCA method. *Chemometrics and Intelligent Laboratory Systems*, 79:10–21, 2005.
- [30] C. Durante, R. Bro and M. Cocchi. A classification tool for N-way array based on SIMCA methodology. *Chemometrics and Intelligent Laboratory Systems*, 106:73–85, 2011.

- [31] S. Verboven and M. Hubert. LIBRA: a Matlab library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75:127–136, 2005.
- [32] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, and R.S. Koch. *PLS_Toolbox 3.5 for use with MATLAB*, 2004. URL <http://software.eigenvector.com/>. Software, Eigenvector Research, Inc., August 2004.
- [33] C.M. Andersen and R. Bro. The N-way toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems*, 52:1–4, 2000.

List of Tables

Table 1: Different types of outlying samples

large	residual outliers	bad leverage samples
small	regular samples	good leverage samples
RD SD	small	large

List of Figures

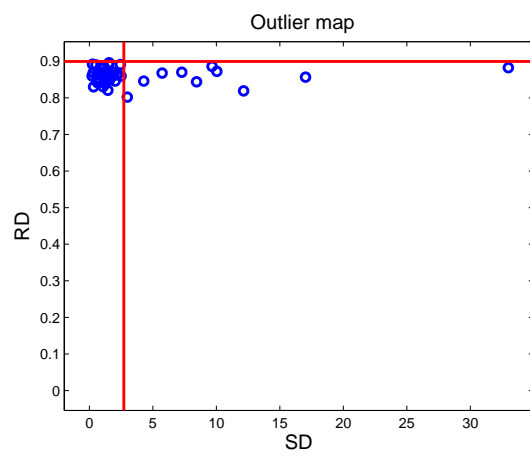


Figure 1: An outlier map for simulated data with 20% good leverage points.

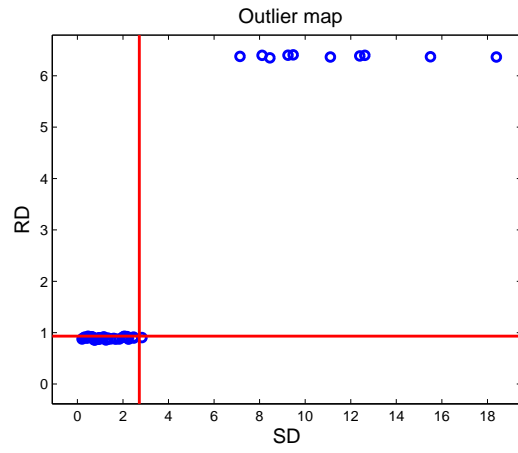


Figure 2: An outlier map for simulated data with 20% bad leverage points.

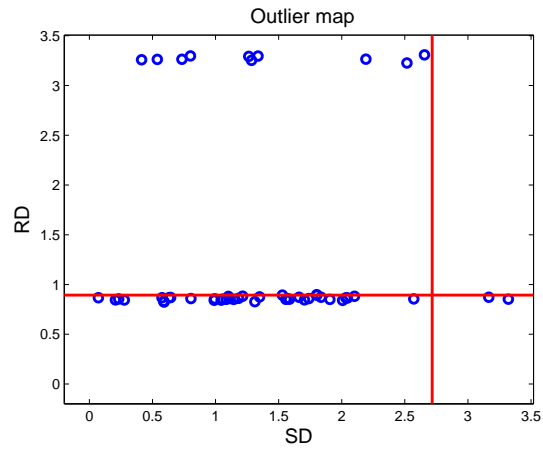


Figure 3: An outlier map for simulated data with 20% residual outliers.

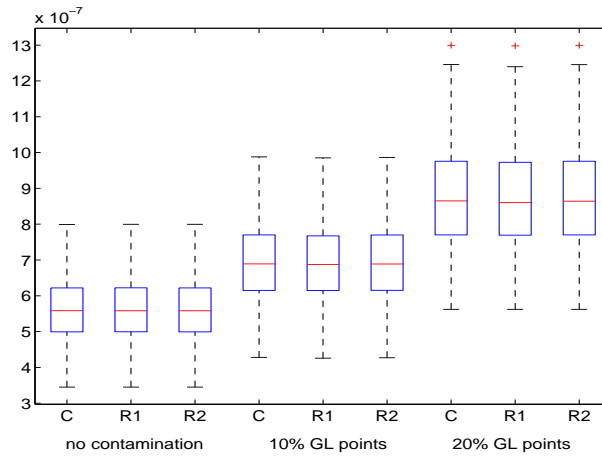


Figure 4: MSE values of classical PARAFAC (C), robust PARAFAC with one initial h -subset (R1) and robust PARAFAC with 100 initial h -subsets (R2) for simulation settings without contamination and with good leverage points.

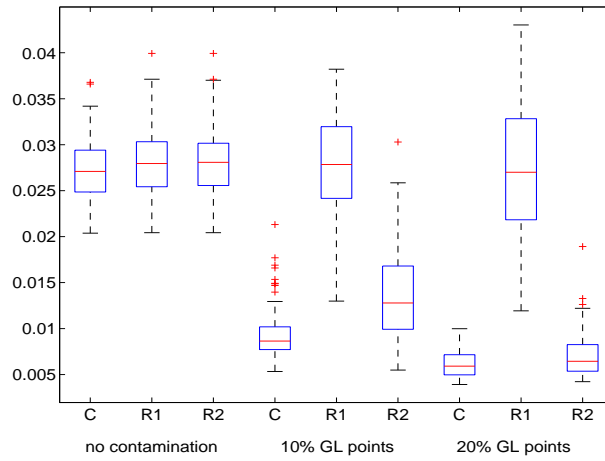


Figure 5: Angle of B -loadings of classical PARAFAC (C), robust PARAFAC with one initial h -subset (R1) and robust PARAFAC with 100 initial h -subsets (R2) for simulation settings without contamination and with good leverage points.

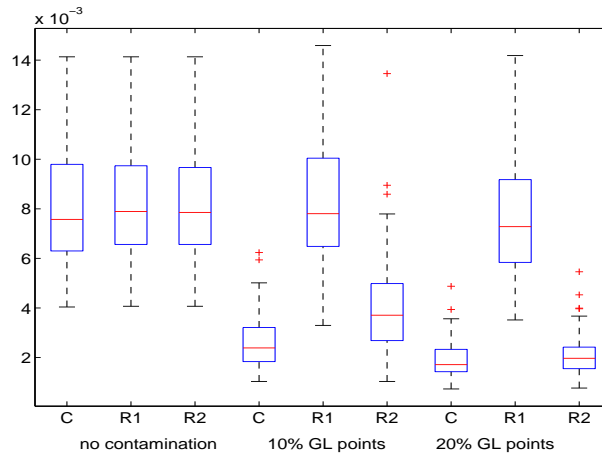


Figure 6: Angle of C -loadings of classical PARAFAC (C), robust PARAFAC with one initial h -subset (R1) and robust PARAFAC with 100 initial h -subsets (R2) for simulation settings without contamination and with good leverage points.

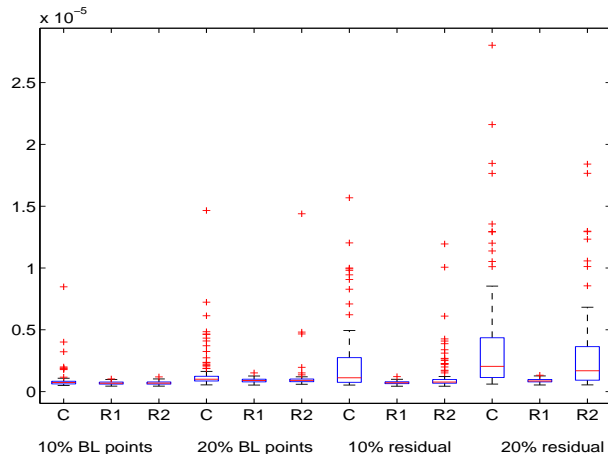


Figure 7: MSE values of classical PARAFAC (C), robust PARAFAC with one initial h -subset (R1) and robust PARAFAC with 100 initial h -subsets (R2) for simulation settings with bad leverage points and residual outliers.

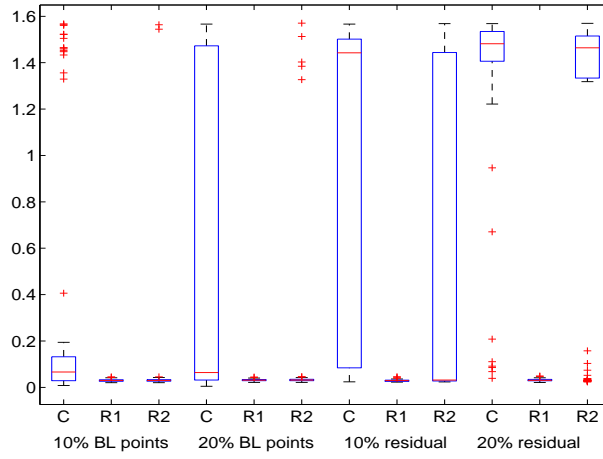


Figure 8: Angle of B -loadings of classical PARAFAC (C), robust PARAFAC with one initial h -subset (R1) and robust PARAFAC with 100 initial h -subsets (R2) for simulation settings with bad leverage points and residual outliers.

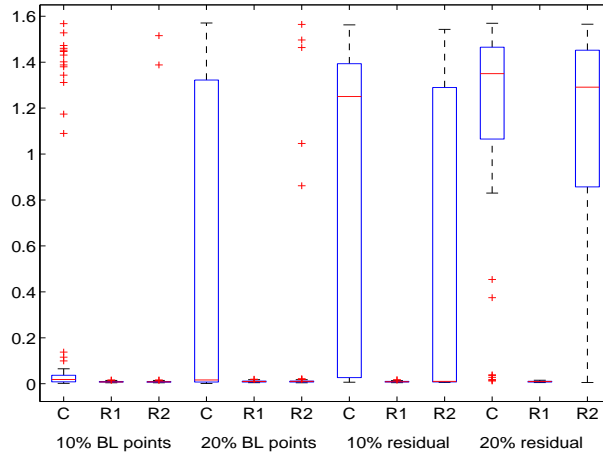


Figure 9: Angle of C -loadings of classical PARAFAC (C), robust PARAFAC with one initial h -subset (R1) and robust PARAFAC with 100 initial h -subsets (R2) for simulation settings with bad leverage points and residual outliers.

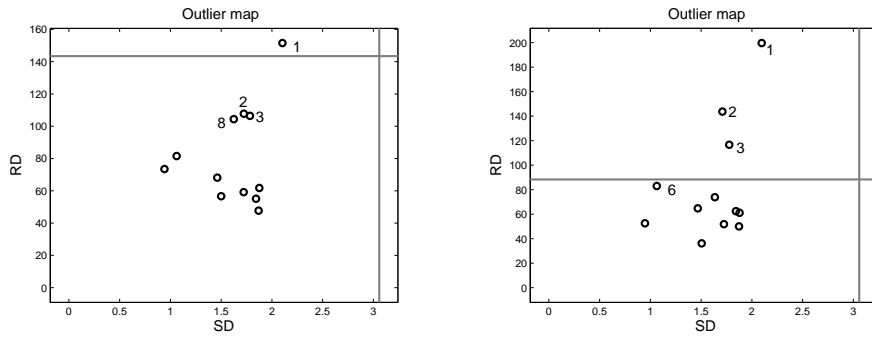


Figure 10: Outlier maps of the Aspirin data set based on classical PARAFAC (left) and robust PARAFAC (right).

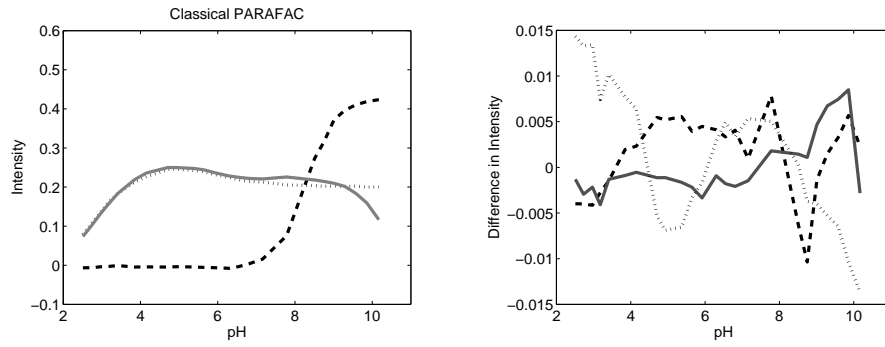


Figure 11: A classical plot of the pH-profiles (left) and the difference between the classical and robust pH-profiles (right) for the aspirin data. The dash-dotted line reflects salicylic acid, the dashed line corresponds to salicylic acid and the gentisic acid is marked with the full line.

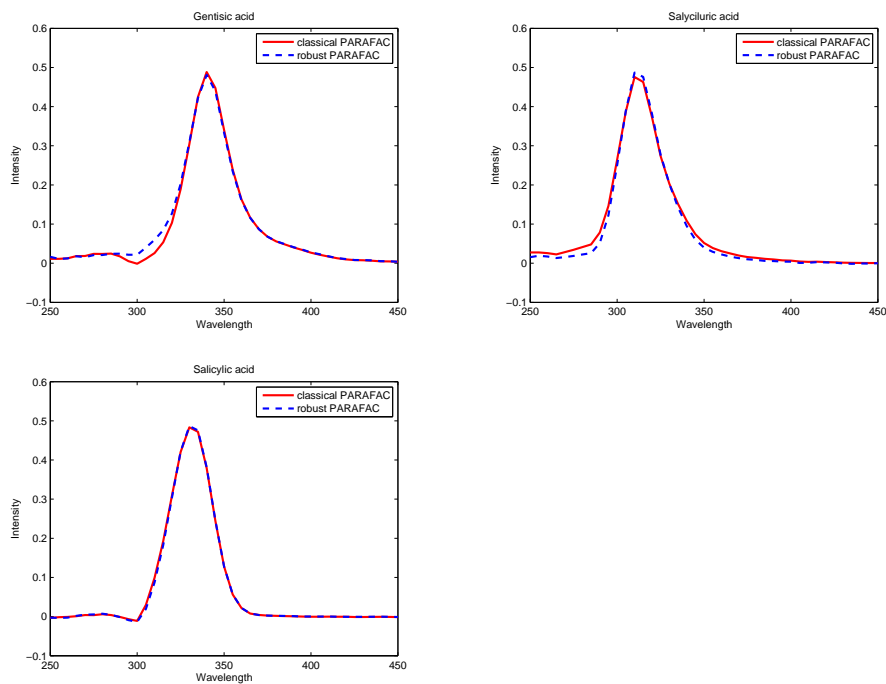


Figure 12: Classical and robust plots of the spectra for the aspirin data.

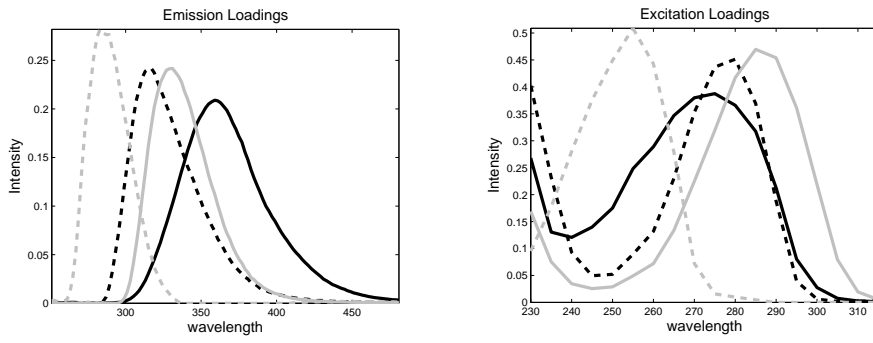


Figure 13: The true emission (left) and excitation (right) loadings for the Dorrit data set.

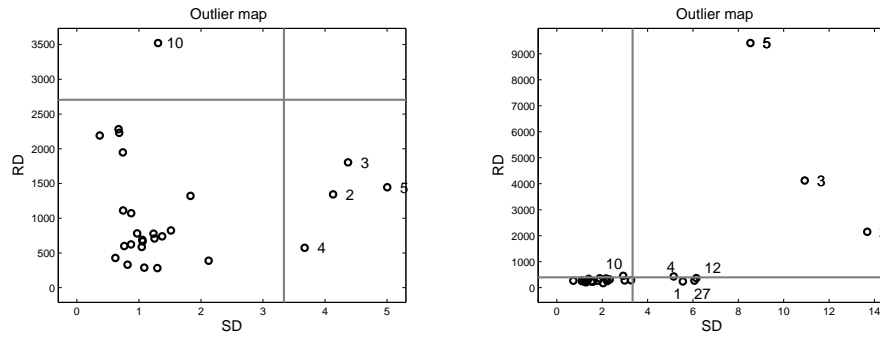


Figure 14: Outlier maps of the Dorrit data set based on classical PARAFAC (left) and robust PARAFAC (right).

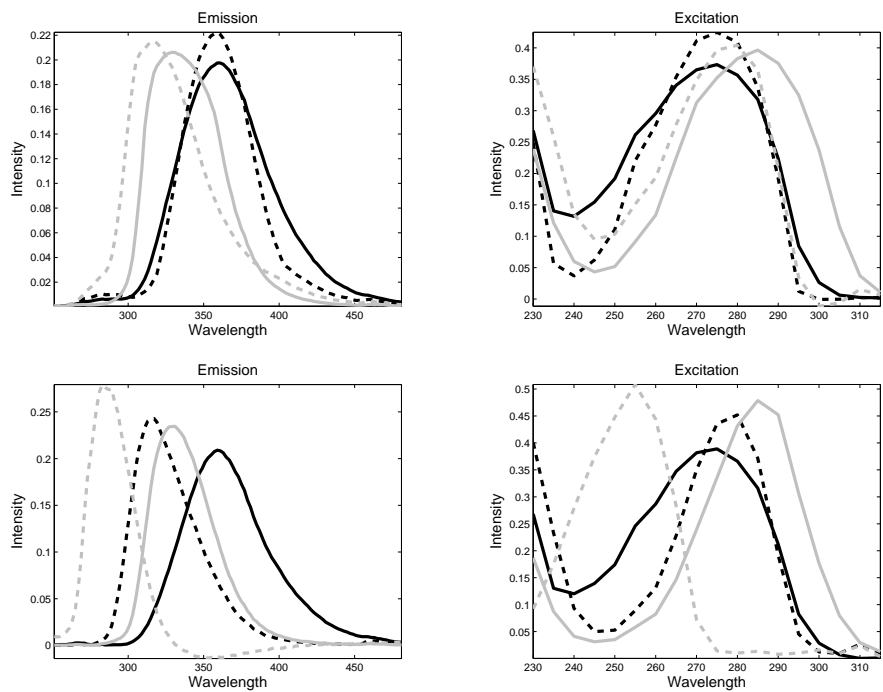


Figure 15: The emission (left) and excitation (right) loadings for the Dorrit data set using the classical (above) and robust PARAFAC (below) algorithm.