

Automatically Identifying Scatter in Fluorescence Data using Robust Techniques.

Sanne Engelen ^{a,*}, Stina Frosch Møller ^b, Mia Hubert ^a

^a*Katholieke Universiteit Leuven, Department of Mathematics, W. De Croylaan 54, 3001 Leuven, Belgium*

^b*Department of Seafood Research, Danish Institute for Fisheries Research, The Technical University of Denmark, Søtofts Plads, Building 221, 2800 Kgs. Lyngby, Denmark.*

Abstract

First and second order Rayleigh and Raman scatter is a common problem when fitting Parallel Factor Analysis (PARAFAC) to fluorescence excitation-emission data (EEM). The scatter does not contain any relevant chemical information and does not conform to the low-rank trilinear model. The scatter complicates the analysis instead and contributes to model inadequacy. As such, scatter can be considered an an example of element-wise outliers. However, no straightforward method for identifying the scatter region can be found in the literature. In this paper an automatic scatter identification method is developed based on robust statistical methods. The method does not demand any visual inspection of the data prior to modeling, and can handle first and second order Rayleigh scatter as well as Raman scatter in various types of EEM data. The results of the automated scatter identification method were used as input data for three different PARAFAC methods. Firstly inserting missing values in the scatter regions are tested, secondly an interpolation of the scatter regions is performed and finally the scatter regions are down-weighted. These results show that the PARAFAC method to choose after scatter identification clearly depends on the data, for example signal to noise ratio and overlap between signal and scatter.

Key words: Raman and Rayleigh scatter, automated method, Robustness, ROBPCA, PARAFAC, Fluorescence

* Corresponding author

Email addresses: sanne.engelen@wis.kuleuven.be (Sanne Engelen), sfr@dfu.min.dk (Stina Frosch Møller), mia.hubert@wis.kuleuven.be (Mia Hubert).

1 Introduction

Fluorescence spectroscopy is a fast, non-destructive technique with high sensitivity and specificity for providing information (quantitative and qualitative) about fluorescent molecules and their environment in a wide variety of biological materials. In fluorescence excitation-emission spectroscopy, each sample is measured by the excitation of the sample at several wavelengths and measuring the emitted light at several wavelengths. The result of such a measurement is an excitation-emission matrix (EEM). When several samples (I) are measured the data can be arranged in a three-way array, $\underline{\mathbf{X}}$ ($I \times J \times K$), where $j = 1, \dots, J$ and $k = 1, \dots, K$ represent the emission and excitation mode respectively. Parallel factor analysis (PARAFAC) [1,2] is a widespread method for modeling such fluorescence excitation-emission landscapes (see e.g. [3–8]). PARAFAC decomposes the fluorescence data into tri-linear components according to the number of fluorophores (F) present in the samples. The structural model can be described as

$$x_{ijk} = \sum_{f=1}^F a_{if} b_{jf} c_{kf} + e_{ijk}$$

where x_{ijk} is the intensity of sample i at emission wavelength j and excitation wavelength k , and where a_{if} , b_{jf} and c_{kf} are parameters describing the importance of the samples/variables to each component f . The residual e_{ijk} contains the variation not captured by the PARAFAC model [9]. For approximate low-rank trilinear data the relative concentrations of analyte f and pure analyte spectra from fluorescence measurements of chemical analytes in mixtures can be extracted, when the correct number of components, equal to the number of fluorophores present in the data, is used.

A common phenomenon, and problem, when fitting PARAFAC to an excitation-emission matrix, is the light scatter effects, such as Raman and first and second order Rayleigh scattering [10–12]. In an EEM-landscape the scatter can be typically found as depicted in Figure 1. This scatter is due to a physical process, which happens when light passes through some kind of medium, like e.g. water. As such, the scatter contains no chemical information and does not conform to the low-rank trilinear model. Therefore it will probably give a model inadequacy, influencing the estimated model parameters [10,12]. Different proposals of how to handle these scatter effects, can be found in the literature; subtracting a standard [13,14], down-weighting the scatter [15–17], inserting missing values [9], avoiding the part containing scatter [4], inserting zeros outside the data area [11] or interpolating the scatter area [18,19]. Unfortunately, all of the proposed methods seem to have some drawbacks. Some of them can only be used in special cases. Others give rise to unacceptable decomposition of the spectra, affect the convergence of the PARAFAC algo-

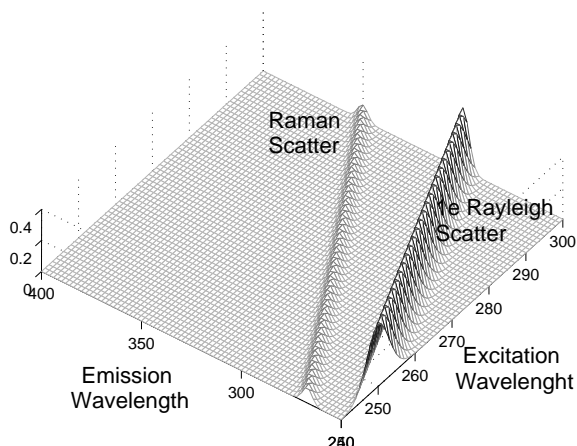


Fig. 1. Raman and Rayleigh scatter in an EEM landscape

rithm or are computational cumbersome [10–12]. A common problem is the visible inspection of the data before the methods can be applied. This makes it difficult to perform all these methods on several data sets at once. It even becomes harder to reduce the effect of scatter when the signal and scatter are overlapping, which is often the case.

In this chapter, we present an automated scatter identification tool using robust statistical methods. Robust statistics overcome badly modeled data due to outliers, i.e. samples that deviate much from the majority of the data points. It is well known that estimates based on a least squares condition are corrupted by outliers in a sense that the models explain the outliers very well, but fit the majority of the data poorly. A lot of research has been done the last decades to adapt known algorithms or to create new ones that can cope with such anomalous observations. For instance in the context of principal components analysis (PCA), the least squares model can be heavily influenced by already one single outlier. Therefore, different robust PCA methods are developed. Among them are the Reflection based Algorithm for PCA (RAPCA) [20] and the Robust PCA (ROBPCA) procedure [21].

The output of these robust multivariate methods is two-fold. Firstly, the provided model fits the majority of the data and is stable in presence of anomalous points. Secondly, each sample is marked as a regular observation or an outlying point for the concerned model, making all these robust procedures useful as outlier identification methods.

An outlying sample can have abnormal values for one or several variables, or it might be deviating from the majority of the samples for almost all of its variables. In three-way data, the latter situation implies that the whole sample landscape is highly different from the others. But outlying values (elements) can also occur in many or all samples. This element-wise contamination often

occurs in multi-way data. A typical example is scattering which affects all samples, and which gives rise to many unexpectedly high values, certainly compared to the other values in the neighborhood.

The correction towards both types of outliers is highly recommended for the PARAFAC model, as an alternating least squares algorithm is used to estimate the scores and loadings [9]. For that reason, the algorithm breaks down when the three-way data contain outlying samples or/and outlying elements. However, a traditional approach, such as the element-wise $L1$ -approach, suggested by [22], for handling outlying elements in combination with scattering will not work well (results not shown). In this $L1$ -approach, PARAFAC estimates are found by minimizing the $L1$ -norm of the residuals instead of the Frobenius norm. This approach would even often lead to discarding chemical information, while keeping the scatter in. Also the robust PARAFAC method for outlying samples proposed in [23] can not handle three-way data with scattering. The major problem with scattering is that it is a systematic corruption of the data within a sample and situated for all the samples in more or less the same area. Randomly placed outlying elements would be easier to handle, but for data sets with systematic deviating parts, it is even not trivial to find for instance robust initial loadings. Nevertheless, robust techniques can still be used in a less conventional way as outlier detection tools to establish an automated identification of the scattering. We focus on ROBPCA, because it can handle high-dimensional two-way data and it is an excellent tool for outlier detection. Moreover, in [21] it is shown that ROBPCA outperforms several other robust PCA methods, such as projection pursuit techniques (e.g. [24,20]) and spherical and elliptical PCA [25].

In the following section we elaborate on this ROBPCA algorithm together with the automated search-engine for element-wise outliers in the form of scattering. Then, we assess the proposed procedure using two laboratory-made data sets in Section 3. The first one is a well-known standard data set, whereas the second one contains highly overlapping components and impurities. Finally, in Section 4 we apply the technique on two real-life examples, where in the first case the algorithm is evaluated for really noisy data. The second data set also is challenging, as the scattering and the signal are very difficult to separate.

All used programs are written in MATLAB. Most of them concerning the robustness are available in the LIBRA toolbox [26], which can be downloaded from <http://www.wis.kuleuven.be/stat/robust.html>. The programs handling multi-way data are available in the PLS-toolbox [27].

2 The automated scatter procedure

The proposed method for automated scatter identification is based on ROBPCA [21]. ROBPCA prevents the corruption of the principal components by outliers through a combination of robust subspace estimation (based on projection-pursuit techniques) and the Minimum Covariance Determinant (MCD) estimator [28] for robust covariance and center estimation. A crucial step in ROBPCA and in the MCD procedure is the search for an outlier-free subset of size h , which will then be used for parameter estimation. The value of h lies between half the number of samples and the total number of samples, n . The higher h , the more accurate, but the less robust the algorithm will be, and vice versa. The default value of h is equal to $\lfloor 0.75n \rfloor$, which often ensures a good compromise between robustness and efficiency.

In the first step of ROBPCA, a preliminary PCA has been performed, such that all the data points are projected in their own space. This means a large dimension reduction for high-dimensional data sets. Secondly, a measure for how far a data point lies from the majority of the other samples, called the outlyingness, is defined for all samples. A further dimension reduction is then obtained by representing all the observations in the space spanned by the d dominant eigenvectors of the h points with smallest outlyingness, with d being the number of principal components to retain. In the next step a reweighted MCD procedure is performed, which provides a robust center and covariance matrix of the d -dimensional data. The principal components are determined as the eigenvectors belonging to the d largest eigenvalues of this covariance matrix. Finally, they are back-transformed to the original data space.

Outlier identification with ROBPCA is obtained by considering two distances for each observation. The orthogonal distance of an observation is defined as the distance between the point and the subspace spanned by the principal components. The second distance, the score distance, can be obtained by computing a robust, Mahalanobis-type distance in the space spanned by the principal components of an observation to the center of the data. If one of these two distances exceeds a certain cut-off value, a sample is flagged as an outlier and receives a zero weight. Other observations obtain a weight equal to 1. The cut-off value for the score distance is based on the assumption that the projected data are normally distributed. As such, the score distances are approximately χ^2 -distributed and the 97.5% quantile of this distribution is taken as cut-off value for the score distances. On the other hand, it can be proven that the orthogonal distances to the power $2/3$ are approximately normal distributed ([29,30]). The 97.5% quantile of the normal distribution to the power $3/2$ is therefore taken as cut-off value for the orthogonal distances. For more information on the cut-off values, we refer to [21]. Hence, a weight vector is given as an extra output when applying ROBPCA, which determines

whether a point is an outlier or not.

To elaborate on the construction of the automated scatter identification method, we first remark that the ROBPCA method can only be performed on two-way data matrices, which should be extracted from three-way data like EEM. In Figure 2 the considered two-way matrices are illustrated. If there is scattering present in the three-way data, it can be found in each observation as shown for example in Figure 3. By slicing these data along the sample mode, the scattering is situated in one or more diagonal lines in each sliced observation (see Figure 2 *B*). This is a problem for ROBPCA, as the scattering might corrupt all or at least a large majority of the variables (element-wise corruption).

On the other hand, slicing the three-way data along the *B*- or *C*-mode establishes useful two-way matrices, in which the scattering is situated in columns for some of these matrices. So, by applying ROBPCA on the transpose of the sliced matrices in the *B*- and *C*-mode leads to the identification of the scattering, because the scattering is now manifesting as an outlying row in the considered two-way matrices (see Figure 2, *C* and *D*).

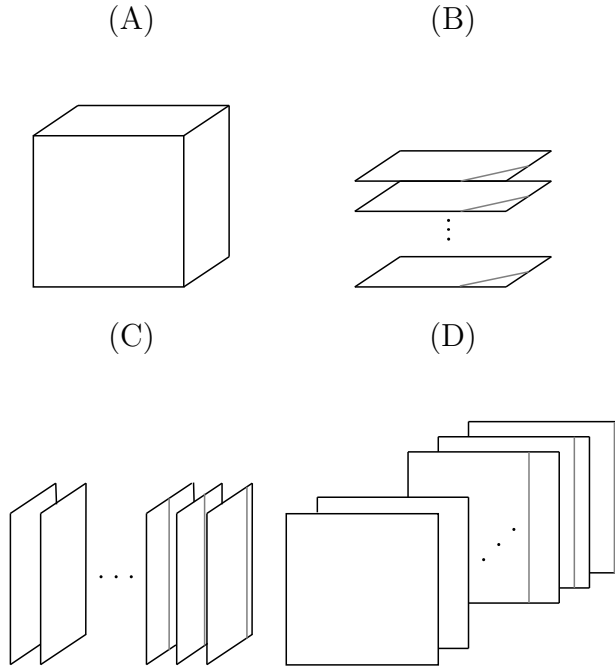


Fig. 2. A visualization of the scattering in three-way data (A) sliced in (B) the first modes, (C) the second mode and (D) the third mode. The grey line represents the scattering.

Thus, in the first step of the identification algorithm, ROBPCA is applied on the transpose of each matrix obtained by slicing the data along the emission mode, noted by $\underline{\mathbf{X}}(:, j, :)$ for the j th slice. So for each $j = 1, \dots, J$, a weight vector $\mathbf{w}_{B,j}$ is created which assigns 1 to a column of $\underline{\mathbf{X}}(:, j, :)$ that is a regular point and 0 to an outlier. All the weight vectors are stored in a $(K \times J)$ weight

matrix \mathbf{W}_B .

In the second step of the algorithm, the same is done for the matrices obtained by slicing the data along the excitation mode, which is $\underline{\mathbf{X}}(:, :, k)$ for the k th excitation wavelength. Again, a weight vector $\mathbf{w}_{C,k}$ is obtained for each $k = 1, \dots, K$ analogously as for $\mathbf{w}_{B,j}$ and the $(J \times K)$ weight matrix \mathbf{W}_C is constructed.

In the next step, both weight matrices \mathbf{W}_B and \mathbf{W}_C are converted to $(I \times J \times K)$ weight arrays $\underline{\mathbf{W}}_B$ and $\underline{\mathbf{W}}_C$, by repeating \mathbf{W}_B and \mathbf{W}_C for each sample and permuting the dimensions of both arrays. Taking the same weight matrices for each observation is justified because the scattering is present in the same area for all observations.

Now, we have two weights $w_{B,ijk}$ and $w_{C,ijk}$ from $\underline{\mathbf{W}}_B$ and $\underline{\mathbf{W}}_C$ respectively for each data element x_{ijk} . The next step consists of merging both weights such that each data element has only one corresponding weight w_{ijk} . This weight w_{ijk} finally defines whether the data element is outlying or not. We take the maximum of both weights $w_{B,ijk}$ and $w_{C,ijk}$ to obtain the final weight $w_{ijk} = \max(w_{B,ijk}, w_{C,ijk})$. This means that a weight w_{ijk} is still assigned a value 1 or 0. Other weighting schemes have been tested, by substituting the minimum instead of the maximum and a smoother version, where values between 0 and 1 are allowed. But the minimum weights nor the smoother weights work well (the results are not included). Mostly they succeed in identifying the scattering, but too much of the signal is also omitted, which leads to inaccurately estimated PARAFAC parameters. The reason that the maximum weight works well to identify scattering, is that scattered elements are outliers that appear in both modes. By taking the maximum, points that are outliers in both modes are only marked as deviating samples for the whole data. So the maximum operator gives the best balance between finding the scattering, without indicating too much of the signal as being outlying.

A final step in the algorithm is turning isolated weights that are assigned a value 0, i.e. weights that are not surrounded by other zero weights, back to 1, as these are not indicating scattering, but parts of the signal.

Note that when applying ROBPCA for each $j = 1, \dots, J$ and each $k = 1, \dots, K$, it should be determined how many components d are retained. In principle, this should be done $J+K$ times by employing common tools such as the scree-plot (see e.g. [31]) or the robust PRESS-curve ([32]). However, this is not advisable here, as this would require many user inputs and hence would result in a highly non-automated method. We thus advice to choose a fixed value d . This has the additional advantage that the sliced data sets $\underline{\mathbf{X}}(:, j, :)'$ and $\underline{\mathbf{X}}(:, :, k)'$ are all investigated on outliers towards a principal components space of the same dimension. From our experience, this optimal dimension d

Table 1

A schematic overview of the scatter identification algorithm

1. For the data sliced along the B -mode :
 - For each $j = 1, \dots, J$:
 - Perform ROBPCA on $\underline{X}(:, j, :)'$
 - Store the weights $\mathbf{w}_{B,j}$ ($1 \times K$)
 - Create the j th row of W_B : $W_B(j, :) = \mathbf{w}_{B,j}$
 - Convert W_B to \underline{W}_B : $\underline{W}_B(i, :, :) = W_B$ for each $i = 1, \dots, I$

2. For the data sliced along the C -mode :
 - For each $k = 1, \dots, K$:
 - Perform ROBPCA on $\underline{X}(:, :, k)'$
 - Store the weights $\mathbf{w}_{C,k}$ ($1 \times J$)
 - Create the k th row of W_C : $W_C(k, :) = \mathbf{w}_{C,k}$
 - Convert W_C to \underline{W}_C : $\underline{W}_C(i, :, :) = W_C'$ for each $i = 1, \dots, I$

3. Define the final weights $w_{i,jk} = \max(\underline{W}_B(i, j, k), \underline{W}_C(i, j, k))$ for each i, j and k .

4. Turn isolated zero-weights back to 1.

lies between 3 and 10. When decreasing d below 3, too many information in the data can be lost, which can lead to not-identified scatter areas. This should be avoided at any time. A too large value of d on the other hand, results in flagging smaller parts of the signal as outlying. This is not a major problem, but it also leads to a computationally more cumbersome ROBPCA algorithm. In our examples of Section 3 and Section 4, we have compared the marked scatter areas for d ranging from 1 to 10. All the results were comparable from $d = 3$ to 10 components for large enough data sets. However, when J or K are really small, d should be taken large enough, such that the signal and the scatter can still be separated by ROBPCA. Taking all these results into consideration, we have set $d = 10$ by default.

To summarize, the proposed method, for which a schematic overview can be found in Table 1, flags areas in the data that are considered as outliers in an automated way, i.e. without visual inspection of the data. To find the final parameter estimates by PARAFAC, the approaches mentioned in the intro-

duction can be performed. One of them is inserting missing values in the areas that are obtained a weight $w_{ijk} = 0$. Another option is to estimate the values in the outlier area by means of interpolation [19]. A third possibility to estimate the PARAFAC loadings can be found in fitting the weighted PARAFAC model of [17,15], where the weights are equal to w_{ijk} . As we are working with fluorescence data which should be strictly positive, non-negativity constraints are used in all modes during this study.

To assess the proposed scatter identification method, we apply it on different kinds of data. We focus on how well the scattering is reduced from the data and how well the signal is preserved with the automated method. Moreover, we investigate the performance of the automated technique in combination with the missing values, the interpolation and the weighted PARAFAC option. The laboratory-made data sets are treated in Section 3 and the environmental data sets are analyzed in Section 4.

3 The analysis of laboratory-made data

3.1 *Dorrit Data*

The method was tested on fluorescence data, containing mixtures of four known fluorophores [33,34]. The four compounds are phenylalanine, 3, 4-dihydroxyphenylalanine (DOPA), 1, 4-dihydroxybenzene and tryptophan. For every sample an excitation-emission matrix was obtained by measuring the emission spectra from 250 to 482 nm at 2 nm intervals, with excitation at every 5 nm from 200 to 315 nm on a Perkin-Elmer LS50 B fluorescence spectrometer.

For both excitation and emission the scan speed was 1500 nm/min. The excitation from 200 to 230 nm and the emission below 250 nm was excluded from the analysis since it is highly influenced by the condition of the xenon lamp as well as by the physical environment and mainly contained missing elements [33]. From previous investigations [33,34], we know that four components are appropriate and that four EEM landscapes can be considered as outliers. Since a classical PARAFAC algorithm is applied on the data after removing the scatter, outlying samples will corrupt the final results. Moreover, the focus in this paper is put on testing the removal of scatter, considered as being in the form of element-wise outliers, not whole samples. In [23], an algorithm to deal with outlying observations is proposed, but this method is not able to withstand the effects of scattering. Handling both types of outliers together, is a challenge for future research. For now, these four observations are therefore removed from the data set. The data set then consists of 23 sam-

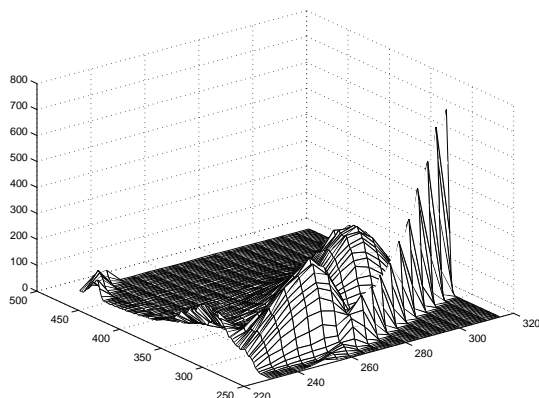


Fig. 3. Example (Sample 4) of a full excitation - emission landscape obtained from the fluorescence measurement. The 1st and 2nd order Rayleigh ridges are clearly seen as diagonal ridges. The 1st order Rayleigh scatter ridge to the right is situated at the diagonal where excitation and emission wavelengths are equal.

ples, 18 excitation wavelengths and 116 emission wavelengths, and will in the following be referred to as the Dorrit data.

An example of a full excitation - emission landscape obtained from the fluorescence measurements is illustrated in Figure 3. The Rayleigh scatter can clearly be seen as diagonal ridges. The scatter seems to be well separated from the chemical signal and no Raman scatter is observed. This is the case for all samples in the Dorrit data set. Therefore this well-known data set appears to be perfect for illustrating the properties of the proposed method for automatic scatter removal.

The emission and excitation loadings from a four component PARAFAC model, fitted to the data set where scatter has been manually removed is shown in Figure 4 (left). This method is based on removing the Rayleigh scatter by inserting a mixture of missing values and zeros. The loadings have a reasonable shape resembling the pure spectra of the four fluorophores. The emission and excitation loadings for the original data set will appear as illustrated in Figure 4 (right). When comparing the emission loadings from the two models, it is clear that the highest peak in the model fitted to the data with Rayleigh scatter is wrong and caused by the scatter. Also the excitation loadings are not fitted accurately. This clearly indicates that the Rayleigh scatter needs to be removed to obtain a good model.

The identification of the scatter by our automated method performs very well as illustrated in Figure 5, where the emission profiles of sample 4 for the 18 excitation wavelengths are shown. The elements flagged as outliers by the algorithm are marked with dots on the X-axis. The scatter corresponding to 2nd order Rayleigh is clearly identified for the first 3 excitation wavelengths (3 first plots) and from excitation wavelength 259 nm on the regions according to the 1st order Rayleigh scatter are clearly identified. The successful detection

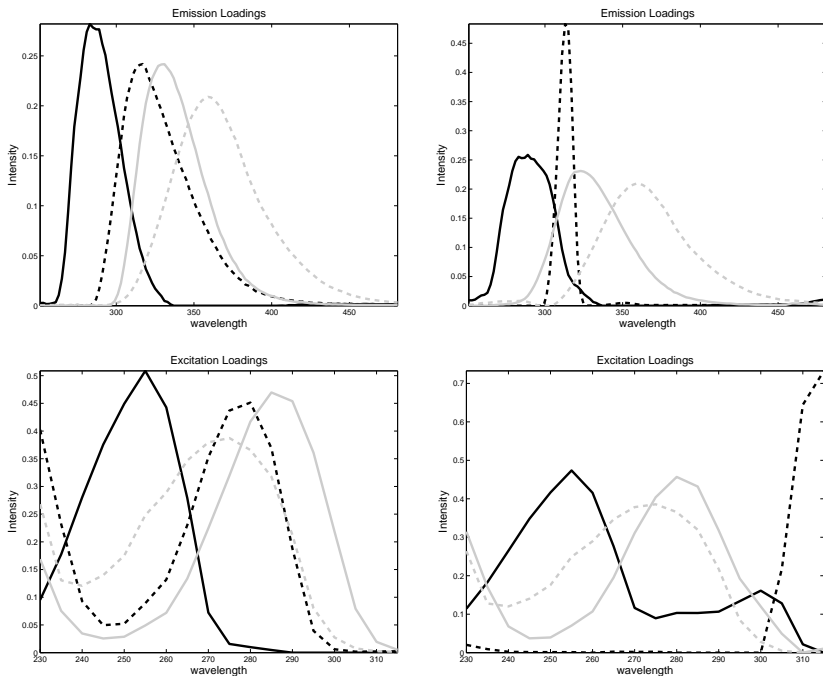


Fig. 4. Left: Emission loadings from a four component PARAFAC model, fitted to the Dorrit data set where scatter has been manually removed. Right: Emission loadings from a four component PARAFAC model, fitted to the full Dorrit data set. of Rayleigh scatter in the remaining samples performs likewise (results not shown).

The three different PARAFAC algorithms; replacing with missing values, interpolation and weighting, were then applied to the data set in combination with the information about outlying elements.

The emission and excitation loadings obtained with the three different PARAFAC algorithms are shown in Figure 6. The three tested algorithms have in common that both emission and excitation loadings are almost identical with the pure spectra of the four fluorophores. This clearly indicates that the automated method for identifying scatter has worked perfectly in marking both 1st and 2nd order Rayleigh scatter, which results in fairly good PARAFAC models. No obvious differences are observed between the three tested PARAFAC algorithms.

3.2 Fluorescence Data

This data set, called Fluorescence data, consists of 35 samples of a larger data set consisting of 405 samples, built by experimental design [35,36]. The Fluorescence data is made from five known analytes; Catechol, Hydroquinone, Indole, Tryptophane and Tyrosine, with two to five analytes present in each

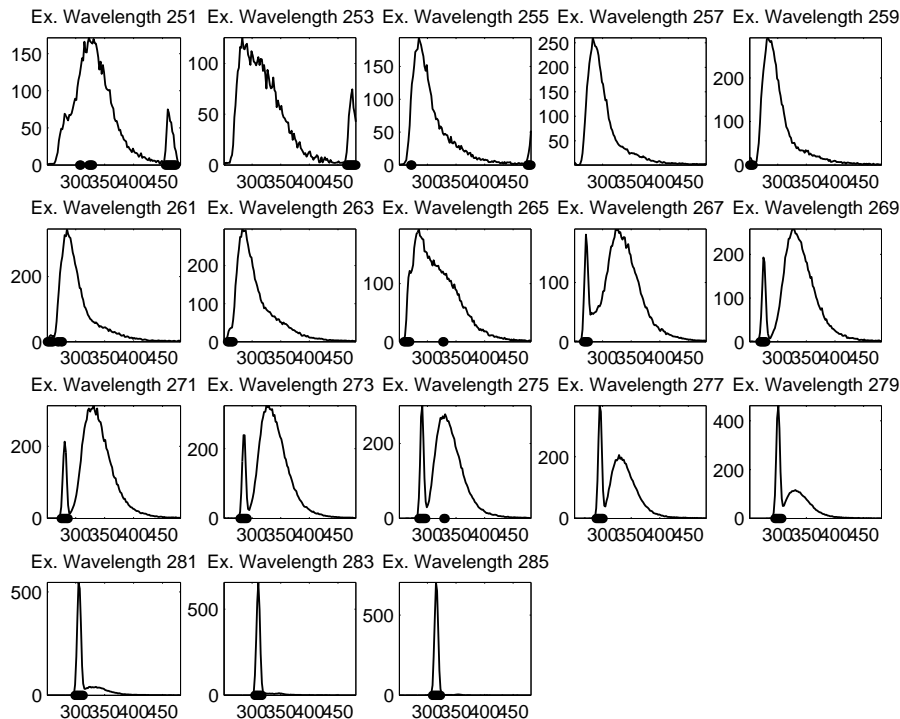


Fig. 5. The emission profiles of the fourth sample of the Dorrit data for the 18 excitation wavelengths. The regions identified as scatter are marked by dots.

sample varying in concentration. These data were chosen on the basis of closeness to the 1st order Rayleigh scatter line and their overlap in both emission and excitation spectra. The prepared samples were measured on a Varian Eclipse Fluorescence Spectrometer with slit widths 5 nm (for both emission and excitation), emission wavelengths 230 - 500 nm (recorded every 2 nm) and excitation wavelengths 230 - 320 nm (recorded every 5 nm), and a scan rate at 1920 nm/min. The sample was left in the instrument and was scanned five consecutive times but only the first measurement is included in this analysis. From the experimental set-up it is known that Catechol contains some impurities, and thus gives rise to an extra component in the data sets. This means that the PARAFAC model should be fitted with 6 components. Furthermore, overlapping emission and excitation profiles [36] might make the modelling part hard and not as simple as for the Dorrit data in the previous section.

A standard, containing only the solvent, exists for this data set. By subtracting this from all other samples the Raman scatter line and possibly the Rayleigh scatter line can be removed or at least reduced [12]. This is not done here since the purpose of this study is to test the possibilities of removing all kinds of scatter by the automated scatter removing method proposed within.

The automated scatter identification method with 6 components was applied on the data set. The scatter region is almost always indicated by the method (Figure 7, left) but sometimes not the whole scatter region is left out (Figure 7,

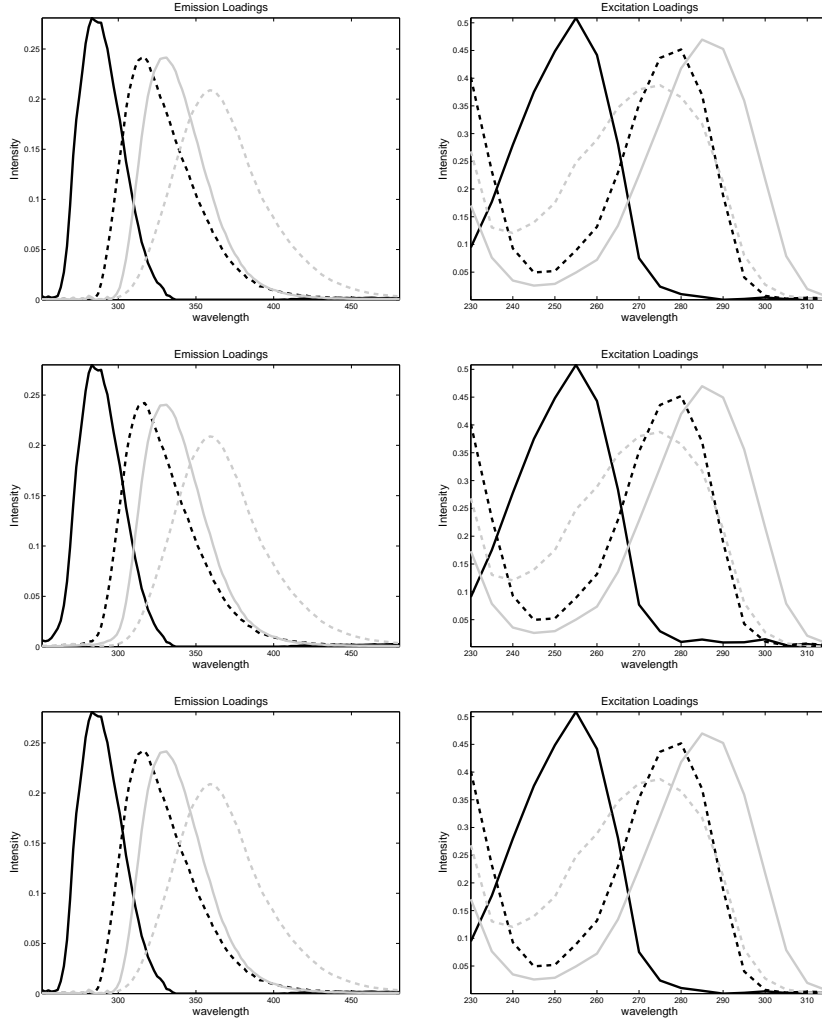


Fig. 6. Four component PARAFAC models ((above) Missing, (middle) Interpolation, and (below) Weighted) fitted to the Dorrit data where the scatter has been detected by the automated method. The left column correspond to emission mode loading and the right column to excitation mode loading.

right). This means that a small part of the scatter is still left in the data set, and consequently included in further computations. This failure of not finding the edges of the scatter region is due to the maximum operator to obtain one weight w_{ijk} . For the data sliced along the C -mode, ROBPCA flags the whole scatter peak as being outlying together with a large part of the signal. But for the sliced data in the B -mode, the scattering appears rather small compared to the signal. The scatter is still found as an outlier by ROBPCA, but the edges are not deviating any more and thus not identified. Taking the maximum over the weights $w_{ijk,B}$ and $w_{ijk,C}$ finally results in a weight $w_{ijk} = 1$ for the edges of the scatter area. The problem will not be solved by taking another operator than the maximum, like e.g. the minimum or smoother weights, as too much of the signal will be omitted then. A better alternative is to enlarge the indicated region with a certain number of zeros. However, we have not

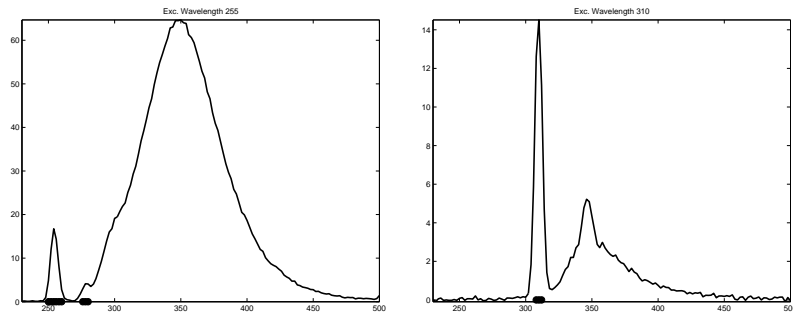


Fig. 7. The emission profiles of sample no. 20 from the fluorescence data at excitation wavelength 255 nm where the scatter region is correctly identified (left) and the emission profile at excitation wavelength 310 nm where not the whole scatter region is identified (right).

done this in this example, because the included scattering is rather small and adding zeros comprises also a greater loss of the signal.

Furthermore, another problem turned up here. A part of the chemical signal is sometimes wrongly identified as scattering as shown in Figure 8. This is caused by the general property of robust methods that the majority of the data determine the final estimates. In this example, for some wavelengths the investigated data contain more than 50% low-value profiles, i.e. profiles that contain no signal, nor scatter. This means that the scatter but also the signal is seen as highly deviating. But, this is only happening for wavelengths 230 nm, 235 nm, 290 nm, 295 nm and 300 nm. Thus only a small part of the chemical signal is deleted. As such, enough signal is left to estimate the loadings correctly and again no changes to the algorithm are made to circumvent this problem.

The estimated excitation and emission loadings when fitting 6 components PARAFAC models in combination with the information from the automated scatter identification method to the Fluorescence data set are illustrated in Figure 9.

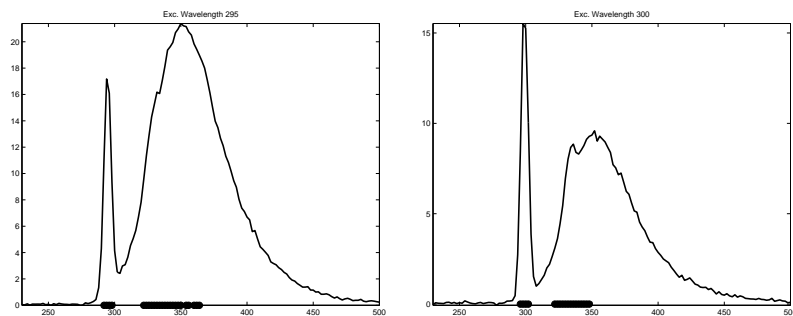


Fig. 8. The emission profiles of sample 20 from the fluorescence data excitation wavelength 295 nm (left) and 300 nm (right), respectively, showing wrongly identification of the chemical signal as outlying.

For the missing and weighted algorithms the estimated excitation and emission profiles are in accordance with profiles of the pure spectra (Figure 9 (first and last row)). The interpolation algorithm has some problems in both the excitation and emission mode (Figure 9, second row). The component indicated by the grey dashed-dotted line is due to the impurities in the samples containing Catechol. The results obtained with these data show that even small inaccuracies in identifying the scatter regions, will establish a good PARAFAC model at the end, when using missing values or the weighted PARAFAC version. The interpolated PARAFAC on the other hand has some problems which is due to the not completely removed scatter, that still causes a small peak in the interpolated data.

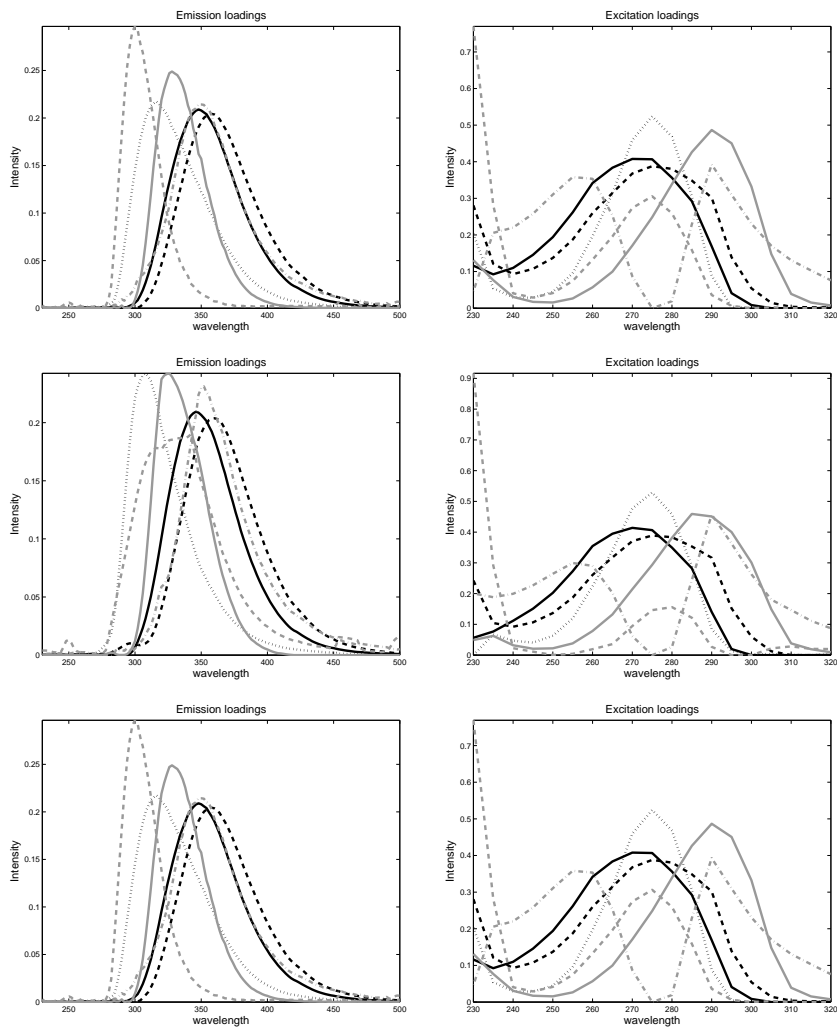


Fig. 9. The emission (left) and excitation (right) loadings for the fluorescence data from a 6 components PARAFAC model for the Missing algorithm (above), the Interpolation algorithm (middle), and the Weighted algorithm (below).

4 Examples

Data created in a laboratory provide an excellent platform for testing newly developed methods, because estimates can be compared to a priori information. But it is also interesting to assess new techniques on environmental data, to find out how well the method can cope with extra difficulties on top of the scattering problem typical for real-life data, such as e.g. noise. Therefore the automated scatter identification algorithm is carried out on the following two examples.

4.1 North Sea data

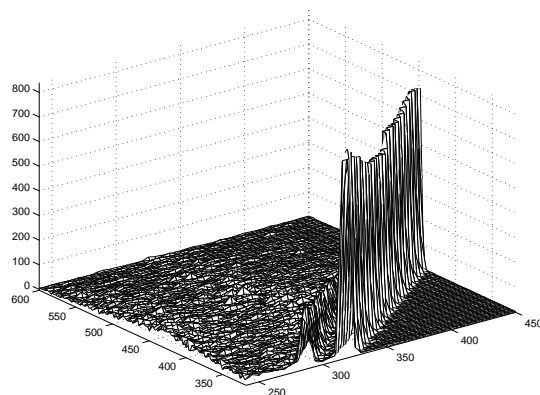


Fig. 10. The 10th sample of the North Sea data with very severe Raman and first order Raleigh scattering. The highest peak corresponds to the Raleigh scattering, the smallest to the Raman scattering.

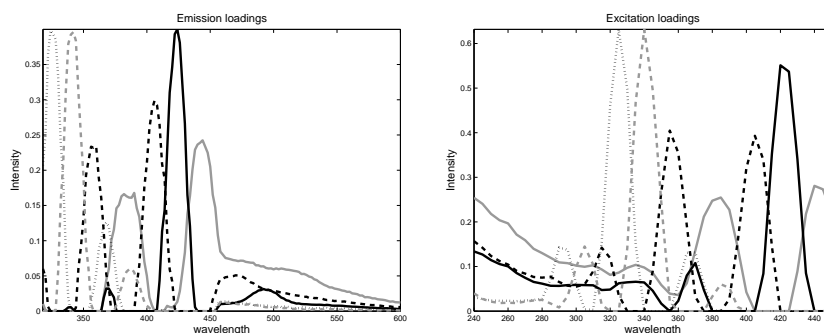


Fig. 11. The emission (left) and excitation (right) loadings of the North Sea data obtained by the classical PARAFAC algorithm.

The 37 samples of the North Sea data, which are kindly provided by Colin A. Stedmon (personal communication), reflect the fluorescence of dissolved organic matter (DOM) of water of the Dogger bank in the North Sea. Measurements were taken with a Varian Eclipse fluorescence spectrophotometer

from 2 vertical profiles at 5 m depth intervals. The excitation wavelengths range from 240 - 450 nm every 5 nm and the emission wavelengths from 240 - 600 nm each 2 nm. This results in a $(37 \times 181 \times 43)$ data cube. No pre-treatment on the data has been carried out, besides a correction for instrument specific effects and a Raman calibration (see [7]). As no blank is subtracted from the samples, severe Raman and Raleigh scattering are present in all the fluorescence measurements. Moreover, some artifacts could be distinguished in the first 39 emission wavelengths. As we focus on removing scattering effects, we delete these artifacts before analysis. This leads to a $(37 \times 142 \times 43)$ data array. It is also known that the signal to noise ratio of the measurements is very low, which means that we have to deal with really noisy data. An EEM landscape is shown in Figure 10. The scattering is so strong that the relevant signal can not be distinguished. A classical PARAFAC analysis therefore fails in estimating useful loadings (see Figure 11).

A split half and residual analysis on the data, obtained after subtracting a blank and after removing manually the scatter, was performed by Colin Stedmon (pers. com.) and indicated that 5 components were suitable for modeling the data. No outlying samples were present in the data.

In Figure 12, the emission profiles of sample 9 for the first 20 excitation wavelengths are shown. In the first 9 plots, there is no scattering present and only small parts of the signal are marked as being outlying. For the other graphics, the scattering is left out in all cases together with minor parts of the signal. So, the identification of the scattering is performed really well by the automated method.

We use the information about the scatter region and perform the PARAFAC algorithm on the data with missing values instead of outlying elements. We have decreased the stop criterium to 10^{-10} to obtain stable PARAFAC estimates in this very noisy data set. The resulting excitation and emission loadings are depicted in Figure 13 (first row). The emission loadings seem chemically relevant, except perhaps the dashed grey loading, that is quite noisy. This is the effect of a high noise level in the data rather than of the scattering.

Furthermore, we also applied the classical PARAFAC algorithm on the interpolated data with the same constraints as for the missing values. We end up with emission and excitation loadings of Figure 13 (second row). The difference between these loadings and the one estimated by inserting missing values, is in contradiction with the uniqueness property of PARAFAC, which states that a unique solution is provided under mild conditions [37]. An explanation for this discrepancy can be given by the different approximations of the final data on which the PARAFAC model is applied. While missing values discard certain issues present in the data, the interpolation technique fits an approxi-

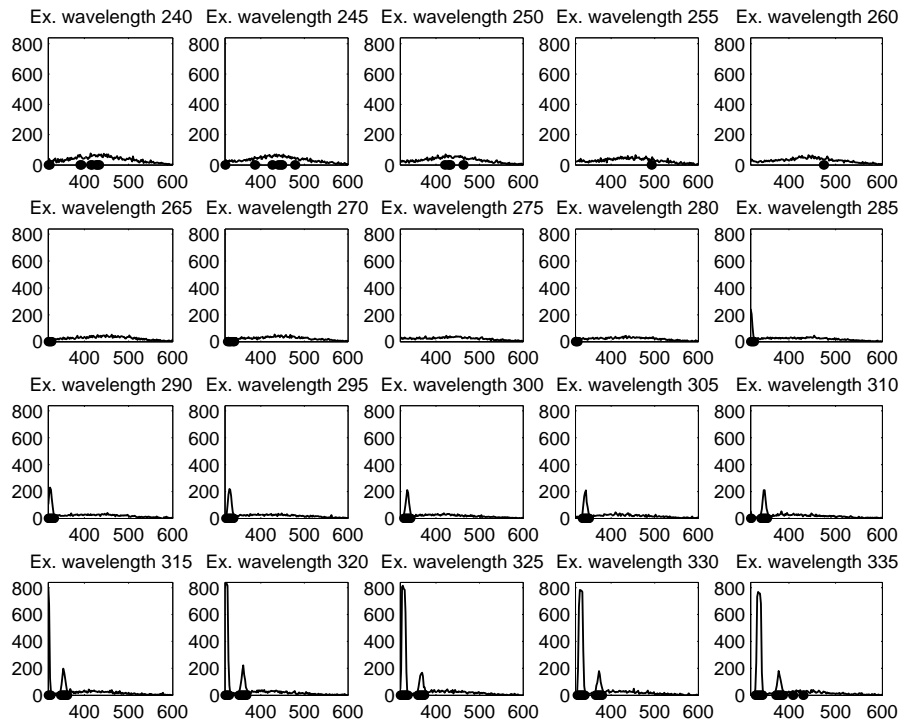


Fig. 12. The emission profile of the ninth observation of the North Sea data for the first 20 excitation wavelengths.

mate value. This leads to quite different data elements at certain areas in the final data set and hence at different estimated loadings. On the other hand, the obtained loadings with the interpolation technique are comparable to the one depicted in [19], which is not surprisingly as in [19] a similar interpolation technique is used. Because this is real life data, it is not known a priori which should be the correct loadings. So, both solutions are possible, and we can not favor one above the other.

Finally, the weighted PARAFAC is carried out and the resulting loadings can be found in Figure 13 (last row). It is obvious that this procedure has broken down because of the scattering. The three loadings with a narrow, but high peak, are fitting the scattering instead of chemically relevant information. The reason why this method failed is a combination of a non-robust initialization of the loadings and too heavy scattering, such that the starting point of the iterative loops in the alternating least squares PARAFAC algorithm, is taken too far from a possible solution.

4.2 Kauai data

The Kauai data, which was kindly provided by the Smithsonian Environmental Research Center in Maryland, USA, consists of 130 seawater EEMs. Of these,

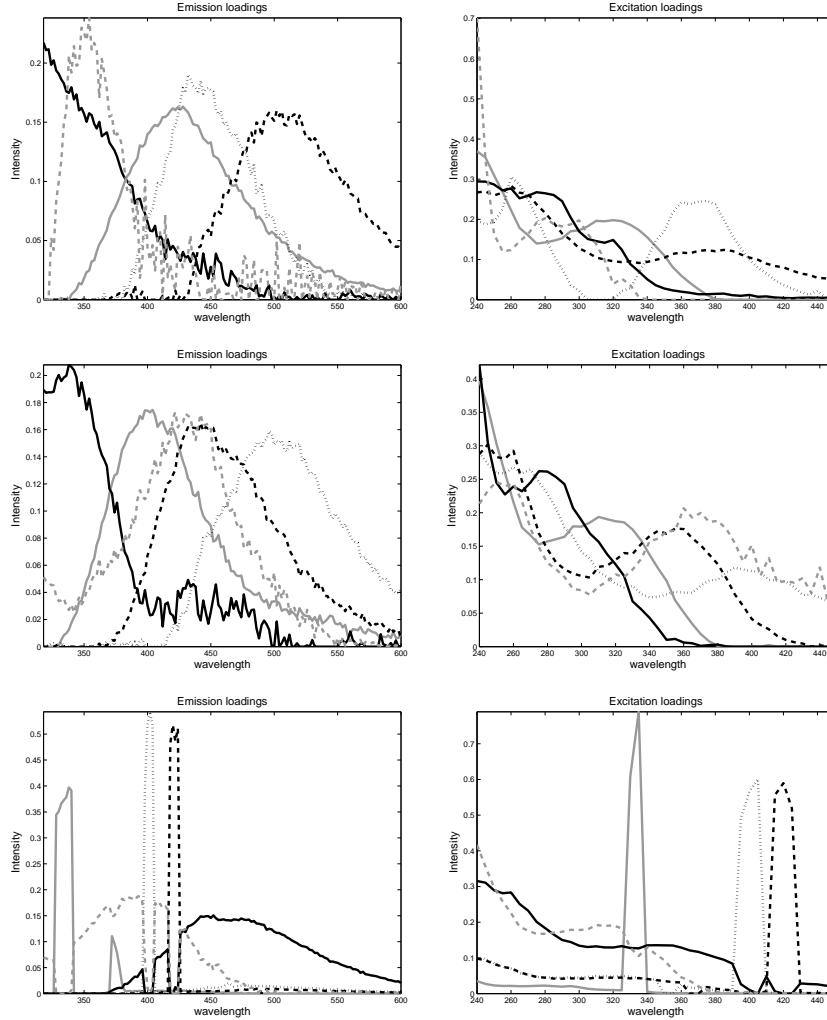


Fig. 13. The emission (left) and excitation (right) loadings for the North Sea data using missing values (above), interpolation (middle) and weighted data (below).

53 were obtained from the ballast tanks of the container ship MV Kauai and 77 were obtained from the Pacific Ocean during the vessel's cruise between Oakland (CA) - Honolulu (HI) and Seattle (WA) in June 2003 [38]. CDOM analysis by excitation-emission matrix spectroscopy (excitation, 240-455 nm in 5-nm intervals; emission, 290-600 nm in 2-nm intervals; 5-nm bandwidths on excitation and emission modes) was performed using a spectrofluorometer at the University of Maine, USA, using a SPEX FluoroMax-2.

Each EEM from the Kauai Dataset consisted of 156 emission wavelengths and 48 excitation wavelengths. As for the Dorrit dataset, the first 4 excitation wavelengths were deleted prior to modeling. In addition, data from the last 26 emission wavelengths ($\lambda_{em} > 518$ nm) were ignored since these were dominated by a spurious signal propagated from intense protein-like fluorescence near $\lambda_{ex}/\lambda_{em} = 270/300$ nm. This resulted in a signal at identical excitation but twice the emission wavelength of the actual fluorescence ($\lambda_{ex}/\lambda_{em} = 275/600$

nm; see Figure 14). Thus modeling was performed upon a $(130 \times 130 \times 44)$ element data array.

First order Rayleigh scatter and first and second order Raman scattering are clearly evident in the data (Figure 14). Figure 15 shows the emission profile of the third observation for the 275 nm and 290 nm excitation wavelength. For the 275 nm excitation wavelength, the right peak is due to the Raleigh scattering, whereas the peak on the left consists partially of the signal and partially of the Raman scattering. The same occurs for the 290 nm excitation wavelength and most of the other wavelengths. Consequently, it is difficult to determine the extent of the scatter region from visual inspection of the data.

From a split half analysis on the data after removing scatter with the interpolation technique of [18], 6 components should be used in the PARAFAC model. Moreover, no outliers could be distinguished in the data.

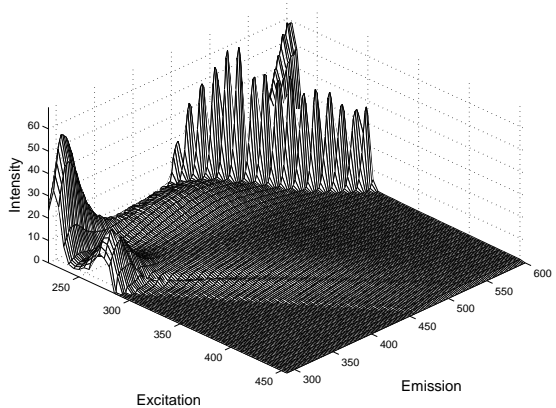


Fig. 14. The third observation of the Kauai data, before the artifacts are removed.

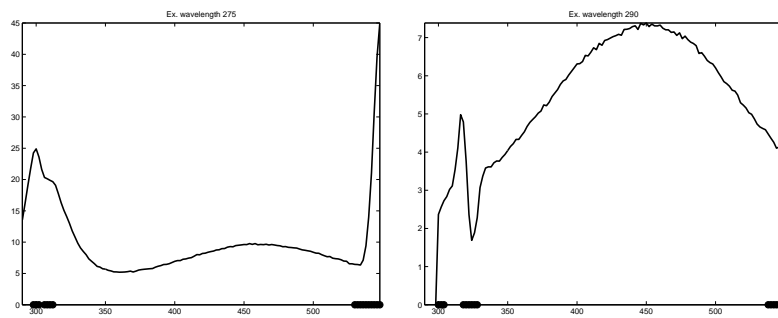


Fig. 15. The emission profile of the third observation of the Kauai data for excitation wavelength 275 nm (left) and 290 nm(right).

Preliminary identification of scatter regions before further analysis of the data, was therefore conducted with 6 components. In Figure 16 the results of the automated scatter identification algorithm can be seen for the first 12 emission profiles of observation 8. The remaining emission profiles of observation 8 gave similar results and are therefore not included.

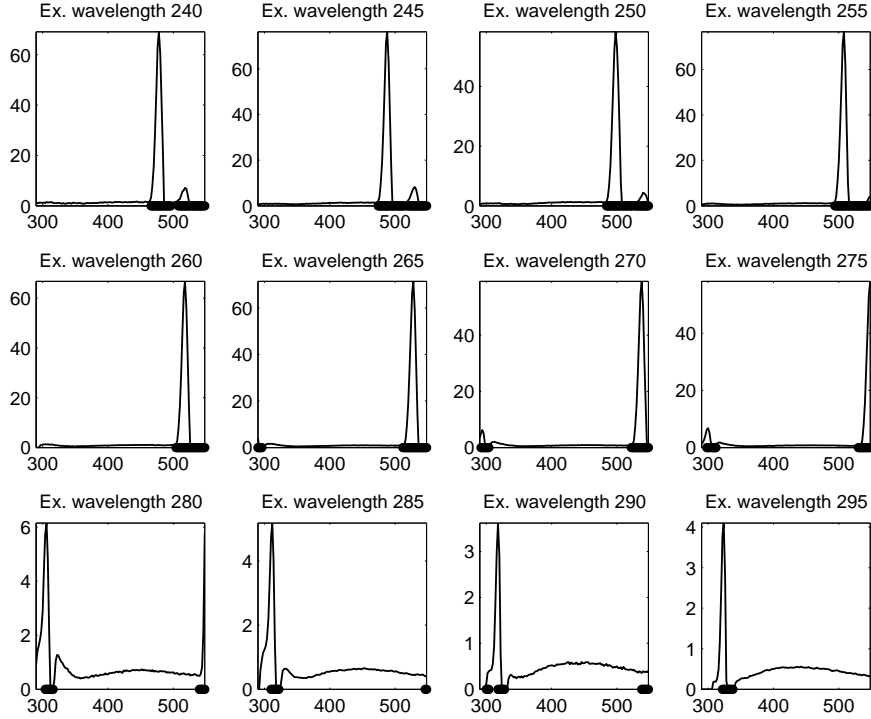


Fig. 16. The emission profiles of the observation 8 of the Kauai data for the first 12 excitation wavelengths.

We see from all these figures that the Rayleigh scattering is captured in the set of the outlying elements. The Raman scattering is indicated for each excitation wavelength, although it is not completely removed, which is caused by the presence of more than 30% zero elements in the Kauai data. The input data sets for the ROBPCA procedure in the automated algorithm therefore contain sometimes more than 50% zero rows, which makes it impossible to identify the scatter accurately. This problem is inherent for the Kauai data and can not be solved by changing the proposed automated technique. Since, at least all the scatter area have partially been detected, we have decided to enlarge the marked outlier regions with two elements at both sides, to be sure that the scattering will not corrupt the final PARAFAC estimates. Small parts of the signal will consequently also be removed. We have depicted results for observation 8 in Figure 16, as it was one of the samples for which the remaining scatter area was the largest. For most of the other samples, this effect is more reduced, where even for some of them, the scattering was almost totally eliminated before the enlargement of the outlier area.

Finally, the PARAFAC model with non-negativity constraints is built for the data with missing, interpolated and down-weighted values. The emission (left) and excitation (right) loadings are placed in Figure 17. It seems that imputing missing values where outliers are marked, did not perform well, because of the strange emission loading (the full grey one) and the excitation loading with two sharp peaks (the full grey one). The profile of the excitation loadings is due

to the missing area going straight through the first large signal peak, which causes a split of 1 signal peak in 2 sharp peaks. For the emission loadings, the area for large emission wavelength containing only missing values or almost zero values, is to blame. Whatever is estimated as loading values instead of the missing values, it will always have no effect in the final model, as it is multiplied by very low values. This results in the narrow, but high peak at the end of the concerned emission loading. Although the model estimates are thus not correct, it is not caused by the scattering. The missing algorithm fails because of omitting crucial data parts.

For the interpolated Kauai data, the estimated loadings for a 6 component PARAFAC model are depicted in Figure 17 (second row). Here, the scattering is out of the model. However, the estimated loadings are not exactly the same as the one obtained in [38]. A reason can be found in a different approximation of the data, due to other interpolation techniques applied on other sets of data elements, as the scatter areas are marked using two different approaches.

The results of the weighted PARAFAC are shown in Figure 17 (last row). The loadings are not good, they try to fit the scattering and not only the chemical information. The same reasons as for the North Sea data cause this breakdown of the model and again confirm that the weighted PARAFAC model is not optimal to use for highly scattered data.

5 Discussion and Conclusions

Despite different existing methods for excluding scattering effects when modeling fluorescence data by PARAFAC, no method for disregarding scattering automatically can be found in the literature. In this article we have established an automated scatter identification method which is based on ROBPCA. The method does not demand any visual inspection of the data. The evaluation of the proposed method clearly shows that the method always succeeds in finding the scatter regions both concerning Rayleigh (1st and 2nd order) and Raman scatter, without marking too much of the signal, due to chemicals under investigation, as outlying. However, smaller parts of the scattering are sometimes hard to detect depending on the data complexity e.g. noise and overlap between scatter and chemical signal. This means that scatter might be included to a minor extent in the PARAFAC modeling step, but also smaller part of the chemical signal might be flagged as outlying and thereby excluded from the analysis.

Nevertheless, this seems not an invincible problem for estimating the final PARAFAC estimates. The three tested PARAFAC methods after removal of the scattering work for the cases they can handle. This means that for the data

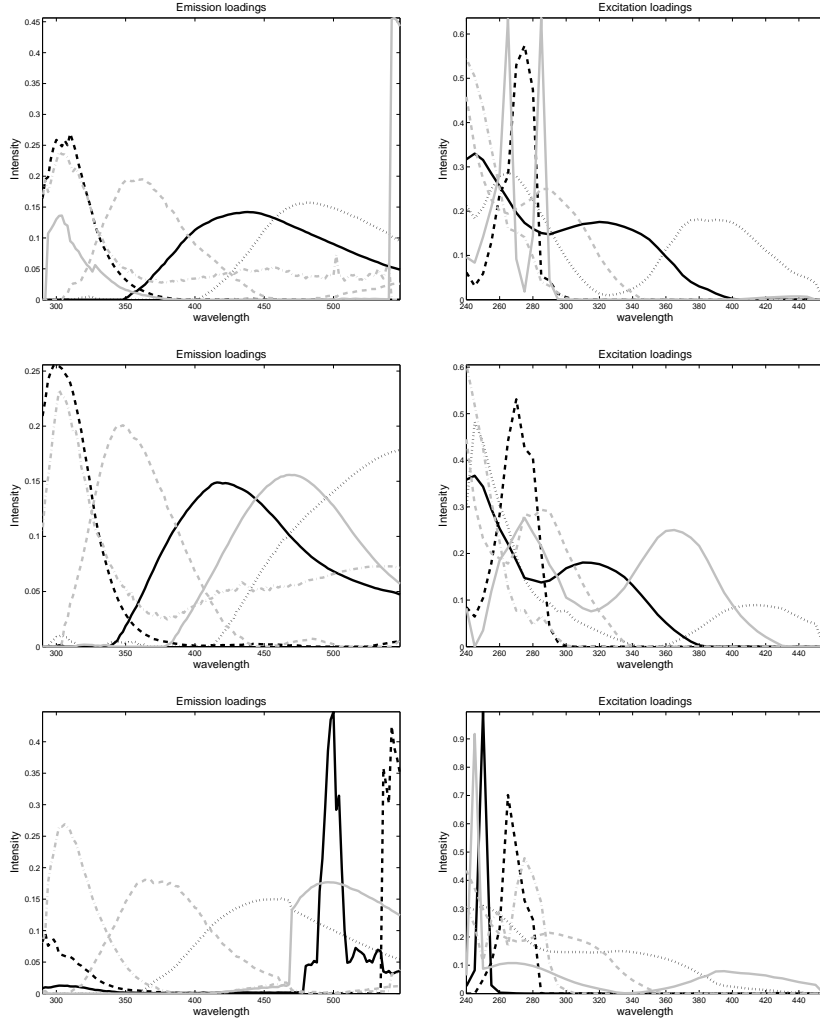


Fig. 17. The emission (left) and excitation (right) loadings of the Kauai data for a PARAFAC model with 6 components using missing values (above), interpolation (middle) and weighted data (below).

with the missing values a fitting problem is only encountered when the signal and the scatter coincide too much, such that essential information vanishes. Secondly, classical PARAFAC applied on interpolated data also performs well, but it is the most subject to the parts of the scattering that are not flagged as outlying. Finally, down-weighting the outlying elements is also a good option, provided that the scattering is in the region of the signal. For too severe scatter, this technique is not useful and actually is the least robust of the three investigated procedures.

We only have considered data sets where the number of components has been known before analysis throughout this paper, because the identification of scatter was the major concern. However, when the optimal value for F is not known, which is mostly the case for real-life data, the following approach could be followed to determine the scatter and F . Start with an initial guess for the

number of components. In the next step, identify and remove the scatter in the data. Then, use existing techniques (like e.g. a split half analysis) to define F on the data without the scatter. Finally, identify again for the known value of F the scatter automatically. This approach is justified, because the scatter is not highly dependent on F , as we have already discussed earlier in Section 2.

Remark that the proposed method can not cope with outlying samples yet. A fully robust procedure handling both sample outlier identification and scatter identification is under development.

References

- [1] J. Carroll, J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition, *Psychometrika* 35 (1970) 283 – 319.
- [2] R. Harshman, Foundations on the PARAFAC procedure: model and conditions for an explanatory multimode factor analysis, *UCLA Working Paper Phonetics* 16 (1970) 1–84.
- [3] R. Ross, C. Lee, C. Davis, B. Ezzeddine, E. Fayyad, S. Leurgans, Resolution of fluorescence spectra of plant-pigment complexes using trilinear models., *Biochimica Biophysica Acta* 1056 (1991) 317320.
- [4] R. Bro, Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis, *Chemometrics and Intelligent Laboratory Systems* 46 (1999) 133–147.
- [5] R. Jiji, G. Andersson, K. Booksh, Application of PARAFAC for calibration with excitation-emission matrix fluorescence spectra of three classes of environmental pollutants, *Journal of Chemometrics* 14 (2000) 171–185.
- [6] D. Pedersen, L. Munck, S. Engelsen, Screening for dioxin in fish oil by PARAFAC and N-PLSR analysis of fluorescence landscapes, *Journal of Chemometrics* 16 (2002) 451–460.
- [7] C. Stedmon, S. Markager, R. Bro, Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy, *Marine Chemistry* 82 (2003) 239–254.
- [8] J. Christensen, E. Miquel Becker, C. Frederiksen, Fluorescence spectroscopy and PARAFAC in the analysis of yoghurt, *Chemometrics and Intelligent Laboratory Systems* 75 (2005) 201–205.
- [9] R. Bro, PARAFAC. Tutorial and applications, *Chemometrics and Intelligent Laboratory Systems* 38 (1997) 149–171.
- [10] C. Andersen, R. Bro, Practical aspects of PARAFAC modelling of fluorescence excitation-emission data, *Journal of Chemometrics* 17 (2003) 200–215.

- [11] L. Thygesen, A. Rinnan, S. Barsberg, J. Møller, Stabilizing the PARAFAC decomposition of fluorescence spectra by insertion of zeros outside the data area, *Chemometrics and Intelligent Laboratory Systems* 71 (2004) 97–106.
- [12] A. Rinnan, C. Andersen, Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation - emission data, *Chemometrics and Intelligent Laboratory Systems* 76 (2005) 91–99.
- [13] P. Wentzell, S. Nair, R. Guy, Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane, *Analytical Chemistry* 73 (2001) 14081415.
- [14] D. M. McKnight, E. Boyer, P. Westerhoff, P. Doran, T. Kulbe, D. T. Andersen, Spectrofluorometric characterisation of dissolved organic matter for indication of precursor organic material and aromaticity, *Limnology and Oceanography* 46 (2001) 38–48.
- [15] R. Bro, N. Sidiropoulos, A. Smilde, Maximum likelihood fitting using ordinary least squares algorithms, *Journal of Chemometrics* 16 (2002) 387 – 400.
- [16] R. JiJi, K. Booksh, Mitigation of Rayleigh and Raman spectral interferences in multi-way calibration of excitation-emission matrix fluorescence data, *Analytical Chemistry* 72 (2000) 718–725.
- [17] G. Andersson, B. Dable, K. Booksh, Weighted parallel factor analysis for calibration of HPLC-UV/Vis spectrometers in the presence of Beer’s law deviations, *Chemometrics and Intelligent Laboratory Systems* 49 (1999) 195–213.
- [18] R. Zepp, W. Sheldon, M. Moran, Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Raleigh and Raman scattering peaks in excitation-emission matrices, *Marine Chemistry* 89 (2004) 15–36.
- [19] M. Bahram, R. Bro, C. Stedmon, A. Afkhani, Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation Submitted.
- [20] M. Hubert, P. Rousseeuw, S. Verboven, A fast robust method for principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems* 60 (2002) 101–111.
- [21] M. Hubert, P. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal components analysis, *Technometrics* 47 (2005) 64–79.
- [22] S. Vorobyov, Y. Rong, N. Sidiropoulos, A. Gershman, Robust iterative fitting of multilinear models, *IEEE Transactions on Signal processing* 53 (2005) 2678–2689.
- [23] S. Engelen, M. Hubert, Detecting outlying samples in a PARAFAC model Submitted.

- [24] G. Li, Z. Chen, Projection-pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, *Journal of the American Statistical Association* 80 (1985) 759–766.
- [25] N. Locantore, J. Marron, D. Simpson, N. Tripoli, J. Zhang, K. Cohen, Robust principal component analysis for functional data, *Test* 8 (1999) 1–73.
- [26] S. Verboven, M. Hubert, LIBRA: a Matlab library for robust analysis, *Chemometrics and Intelligent Laboratory Systems* 75 (2005) 127–136.
- [27] B. Wise, N. Gallagher, R. Bro, J. Shaver, W. Windig, R. Koch, PLS_Toolbox 3.5 for use with MATLAB, software, Eigenvector Research, Inc., August 2004 (2004).
URL <http://software.eigenvector.com/>
- [28] P. Rousseeuw, Least median of squares regression, *Journal of the American Statistical Association* 79 (1984) 871–880.
- [29] G. 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 (1954) 33–51.
- [30] P. Nomikos, J. MacGregor, Multivariate SPC charts for monitoring batch processes, *Technometrics* 37 (1995) 41–59.
- [31] I. Jolliffe, *Principal Component Analysis*, Springer, New York, 1986.
- [32] M. Hubert, S. Engelen, Fast cross-validation for high-breakdown resampling algorithms for PCA, under Revision (2006).
- [33] D. Baunsgaard, Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes, Ph.D. thesis, Royal Veterinary and Agricultural University, Department of Dairy and Food technology, Frederiksberg, Denmark (1999).
- [34] J. Riu, R. Bro, Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models, *Chemometrics and Intelligent Laboratory Systems* 65 (2003) 35–49.
- [35] A. Rinnan, Application of PARAFAC on spectral data, Ph.D. thesis, Royal Veterinary and Agricultural University, Denmark (2004).
- [36] R. Bro, A. Rinnan, N. Faber, Standard error of prediction for multilinear PLS2. Practical implementation in fluorescence spectroscopy, *Chemometrics and Intelligent Laboratory Systems* 75 (2005) 69–76.
- [37] A. Smilde, R. Bro, P. Geladi, *Multi-way Analysis with Applications in the Chemical Sciences*, Wiley, England, 2004.
- [38] K. Murphy, G. M. Ruiz, W. T. M. Dunsmuir, T. D. Waite, Optimized parameters for rapid fluorescence-based verification of ballast water exchange by ships., *Environmental Science and Technology* 40, in press.