JOURNAL OF CHEMOMETRICS
J. Chemometrics (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/cem.978

Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation

Morteza Bahram[1], Rasmus Bro[2]*, Colin Stedmon[3] and Abbas Afkhami[1]

[1]Department of Chemistry, Faculty of Sciences, Bu-Ali Sina University, Hamadan, Iran
[2]Chemometrics Group, Food Technology, Department of Dairy & Food Science, Royal Veterinary & Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark
[3]Department of Marine Ecology, National Environmental Research Institute, P.O. Box 358, Frederiksborgvej 399, DK-4000 Roskilde, Denmark

Received 17 February 2006; Revised 23 June 2006; Accepted 23 June 2006

Fluorescence excitation-emission matrix (EEM) measurements are useful in fields such as food science, analytical chemistry, biochemistry and environmental science. EEMs contain information which can be modeled using the parallel factor analysis (PARAFAC) model but the data analysis is often complicated due to both Rayleigh and Raman scattering. There are several established ways to deal with scattering effects. However, all of these methods have associated problems. This paper develops a new method for handling scattering using interpolation in the areas affected by first- and second-order Rayleigh and Raman scatter in such a way that the interfering signal is, at best, removed. The suggested method is fast and requires no additional input other than specifying the scattering region. The results of the proposed method were compared with those obtained from common alternative approaches used for preprocessing fluorescence data before analysis with PARAFAC and were shown to be equally good for various types of EEM data. The main advantage of the interpolation method is in its lack of additional metaparameters, its algorithmic speed and subsequent speed-up of PARAFAC modeling. It also allows for using EEM data in software not able to handle missing data. Copyright © 2007 John Wiley & Sons, Ltd.

KEYWORDS: PARAFAC; Rayleigh scatter; Raman scatter; missing values; interpolation

1. INTRODUCTION

Fluorescence spectroscopy has become a widely used technique in many fields of physical, chemical, biological and medical sciences[1]. In simple applications, the fluorescence of a fluorophore at a specific pair of excitation and emission wavelengths is used to measure its concentration. However, more often detailed fluorescence scans are used to measure the combined signal from a mixture of known or unknown fluorophores. The mixtures of fluorophores analyzed can vary from simple laboratory mixtures [2] to complex environmental samples [3,4]. When measuring the fluorescence of mixtures the fluorescence properties are usually measured by recording measurements spanning both the excitation and emission properties of the mixture. Collating a series of emission scans from a range of excitation wavelengths produces an excitation-emission matrix (EEM) of the sample, representing a detailed map of the fluorescence properties of the mixture.

Such EEMs contain a large amount of data, which can in turn hinder the ability of the analyst to utilize all the information collected. Parallel factor analysis (PARAFAC) [2,5] is a powerful technique for analyzing the data contained within EEMs, separating the fluorescence signal of the underlying fluorophores mathematically, in much the same way as physical chromatography [2]. Fluorescence PARAFAC analysis has been shown to be useful with a wide variety of mixtures of fluorophores [6,7,8], however, fluorescence data are often plagued by scattering effects; mainly Rayleigh and Raman scatter, which can hamper PARAFAC modeling of the data unless care is taken. Since PARAFAC decomposes the fluorescence signal into a series of tri-linear structures and the scatter peaks do not behave trilinearly (e.g. shape and position of the scatter peaks changes with excitation wavelength), the scatter signal causes some mathematical difficulties in the decomposition. It is, therefore, beneficial to remove this effect, or at least to reduce its

*Correspondence to: R. Bro, Chemometrics Group, Food Technology, Department of Dairy & Food Science, Royal Veterinary & Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark.
E-mail: rb@kvl.dk

influence as much as possible. Several ways of handling scatter effects in relation to PARAFAC modeling have been proposed in literature:

- Down-weighting of the scatter region (MILES) [9,10,11],
- specific modeling of scatter [12],
- subtraction of a standard [13],
- inserting missing values [14],
- constraints in the PARAFAC decomposition [15,16],
- inserting zeros outside the data area [11],
- or plainly avoiding the part of the matrix that includes the scatter.

At present, it seems that the best method for handling scatter, in general, is to combine the above in the following way:

- Subtract an EEM of a solvent blank if such is available to minimize Raman scattering,
- replace Rayleigh bands with a band of missing values or alternatively use MILES (Maximum likelihood via Iterative Least squares EStimation) to down-weigh these,
- furthermore introduce a lower triangular set of zeros in the 'emission far below excitation' if this area has not been measured.

In some situations it is desired to avoid the use of missing values for various reasons (some algorithms or visualization tools do not handle missing data, sometimes handling missing data is extremely slow). In those situations it is possible to do the same; essentially, by down-weighing the area that is supposed to be set to missing. However, if the scatter signal is orders of magnitude larger than the relevant data then such a weighing approach is likely to be inadequate. By replacing huge scatter peaks with data in accordance with the rest of the EEM, fitting of weighted or least squares models is facilitated as well as simpler visualization. In this paper, an approach for providing values to replace the scatter variation is treated. The subject of weighted fitting has already been described in the literature [9,10,12].

Recently, a method based on replacing the scatter areas by three-dimensional Delaunay interpolation using splines has been proposed as a way to replace scatter peaks in emission spectra [17]. This method appears promising for providing a way to minimize the influence of scatter, but the method is hampered by the need to define a number of metaparameters whose optimal settings are not obvious.

In this paper, a simple way of treating the first-order Rayleigh, Raman and second-order Rayleigh scatter is proposed, involving removing and replacing the values with interpolated values. The effect of this procedure on the subsequent PARAFAC analysis is assessed and compared to other approaches, although we stress that the main aim with this method is to be able to provide more suitable values than missing data for example for weighted least squares fitting as well as for providing data that are simpler to visualize. The method was evaluated on three data sets of varying complexity and compared with modeling original data and non-interpolated data decomposition by PARAFAC modeling as well as with weighted regression.

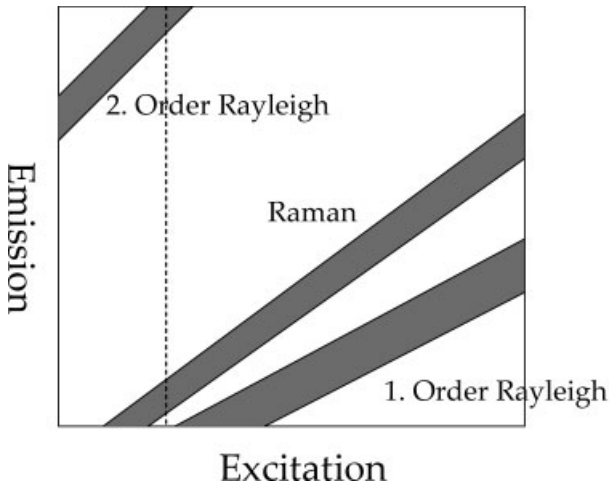

Figure 1. A sketch of the scattering occurring in a fluorescence EEM.

2. THEORY

Rayleigh and Raman scatterings are situated diagonally in the EEM as shown in Figure 1. The purpose of the interpolation is to replace the data in the gray areas with new data consistent with the data in the remaining parts of the EEM. Two-dimensional interpolation would seem to be an obvious choice but initial studies indicated that it was too complex and often led to overfitting. Using one-dimensional interpolation on individual emission spectra overcame this problem and is described in the following.

There are several ways to implement interpolation and in this study a shape-preserving piecewise cubic polynomial was chosen [18,19]. This function is directly available in MATLAB and in contrast to for example spline functions, it seeks to preserve local minima and other features of the data such that extreme artifacts are not introduced by the interpolation (see Figure 2).

In Figure 3, an example is given on how this interpolation works on a simple single emission spectrum. The interpolation has been implemented in the following way. A window width is defined for first-order and, if necessary, for second-order Rayleigh as well as for Raman scatter. These three widths are the only user-defined parameters in the interpolation. The specific position of the Rayleigh scatter is defined as the diagonal where the excitation (or two times excitation for second-order scatter) equals the emission. For Raman, the centerline is at a constant energy shift compared to the Rayleigh scatter.

For every emission spectrum, the measured signal in a window, defined by the width is removed around the scatter lines. The whole spectrum except the window is used for interpolation and the window is replaced with the interpolated values. Special care is taken for two parts of the EEM. For low-excitation wavelengths, there can be situations where there is no emission below the window. In this case, an artificial lower emission of zero is added during interpolation 30 nm below the window of interpolation (see Figure 4). Another special case is the second-order Rayleigh scatter. For first-order scatter it can be safely assumed that the

DOI: 10.1002/cem

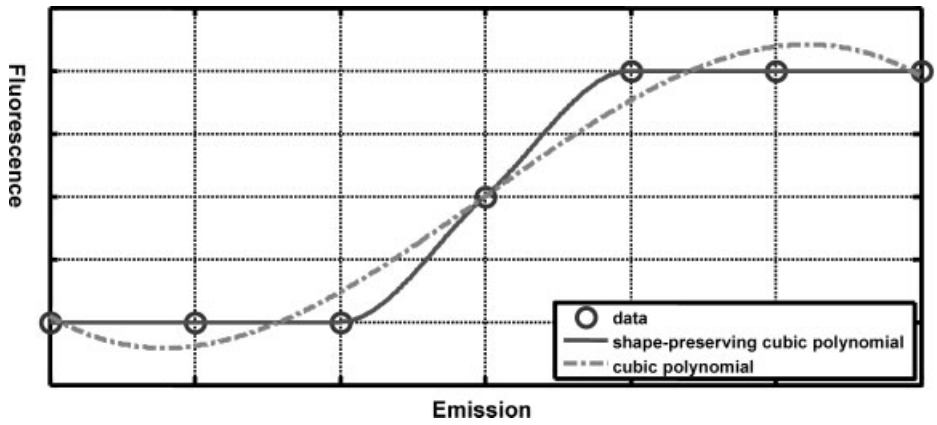

Figure 2. A simple example on how the shape-preserving interpolation avoids artifacts compared to a cubic interpolation.

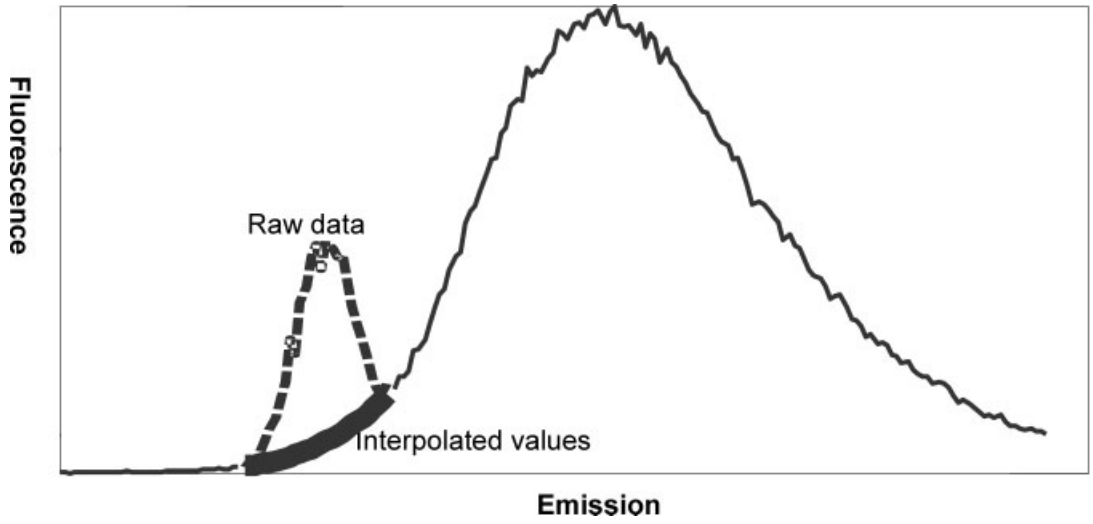

Figure 3. Removal of scatter peaks (dotted line) from an emission spectrum. The thick line is the interpolation based on the emission spectrum with a window corresponding to the thick line removed.

emission below the excitation is zero, but this is not the case for emission beyond the second-order Rayleigh because fluorescence may occur in this area (the upper left triangular part of the EEM in Figure 1). In the situation where there are no emission values at greater wavelengths than the window to be interpolated (see e.g. upper part of dashed line in Figure 1), the missing values in the last *excitation* spectrum are interpolated in order to provide end values for the emission interpolation.

The complexity of the interpolation algorithm is insignificant compared to PARAFAC. For the data used in this paper, the interpolation never exceeded 1 min of computations and as the interpolation is usually a one-shot preprocessing prior to analyzing the data, the computation time is not significant.

3. EXPERIMENTAL

3.1. The data

For analyzing the ability of the proposed new method, three different fluorescent data sets were used. Two of these can be considered standard EEM data with moderate scattering. One was made up of lab-prepared mixtures of four fluorophores in varying concentrations and the other was a set of 268 samples from a sugar factory. The two first data sets mainly help to show that the principle works while the last dataset III is more demanding.

3.1.1. Dataset I

Mixtures of four fluorophores were measured (L-phenylalanine, L-3,4-dihydroxyphenylalanine (DOPA), 1,4-dihydroxybenzene and L-tryptophan). A stock solution was made of each compound using Milli-Q water as the solvent (pH was not adjusted). Fluorescence landscapes were measured of 27 samples of varying concentrations of the four fluorophores. A Perkin-Elmer LS50B fluorescence spectrometer was used to measure fluorescence landscapes using excitation wavelengths between 200 and 350 nm with 5 nm intervals. The emission wavelength range was 200–750 nm. Excitation and emission monochromator slit widths were set to 5 nm, respectively. Scan speed was 1 500 nm/min. There are four components in data set I and a four-component PARAFAC model was therefore the most suitable [20,21].

DOI: 10.1002/cem

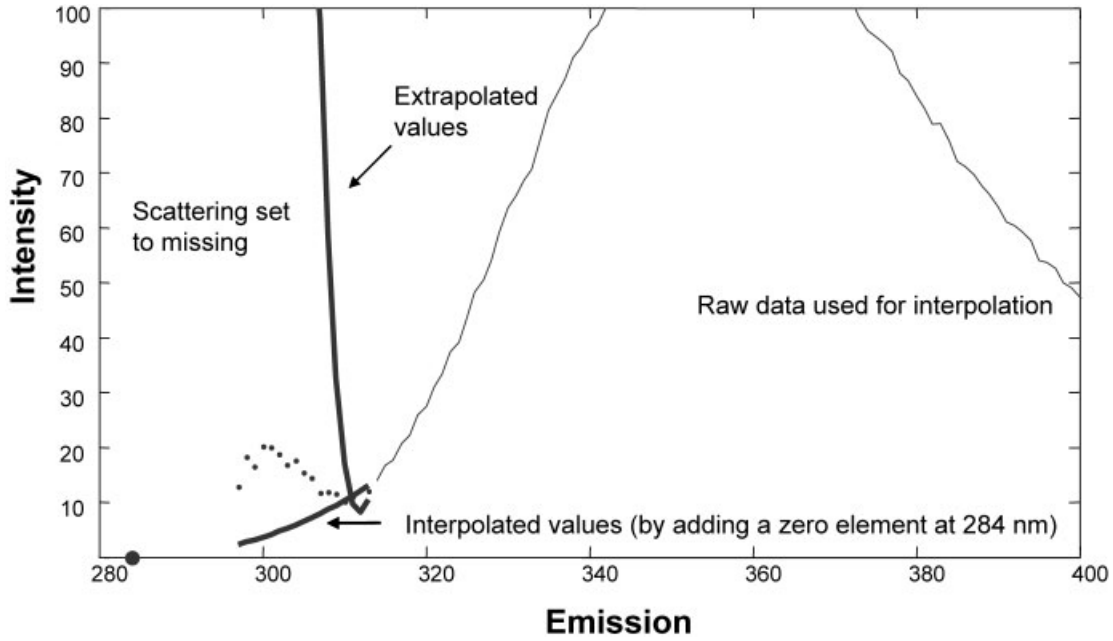

Figure 4. Handling of extrapolation in extreme cases where there is no signal below the removed scattering. A zero is inserted below the real emissions to enable interpolation and avoid artifacts.

3.1.2. Dataset II

Sugar was sampled continuously every 8 h during the 3 months of operation from a sugar plant in southern Scandinavia giving a total of 268 samples. The sugar was sampled directly from the final unit operation (centrifuge) of the process. The sugar was dissolved in un-buffered water 2.25 g/15 mL and the solution was measured spectrofluorometrically in a 10×10 mm cuvette on a PE LS50B spectrofluorometer. Raw non-smoothed data were output from the fluorometer. For every sample, the emission spectra from 275 to 560 nm were measured in 0.5 nm intervals (571 wavelengths) at seven excitation wavelengths 230, 240, 255, 290, 305, 325, 340 nm. There are four distinct PARAFAC components in data set II [7].

3.1.3. Dataset III

The fluorescence of dissolved organic matter (DOM) in seawater samples taken from the Dogger bank in North Sea was measured. A total of 21 samples were taken from two vertical profiles, sampling at 5 m depth intervals. Immediately after sampling the samples were filtered through a 0.2 μm filter, and then stored refrigerated until analysis in the laboratory (within 5 days). In addition to the profile samples, samples from zooplankton grazing experiments carried out onboard using the same sampled water were included. The fluorescence was measured on a Varian Eclipse fluorescence spectrophotometer with excitation and emission slit widths both set to 5 nm. The emission scans were from 240 to 600 nm every 2 nm and the excitation wavelength range was 240–450 nm every 5 nm. The fluorescence spectra were corrected for instrument-specific effects and Raman calibrated using the techniques described in reference [4]. However, the procedure differed slightly in that a Milli-Q blank was not subtracted from the data as the focus of this study is to examine procedures to remove scatter effects from the fluorescence signal. Spectra were measured with the maximum scan speed and highest PMT voltage setting. The measurements were made as part of an internal exercise to test the performance of the PARAFAC algorithm on samples with a low signal-to-noise ratio, and as a result the spectra are very noisy. Five PARAFAC components were identified using split half and residual analysis.

3.2. Software

MATLAB (The MathWorks, Natick, MA), version 7 was used during the calculations. The algorithms in use were from PLS_Toolbox ver. 3.5.3 (Eigenvector Research, Inc., WA). A dedicated program was written for interpolation of Rayleigh and Raman scatter area at EEM landscape which is available at www.models.kvl.dk.

4. RESULTS AND DISCUSSION

In using the interpolation method as well as the alternative approaches, the width of the scatter areas must be assessed. This was done by visual inspection of the data and confirmed subsequently by plotting parts of the preprocessed data. Very large widths will cause some uncertainty in the interpolated area whereas too narrow widths will bias the solution because scatter will be included. Approximately 1.5 times of the visually assessed scatter area was removed in order to completely remove the scatter values. For data set I ±10, ±10 and ±20 nm, for data set II ±20, ±10 and ± 0 nm (no. 2 order scatter) and for data set III ±20, ±12 and ±0 nm (no. 2 order scatter) were used for first-order Rayleigh, Raman and second-order Rayleigh scatter areas, respectively. For dataset I, typical interpolation results are shown in Figure 5 showing that even with quite wide interpolation bands the interpolation seems to provide visually sound results.

DOI: 10.1002/cem

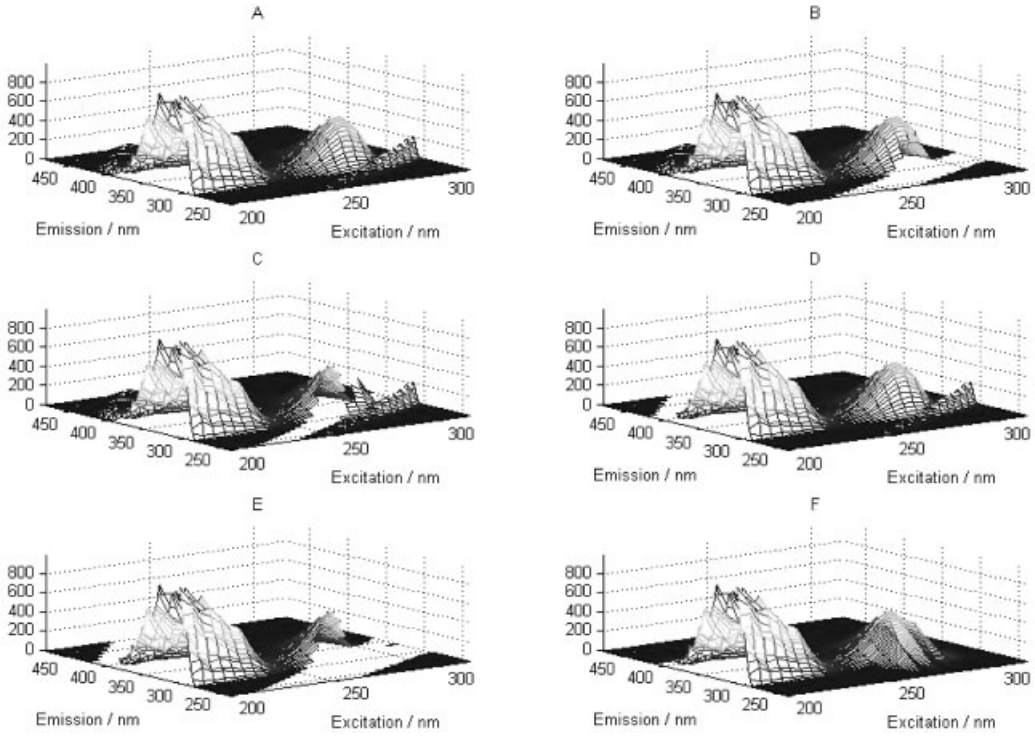

Figure 5. Sample 1 from data set I (A), first-order Rayleigh area removed (B), Raman area removed (C), second-order Rayleigh area removed (D), all three scatter areas removed (E), landscape with interpolated values (F).

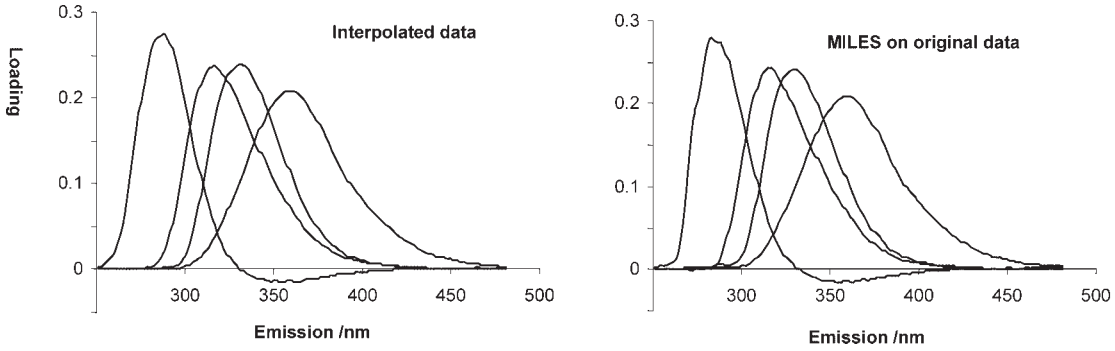

Figure 6. Example of estimated emission loadings from a PARAFAC model of dataset I.

The results of fitting PARAFAC to dataset I with interpolated values are shown in terms of the estimated emission spectra in Figure 6. In the figure, the loadings are shown together with the result of using PARAFAC–MILES which uses weighted least squares fitting to down-weigh the areas corresponding to scattering. Similar results were also obtained by setting the scatter areas as missing. Thus, interpolation is shown to provide results similar to the standard approaches with the only significant difference being that the speed of analysis was two to three times faster using interpolated data.

The second data set was originally difficult to handle because of scattering [7]. Non-negativity was applied but deemed not to be sufficient for avoiding problems pertaining

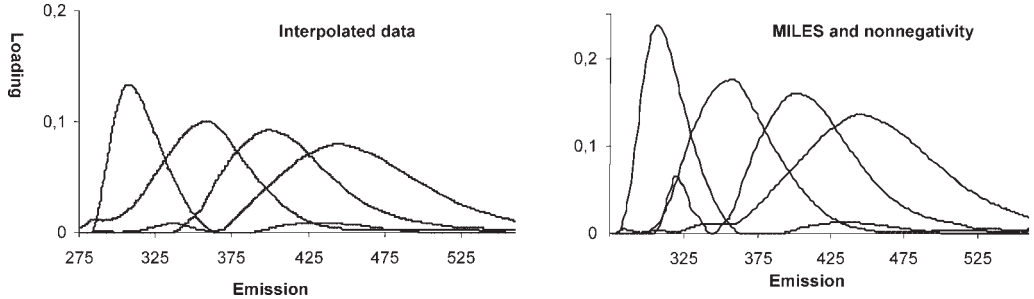

Figure 7. Loadings from PARAFAC models of dataset II.

DOI: 10.1002/cem

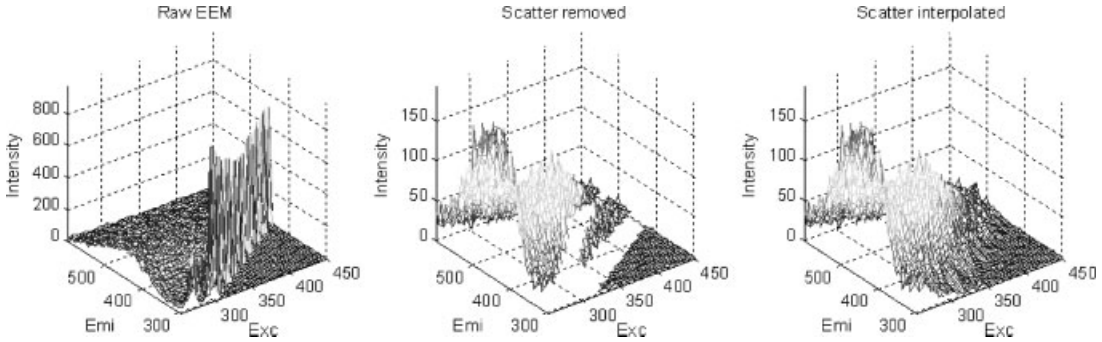

Figure 8. Leftmost, an example of the raw EEM of one sample. Next, the same EEM after scattering areas has been set to missing. The following EEM is the same but after interpolation. (Emi: Emission/nm, Exc: Excitation/nm).

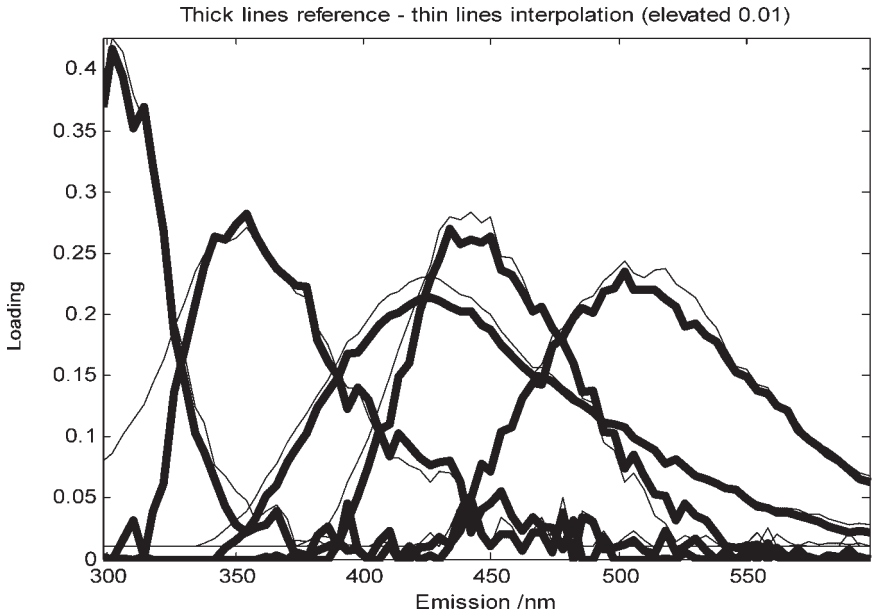

Figure 9. Comparison of emission loadings of dataset III.

to the scattering. An algorithm for imposing unimodality was developed and applied as a constraint on the emission loadings to avoid the scattering areas to influence the estimated emission loadings. The emission loadings using MILES and non-negativity are shown in right part of Figure 7. A small additional peak in the emission loadings of one of the components at 325 nm is apparent and due to scatter effects. This was originally solved by applying unimodality constraints but the results here show that the use of interpolated data rather than down-weighing (or setting scattering to missing) provides good results and, as was the case for dataset I, also speeds up the analysis.

The third dataset was the most difficult one to handle with the traditional methods. In this dataset, the amount of scatter is huge compared to the chemical signals. In Figure 8, one example is given where it is seen that the chemical information in the EEM is almost invisible compared to the scatter.

Upon removing the scatter and replacing it with interpolated values, the chemical part of the EEM is much more apparent. Still, the chemical information is weak (low signal-to-noise ratio) so some uncertainty in the estimated parameters is expected. However, both with least squares fitting or weighted least squares fitting, the results obtained are as good as the hitherto best results (see Figure 9 where the currently best results are the reference loadings) and obtained much faster.

5. CONCLUSION

It has been shown that the proposed method is suitable for extrapolating across missing values in the scatter regions in EEMs. The approach is advantageous as it only requires the width of the scatter signals as input, as opposed to additional metaparameters required in alternative approaches. In general, the results with this interpolation method are as good as existing approaches, and have the added advantage of speeding up the PARAFAC analysis. Furthermore, it enables handling of EEM data in data analysis tools that for example do not handle missing values.

DOI: 10.1002/cem

REFERENCES

1. Lakowicz JR. *Principles of Fluorescence Spectroscopy*. Kluwer Academic: New York, 1999.
2. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst*. 1997; **38**: 149–171.
3. Christensen JH, Hansen AB, Mortensen J, Andersen O. Characterization and matching of oil samples using fluorescence spectroscopy and parallel factor analysis. *Anal. Chem*. 2005; **77**: 2210–2217.
4. Stedmon CA, Markager S, Bro R. Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy. *Mar. Chem*. 2003; **82**: 239–254.
5. Harshman RA, Lundy ME. PARAFAC: parallel factor analysis. *Computat. Stat. Data Anal*. 1994; **18**: 39–72.
6. Andersson CA. Analysis of 3- and 4-way data from low-pressure liquid chromatography and spectrofluorometry. Resolving pure spectral profiles with 3-way PARAFAC and classifications with 4-way PCA, *Conference on Applied Statistics and Chemometrics*, (Høskuldsson, Agnar and Nørgaard, Lars (eds.)), Copenhagen, Thor Publishing, 1997; Applied Chemometrics, 95–104.
7. Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics Intell. Lab. Syst*. 1999; **46**: 133–147.
9. Bro R, Sidiropoulos ND, Smilde AK. Maximum likelihood fitting using simple least squares algorithms. *J. Chemometrics* 2002; **16**: 387–400.
8. Moberg L, Robertsson G, Karlberg B. 3d Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis. *Talanta* 2001; **54**: 161–170.
10. Jiji RD, Booksh KS. Mitigation of Rayleigh and Raman spectral interferences in multiway calibration of excitation-emission matrix fluorescence spectra. *Anal. Chem*. 2000; **72**: 718–725.
11. Rinnan A, Andersen CM. Handling of first-order Rayleigh scatter in PARAFAC modelling of fluorescence excitation-emission data. *Chemometrics Intell. Lab. Syst*. 2005; **76**: 91–99.
12. Rinnan A, Booksh K, Bro R. First order Rayleigh as a separate component in the decomposition of fluorescence landscapes. *Anal. Chim. Acta* 2005; **537**: 349–358.
13. McKnight DM, Boyer EW, Westerhoff PK, Doran PT, Kulbe T, Andersen DT. Spectrofluorometric characterisation of dissolved organic matter for indication of precursor organic material and aromaticity. *Limnol. Oceanogr*. 2001; **46**: 38–48.
14. Christensen J, Povlsen VT, Sørensen J. Application of fluorescence spectroscopy and chemometrics in the evaluation of processed cheese during storage. *J. Dairy Sci*. 2003; **86**: 1101–1107.
15. Andersen CM, Bro R. Practical aspects of PARAFAC modelling of fluorescence excitation-emission data. *J. Chemometrics*. 2003; **17**: 200–215.
16. Bro R, Sidiropoulos ND. Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometrics* 1998; **12**: 223–247.
17. Zepp RG, Sheldon WM, Moran MA. Dissolved organic fluorophores in southeastern US coastal waters: correction method for eliminating Rayleigh and Raman scattering peaks in excitation–emission matrices. *Mar. Chemistry* 2004; **89**: 15–36.
18. Carlson RE, Fritsch FN. An algorithm for monotone piecewise bicubic interpolation. *SIAM J. Num. Anal*. 1989; **26**: 230–238.
19. Kahaner D, Moler C, Nash S. *Numerical Methods and Software*. Prentice Hall, Upper Saddle River, NJ, USA, 1989.
20. Baunsgaard D. Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes. *Internal report, Dept. Dairy and Food Science, The Royal Veterinary and Agricultural University Denmark*. 1999.
21. Riu J, Bro R. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. *Chemometrics Intell. Lab. Syst*. 2003; **65**: 35–49.

DOI: 10.1002/cem