ELSEVIER

Chemometrics and Intelligent Laboratory Systems 46 (1999) 133–147

Chemometrics and intelligent laboratory systems

# Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis

Rasmus Bro [a,b,*]

[a] *Chemometrics Group, Food Technology, The Royal Veterinary and Agricultural University, Frederiksberg, Denmark*
[b] *Matforsk, Oslovеien 1, 1430 Ås, Norway*

Received 5 March 1998; revised 23 April 1998; accepted 20 May 1998

**Abstract**

This paper is concerned with the possibility of obtaining chemically meaningful models of complicated processes by the use of fluorescence spectroscopy screening and the unique parallel factor analysis (PARAFAC) model. The second-order nature of fluorescence excitation emission data and the fact that the PARAFAC model has no rotational indeterminacy mean that in certain cases, it is possible to decompose complex mixture signals into contributions from individual chemical components. Relating the thus obtained information to, e.g., important quality parameters, it is possible to analyze, understand, predict and monitor the quality based on a chemical foundation. The proposed approach thus gives a direct link between process analytical chemistry and multivariate statistical process control. © 1999 Elsevier Science B.V. All rights reserved.

*Keywords:* PARAFAC; Multilinear PLS; N-PLS; Uniqueness; Non-negativity; Unimodality; MSPC; PAC

## Contents



* Department of Dairy and Food Science, Chemometrics Group, Food Technology, The Royal Veterinary and Agricultural University, Rolighedsvej 30, 1958 Frederiksberg, Denmark. E-mail: rb@kvl.dk

0169-7439/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0169-7439(98)00181-6

## 1. Introduction

There is a need in the sugar industry to rationalize and improve quality and process control. One aspect of this is to gain a better understanding of the chemistry involved in the process. This can lead to better guidance of the sugar-beet growers and to a better foundation for controlling the process. Earlier investigations have primarily focused on establishing which chemical analytes are present in the sugar and intermediate products. Winstrøm-Olsen et al. [1,2] were able to separate a dozen catecholamines from raw juice of sugar and found that the typical concentration of norepinephrine was about 1–2 ppm, while the color-forming precursor dopa (3,4-dihydroxyphenylalanine) typically appeared in the concentration range 1–5 ppm. In Refs. [3,4], the results were further elaborated on by isolating from beets enzymatic material which was then characterized with respect to how it affected catecholamines in pure solutions producing colored components (melanins). It was noted that if more than one catecholamine was present, the enzymatic effect was difficult to predict due to the interrelationship between the catecholamines and the enzymes.

This type of information seldom leads to conclusive suggestions regarding a complicated process like the sugar manufacturing. It is an extremely expensive and reductionistic approach to learning about technological aspects of sugar production. In this case, it is additionally well-known that the enzymatic miscoloring of sugar is but one of several ways that miscoloring occurs. For example, non-enzymatic colorforming processes due to reaction of amino acids and phenols with reducing sugars creating melanoidines (Maillard products) are also known to be important.

An attempt to use a more exploratory approach can be based on the following alternative strategy, much in line with the exploratory approach suggested by Munck et al. [5] in which fluorescence analysis is used to monitor the beet sugar process from the beet raw-material to the intermediate products and the final product sugar:

- Measure sugar samples spectrofluorometrically;
- Decompose the spectra using the parallel factor analysis (PARAFAC) model;
- Identify correlations between PARAFAC scores and quality and process parameters;
- Identify the underlying chemical components generating the relevant PARAFAC components; and
- Utilize these components as indicator substances to monitor the process throughout the different production steps as a screening analysis for chemical, physical and process parameters.

Thus, the sugar samples are observed and measured non-selectively and almost directly in the process, instead of being 'dissected' in a chemical laboratory analysis. This can bring about more relevant information.

The first hypothesis initiating such an analysis is that fluorescence data contain valid information about the sugar production, hence a very broad, non-focused, and little biased hypothesis. From the preliminary results and *indications* obtained, new hypotheses and new problems can be formulated gradually leading to stronger provisional hypotheses based mainly on the experience from the real world. Validation is to be used throughout to ensure the quality of the results. Validation can be based on both statistical and numerical validation *as well as* external validation, i.e., assessing and comparing the model to the real world.

In the following, the data will first be described, then the PARAFAC model as well as the multilinear partial least squares regression model (N-PLS) is described. A PARAFAC model of the fluorescence data will be developed and validated with respect to numerical as well as exploratory aspects. This part of the paper is quite central and will be elaborated on in some detail. The results obtained solely from this model will first be elaborated on. The results lead to the secondary hypothesis that fluorescence data re-

flect chemical variation and are related to the quality and other external parameters. Note that this *is not* an a priori hypothesis arising from theoretical considerations, but a result of the general primary hypothesis initiating the exploration. The results from the secondary hypothesis will be investigated by showing how the fluorescence information compares to parameters relevant in the sugar production.

## 2. Data

Sugar was sampled continuously during 8 h to make a mean sample representative for one 'shift' (8-h period). Samples were taken during the 3 months of operation (the so-called campaign) in late autumn from a sugar plant in Scandinavia, giving a total of 268 samples of which three were discarded as extreme outliers in this investigation. 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 $10 \times 10$ mm cuvette on a PE LS50B spectrofluorometer. Raw non-smoothed data was 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). To the left in Fig. 1, a typical sample is shown.

The data of all the 265 samples can be arranged in an $I \times J \times K$ three-way array of specific size $265 \times 571 \times 7$. The first mode refers to samples, the second to emission wavelengths, and the third to excitation wavelengths. The $ijk$th element in this array corresponds to the measured emission intensity from sample $i$, excited at wavelength $k$, and measured at wavelength $j$. The reason for the large number of emission wavelengths is that the fluorometer used in this study only allows emission in half nanometer steps to be measured. One may, of course, simply use a subset of the wavelengths if physical computer memory is limited.

Also available were laboratory determinations of the quality of the produced sugar sampled at the same rate. These quality measures are ash content and color. Ash content is determined by conductivity and is a measure of the amount of inorganic impurities in the refined sugar. It is given in percentages. Color is determined as the absorption at 420 nm of a membrane-filtered solution of sugar adjusted to pH 7. The color is given as a unit derived from the absorbance where 45 is the maximum allowed color of standard sugar. It gives an indication of the miscoloring of the sugar. This color is by far so low, that it is of no importance for the consumer, but it is of interest for process control and for retailers.

Finally, 67 automatically sampled process variables were available of which 10 were sampled so infrequently that they were not included in this investigation. The process variables include temperature, flow, and pH determinations at different points

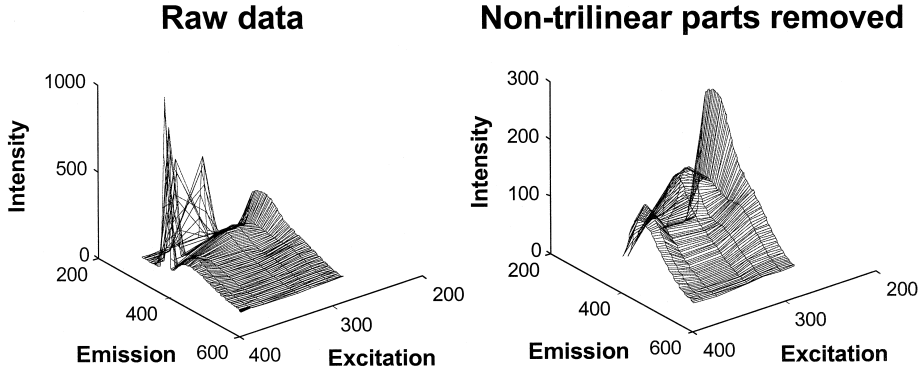


Fig. 1. Fluorescence data from one sugar sample. To the left, the raw data are shown and to the right, the data are shown after removal of emission below excitation wavelength as well as Rayleigh scatter (upper left part).

in the process. Typically, these variables are noisy and sampled at quite different rates. For this investigation, all process measurements have been resampled to the same frequency as the above variables by simply ignoring additional measurements. This simple approach is justified, by the fact that the process data are not of primary concern at this stage of the exploratory analysis. Only indications of patterns and relationships which are technologically and chemically explanatory are sought. For similar reasons, 22 of the 67 process variables were selected in this investigation. This was done to eliminate the irrelevant variables simply by removing those variables that were almost orthogonal to the ash and color determinations after removal of gross outliers. The 22 selected variables are primarily pH measurements from different parts of the process, but also some temperatures, flows and other variables.

## 3. Theory

Scalars including elements of vectors and matrices are indicated by lower-case italics, vectors by bold lower-case characters, and bold capitals are used for two-way matrices. The letters $I$, $J$, and $K$ are reserved for indicating the dimension of the first, second, and third mode of a three-way array, respectively, and $i$, $j$, and $k$ are used as indices for each of these modes. An $I \times J \times K$ array is defined by its elements as $x_{ijk}$, implicitly assuming that the indices run from one to the respective dimensionalities.

### 3.1. The PARAFAC model

Parallel factor analysis is a decomposition method for multi-way data [6]. In case of a three-way analysis, a decomposition of the data is made into triads or trilinear components. The structural model behind two-way PCA is a bilinear model:

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf}, \qquad (1)$$

ignoring the residual term for simplicity. A PARAFAC model of a three-way array is given by three loading matrices, **A**, **B**, and **C** with typical elements $a_{if}$, $b_{jf}$, and $c_{kf}$. The PARAFAC model is defined by the structural model:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}. \qquad (2)$$

Algorithms for estimating the PARAFAC model can be found in Refs. [7–9]. All algorithms used here are implemented in MATLAB™ and are available from http://newton.foodsci.kvl.dk/foodtech.html.

An obvious advantage of the PARAFAC model is the uniqueness of the solution. In bilinear models, there is the well-known problem of rotational freedom. The loadings in a spectral bilinear decomposition reflect the pure spectra of the analytes measured, but it is not possible without external information to actually estimate the pure spectra because of the rotation problem. In most circumstances, the PARAFAC model is uniquely identified up to scaling and permutations of components. Hence, no postprocessing is necessary as the model is *the* best model in the least squares sense. If the data are approximately trilinear, the true underlying parameters will be approximated if the right number of components is used and the signal-to-noise ratio is appropriate [10–12]. The conditions under which the PARAFAC decomposition is unique can be found in the literature [10,12,13]. In practice, though, one may expect a decomposition to be unique unless some loading vectors are exactly identical.

In order to find the proper complexity of a PARAFAC model, several approaches are often used. The main technique used here to determine the optimal number of components is the so-called *split-half analysis* [14,15]. In the split-half analysis, different subsets of the data are analyzed independently. Due to the uniqueness of the PARAFAC model, the same loadings will be obtained in the non-splitted modes from models of any suitable subset of the data if the correct number of components is chosen. If too many or too few components are chosen, the model parameters will differ if a different data set is used for estimating the parameters. Even though the model may be unique, the model parameters will be dependent on the specific samples as the amount of underlying phenomena present in the data set determines which linear combination of the intrinsic set of profiles and the noise will give a unique solution for the specific

model at hand. To judge if two models are equal, the indeterminacy in multilinear models has to be respected: the order and scale of components may change if not fixed algorithmically. If a model is stable in a split-half sense, it is a clear indication that the model is real; that it captures essential variation, that not only pertains to the specific samples.

In order to avoid that, an unlucky splitting of the samples causes some phenomena to be absent in certain groups, the following approach is often adopted in split-half analysis. The samples are divided into two groups: (A) and (B). If the samples are presumed to have some kind of correlation in time, the sets are constructed contiguously, i.e., (A) consists of the first half of the samples and (B) of the last. Accidentally, it may happen that one of these sets does not contain information on all latent phenomena. To ensure or at least increase the possibility, that the sets to be analyzed cover the complete variation two more sets are generated, (C) and (D). The set (C) is made from the first half of (A) and (B) and the set (D) consists of the last half the of samples in (A) and (B). These four sets are pairwise independent. If the solution replicates when estimated from set (A) and (B) *or* from set (C) and (D), correctness of the solution is empirically verified.

Constraining a model can sometimes be helpful. For example, resolution of spectra may be wanted. To ensure that the estimated spectra make sense, it may be reasonable to estimate the spectral parameters under non-negativity constraints as most spectral parameters are known to be non-negative. Constraints can for example help to: (i) obtain parameters that do not contradict with a priori knowledge (e.g., require chromatographic profiles to have but one peak); (ii) obtain unique solution where otherwise a non-unique model would be obtained (e.g., use selective channels in data to obtain uniqueness); (iii) test hypotheses (e.g., investigate if tyrosine is present in sample); (iv) avoiding degeneracy and numerical problems (e.g., enabling a PARAFAC model of data otherwise inappropriate for the model), (v) speed up algorithms (e.g., use truncated bases to reexpress problem by a smaller problem), (vi) enable quantitative analysis of qualitative data (e.g., incorporate sex and job type in a model for predicting income).

It may be argued that constraining the PARAFAC model is superfluous, as the structural model in itself should be unique. However, even though the model is unique, the model may not provide a completely satisfactory description of the data. Rayleigh scatter in fluorescence spectroscopy is but one instance where slight model inadequacy can cause the estimated model parameters to be misleading. Constraints can be helpful in preventing that. In other situations, numerical problems or intrinsic ill-conditioning of the component matrices can make a model problematic to estimate. At a more general level, constraints may be applied simply because they are known to be valid. This can give better estimates of model parameters and of the data (see Ref. [7]).

A constrained model will fit the data poorer than an unconstrained model, but if the constrained model is more interpretable and realistic, this may justify the decrease in fit. Applying constraints should be done carefully considering the appropriateness beforehand, considering why the unconstrained model is unsatisfactory, and critically evaluating the effect afterwards. In Refs. [7,16,17], it is discussed in detail how to implement and assess constraints properly.

### *3.2. The multilinear partial least squares regression model*

In 1989, Ståhle [18] extended the PLS model to three-way data by extending the two-way algorithm in a straightforward manner. The optimality of the proposed algorithm, however, was not substantiated. Later, Bro [19] developed a general multi-way PLS regression model (N-PLS) which was shown to be optimal according to the underlying theory of two-way PLS regression. Smilde [20] and de Jong [21] further elaborated on the properties of N-PLS.

In the three-way version of PLS, the three-way array of independent variables is decomposed into a trilinear model similar to the PARAFAC model, only for N-PLS, the model is not a least squares model of the data, but seeks in accordance with the philosophy of PLS to describe the covariance of the dependent and the independent variables. This is achieved by simultaneously estimating a multilinear model of the dependent variables and a multilinear model of the independent variables. There are several delicate points in actually implementing an algorithm for this model. These can be found in Refs. [19–21].

## 4. A model of the fluorescence data

The fluorescence emission intensity in a sample $i$ measured at emission wavelength, $j$, excited with light at wavelength $k$, can be expressed:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf}, \tag{3}$$

plus some additional residual variation. The parameter $b_{jf}$ is linearly related to the relative fluorescence intensity of the $f$th fluorophore at emission wavelength $j$. The parameter $c_{kf}$ is linearly related to the extinction coefficient of analyte $f$ at excitation wavelength $k$ and the relative concentration of analyte $f$ in sample $i$ is $a_{if}$. This relation holds approximately for diluted solutions [22,23]. For weak solutions, fluorometric data can thus be approximated by a PARAFAC model, with the exception that for each sample the measured excitation–emission matrix (size $J \times K$, specifically $571 \times 7$) has a part that is systematically missing in the context of the trilinear model [22]. Emission is not defined below the excitation wavelength and due to Rayleigh scatter, emission slightly above the excitation wavelength does not conform to the trilinear PARAFAC model. As the PARAFAC model only handles regular three-way data, it is necessary to set the elements corresponding to the 'non-trilinear' areas to missing, so that the estimated model is not skewed by these data points (see Fig. 1). The PARAFAC loss function defining the optimization criterion of the fitting algorithm is then only optimized over non-missing elements using expectation maximization. Specifically, the missing elements are iteratively replaced by their model estimates during fitting [7]. It is very important to note, that the elements in this triangular part of the matrix holding the data of each sample cannot be replaced with, e.g., zeros. Even though emission well below the excitation wavelength is approximately zero, this part *does not* conform to the trilinear model. Therefore, no matter if these data are absent or not, they should be treated as missing. In this case, a large part of the data is missing in the emission area from 275 to 340 nm, hence making the model prone to some instability in this area.

### 4.1. An unconstrained model

Unconstrained PARAFAC models of the fluorescence data were fitted using from one to eight components. In Table 1, the fit-values are given as:

$$\text{Fit\%} = 100\left(1 - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} e_{ijk}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}^2}\right), \tag{4}$$

where $e_{ijk}$ is the residual of the $ijk$th element and $x_{ijk}$ the corresponding data element. The summations are only performed over non-missing elements (corresponding to setting missing residual and data elements to zero in the above equation). Evidently, it is difficult to decide upon the complexity of the model from fit values alone. The primary conclusion must be that two or more components are needed. An additional indication of the appropriate number of components stems from the fact that for more than four components, the model parameters were extremely difficult to estimate. The fitting took very long time and an increased number of refitted models ended up in local minima. Though, this provides no conclusive evidence, it does indicate that four components may be appropriate.

A split-half analysis also confirms that four components is most suitable for the PARAFAC model since for higher number of components several loading vectors change considerably depending on the subset used (Fig. 2) and regardless of constraints used.

In Fig. 3a, the estimated emission loading vectors of a four-component model are shown. Some of the

Table 1
Fit values in percentages vs. the number of components in a PARAFAC model of the fluorescence data

| Number of components | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Fit (%) | 94.22 | 98.99 | 99.63 | 99.91 | 99.94 | 99.96 | 99.97 | 99.98 |

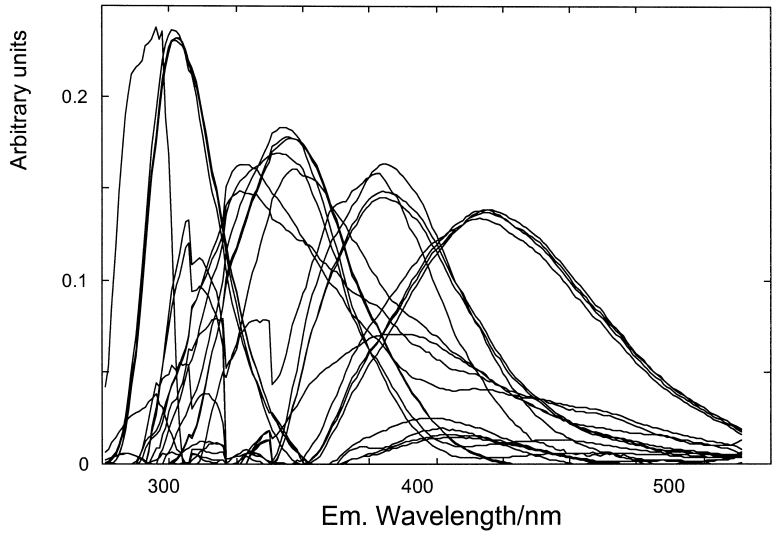


Fig. 2. Emission loading vectors resulting from a split-half analysis of a five-component non-negativity constrained PARAFAC model. The resulting loadings of the four models of different samples are shown together in the plot.

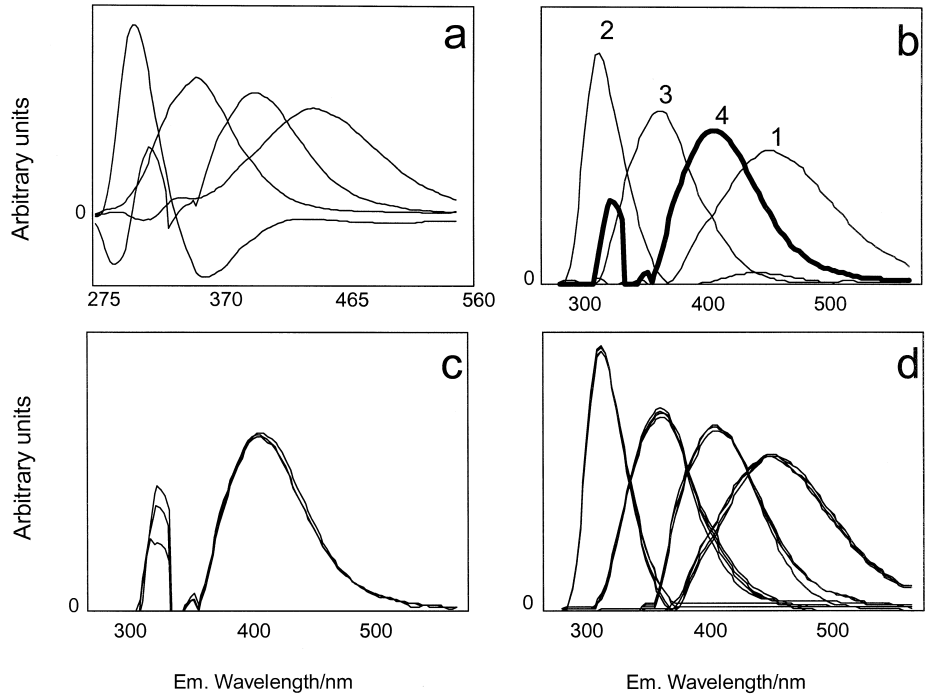


Fig. 3. Emission mode loading vectors estimated from four-component PARAFAC model of fluorescence data of the sugar samples. (a) Unconstrained model. (b) Non-negativity constrained model. The 'suspicious' spectrum, four, is marked with a thicker line. (c) Split-half estimates of component four of non-negativity constrained model. (d) Split-half estimates of all components using non-negativity and unimodality.

parameters in the left part of the plot are negative. Since the emission spectra of pure analytes should not be negative, it is worthwhile to investigate if a non-negativity constrained model can help remedying this.

### 4.2. Using non-negativity

It may be inferred, that non-negativity should not be necessary, since the model should be identifiable even without using non-negativity. The PARAFAC model, however, is only an approximation of how the fluorescence data arise. There is a large portion of the data missing in the low wavelength area. Since only few excitations are hence available here, only slight model-errors may strongly bias the parameter estimates. Also very likely some of the elements that have not been set to missing may be influenced by Rayleigh scatter to a slight degree. Furthermore, heteroscedasticity, quenching and other deviations from the model can cause problems.

It is preferable that estimating the data by an unconstrained and a constrained model would i) give similar results, and ii) that any deviations between the models should be explainable and plausible. Indeed, similar results are obtained from an unconstrained and a non-negativity constrained model. In the sample and excitation modes the loadings of the two models are very correlated ($\approx 0.99$). In Fig. 3b the estimated emission spectra are shown. From visual inspection the spectra seem mainly reasonable, but for spectrum four, the bump slightly above 300 nm is not satisfactory. This may implicitly be caused by the low number of emission wavelengths in this area. From 275 to 312 nm only three of the seven excitations are present, hence almost 60% is missing. From 360 nm and above no variables are missing.

To possibly substantiate the visual judgement, a split-half experiment was performed by dividing the samples into four groups as described earlier. The resulting model estimates of the problematic emission spectrum are shown in Fig. 3c. The area around 300 nm is seen to be unstable in a split-half sense. The estimated parameters in this region change depending on which subset of samples is used for estimating the model, whereas the remaining parameters are more or less insensitive to subset variations. The split-half experiment thus confirms that the area is ill-modeled. The following features all indicate that the 300 nm area is unreliable:

- The parameters are even visually off-the-mark, in the sense that wavelength-to-wavelength changes are not smooth;
- The split-half experiment shows that the parameters cannot be identified in a stable fashion;
- The fact that the data contain many missing values (60%) in the area of the unstable region indicates one reason why the instability occurs.

It is important to note that each of the four submodels in the split-half analysis is uniquely estimated. However, the uniqueness of each model is governed partially by random or non-trilinear variation and hence not attributable to the real underlying systematic trilinear variation. Therefore, the solution for a given subset of data does not generalize to other sets. The question then is what to do? As the most probable cause for the problem is that too few excitation wavelengths have been used (seven), the best thing to do would probably be to remeasure the samples using more excitation wavelengths. However, the measurements as they are currently being performed require a substantial amount of work, and remeasuring is therefore not realistic.

### 4.3. Using non-negativity and unimodality

One possible cause for the unsatisfactory four-component model could be that there should indeed be five components in the model. However, as noted earlier, split-half analysis and the general convergence problems indicated that five components could not be modeled with the PARAFAC model. A four-component model seems more appropriate.

Even though unimodal spectra are not in general common, several aspects indicate that the 'problematic' spectrum should be unimodal.

- The spectrum *is* unimodal apart from the unstable part.
- The remaining estimated emission spectra are almost unimodal.
- The most likely fluorophores in sugar (amino acids, simple phenols, and derivatives) have unimodal emission spectra.
- The Kasha rule [24] states that a fluorophore will emit light under the same ($S_1-S_0$) transition regardless of excitation, i.e., an excited molecule will

drop to the lowest vibrational level through radiationless energy transfer, and then from the excited singlet level $S_1$ return to the ground state $S_0$ by fluorescence [22]. Even though there are exceptions to this rule, it often holds especially for simple molecules. The fact that the emission occurs from the same transition, mostly implies that the corresponding emission spectrum will be unimodal.

The above reasoning led to specifying a new model where all emission spectra were estimated under non-negativity [16] and unimodality [17] constraints and remaining parameters under non-negativity constraints. The estimated model was stable in a split-half sense (Fig. 3d) and the estimated excitation spectra and relative concentrations did not change much from that of the non-negativity constrained model. This confirms that the cause of the artifact in Fig. 3c is mainly due to the amount of missing data in the specific region. Otherwise, if the bump was caused by a hypothetical fifth component, it would be expectable that the scores and excitation mode loadings would change considerably when imposing unimodality, since the unimodality apparently filters off the influence of this fifth component in the emission mode. Since this does not happen, then either there is no fifth trilinear component or this fifth component has such a low variation that eliminating it does not affect the remaining components. Since a five-component model cannot be reliably estimated, the four-component model is preferable and not likely to be affected even if five components are present. Unimodality is therefore a valid constraint and mainly necessary for improving the visual appearance of the emission loadings, hence enabling better identification of the underlying analytes.

### 4.4. Exploring the PARAFAC model

Selected estimated spectra are shown in Fig. 4 together with the emission and excitation spectra of tyrosine and tryptophan, two substances of known technological importance. The spectra of tyrosine and tryptophan were acquired in experiments unrelated to this study. Still, the similarity confirms that the PARAFAC model is capturing chemical information. In order to verify with more confidence the identity

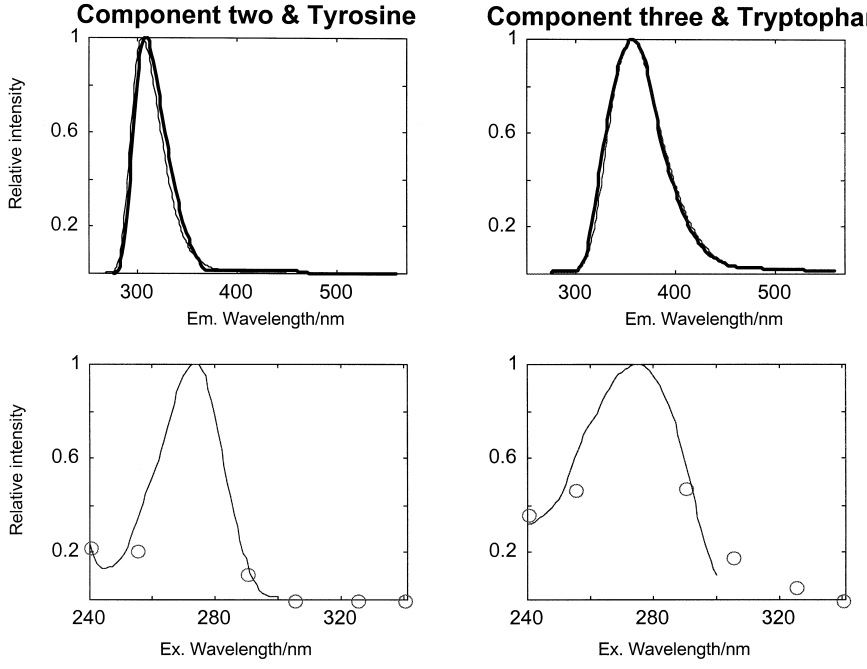


Fig. 4. Comparing estimated spectra from sugar samples with selected spectra of pure analytes. PARAFAC emission parameters shown with thick lines (emission)/circles (excitation). Note the good estimates of the emission spectra, while the excitation spectra are less well estimated especially due to lack of excitation measurements between 255 and 290 nm.

of the (additional) underlying analytes, standard addition or chromatographic analysis could be used. The similarities observed here can be taken as preliminary indications.

The scores of the model of the fluorescence data are estimates of relative concentrations, or in a more general setting, a window into the chemistry of the sugar beets. In the score plot (Fig. 5), several interesting features are revealed. All four scores seem to follow the same overall pattern. The first half of the campaign they vary with the same low frequency. This frequency follows the weeks—6 weeks, six periods. From shift number 150 to 200, the variation is more modest and in the final period from shift 200, only minor variations are seen with the exception that component four increases steadily. These distinct patterns of variation must be reflecting a variation in the chemistry of the sugar during the campaign. A preliminary hypothesis—to be investigated—that may explain these variations is based on the following observations. The beets are stored before entering the factory. The storage time differs, and there is a pile up of beets during the weekend. During storage, a significant increase in temperature is likely to occur possibly leading to increased enzymatic activity which can then be reflected in the weekly patterns of the fluorescence scores seen in the first half of the campaign. In this part of the campaign, the weather was relatively warm and all scores follow the same overall pattern. This also explains why the variations generally decrease in time as the outdoor temperature influencing the stored beets steadily decreases during the campaign. The increase in the amount of compound four from shift 200 (15 November) seems to be correlated with the onset of the frost according to the process records, hence again a temperature phenomenon. During this time, the variation in component four is highly correlated to color [5]. To the extent that these provisional ideas and hypotheses are correct, they indicate that controlling the temperature of the incoming beets is *the* most important factor for maintaining a well-controlled process on a chemical level. To the extent that the chemical information from fluorescence carries information on other parameters, this conclusion carries over to the process quality as well.

## 5. Using PARAFAC scores for modeling process parameters and quality

In order to see whether there is any connection between the variation in the chemistry of the sugar as given by the four PARAFAC scores, the variation in the quality parameters of sugar, and the process variables, several models were investigated. In the sequel, only the scores of the PARAFAC model will be used in the models investigated. Even though it may be feasible to use the raw fluorescence data directly, the idea here is to explore whether the chemical representation of the fluorescence data, i.e., the PARAFAC model, can be related to process and laboratory data.

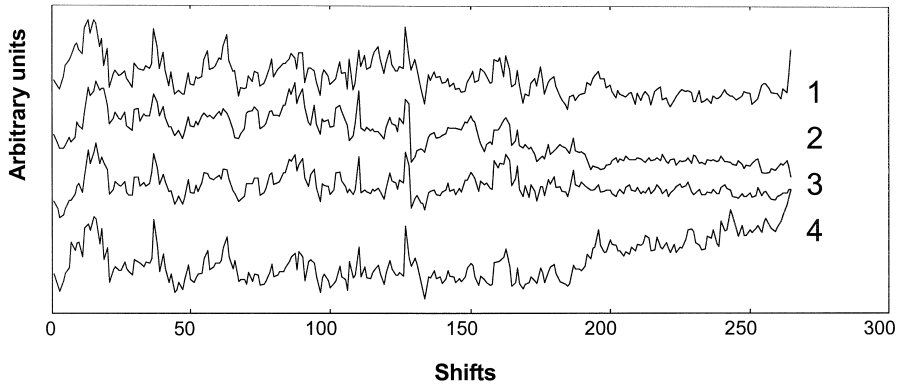


Fig. 5. The scores of the four-component PARAFAC model of the sugar fluorescence data. The score vectors have been shifted vertically for visual appearance. As the samples are ordered according to time, the line plots represents the variation of each component over the campaign, each sample representing an 8-h period. Component two is the tyrosine-like and component three, the tryptophan-like component (see Fig. 4).

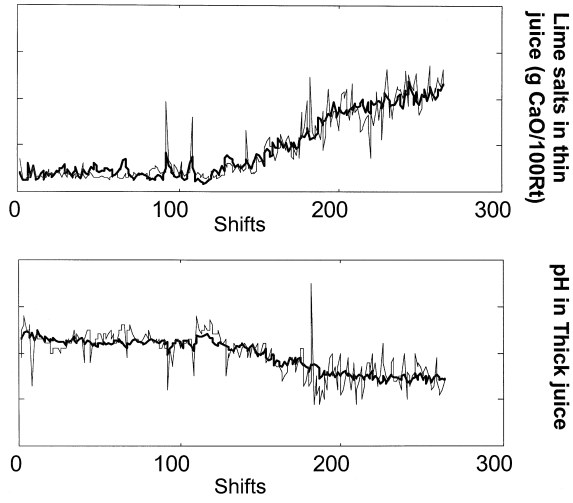


Fig. 6. Predictions of selected process parameters from the four PARAFAC scores of the fluorescence data model. Thin lines are reference values. Notice the smoothing effect of the predictions due to the combined effect of noisy process measurements and the sugar samples being average samples collected over 8 h.

Initially, the correlation between the PARAFAC scores and the process variables was investigated. For some process variables, there were almost no correlation, but for a large number of process variables excellent correlations were obtained. For examples of this, see Fig. 6. Here, the fitted values obtained using multiple linear regression (MLR) are shown. MLR was chosen because the condition number of the matrix of independent variables (265 samples $\times$ 4 PARAFAC scores) is low, hence no problems arising from collinearity are expected. Secondly, because the aim here is not to establish the exact predictability, little attention was paid to assessing predictability in terms of cross-validation, etc. Instead, the statistics of the MLR models were observed in order not to obtain misleading results.

Multiple linear regression models were also made for predicting the quality parameters ash content and color from PARAFAC scores. The models for predicting ash and color of the sugar were excellent. The

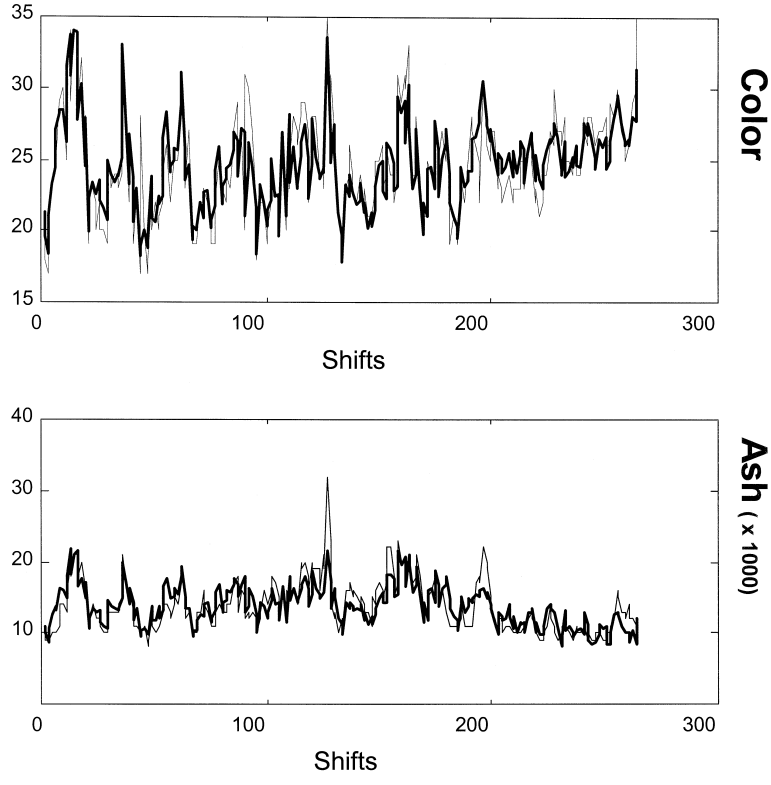


Fig. 7. The MLR predictions of color and ash from PARAFAC scores. Thin lines are reference values.

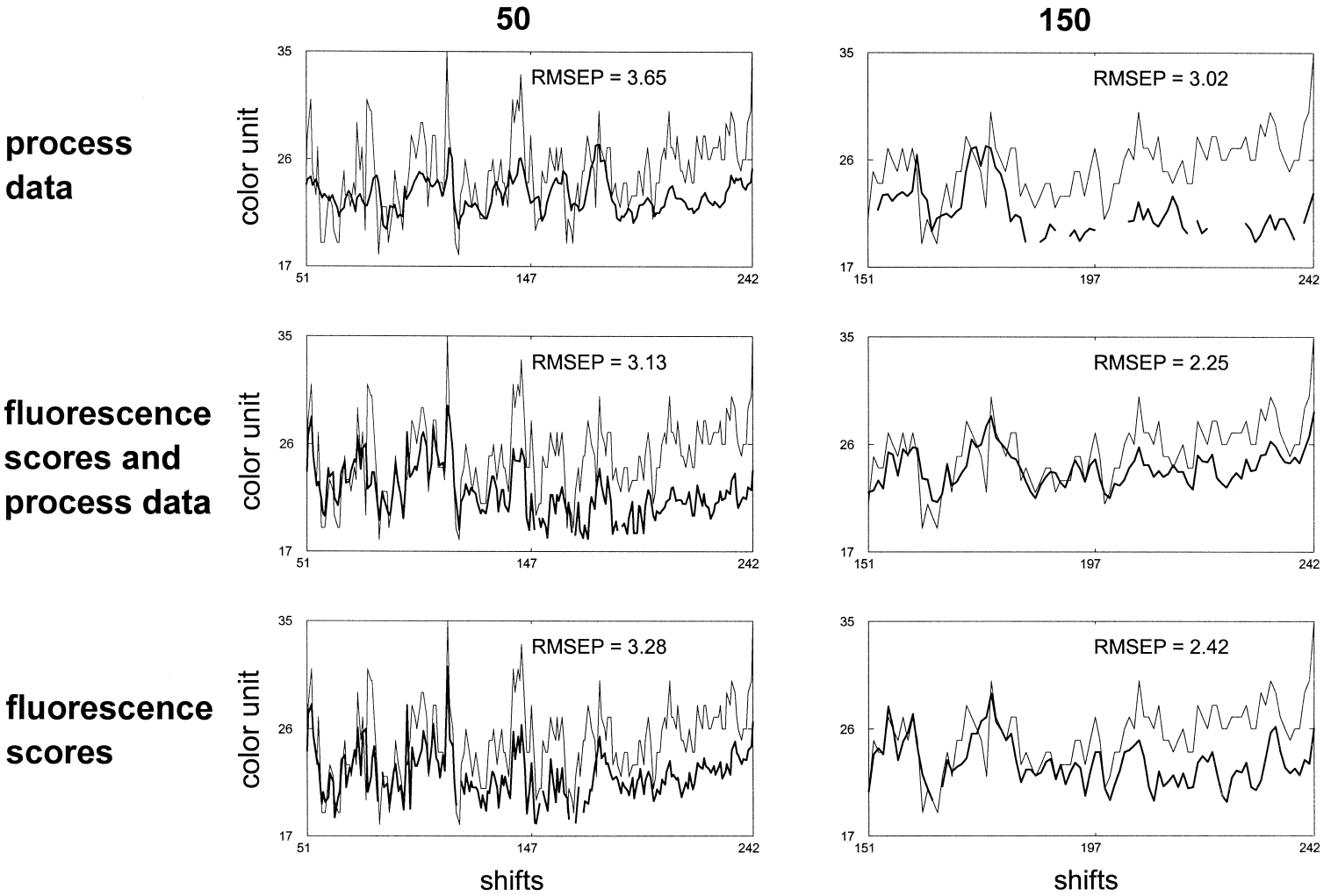


Fig. 8. Predictions of color using N-PLS of lagged data. The reference values are shown with thin lines and the predictions with thick lines. The top two figures show the predictions obtained from using only the process data as independent variables; the middle figures show the predictions obtained using both process variables and fluorescence PARAFAC scores as independent variables; and the lower ones show the predictions obtained using only the fluorescence PARAFAC scores as independent variables. The leftmost three figures give the results using the first 50 samples and predicting the remaining ones. The rightmost give the predictions using the first 150 samples and predicting the rest. Above, in each figure, the root mean squared error of predictions for the first 30 (10 days) samples is shown. Predictions outside the ordinate scale have been removed for consistency.

predicted values are shown together with the reference values in Fig. 7. Though no cross- or test set-validation has been performed, the prediction models are only based on *four* regression coefficients each. With these models it is confirmed, that it is possible to use fluorescence data for on-line or at-line monitoring of sugar quality. This is important as currently these parameters are only determined every eighth hour and with a certain lag as the laboratory analysis takes some time. Furthermore, by scrutinizing the calibration models, important clues to the chemical background for the variations in color and ash can be obtained. For example, component one ($r = 0.60$) and component four ($r = 0.81$) seems to be correlated to color [5]. Only component four, as well as color, seems to be influenced by the increasing amount of frozen beets during the last part of the campaign. Component one ($r = 0.69$) and the tryptophan-like component three ($r = 0.64$) shows some correlation with ash. The tyrosine-like component two does not correlate significantly with neither ash nor color. Thus, the four PARAFAC components show different patterns in connection with the two important quality parameters color and ash.

It is interesting to compare the predictions of ash and color from fluorescence data with predictions obtained using the process variables. Calibration models were constructed for ash and color separately using more thorough validation than above. The following three sets of independent variables were tested: process data of size $265 \times 22$, the fluorescence data given by the four score vectors of the PARAFAC model ($265 \times 4$), or both ($265 \times 26$). [1]

These independent variables were autoscaled and lagged twice (one and three lags were also tried), thereby giving a three-way array of size 265 (samples) $\times$ 22 (variables) $\times$ 3 (lags) in case of using the process data only.

For each data set, a model was estimated using either the first 50 or 150 samples. Recall that the samples are obtained contiguously every eighth hour so that, e.g., 50 samples correspond to approximately 17 days. The model used for calibration was N-PLS regression (unfold-PLS was also tried giving similar predictions though much less parsimonious models).

For, e.g., the model of ash predicted from process data using the first 50 samples, the size of the independent calibration data array is thus 50 (samples) $\times$ 22 (variables) $\times$ 3 (lag mode). The first two samples were excluded as two thirds of the data elements were missing due to lagging. The number of latent variables was determined by minimum cross-validation error and the model was then used for predicting the remaining left-out samples. For the calibration models using the first 50 samples, this means that the models are based on the first half month of the campaign and tested on the last two and a half month.

It is worth mentioning that the models are probably suboptimal with respect to variable selection and lagging. More ingenious variable selection and more in-depth analysis of the variation in time of the variables may lead to better models. However, as stated before, the goal is not to find *the* model, but to explore the relevance in and the patterns of the variations in the data. In essence, this is an example of a first shot analysis. Depending on the relevance of the results, further experiments, data, and analyses may lead to better models.

The results of the predictions are shown in Fig. 8 for the color determinations. The ash determinations were similar. The models are seen to be valid in up to 50 days in some cases. For ash, there is little difference in the quality of the predictions obtained from the process data and the fluorescence data, while for color it seems that the fluorescence data provide more information than the process data. This is very clearly seen from the quality of the predictions in the first thirty to fifty days ahead. It suggests that the fluorescence data give information not already present in the process data, but rather supplements the process data with extra chemical information important for predicting the color of sugar. This is expectable and illustrates that spectroscopic techniques provide a general source of information in process control and quality monitoring.

## 6. Conclusion

The models described in this application are quite extraordinary. They give a direct connection between

[1] It was also tried to smooth the process variables with both median filters and wavelets to remove spikes, but this had no significant influence on the results.

the state of the process, the product (from a scientific point of view the chemistry of the sugar), and the quality (as defined by laboratory measurements defining the internal as well as the external consumer quality). As such, the conceptual idea behind the results reaches far beyond the specific data treated here. It provides means for combining process analytical chemistry and multivariate statistical process control by the use of fluorescence screening and PARAFAC modeling.

In this work, the trust that the fluorescence analysis may work is the primary hypothesis. By data and model selection [5], four components are found of which two have already been chemically identified. Together, they give representative information regarding both sugar quality and process parameters. Thus, four valid indicator substances for quality and process parameters are identified in this preliminary screening. The results have to be checked and followed up in several product seasons if one wants to develop a process control system based on fluorescence.

That it is possible to capture and identify specific chemical variation is somewhat surprising considering the fact that the samples are simply taken directly from the process and dissolved in water. Further, the sample matrix is very complex. Approximately 99.999% of the sugar is sucrose, which is not fluorescent. It is the very small fraction of impurities such as amino acids, phenols and their reaction products that are detected by the fluorometer. Hence, it is the complex mixture of a very small amount of natural and process related components that are being measured with the sensitive fluorescence method.

When more certain conclusions have been drawn, it becomes relevant from an academic and technological point of view to identify with more confidence what the estimated fluorescence spectra represent using, e.g., standard addition or chromatography. In this work, this has only been hinted at, since there is little sense in spending too much efforts in elucidating the chemical background until the relevance and usefulness of the model is established. This is one of the key benefits of using exploratory analysis, already though it has been established that the variations in the fluorescence data are closely correlated to the variations in the quality of the sugar as well as important process parameters. Coupling this with the results from studying the fluorescence data alone, this could indicate that by controlling the temperature of the beet stores more precisely it may be possible to avoid large fluctuations in the sugar quality.

The predictive models obtained from using the process variables in conjunction with the chemical information are important and should be further elaborated on, when new data are available. In the estimated PARAFAC models, there are signs of more components in the data. It is also plausible from a chemical point of view that more fluorophores may be represented. However, it is not possible with the given data to estimate more components reliably. In the near future, samples from the 1996 campaign will be measured using more excitation wavelengths, in the hope, that more components can be extracted.

## Acknowledgements

The author is indebted to Ellen M. Færgestad and Tormod Næs at Matforsk, Norway for inviting me on a study leave during which this manuscript was prepared. Financial support was obtained to Prof. Lars Munck, Food Technology, Department of Dairy and Food Science, Royal Veterinary and Agricultural University, Denmark, through the Nordic Industry Foundation project P93149 and the FØTEK foundation for part of the work reported here. Lars Bo Jørgensen and John Jensen, Danisco Sugar Development Center, Nakskov, Denmark as well as Lars Munck have been extremely helpful in discussing and criticizing the work reported.

## References

[1] B. Winstrøm-Olsen, R.F. Madsen, W.K. Nielsen, Sugar beet phenols: I, Int. Sugar J. 81 (1979) 332.
[2] B. Winstrøm-Olsen, R.F. Madsen, W.K. Nielsen, Sugar beet phenols: II, Int. Sugar J. 81 (1979) 362.
[3] B. Winstrøm-Olsen, Enzymic color formation in sugar beet. Characterization of enzymes using catecholamines: I, Int. Sugar J. 83 (1981) 102.
[4] B. Winstrøm-Olsen, Enzymic color formation in sugar beet. Characterization of enzymes using catecholamines: I, Int. Sugar J. 83 (1981) 137.
[5] L. Munck, L. Nørgaard, S.B. Engelsen, R. Bro, C.A. Andersson, Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strat-

egy of fundamental significance, Chemom. Intell. Lab. Syst. (1998) in press.
[6] R.A. Harshman, Foundations of the PARAFAC procedure: model and conditions for an explanatory multi-mode factor analysis, UCLA Working Papers in Phonetics 16 (1970) 1.
[7] R. Bro, Multi-way analysis in the food industry. Theory, algorithms and applications, Doctoral dissertation, University of Amsterdam (1998).
[8] R. Bro, PARAFAC: tutorial and applications, Chemom. Intell. Lab. Syst. 38 (1997) 149.
[9] P. Paatero, A weighted non-negative least squares algorithm for three-way PARAFAC factor analysis, Chemom. Intell. Lab. Syst. 38 (1997) 223.
[10] R.A. Harshman, Determination and proof of minimum uniqueness conditions for PARAFAC1, UCLA Working Papers in Phonetics 22 (1972) 111.
[11] J.B. Kruskal, More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling, Psychometrika 41 (1976) 281.
[12] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decomposition, with application to arithmetic complexity and statistics, Linear Algebra and its Applications 18 (1977) 95.
[13] J.B. Kruskal, Rank, decomposition, and uniqueness for 3-way and N-way arrays, in: R. Coppi, S. Bolasco (Eds.), Multiway Data Analysis, Elsevier, North-Holland (1989) p. 7.
[14] R.A. Harshman, M.E. Lundy, The PARAFAC model for three-way factor analysis and multidimensional scaling, in: H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York (1984) p. 122.
[15] R.A. Harshman, W.S. de Sarbo, An application of PARAFAC to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques, in: H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald (Eds.), Research Methods for Multimode Data Analysis, Praeger, New York (1984) p. 602.
[16] R. Bro, S. de Jong, A fast non-negativity constrained linear least squares algorithm for use in multi-way algorithms, J. Chemom. 11 (1997) 393.
[17] R. Bro, N. Sidiropoulos, Least squares algorithms under unimodality and non-negativity constraints, J. Chemom. 12 (1998) 223.
[18] L. Ståhle, Aspects of analysis of three-way data, Chemom. Intell. Lab. Syst. 7 (1989) 95.
[19] R. Bro, Multi-way calibration. Multi-linear PLS, J. Chemom. 10 (1996) 47.
[20] A.K. Smilde, Comments on multilinear PLS, J. Chemom. 11 (1997) 367.
[21] S. de Jong, Regression coefficients in multi-linear PLS, J. Chemom. 12 (1998) 77.
[22] G.W. Ewing, Instrumental Methods of Chemical Analysis, McGraw-Hill, New York, NY, 1985.
[23] R.T. Ross, S. Leurgans, Component resolution using multilinear models, Methods in Enzymology 246 (1995) 679.
[24] J.W. Verhoeven, Glossary of terms used in photochemistry, Pure Appl. Chem. 68 (1996) 2223.