JOURNAL OF CHEMOMETRICS
J. Chemometrics 2005; **19**: 236–252
Published online 2 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.924

Mathematical improvements to maximum likelihood parallel factor analysis: experimental studies

Lorenzo Vega-Montoto and Peter D. Wentzell*
Trace Analysis Research Centre, Department of Chemistry, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J3

Received 15 February 2005; Revised 12 July 2005; Accepted 12 July 2005

In this paper, the application of a number of simplified algorithms for maximum likelihood parallel factor analysis (MLPARAFAC) to experimental data is explored. The algorithms, described in a companion paper, allow the incorporation of a variety of correlated error structures into the three-way analysis. In this work, three experimental data sets involving fluorescence excitation-emission spectra of synthetic three-component mixtures of aromatic compounds are used to test these algorithms. Different experimental designs were employed for the acquisition of these data sets, resulting in measurement errors that were correlated in either two or three modes. A number of data-analysis methods were applied to characterize the error structures of these data sets. In all cases, the introduction of statistically meaningful information translated to estimates of better quality than the conventional PARAFAC estimates of concentrations and spectra. The use of the algorithms that employ the error structure suggested by the analysis of the error covariance matrix yielded the best results for each data set. Copyright © 2005 John Wiley & Sons, Ltd.

KEYWORDS: MLPARAFAC; PARAFAC; weighted PARAFAC; maximum likelihood; three-way data; trilinear data; measurement errors; error covariance; parallel factor analysis; data compression; fluorescence excitation-emission spectra

1. INTRODUCTION

In 1980, Hirschfeld [1] presaged the current state of analytical instrumentation when he made a very complete compilation of all feasible combinations of techniques capable of providing second-order data at that time. Nowadays, many of these combinations are commonplace in the analytical laboratory and they have been extended a step further by adding other orders to produce three-way and multi-way data in general. The vast majority of these combinations involve a spectroscopic domain, where measurements are made as a function of wavelength. The spectroscopic order can be combined with a broad selection of techniques exploiting different spectroscopic, chromatographic, kinetic, and physicochemical characteristics of the analyzed samples. Even though the combination of spectroscopic information with chromatographic, kinetic, and physicochemical attributes have a number of drawbacks, such as poor reproducibility of retention times for chromatography, poor sensitivity in the spectroscopic order with respect to changes in physicochemical properties, and important deviations from the bilinear structure in kinetic experiments, these combinations have been extensively used in the chemical literature [2–23].

Three-way data obtained by pairing fluorescence excitation and emission spectra to produce fluorescence excitation-emission matrices (EEMs) is perhaps the most common combination used in chemistry due to the wide availability of spectrofluorometers and a number of useful features. First, the measurements can be made on a single instrument with consistent channel registration. Second, EEMs are characterized by excellent sensitivity, selectivity, and bilinearity. Finally, a wide variety of different options can be used to produce trilinear data [17–23]. However, real EEMs can give rise to nonideal behavior that can disturb the trilinearity of the data. Among the most common cases are nonlinear effects caused by high concentration of the analytes and the presence of instrumental effects such as scattering within the measurements.

A common problem that arises in the analysis of experimental fluorescence data is related to primary absorption due to high concentration of chromophores. As the concentration of the compounds increases, their absorptions become more significant at the edge of the cuvette and it will reduce the amount of light reaching the fluorophores in the rest of the cell. This will decrease the emission intensity in a nonlinear fashion. In order to avoid this situation,

*Correspondence to: Peter D. Wentzell, Department of Chemistry, Trace Analysis Research Centre, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J3.
E-mail: Peter.Wentzell@Dal.Ca

Contract/grant sponsors: Natural Sciences and Engineering Research Council (NSERC) of Canada; Dow Chemical Company (Midland, MI).

fluorescence excitation-emission measurements of dilute samples are usually preferred, or in cases where this is not possible, some corrections can be applied [24,25]. A second problem is the inadequacy of the mathematical model to represent scattering effects in the samples (i.e., Rayleigh and Raman scatter). Unfortunately, corrections for scattering effects cannot be implemented as easily as the previous case from an experimental point of view and, in general, corrections have to be made in the estimation step. Further scrutiny of this problem has been done and thus far the only real applications use some kind of weighted decomposition [9,24,26] to eliminate this problem by considering the scattering as noise rather than model deviations. In this work, special attention has been given to the selection of a range of concentration profiles and excitation and emission wavelengths to produce data sets that are not affected by these deviations of the model.

Deviations apart, the physical model describing this type of measurement is equivalent to the well-known structural model called PARAFAC [27,28]. Many different algorithms [28–35] based on different optimization strategies have been formulated to estimate the parameters describing the model. However, the PARAFAC algorithm, based on an alternating least squares optimization technique, accounts for the majority of the applications reported in the chemical literature due to its excellent convergence characteristics and simplicity. In addition, the statistical optimality of the least-squares solutions obtained by PARAFAC was empirically proven by Faber *et al.* [36] using a comparative study with the other methods, and from a more theoretical point of view by Liu and Sidiropoulos [37] by comparing the performance of the PARAFAC solutions obtained for simulated data affected by *iid* noise with the Cramer–Rao lower bound. A few examples cover areas as dissimilar as the estimation of sugar quality and process parameters in the food industry and the determination of polycyclic aromatic compounds, pesticides, and dioxins in different matrices [38–43].

In general, even though the characteristics of the noise affecting fluorescence EEMs are well documented [44], they are disregarded in favor of the more simplistic and therefore unrealistic features characterized by an identical distribution of independent errors from channel to channel, since this provides optimal estimates when algorithms based on simple least squares optimization are used. Recently, two methods, called maximum likelihood via iterative least squares estimation (MILES) and maximum likelihood parallel factor analysis (MLPARAFAC), have been introduced to the chemometrics literature [26,45] to optimally estimate the model using measurement error information. The main difference between MILES and MLPARAFAC is that MLPARAFAC is a method based solely on ALS optimization, while MILES works as an iterative preprocessing tool to condition the data from a maximum likelihood perspective in order that least squares methods such as PCA and PARAFAC can optimally handle the estimation process.

In an earlier companion paper [46], a number of important simplifications of the general MLPARAFAC [45] methodology for cases where the error covariance matrix is dominant along one or two orders, and a compression step prior to the use of general MLPARAFAC for cases where the error was corrupting more than two orders, were introduced. These simplifications complete the theoretical background of the general methodology presented in the original work [45] by introducing a new approach to obtain the estimation equations.

Traditionally, the estimation equations for the standard PARAFAC model and for its derived errors-in-variables model, general MLPARAFAC, were obtained by switching among different mathematical arrangements of the same objective function, expressed differently for each mode. This strategy is used because, due to the symmetry of the PARAFAC model, the implementation is not only efficient but also extremely simple, making the normal equations very similar from one mode to the other. However, when the characteristics of the noise are taken into account, this symmetry is lost, making it necessary to express the estimation problem as the general problem, since the existence of a simplified version of the error covariance matrix in the given space is not possible or extremely difficult to find. Therefore, a new approach was introduced in which the data are initially arranged in order to have the major source of correlated noise along the mode B, followed by the second major source of correlation along mode C, leaving mode A as the mode not affected by correlated noise. After the data are arranged, the estimation equations are obtained by expressing all of the substeps as minimization problems of the same objective function written by preserving mode A alone.

The simplifications obtained by using this approach were tested using simulations. These simulations showed the statistical characteristics of these new algorithms and the improvements in terms of performance and quality of the estimates when the proper simplifications given the available data were used. However, they also illustrated the importance of a thorough characterization of the error covariance matrix in order to use the most suitable algorithm. Unfortunately, the simulations had a very well-defined error structure, making the process of choosing the appropriate simplification extremely simple, since information about the number of orders affected by the correlated noise and its structure were accurately known in advance. Real-life applications are not characterized by this simplicity, making the decision process a more complex task. Therefore, the objective of the present paper is threefold. First, a set of guidelines are introduced to thoroughly characterize the error structure and rationalize the way in which the different orders are arranged and the simplifications used. Second, the different simplifications are applied to experimental EEM data sets to test whether the improvement observed in simulations can translate to experimental data. Finally, the effect of using the different simplifications is explored with variations in the way the orders are arranged.

2. THEORY

The companion to this paper showed the relationship between the optimal representation of the error covariance matrix (the one including all the information about the variance and the covariance among the elements) for different scenarios and the different simplifications used in each case, reducing to a considerable degree the computational

burden for MLPARAFAC. Unfortunately, for all cases, it was assumed that the error covariance matrix describing the given system was completely known in its structure as well as its numerical value. In reality, the situation is more complex. For a given application, it is necessary to initially characterize the structure of the error covariance matrix to choose the proper representation and, once this is established, its numerical estimation has to be performed. Until recently, the literature on characterizing error covariance matrices was virtually nonexistent but a recent paper by Leger *et al.* [47] has shed some light on this topic. A number of two-way data sets were analyzed in this work using those tools developed by the authors, and these tools can be extended to three-way data in a straightforward manner, as suggested by the authors and to be demonstrated here. A principal objective of this work is to develop a set of tools for understanding and classifying the measurement error structure of a given multi-way system through an analysis of the error covariance matrix. This knowledge will then be used in conjunction with the different simplifications of general MLPARAFAC introduced in the companion paper. There are two immediate benefits to such an analysis. First, the analysis can provide insight into the main sources of error affecting the measurement. This can potentially be used to redesign experiments to minimize these error sources, since the error structure is directly related to the experimental design as well as the detection technique used to collect the data. Most importantly, it can help the practitioner choose the proper estimation method to accommodate the error structure in an optimal way. Leger *et al.* [47] also speculated on the idea of using this information to produce a deterministic model of the error covariance matrix in order to eliminate the need for extensive replication in order to estimate the error covariance matrix. However, in this work, this possibility will not be explored.

In order to put into context the motivation behind these tools, a brief description of the structure of error covariance matrices will be given. The tools will then be described, devoting some attention to the pieces of information provided by them. Finally, a flow chart will be presented to choose the optimal representation of the error covariance matrix and, in turn, the algorithm needed to estimate the PARAFAC model.

2.1. Analysis of the error covariance matrix

A few important pieces of information are needed to construct an optimal representation of the error covariance matrix. The first one is the answer to the following question: How many orders are affected by correlated noise? Second: Which are the orders affected by correlated noise? Once these two questions are answered and the data are reorganized by using permutations in a way that the order affected by correlation is located in mode B if the errors are only affecting one order, or in modes B and C if the errors are only affecting two orders. At this point, another important issue must be addressed by answering the following question: is the correlation structure the same for all the objects used in the construction of the error covariance matrix? (In other words, is pooling of the individual error covariance matrices statistically correct?)

Figure 1 shows a schematic representation of the structure of the full error covariance matrix and its equivalent simplified representations for each case in order to understand the characterization of the error covariance matrix and the tools used to do it.

It is clear from Figure 1 that the errors can be correlated along one, two, or three orders, giving rise to different representations of the full error covariance matrix. For the cases where the errors are correlated along only one or two orders, more simplified representations exist. Unfortunately, the analysis of the full-error covariance matrix is usually precluded by its size. Therefore, this case has to rely on alternative representations providing similar information. A substantial amount of information about the measurement error structure can often be gleaned through a visual examination of the pooled experimental error covariance matrix for each mode. As already noted elsewhere [47–49], this matrix is typically obtained through the use of replicate measurements. Normally, a series of R replicates of each object order is obtained. The definition of a replicate can vary for different fields, applications and experiments, but in the present context it is defined as the measurement realization made to capture the relevant sources of variation while the underlying chemical information defining the unattainable true signal is kept constant.

Operationally, the process to construct the pooled error covariance starts by unfolding the replicate r of the three-way data ${}^{r}\underline{\mathbf{X}}$ retaining the order to be analyzed. This operation is repeated for the R replicates. For example, to calculate the error covariance matrix for mode A, ${}^{r}\underline{X}$ $(I \times J \times K)$ is unfolded while retaining mode A, producing ${}^{r}\mathbf{X}_{\mathrm{a}}$ $(I \times JK)$. Then, ${}^{r}\mathbf{X}_{\mathrm{a}}$ is transposed and used to calculate the individual experimental error covariance matrix for each object included in mode B and C via Equation 1.

$$\mathbf{\Sigma}_o = \frac{1}{(R-1)} \sum_{r=1}^{R} ({}^{r}\boldsymbol{x}_o - \bar{\boldsymbol{x}}_o)^{\mathrm{T}} ({}^{r}\boldsymbol{x}_o - \bar{\boldsymbol{x}}_o) \tag{1}$$

where ${}^{r}\boldsymbol{x}_o$ is the oth $1 \times I$ row vector of replicate r taken from ${}^{r}\mathbf{X}_{\mathrm{a}}^{\mathrm{T}}$ and $\bar{\boldsymbol{x}}_o$ is the $1 \times I$ mean vector of the replicate measurements. The subscript 'o' is used in a generic way to represent objects from mode B and C. The degrees of freedom used in this equation are analogous to the calculation of variance (which will be represented by the diagonal elements of $\mathbf{\Sigma}$) and, as with the calculation of variance, the estimated error covariance matrix will have a high degree of uncertainty unless a large number of replicates are used. In many cases, as we will see shortly, the error covariance matrices estimated for several objects can be combined to give a pooled error covariance matrix, $\mathbf{\Sigma}_{\mathrm{avg}}$. For this example $\mathbf{\Sigma}_{\mathrm{avg}}$ can be calculated as follows:

$$\mathbf{\Sigma}_{\mathrm{avg}} = \frac{1}{JK} \sum_{o=1}^{JK} \mathbf{\Sigma}_o \tag{2}$$

Of course, such a pooling is statistically valid only if it can be assumed that the row error covariance structure is the same for all the objects in the other modes. This situation will be rigorously analyzed in the next step of the characterization process, but here only a subjective analysis will be carried

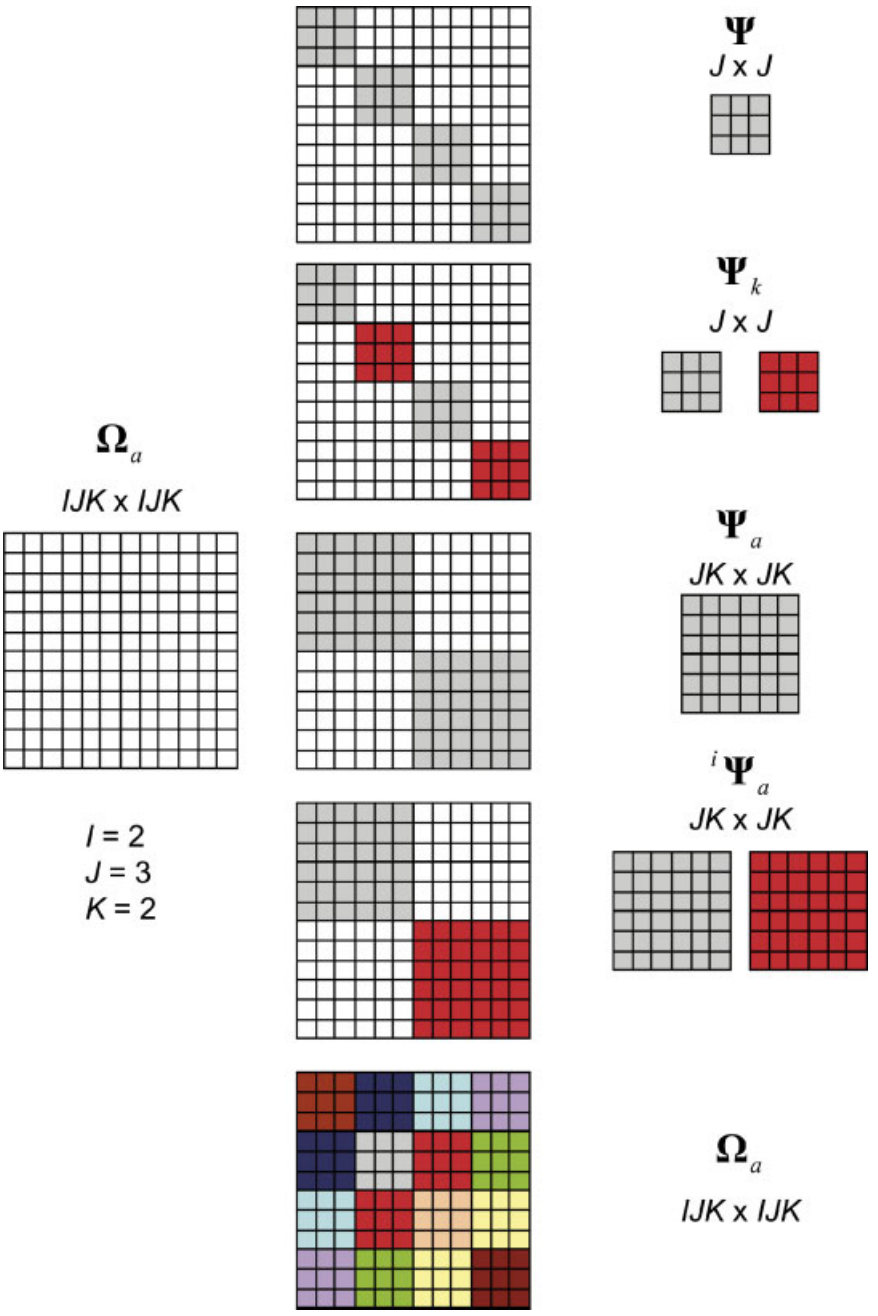

Figure 1. Illustration of the possible scenarios in which a full error covariance matrix can be expressed using different simplified representations of the error structure to describe all of the sources of variation.

out to determine the extent of the error correlation effect. Mathematically, Equations 1 and 2 can be combined to give a clearer view of this calculation. This is done by considering the $I \times JKR$ matrix of residuals for all replicates of all objects, $\mathbf{E}_a$. The equation can then be written as:

$$\mathbf{\Sigma}_{\text{avg}} = \frac{1}{JK(R-1)} \mathbf{E}_a \mathbf{E}_a^T \tag{3}$$

It is important to emphasize that, despite the use of mode A for the example, this process is exactly the same for the rest of the modes, but with the given equation adapted accordingly.

Despite the central role of error covariance matrices in maximum likelihood estimation, their visual interpretation may be of limited utility since, in the presence of heteroscedastic errors, a few elements with a high variance can obscure the interactions among other elements. A more complete understanding of the interactions of the elements in the error structure can be gained through inspection of error correlation matrices. Error correlation matrices can be calculated by dividing each element of the covariance matrix by the two contributing standard deviations

$$\rho_{rs} = \frac{\sigma_{rs}}{\sigma_r \sigma_s} \tag{4}$$

In this equation, ρ_{rs} and σ_{rs} represent the elements in the rth row and sth column of the correlation and covariance matrices, respectively, and σ_r and σ_s are the standard deviations at elements r and s, calculated from the square root of the

corresponding elements of the diagonal of the covariance matrix. In matrix notation, this can be given as:

$$\mathbf{S}_{\text{corr}} = \boldsymbol{\Sigma}. \Big/ \sqrt{diag(\boldsymbol{\Sigma}) \cdot diag(\boldsymbol{\Sigma})^T} \tag{5}$$

where the notation './' indicates an element-wise division (Hadamard quotient), the function '*diag*' converts the diagonal of $\boldsymbol{\Sigma}$ into a column vector, and the square root is taken to be an element-wise operation. By definition, the diagonal elements of the correlation matrix will be unity. The off-diagonal elements will indicate the degree of error correlation among elements, although information about the absolute magnitude of the covariance is lost.

Once these correlations matrices are constructed for each mode, a conclusion regarding what orders are affected by correlated errors can be drawn. Based on this conclusion, the three-way arrays can be permuted in order to have either the uncorrelated orders in mode A and C for cases where correlation is only affecting one order, or the uncorrelated order in mode A for cases where correlation is affecting two orders.

It is worth noting that the construction of error covariance matrices for cases where correlated noise is affecting two orders is extended in a straightforward manner as shown in the following equation where the correlated orders are B and C

$$\boldsymbol{\Sigma}_{\text{a}}^{i} = \frac{1}{(R-1)} \sum_{r=1}^{R} ({}^{r}\boldsymbol{x}_{\text{a}}^{i} - \bar{\boldsymbol{x}}_{\text{a}})^{T} ({}^{r}\boldsymbol{x}_{\text{a}}^{i} - \bar{\boldsymbol{x}}_{\text{a}}) \tag{6}$$

where ${}^{r}\boldsymbol{x}_{\text{a}}^{i}$ is the $1 \times JK$ row vector of replicate r and $\bar{\boldsymbol{x}}$ is the $1 \times JK$ mean vector of the replicate measurements. The pooled error covariance matrix is calculated as shown

$$\boldsymbol{\Sigma}_{\text{a}}^{\text{avg}} = \frac{1}{I} \sum_{i=1}^{I} \boldsymbol{\Sigma}_{\text{a}}^{i} \tag{7}$$

2.2. Homogeneity among different error covariance matrices

The visual analysis of the average error covariance and correlation matrices treats the error structure as a pooled entity. The pooling of individual error covariance matrices is permitted by an *a priori* assumption that the sources giving rise to this error structure are constant from object to object, and that each object's own contribution to the error structure is fairly constant. Even though a few statistical tests, such as Wilks' Λ and Box's M tests [50], have been designed to test the similarity and homogeneity of covariance matrices, the approximations used for these tests are only valid when the number of replicates is larger than 20 and the number of objects/variables is less than 5. Usually, for multi-way data, these assumptions are violated. Therefore, since the assessment of structure and homogeneity of error covariance matrices is an important subject, a decomposition tool will be introduced here taking into account the special requirements of this type of data.

To understand the theoretical idea behind the decomposition tool used in this work, we will initially assume that the measured error covariance matrix can be factorized according to a low-rank bilinear model. This assumption is obviously limiting in the context of a general model for error covariance. The authors recognize the limited scope of this assumption. For instance, the simplest error structure, *iid*-normal errors, cannot be represented by this low-rank bilinear model, and neither can certain sources of covariance arising from cosmetic manipulations, such as digital filtering. Nevertheless, reference 47 demonstrated the validity of these simplified assumptions using a number of examples.

The theoretical foundation supporting this tool will be illustrated using fluorescence emission spectroscopy, which is the simplest case of EEMs, since a set emission measurements is recorded at a fixed excitation wavelength. Two sources of error that have been identified in fluorescence emission spectroscopy are offset noise and multiplicative offset noise [47]. In the first case, which can arise, for example, from variable cell positioning, the entire spectrum is offset by a fixed amount. In the second case, the offset depends on the magnitude of the square root of the signal in a multiplicative way. This square root dependence might be expected due to the shot noise characteristics of emission measurements, which follow Poisson statistics. Therefore, a structural component similar to the square-root of the mean emission spectra can be anticipated. If we consider a series of spectra, $\mathbf{X}$ (R replicates by J wavelength channels), the errors of these types in the spectra, $\mathbf{E}$ ($= \mathbf{X} - \mathbf{X}^{o}$), could be represented as:

$$\begin{aligned}\mathbf{E} &= \begin{bmatrix} \mathbf{x}_{1\bullet} - \mathbf{x}^{o} \\ \mathbf{x}_{2\bullet} - \mathbf{x}^{o} \\ \vdots \\ \mathbf{x}_{R\bullet} - \mathbf{x}^{o} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_{1\bullet} \\ \mathbf{e}_{2\bullet} \\ \vdots \\ \mathbf{e}_{R\bullet} \end{bmatrix} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1R} \end{bmatrix} [1, 1, \ldots, 1] \\ &+ \begin{bmatrix} e_{21} \\ e_{22} \\ \vdots \\ e_{2R} \end{bmatrix} \left[\sqrt{x_1^{o}}, \sqrt{x_2^{o}}, \ldots, \sqrt{x_J^{o}} \right] = \mathbf{e}_1 \cdot \mathbf{1}_J^{\mathrm{T}} + \mathbf{e}_2 \cdot \sqrt{\mathbf{x}^{o}}\end{aligned} \tag{8}$$

In this equation, $\mathbf{x}_{i\bullet}$ is a row vector (replicate spectrum) from $\mathbf{X}$, $\mathbf{e}_{i\bullet}$ is a row vector (residuals) from $\mathbf{E}$, $\mathbf{1}_J$ is a $J \times 1$ vector of ones, and $\mathbf{x}^{o}$ is a row vector representing the error-free spectrum. The $R \times 1$ vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ contain the individual realizations of the offset error and the multiplicative offset error for each replicate, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are assumed to be normal random variables with standard deviations of σ_1 and σ_2. Taking the expectation for the error covariance matrix, we can write

$$\mathbf{S} = E(\mathbf{e}^T \cdot \mathbf{e}) = \mathbf{S}_1 + \mathbf{S}_2 = \sigma_1^2 \cdot \mathbf{1}_J \cdot \mathbf{1}_J^{\mathrm{T}} + \sigma_2^2 \left(\sqrt{\mathbf{x}^{o}} \right)^T \cdot \sqrt{\mathbf{x}^{o}} \tag{9}$$

It is important to mention that the structural model shown in Equation 1 will describe most, but not all, of the variation for this type of data, since the contributions of other sources, most notably independent errors (either homoscedastic or heteroscedastic), are not included. This will have an impact when the methodology is employed to obtain a deterministic model for the error covariance matrix, but since our main objective is the characterization of the homogeneity error covariance matrix these contributions will be neglected here.

Equation 9 represents the physical model behind the error structure for a particular object. When different objects are

considered, this physical model can be mimicked by the INDSCAL structural model, introduced by Carroll and Chang [27]. Mathematically, this can be done by collating individual error covariance matrices into a three-way array consisting of symmetric slices $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_O$. The model decomposes the slices as:

$$\mathbf{\Sigma}_o = \mathbf{F}\mathbf{D}_o\mathbf{F}^{\mathrm{T}} + \mathbf{E}_o \tag{10}$$

where $\mathbf{F}$ is a $J \times P$ matrix representing the sources of variation (i.e., structural factors) and $\mathbf{D}_o$ is the $P \times P$ diagonal matrix whose elements represent the contribution of each source of variation to the error covariance of object o.

Often, as noted previously, error covariance matrices from different objects are pooled to give a better estimate of the error covariance matrix. In these cases, it is expected that the decomposition of the individual error covariance matrices (different objects) can be factorized using common structural factors with contribution vectors that share the same statistical properties of the specific model. Therefore, the homogeneity of the individual error covariance matrices can be reduced to the homogeneity of the structural factors describing the error sources and the similarity in the statistical properties of the contribution of each individual object. For instance, in the example presented, this would mean that the spectra for individual samples show a strong similarity (structural factors) and $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ (contribution vectors) share the same statistical characteristics for all samples (i.e., same σ_1 and σ_2). As explained in Reference [47], this model is solved using the PARAFAC algorithm [28], which is simpler and less constrained, but mathematically equivalent in terms of the solution produced by Equation 10. It is recommended that the PARAFAC algorithm be run in a split-half [51] fashion to make sure of the validity of the estimates.

2.3. Assessment of the error structure

Figure 2 depicts a flow chart indicating the important steps and metrics to direct the user in the optimal construction, characterization, and calculation of the error covariance matrix. This will lead to the use of the optimal estimation method given the available data.

The first step uses the information obtained through a subjective analysis of the pooled error correlation matrices for each mode to make a decision about the number of modes affected by correlated errors and to sort the modes in a way that the permuted array will have the uncorrelated orders in modes A and C for cases where correlation is only affecting one order, or the uncorrelated order in mode A for cases where correlation is affecting two orders. This step will also provide the necessary information to decide whether a $J \times J$, $JK \times JK$, or $IJK \times IJK$ error covariance matrix will be needed. Matrices with a majority of their elements showing significant correlation will be considered to describe important correlation in this mode. As mentioned before, this interpretation will be largely subjective as the different error covariance matrices are visually analyzed. However, some numerical interpretation can be added by considering that the decomposition of pooled error covariance matrices describing important sources of correlation will produce a low-rank model with few components accounting for a large proportion of the variance. It is important to mention that this interpretation must be treated carefully, since on many occasions the structure in other modes will produce some artificial structure in the analyzed mode, as was described by Leger *et al.* [47].

Once the form of the error covariance matrix is decided, an analysis of the homogeneity is necessary, regardless the form. For cases where correlation is important along only one dimension, it is important to assess whether the objects in the other two orders will contribute to the structure equally. This is also true for cases where the correlation is affecting two orders, with the only difference being that the equivalence of the contribution is only tested for objects within the one remaining order. The general procedure starts by calculating the individual error covariance matrices of order determined in step 1 of the flowchart. Different split-half data sets are created to assess the contribution of different objects to the structural factors when the INDSCAL model is estimated. In the present context, the split-half method [51] is a type of cross-validation method in which the homogeneity of the structural factors in one or more modes is examined by partitioning the data in half along a remaining mode and analyzing each half individually. The partitioning is typically done in such a way to examine variations in the structural factors that depend systematically on the other mode. For example, it is advisable to use partition strategies that provide information about short range (e.g. by taking alternate objects) and long range (e.g. by taking consecutive blocks of objects) differences in the contribution of the objects to the error covariance matrix.

The number of factors describing the structural model of the error covariance model will be chosen by using information such as variance accounted for the models, corcondia values, and visual appearance of the factor [47]. Once this number is established, the structural factors obtained by different split-half models are aligned to eliminate the permutation indeterminacy, and then the average structural factors are calculated. These average structural factors are used as reference values to calculate the similarity of the corresponding structural factors obtained from different split-half models via the average vector angle. The decomposition of the INDSCAL model also provides information about object contributions. Low-average vector angles and statistically homogeneous sample contribution values will indicate that that pooling of the error covariance matrices for different objects is correct from a statistical point of view.

It is important to note that, for cases where the correlation affects only one dimension, an additional homogeneity test separating objects from different modes has to be carried out if the first test fails to indicate global homogeneity. In the second test, the homogeneity in the other two modes is examined individually. If the second test also fails, the data must be treated with an algorithm that is also used to treat cases where correlation is affecting two modes, as shown in the flowchart. These approaches will be illustrated with real samples in Section 4.

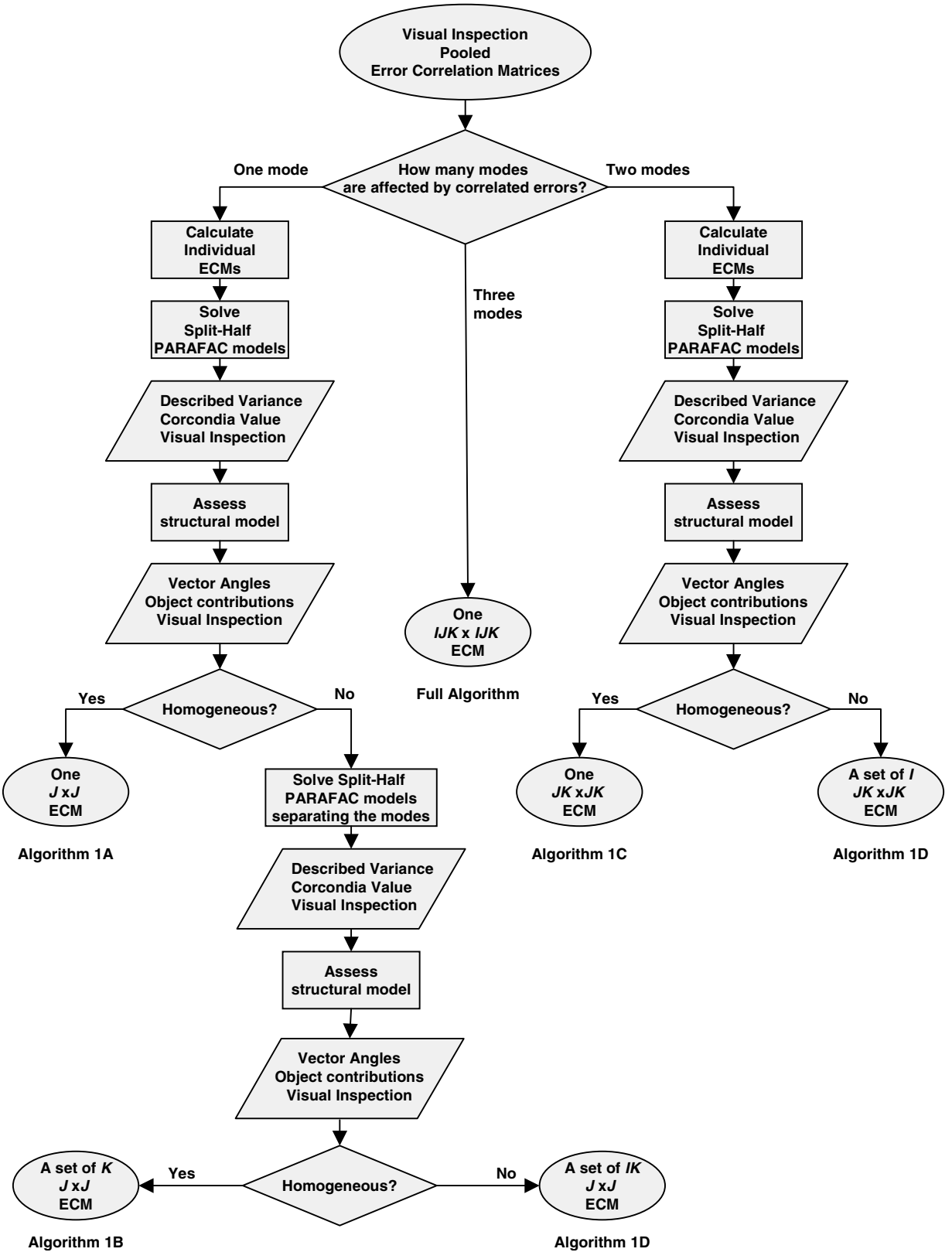

Figure 2. Flow chart employed to characterize the error structure.

3. EXPERIMENTAL

3.1. Methods

3.1.1. Reagents and samples

Naphthalene (Fisher) was used as received. Acenaphthylene (Aldrich) and phenanthrene (BDH) were recrystallized prior to use. Stock solutions of the individual samples were prepared by mass in acetonitrile (Anachemia, spectrophotometric grade, 99.9%). The final concentration ranges were approximately 0.10–0.34 μg/g (ace), 0.018–0.063 μg/g (nap), and 0.0072–0.027 μg/g (phe).

3.1.2. Instrumentation

Fluorescence emission spectra were obtained from samples in a 1 cm quartz cuvette on a Shimadzu RF-301PC spectrofluorometer with a xenon lamp excitation source. The excitation wavelength range was between 250 and 305 nm using intervals of 5 nm. The emission wavelength was scanned between 309 and 415 nm in steps of 1 nm. A medium scan speed was used and the slits for both excitation and emission were set at 5 nm. The pure excitation and emission spectra for each component are the average of 10 replicates using the same experimental conditions. These are shown in Figure 3.

3.1.3. Procedure

Fluorescence emission spectra were obtained from mixtures of three polycyclic aromatic hydrocarbons (PAHs): acenaphthylene (ace), naphthalene (nap), and phenanthrene (phe). Five replicate sets of spectra were obtained from each of 27 mixtures. A three-level, three-factor factorial design was used to prepare the mixtures and a blank containing only the solvent (acetonitrile) was run before and after each block.

It is well known that the error structure affecting spectroscopic data depends on both the spectroscopic technique and the experimental design used to record the data. Since the main objective of this work is testing the performance of different simplifications of the general MLPARAFAC algorithm in the presence of different error structures, the procedure described was used to produce three different data sets through changes in the data acquisition protocols.

Data Set 1 was obtained by scanning all of the samples in each replicate block in a randomized order. Also, in order to decrease the possibility of correlated errors, the excitation wavelengths were also randomized for each replicate block. Emission spectra were obtained in a consecutive fashion.

Data Set 2 was also obtained by scanning all of the samples in each replicate block in a randomized order. In this case, the excitation and emission were scanned in a consecutive fashion to see if some additional correlation is introduced by the non-randomized use of the excitation range. The excitation range was scanned from the highest to the lowest excitation wavelength to decrease the potential effects of photodecomposition.

Data Set 3 represents the most complex error structure since the objects in all modes were scanned in a consecutive fashion (i.e., samples were run in a sequential order and excitation and emission wavelengths were scanned consecutively). This experimental design is generally avoided by practitioners, since it can introduce temporal correlation from different sources [44]. Again, the excitation range was scanned from the highest to the lowest excitation wavelength. These different designs are represented pictorially in Figure 4.

3.2. Computational aspects

All the calculations performed in this work were carried out on a Sun Ultra 60 workstation with 2×300 MHz processors and 512 MB of RAM and a 3.2 GHz Pentium-IV PC with 1 GB of RAM. All programs were written in-house using Matlab 6.0 (The MathWorks, Inc., Natick, MA) with the exception of the PARAFAC and TUCKER3 functions that were run using the N-Way Toolbox [52].

4. RESULTS AND DISCUSSION

4.1. Analysis of the error covariance matrices

Figures 5–7 show the pooled correlation matrices of each mode for Data Sets 1, 2, and 3, respectively. They are plotted

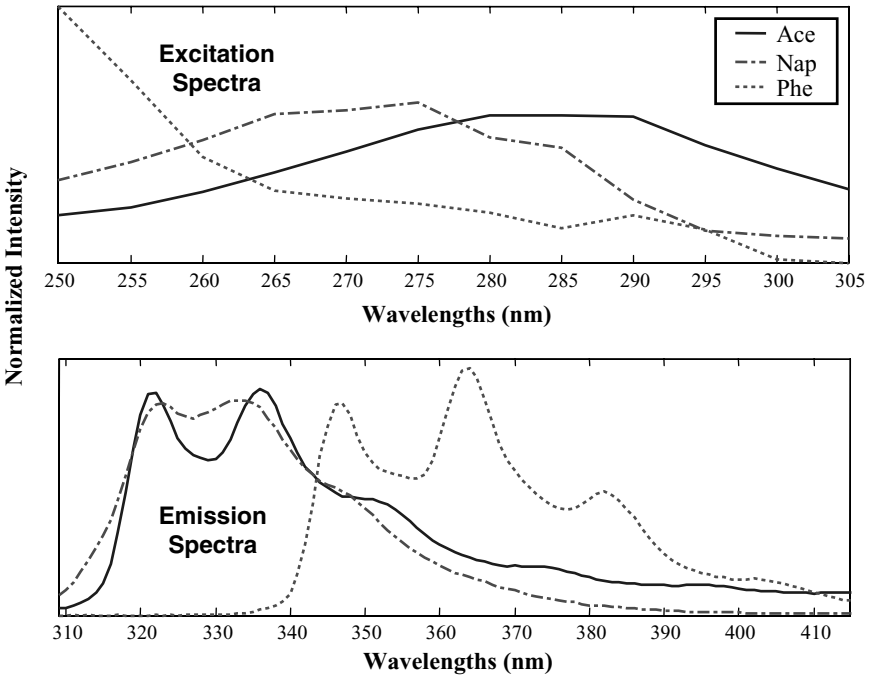

Figure 3. Pure excitation (top panel) and emission (bottom panel) normalized spectra of the compounds employed in this work. Each spectrum is the average of ten replicate measurements.

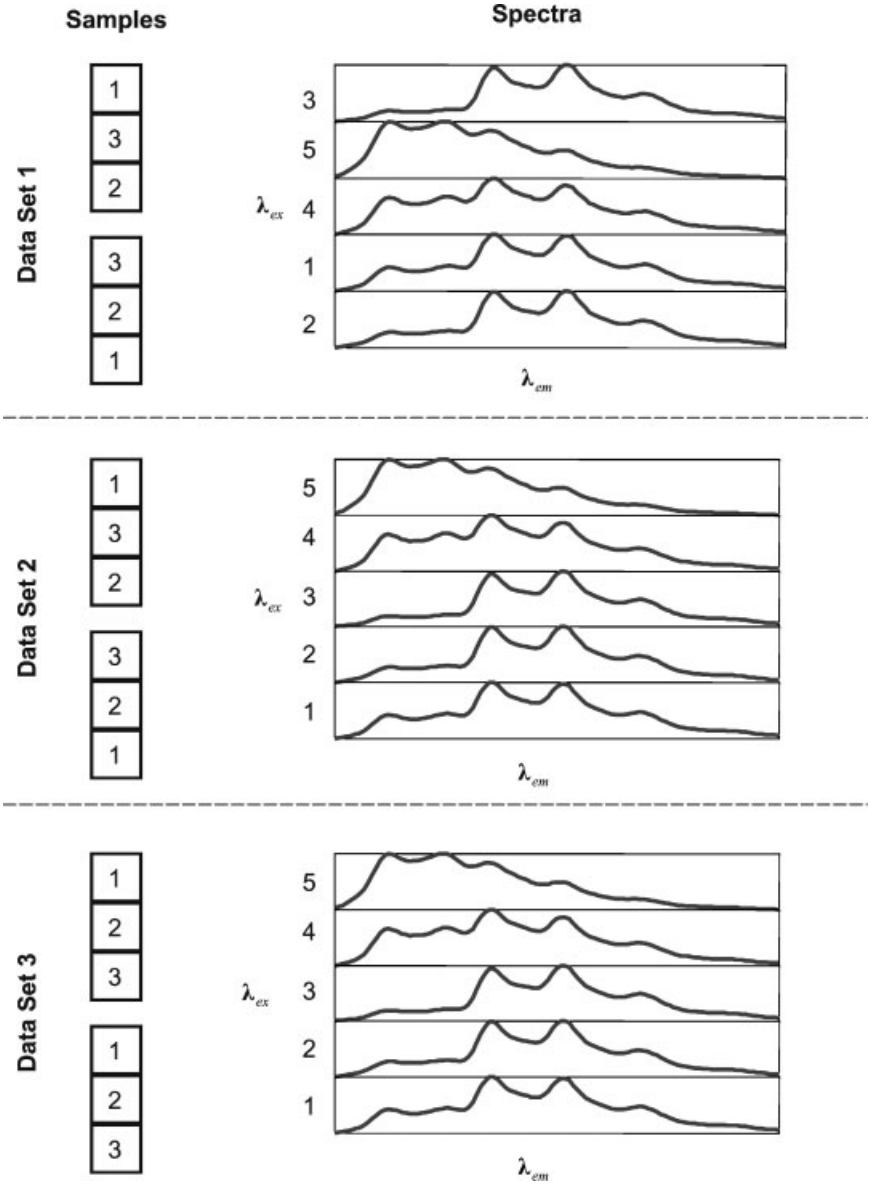

Figure 4. Simplified pictorial representation of the experimental designs employed to acquire Data Sets 1, 2 and 3.

using an intensity map in which a darker tonality indicates an absolute correlation value closer to one and a paler tonality indicates a correlation value closer to zero. The three cases present a very strong pattern of correlation for the emission modes, as was expected due to the consecutive fashion in which this mode was recorded in every case. It is important to note that the correlation patterns were very similar for Data Sets 1 and 2 but some differences were observed for Data Set 3. The physical reason for this difference is not entirely clear, but is undoubtedly linked to the sequential order of the samples in the third data set and indicates the close relationship between experimental design and error structure. The excitation mode was also highly affected by correlation in all cases, even though Data Set 1 was scanned in a random manner in the excitation mode. This result is not completely surprising since, for a given sample, the emission spectra at each excitation wavelength were recorded without removing the sample from the spectrometer. Therefore, the cuvette positioning will produce an offset, which is one of the most common sources of correlated errors. This will carry through all the excitation wavelengths, and is likely an important source of correlation affecting this mode. Another expected result was related to the correlation affecting the sample orders. Data Sets 1 and 2 showed a very random distribution of tonalities, indicating the lack of important sources of correlation affecting these data sets. However, Data Set 3 was characterized by a very dark correlation map, indicating important sources of correlation that need to be taken into account in the sample mode.

Conclusions about the necessary permutations and the optimal representation of the error covariance matrices for each data set can be drawn based on these plots. For Data Sets 1 and 2, the correlation pattern suggests that the emission and excitation orders should be located in modes B and C and the use of a $JK \times JK$ format for the error covariance matrix. Order permutations are not necessary for Data Set 3,

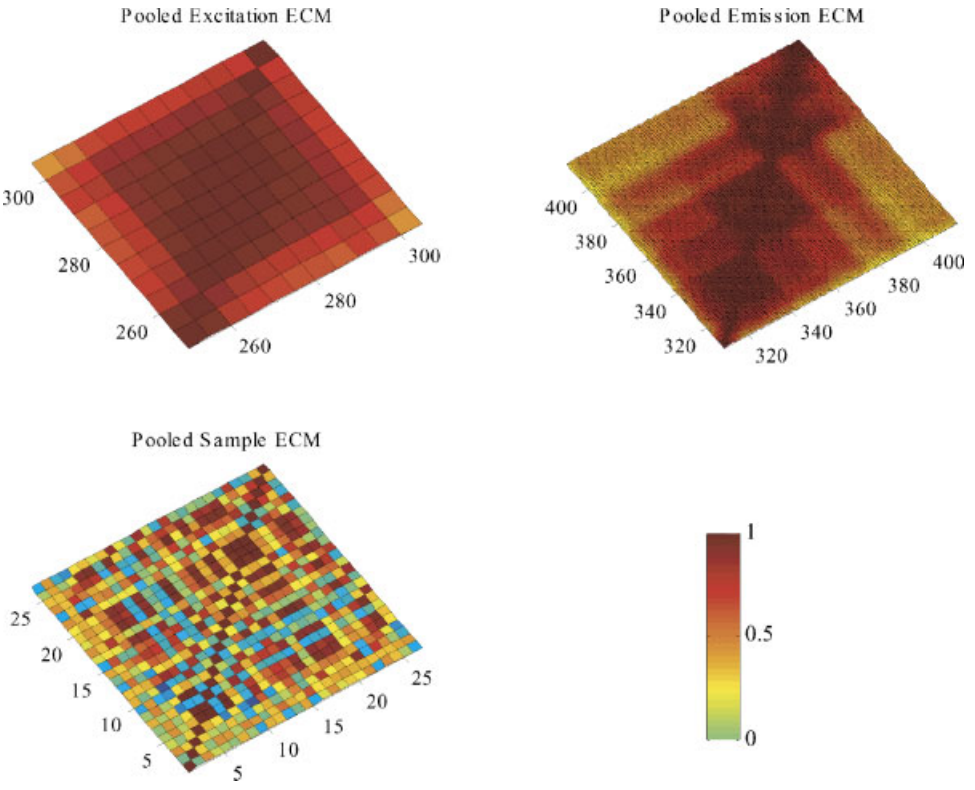

Figure 5. Pooled correlation matrices for each mode of Data Set 1 using intensity maps.

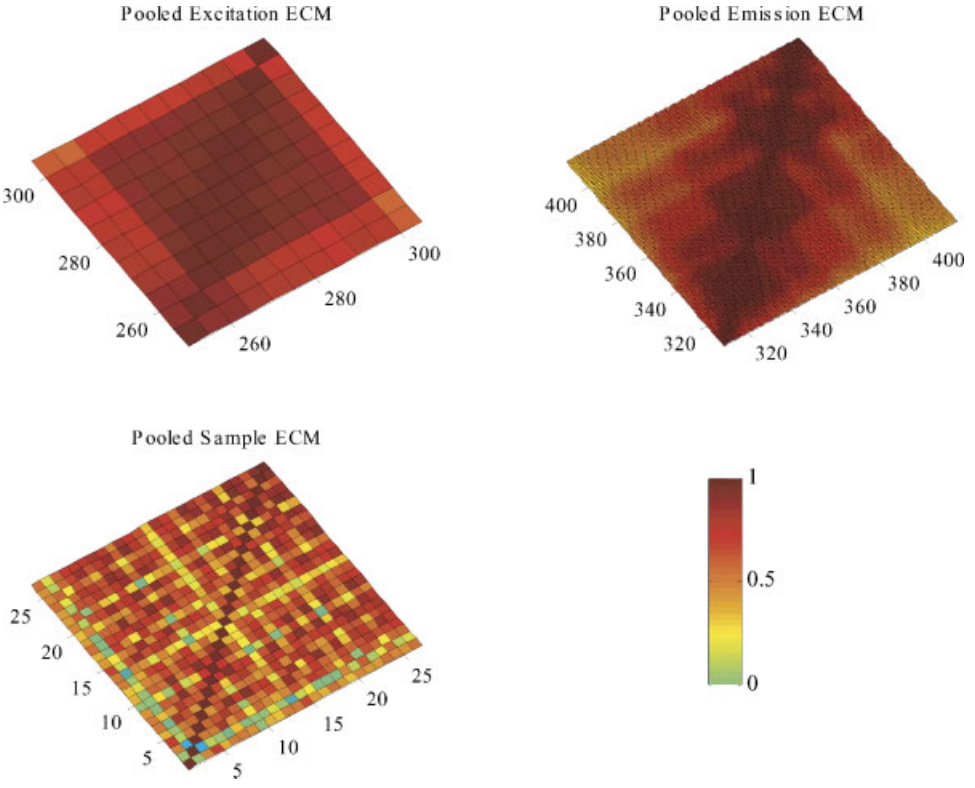

Figure 6. Pooled correlation matrices for each mode of Data Set 2 using intensity maps.

since the general MLPARAFAC algorithm will be needed to provide optimal estimates, requiring a full $IJK \times IJK$ error covariance matrix and the use of compression in order to use the algorithm.

Although previous results indicate that the use of $J \times J$ error covariance matrices was unjustified since correlation is affecting more than one order, a structural decomposition of the individual error covariance matrices for each mode was done. In all cases, it was clear that the different objects pooled produce different sources of structure (results not shown) indicating again that the use of a pooled $J \times J$ error covariance matrix would be sub-optimal.

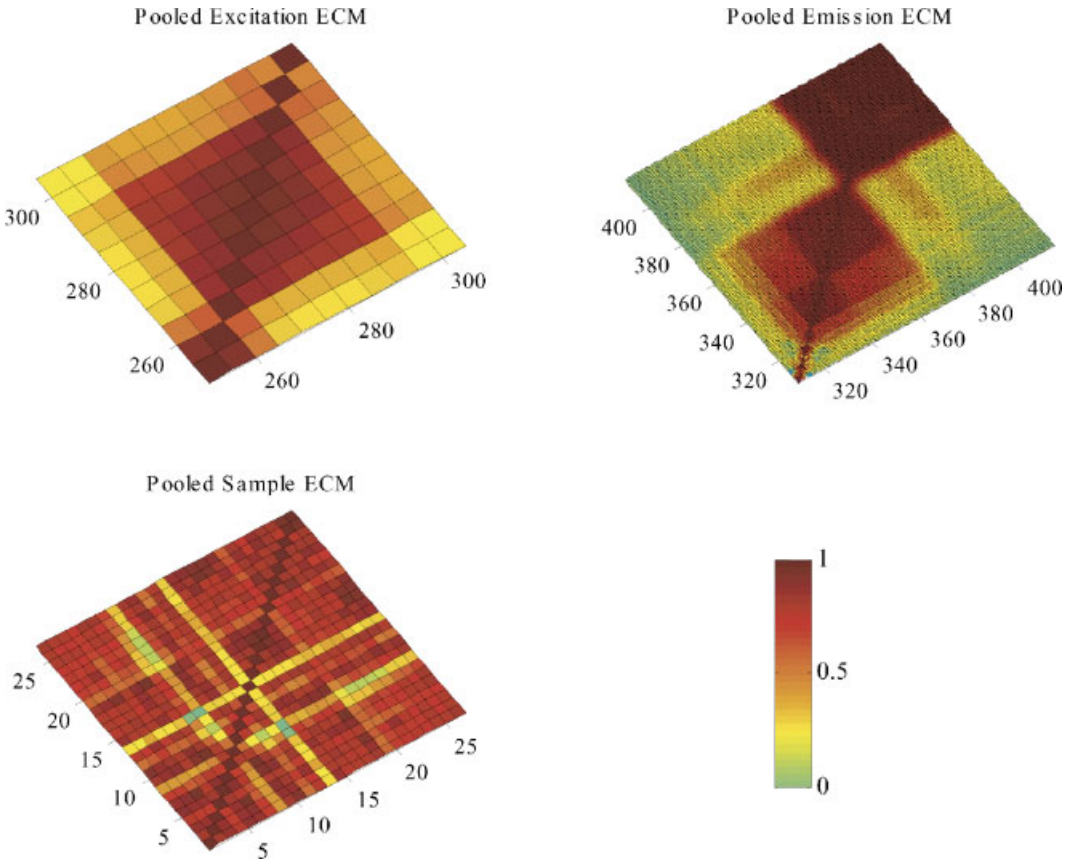

Figure 7. Pooled correlation matrices for each mode of Data Set 3 using intensity maps.

The flowchart in Figure 2 indicates that the next step in the characterization process is the assessment of the homogeneity of the individual error covariance matrices to determine whether or not pooling is theoretically justified. This step was carried out for Data Sets 1 and 2, but was not necessary for Data Set 3 since the full error covariance matrix was required in this case.

Figure 8 shows the average structural factors and sample contributions obtained for Data Sets 1 and 2 when a PARAFAC model is used. In both cases, the decomposition was carried out on the error covariance matrices characterizing the composite mode formed by the excitation and emission modes. Consequently, the plot of structural factors exhibits a repeating pattern of features corresponding to each of the excitation wavelengths. As discussed in Reference [47], different pieces of information, such as the variance accounted for the model, the corcondia value [53], and the shape of the structural factors, are used to identify the structural model that describes the array of error covariance matrices. Different split-half models suggest that the error structure in both cases can be decomposed using two factors, since the models accounted for more than 90% of the variance and gave corcondia values of 100%. When a third component was added, the corcondia values decreased in all cases to values below 70%. Furthermore, additional extracted components explained little variation (less than 2% in all cases), were very noisy, and similar in shape to the preceding components. The first structural factor resembles the average emission profile for different excitation wavelengths, as anticipated in the theory section, and it also describes more than 90% of the variation of the model. This component is characterized by a very low vector angle (2.7° and 8.9° for Data Sets 1 and 2, respectively), indicating a high similarity among the estimates for different split-half models. The second structural component is more heterogeneous than the first, as the analysis of vector angles indicates (16.7° and 10.8° for Data Sets 1 and 2, respectively). However, the contribution of this component to the error covariance matrix structure is smaller, as is the variance that it describes. In addition, some split-half models indicate that this high variability arises from a few odd-numbered samples in the first half of the data set (i.e., samples 1–13). It is also localized in the long-wavelength region of the emission spectrum where chemical information is likely minimal, as can be seen in Figure 3.

The sample contributions for the first component are quite variable, which is expected since these contribution values represent the stochastic contributions of the structural factors, as represented in Equations 8 and 10. Assuming the original contributions satisfy a normal distribution, the contributions extracted from the error covariance matrices should follow a squared normal distribution if pooling is acceptable. This is consistent with the pattern observed for the first component. However, the sample contributions for the second component are characterized by substantial deviations in the contributions of samples 1, 5, and 11. Problems with these samples are the likely cause of disturbances in the estimation of the second structural factor described in the previous paragraph. Because this disturbance appears in both data sets, it may be related to the preparation of these samples. However, as noted in earlier work [47], the departure from homogeneity due to sample contribution will not

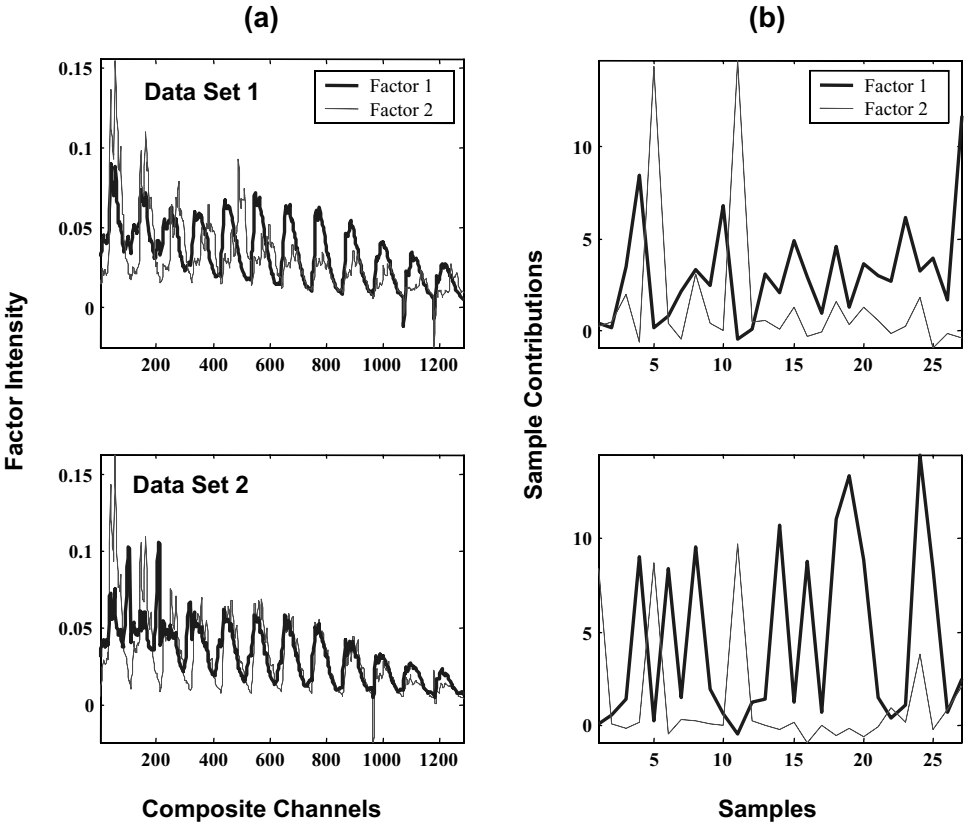

Figure 8. Results of two-component PARAFAC decomposition of the individual error covariance matrices for the composite mode formed by excitation and emission modes for Data Sets 1 and 2: (a) structural factors, (b) sample contributions.

preclude the pooling of the error covariance matrices. This violation is not as important as the violation of structural similarity, and in these cases the structural differences are not really considerable, since the second component has a small contribution to the error structures.

Summarizing all of the information presented, it can be said that Data Sets 1 and 2 are affected by correlated noise that permeates though the excitation and emission modes, while in Data Set 3, the correlated noise is also affecting the sample mode. These results indicate that Data Set 3 will need the use of general MLPARAFC to produce optimal results. The homogeneity analysis of the error covariance matrices for Data Sets 1 and 2 using a number of split-half models indicates that pooling is advisable, since the model was well described by two structural factors and followed an expected distribution in the sample contributions.

4.2. Estimation assessment

4.2.1. Figures of merit

Due to the intrinsic differences in the experimental orders estimated (concentrations and spectra), two different figures of merit will be used to assess the performance of the methods. The figure of merit used to measure the quality of the concentration estimates is the root-mean-square error of the estimation (RMSEE) calculated as follows:

$$\mathrm{RMSEE}_p^r = \sqrt{\frac{\sum (\hat{\mathbf{y}}_p^r - \mathbf{y}_p^{\mathrm{o}})^2}{N_\mathrm{s} - 1}} \tag{11}$$

where $\hat{\mathbf{y}}_p^r$ represents the estimated $N_\mathrm{s} \times 1$ vector of concentrations for component p and replicate block r, $\mathbf{y}_p^{\mathrm{o}}$ is the corresponding $N_\mathrm{s} \times 1$ vector of standard concentrations, and N_s is the number of samples. The use of $(N_\mathrm{s} - 1)$ degrees of freedom for RMSEE is justified by the fact that PARAFAC model has a well-known scaling indeterminacy that has to be estimated using at least a reference sample. This equation is applied to the R replicate blocks and the average value is obtained using Equation 12:

$$\overline{\mathrm{RMSEE}}_p = \frac{\sum_{r=1}^{R} \mathrm{RMSEE}_p^r}{R} \tag{12}$$

In order to make the interpretation of this value more meaningful, a relative average root-mean-square error of the estimation ($\overline{\mathrm{RRMSEE}}_p$) is calculated. This is determined with respect to the average concentration for component p, symbolized by $\bar{y}_p$, yielding:

$$\overline{\mathrm{RRMSEE}}_p = \frac{\overline{\mathrm{RMSEE}}_p}{\bar{y}_p} \tag{13}$$

For the excitation and emission modes, vector angles are preferred as a figure of merit, since they describe the quality of the estimates more clearly from a geometric point of view. The expression used to calculate this figure of merit is given in Equation 14

$$\theta_p^r = \cos^{-1}\left(\frac{\mathbf{f}_p^{\mathrm{T}} \mathbf{f}_p^r}{\|\mathbf{f}_p\| \|\mathbf{f}_p^r\|}\right) \tag{14}$$

Here, $\mathbf{f}_p^r$ represents the estimated emission or excitation profile for component p using replicate block r and $\mathbf{f}_p$ represents the corresponding reference emission or excitation profile, obtained from separate scans of the pure components. As in the case of the RMSEE, the vector angle is also averaged over R replicates:

$$\bar{\theta}_p = \frac{\sum_{r=1}^{R} \theta_p^r}{R} \tag{15}$$

In addition to these two figures of merit that can be used individually to assess the performance of each algorithm for each component and mode, a global indicator of the relative performance of each algorithm with respect to the corresponding standard, PARAFAC estimation, was used. This magnitude will be referred to as the *performance ratio*, PR, and is calculated as follows for the spectral and concentration modes:

$$\begin{aligned} \mathrm{PR}_{\mathrm{spec}} &= \frac{\left(\sum_{p=1}^{P} \bar{\theta}_p^{\mathrm{X}}\right)}{\left(\sum_{p=1}^{P} \bar{\theta}_p^{\mathrm{PARAFAC}}\right)} \\ \mathrm{PR}_{\mathrm{conc}} &= \frac{\left(\sum_{p=1}^{P} \overline{\mathrm{RRMSEE}}_p^{\mathrm{X}}\right)}{\left(\sum_{p=1}^{P} \overline{\mathrm{RRMSEE}}_p^{\mathrm{PARAFAC}}\right)} \end{aligned} \tag{16}$$

In this equation the superscript 'X' represents any of the possible algorithms that will be used in this work. PR values lower than unity will indicate superior performance of the given method over PARAFAC for the same data set, while values greater than unity will indicate inferior performance. The authors are aware of the drawbacks of such a summary statistic, which can be significantly biased by extreme values of any of the components. However, if the indicator is used with caution, it has the ability to simplify the analysis considerably.

4.2.2. *Performance of the algorithms*

Although a proper permutation arrangement and format for the error covariance matrix were suggested for each data set in Subsection 4.1, in this section, the results obtained for all possible permutations and with different error covariance matrix formats are presented. This was done to compare the results obtained with different formulations and permutation orders. There were two objectives in doing this: (1) to demonstrate that the incorporation of measurement error information can yield improved results over PARAFAC, even if it is done in a sub-optimal manner, and (2) to show that best results are obtained with the proper error covariance structure. It is important to note, before starting the description and discussion of the results, that no cross-comparisons among different data sets were done since a number of experimental factors such as photodecomposition, solvent volatilization, and other factors associated with the temporal stability of the samples cannot be controlled. A good indication of these effects is the fact that the performance of the optimal method for each data set decreases from Data Set 1 (first data set recorded) to Data Set 3 (last data set recorded).

Tables 1, 2, and 3 summarize the results for Data Sets 1, 2, and 3, respectively. Each table shows the performance for each component when different structures of the error covariance matrix and the corresponding algorithms were used. The first column of each table gives the algorithm used and, by implication, the format of the error covariance matrix assumed. The second column specifies which mode(s) were considered to be affected by correlated errors. The performance is measured as $\overline{\mathrm{RRMSEE}}_p$ for the concentration profiles and as $\bar{\theta}_p$ for the emission and excitation profiles. In addition, the performance ratios (PR) with respect to the PARAFAC estimates are also reported as a global indicator of performance. The rows of the tables shown in bold indicate the best conditions found in this study for each data set.

For all of the data sets, the use of error information translated into a superior performance of the algorithms tested over PARAFAC as a general trend, with the only exceptions being Data Sets 1 and 2 when analyzed using error covariance matrices assuming only sample correlation. This result is expected, since the previous analysis of the measurement errors indicated that, for Data Sets 1 and 2,

Table 1. Results obtained by different algorithms when applied to different arrangements of Data Set 1

		$\overline{\mathrm{RRMSEE}}$				$\bar{\theta}$ (Emission profiles)				$\bar{\theta}$ (Excitation profiles)			
Method	Correlated orders	Ace	Nap	Phe	PR	Ace	Nap	Phe	PR	Ace	Nap	Phe	PR
PARAFAC	—	0.0713	0.0827	0.0533	1.00	2.82	1.79	4.38	1.00	1.80	2.79	7.30	1.00
Compress PARAFAC	—	0.0673	0.0785	0.0510	0.95	2.16	1.67	3.98	0.87	1.21	0.95	3.85	0.50
1A	Emission	0.0516	0.0655	0.0399	0.76	1.71	1.30	3.17	0.69	0.94	0.76	2.95	0.39
1A	Excitation	0.0606	0.0644	0.0428	0.81	1.85	1.44	3.31	0.73	0.99	0.80	3.21	0.42
1A	Samples	0.0754	0.0904	0.0584	1.08	2.61	1.95	4.61	1.02	1.40	1.04	4.31	0.57
1B	Emission	0.0575	0.0639	0.0433	0.79	1.84	1.36	3.27	0.72	1.00	0.77	3.16	0.41
1B	Excitation	0.0596	0.0688	0.0414	0.82	1.98	1.45	3.44	0.77	1.02	0.81	3.28	0.43
1B	Samples	0.0786	0.0944	0.0582	1.11	2.55	1.98	4.47	1.00	1.39	1.09	4.25	0.57
1C	**Emission excitation**	**0.0488**	**0.0548**	**0.0338**	**0.66**	**1.61**	**1.13**	**2.81**	**0.62**	**0.85**	**0.63**	**2.57**	**0.34**
1C	Emission-samples	0.0570	0.0686	0.0419	0.81	1.95	1.41	3.27	0.74	1.03	0.77	3.14	0.42
1C	Excitation samples	0.0693	0.0795	0.0492	0.95	2.19	1.64	3.95	0.87	1.17	0.91	3.80	0.49
1D	**Emission excitation**	**0.0478**	**0.0525**	**0.0348**	**0.65**	**1.58**	**1.15**	**2.60**	**0.59**	**0.83**	**0.61**	**2.55**	**0.33**
Compress full MLPARAFAC	Sample emission excitation	0.0642	0.0719	0.0491	0.89	2.06	1.60	3.63	0.81	1.13	0.88	3.58	0.47

Row(s) in bold represent(s) best case scenario.

Table 2. Results obtained by different algorithms when applied to different arrangements of Data Set 2

		$\overline{\text{RRMSEE}}$				$\bar{\theta}$(Emission profiles)				$\bar{\theta}$ (Excitation profiles)			
Method	Correlated orders	Ace	Nap	Phe	PR	Ace	Nap	Phe	PR	Ace	Nap	Phe	PR
PARAFAC	—	0.0740	0.1019	0.0618	1.00	2.55	2.62	7.45	1.00	2.12	2.60	7.20	1.00
Compress PARAFAC	—	0.0721	0.0965	0.0572	0.95	1.94	1.98	6.71	0.84	1.21	1.12	3.45	0.48
1A	Emission	0.0580	0.0798	0.0488	0.78	1.61	1.65	5.58	0.70	1.02	0.92	2.96	0.41
1A	Excitation	0.0651	0.0861	0.0542	0.86	1.72	1.76	6.17	0.76	1.09	1.05	3.17	0.45
1A	Samples	0.0810	0.1105	0.0694	1.10	2.32	2.30	7.94	0.99	1.47	1.34	4.14	0.58
1B	Emission	0.0615	0.0796	0.0506	0.81	1.66	1.66	5.84	0.73	1.10	0.98	3.01	0.43
1B	Excitation	0.0620	0.0886	0.0523	0.85	1.80	1.77	6.11	0.77	1.15	1.02	3.17	0.45
1B	Samples	0.0846	0.1053	0.0708	1.10	2.26	2.20	8.08	0.99	1.42	1.34	4.11	0.58
1C	**Emission excitation**	**0.0433**	**0.0625**	**0.0372**	**0.60**	**1.32**	**1.31**	**4.47**	**0.56**	**0.79**	**0.72**	**2.36**	**0.32**
1C	Emission samples	0.0618	0.0845	0.0519	0.83	1.68	1.71	6.02	0.75	1.11	0.99	3.17	0.44
1C	Excitation samples	0.0713	0.0950	0.0588	0.95	1.92	1.88	6.62	0.83	1.22	1.12	3.48	0.49
1D	**Emission excitation**	**0.0454**	**0.0641**	**0.0387**	**0.62**	**1.30**	**1.26**	**4.51**	**0.56**	**0.77**	**0.77**	**2.19**	**0.31**
Compress full MLPARAFAC	Sample emission excitation	0.0606	0.0828	0.0524	0.82	1.62	1.72	5.88	0.73	1.10	1.00	2.89	0.42

Row(s) in bold represent(s) best case scenario.

Table 3. Results obtained by different algorithms when applied to different arrangements of Data Set 1

		$\overline{\text{RRMSEE}}$				$\bar{\theta}$ (Emission profiles)				$\bar{\theta}$ (Excitation profiles)			
Method	Correlated orders	Ace	Nap	Phe	PR	Ace	Nap	Phe	PR	Ace	Nap	Phe	PR
PARAFAC	—	0.1119	0.1801	0.0700	1.00	8.12	2.41	10.56	1.00	4.53	7.24	8.52	1.00
Compress PARAFAC	—	0.1032	0.1650	0.0666	0.92	4.18	1.90	8.82	0.71	3.16	2.32	4.07	0.47
1A	Emission	0.0944	0.1478	0.0635	0.84	3.89	1.78	8.33	0.66	2.85	2.07	3.69	0.42
1A	Excitation	0.0955	0.1587	0.0636	0.88	4.09	1.89	8.38	0.68	3.04	2.22	3.93	0.45
1A	Samples	0.0939	0.1477	0.0605	0.83	4.06	1.89	8.22	0.67	2.93	2.12	3.73	0.43
1B	Emission	0.0937	0.1546	0.0616	0.86	3.99	1.78	8.34	0.67	2.88	2.16	3.72	0.43
1B	Excitation	0.1004	0.1585	0.0605	0.88	4.14	1.86	8.41	0.68	2.98	2.29	3.92	0.45
1B	Samples	0.0893	0.1570	0.0610	0.85	3.93	1.79	8.60	0.68	2.91	2.19	3.67	0.43
1C	Emission excitation	0.0979	0.1402	0.0560	0.81	3.87	1.69	7.95	0.64	2.73	2.04	3.51	0.41
1C	Emission samples	0.0944	0.1467	0.0625	0.84	3.93	1.72	8.37	0.66	2.92	2.13	3.58	0.43
1C	Excitation samples	0.0931	0.1398	0.0598	0.81	3.78	1.80	7.82	0.64	2.86	2.01	3.57	0.42
1D	Emission excitation	0.0956	0.1476	0.0560	0.83	3.65	1.75	7.66	0.62	2.68	2.08	3.53	0.41
Compress full MLPARAFAC	**Sample emission excitation**	**0.0870**	**0.1319**	**0.0542**	**0.75**	**3.46**	**1.55**	**7.50**	**0.59**	**2.56**	**1.81**	**3.29**	**0.38**

Row(s) in bold represent(s) best case scenario.

there was no correlation affecting the sample domain. Therefore, the use of an erroneous error covariance matrix with spurious correlations will only have a negative effect on the performance. Comparatively, introduction of error information related to the emission order produces marginally better performance than the use of error information describing the excitation mode. Different levels of improvement were found when information about the correlated error affecting the emission and excitation orders as a composite mode was utilized by different algorithms. In other words, the performances of algorithms such as 1C and 1D, which include error covariance information about the composite mode formed by emission and excitation profiles, were significantly better than the performance of algorithms using error covariance information of either emission or excitation profiles alone (e.g., algorithms 1A and 1B). Further to this argument, insignificant advantages in terms of performance were found by introducing more localized information about the error structure, as can be seen by comparing the results obtained by algorithms 1A and 1B when the same spectroscopic order (emission or excitation) was considered. This is a clear indication that the sources of variation contributing to the error structure are a combination of effects, such as the multiplicative and offset contributions anticipated in the analysis of measurement errors, that permeate through the composite mode formed by the excitation and emission modes. Similar levels of improvement were also observed for Data Set 3 when information about the error covariance affecting the sample mode was introduced. This was an important confirmation that the sources of correlation found for the sample mode in the analysis of the measurement errors for Data Set 3 were real and the inclusion of them will translate in a better performance.

The results for Data Sets 1 and 2 were very similar. This was anticipated due to the similar error structure found in both data sets. Methods 1C and 1D using error covariance

matrices of a composite mode formed by the emission and excitation orders yielded the best results for both data sets, as was anticipated by the analysis of the error covariance. Improvements in performance in the range between about 60 and 80% were observed for different modes. There were not significant differences in performance observed between algorithms 1C, which use a pooled $JK \times JK$ error covariance matrix, and 1D, which use a set of $JK \times JK$ error covariance matrices. However, for method 1D, only two pooled error covariance matrices were used instead of a set of I individual error covariance matrices. This simplification was carried out to reduce the computational load of the algorithm, which would have been prohibitive. One of the error covariance matrices was constructed by pooling the odd-numbered samples and the other was constructed pooling the even-numbered samples. This partitioning was based on evidence found during the analysis of error covariance that suggested anomalous behavior of some odd-numbered samples in the first half of the data set (see Subsection 4.1). Although no significant differences were found using algorithm 1D with this approach, it is difficult to generalize this conclusion to the case of I covariance matrices.

The relative improvement in predictive ability of the compressed general MLPARAFAC algorithm was the most important difference between Data Set 3 and Data Sets 1 and 2. For the three data sets, a Tucker3 compression basis set formed by 12 components for the sample and excitation modes and 20 components for the emission mode was employed. These parameters were selected on the basis of principles developed in a companion paper [46], but results were not especially sensitive to them as long as a sufficient number of components were used. Even though this alternative produced an improvement over the PARAFAC model for Data Sets 1 and 2, the results were worse than the those produced by most other algorithms. This situation can be explained by considering that, for Data Sets 1 and 2, the introduction of error information about the sample domain is likely to make the error covariance matrix less reliable due to the introduction of spurious correlations and a reduction in the number of replicates in the estimation process. On the other hand, the existence of an important source of error structure in the sample order for Data Set 3 makes the estimation of the error covariance matrix essential and more than makes up for a reduction in the number of replicates.

In the application of the general MLPARAFAC methodology to compressed data sets, performance enhancement can result not only from the use of error covariance information but also from the compression procedure itself. To dissect the improvements from each of these sources, PARAFAC was also applied to the compressed data. As can be seen from the results in Tables 1–3, the use of PARAFAC on the compressed data produced some improvements, but these are not as large as the improvements observed by using MLPARAFAC on the same data, indicating the benefit of using a weighted estimation method.

Some interesting details emerge when the prediction performances are analyzed for each component. In all cases the concentration profile of phenanthrene yields the lowest error followed by acenaphthylene and naphthalene. However, the emission and excitation profiles of phenanthrene are poorly predicted in comparison to the other two compounds. This may be indicative of a trade-off trend in the estimation process that needs to be studied more thoroughly. It is also worth noting that poor performance exhibited by Data Sets 1 and 2 when error information describing the sample domain was used mainly affected the estimation of the concentration profiles, indicating again the irrelevant information carried by these error covariance representations.

Even though the time involved in the calculations of these models was not specifically tabulated, it typically ranged from one to a few hours. By comparison, PARAFAC models were computed in time windows of a few minutes to an hour, depending on the size of the data set and initial estimates. Therefore, the construction of a table similar to the ones presented here to choose the best arrangement and algorithm for a given data set is not recommended. However, the results presented here validate the analysis of error covariance as an exploratory strategy to choose the best arrangement and algorithm given the available data.

5. CONCLUSIONS

In this work, a number of practical aspects related to the application of the different simplifications of MLPARAFAC to experimental data have been explored. The algorithms employed were described in an earlier companion paper [46] and these were applied to three sets of fluorescence EEM data from mixtures of three polycyclic aromatic hydrocarbons. A number of important tools, previously introduced for the analysis of the error structure affecting two-way data [47], were extended to three-way data in this work. These tools were applied to the three different data sets to characterize the error structure. Two of the data sets exhibited error covariance along the composite mode consisting of excitation and emission modes, while the third exhibited error covariance along all three modes. These characterizations allowed estimation of an optimal representation of the error covariance matrix for each data set. When used with the corresponding algorithm, these error covariance matrices yielded the best models in each case. Different error structures and algorithms were employed, showing that the inclusion of statistically meaningful error information always produced an improvement in the estimates over conventional PARAFAC, even in cases where the error covariance information was incomplete. The level of improvement depends on the quality and importance of the error information, but in this work, improvements over PARAFAC by as much as a factor of three were observed.

REFERENCES

1. Hirschfeld T. The hy-phen-ated methods. *Anal. Chem.* 1980; **52**: 297A.
2. Beltran JL, Ferrer R, Guiteras J. Multivariate calibration of polycyclic aromatic hydrocarbon mixtures from excitation-emission fluorescence spectra. *Anal. Chem. Acta* 1998; **373**: 311–319.
3. de Juan A, Rutan SC, Tauler R, Massart DL. Comparison between the direct trilinear decomposition and the

multivariate curve resolution-alternating least squares methods for the resolution of three-way data sets. *Chemometrics Intell. Lab. Syst.* 1998; **40**: 19–32.
4. Beltran JL, Guiteras J, Ferrer R. Three-way multivariate calibration procedures applied to high-performance liquid chromatography coupled with fast-scanning fluorescence spectrometry detection: determination of polycyclic aromatic hydrocarbons in water samples. *Anal. Chem.* 1998; **70**: 49–1955.
5. Beltran JL, Guiteras J, Ferrer R. Parallel factor analysis of partially resolved chromatographic data: determination of polycyclic aromatic hydrocarbons in water samples. *J. Chromatogr. A* 1998; **802**: 3–275.
6. Wu HL, Shibukawa MOK. An alternating trilinear decomposition algorithm with application to calibration of HPLC-DAD for simultaneous determination of overlapped chlorinated aromatic hydrocarbons. *J. Chemom.* 1998; **12**: 1–26.
7. Bro R. PARAFAC. Tutorial and applications. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 149–171.
8. Xie YL, Baeza-Baeza JJ, Ramis-Ramos G. Second-order tensorial calibration for kinetic spectrophotometric determination. *Chemometrics Intell. Lab. Syst.* 1996; **32**: 215–232.
9. 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.
10. Gui M, Rutan SC, Agbodjan A. Kinetic detection of overlapped amino acids in thin-layer chromatography with a direct trilinear decomposition method. *Anal. Chem.* 1995; **67**: 3293–3299.
11. Karukstis KK, Krekel DA, Weinberger DA, Bittker RA, Naito NR, Bloch SH. Resolution of the excited states of the fluorescence probe TNS using a trilinear analysis technique. *J. Phys. Chem.* 1995; **99**: 449–453.
12. Henshaw JM, Burgess LW, Booksh KS, Kowalski BR. Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. 1. Multivariate calibration of Fujiwara reaction products. *Anal. Chem.* 1994; **66**: 3328–3336.
13. Booksh KS, Lin Z, Wang Z, Kowalski BR. Extension of trilinear decomposition method with an application to the flow probe sensor. *Anal. Chem.* 1994; **66**: 2561–2569.
14. Lin Zo, Booksh KS, Burgess LW, Kowalski BR. A second-order fiber optic heavy metal sensor employing second-order tensorial calibration. *Anal. Chem.* 1994; **66**: 2552–2560.
15. Sanchez E, Kowalski BR. Tensorial resolution: a direct trilinear decomposition. *J. Chemom.* 1990; **4**: 29–45.
16. da Silva JC, Novais SA. Trilinear PARAFAC decomposition of synchronous fluorescence spectra of mixtures of the major metabolites of acetylsalicylic acid. *Analyst* 1998; **23**: 2067–2070.
17. Appellof CJ, Davidson ER. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Anal. Chem.* 1981; **53**: 2053–2056.
18. Tauler R, Marques I, Casassas E. Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths. *J. Chemom.* 1998; **12**: 55–75.
19. Millican DW, McGown LB. Fluorescence lifetime resolution of spectra in the frequency domain using multiway analysis. *Anal. Chem.* 1990; **62**: 2242–2247.
20. Phillips GR, Georghiou S. Global analysis of steady-state polarized fluorescence spectra using trilinear curve resolution. *Biophys. J.* 1993; **65**: 918–926.
21. Lee JK, Ross RT, TS, Leurgans S. Resolution of the properties of hydrogen-bonded tyrosine using a trilinear model of fluorescence. *J. Phys. Chem.* 1992; **96**: 9158–9162.
22. Martins JA, Sena MM, Poppi RJ, Pessine FBT. Fluorescence piroxicam study in the presence of cyclodextrins by using the PARAFAC method. *Appl. Spectrosc.* 1999; **53**: 510–522.
23. Ross RT, Lee CH, Davis CM, Ezzeddine BM, Fayyad EA, Leurgans SE. Resolution of the fluorescence spectra of plant pigment-complexes using trilinear models. *Biochim. Biophys. Acta* 1991; **1056**: 317–320.
24. Wentzell PD, Nair SS, Guy RD. Three-way analysis of fluorescence spectra of polycyclic aromatic hydrocarbons with quenching by nitromethane. *Anal. Chem.* 2001; **73**: 1408–1415.
25. Anderson GG, Dable BK, Booksh KS. Weighted parallel factor analysis for calibration of HPLC-UV/Vis spectrometers in the presence of Beer's law deviations. *Chemometrics Intell. Lab. Syst.* 1999; **49**: 195–213.
26. Bro R, Sidiropoulos ND, Smilde AK. Maximum likelihood fitting using simple least squares algorithms. *J. Chemom.* 2002; **16**: 387.
27. Carroll JD, Chang J. Analysis of individual differences in multidimensional scaling via an N-way generalization of ''Eckart-Young'' decomposition. *Psychometrika.* 1970; **35**: 283.
28. Harshman RA. Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multimode factor analysis. *UCLA Working Papers in phonetics* 1970; **16**: 1.
29. Paatero P. A weighted non-negative least squares algorithm for three-way ''PARAFAC'' factor analysis. *Chemometrics Intell. Lab. Syst.* 1997; **38**: 223.
30. Wu HL, Shibukawa M, Oguma K. Second-order calibration based on alternating trilinear decomposition: a comparison with the traditional PARAFAC algorithm. *Anal. Sci.* 1997; **13**: 53–58.
31. Chen ZP, Wu HL, Jiang JH, Li Y, Yu RQ. A novel trilinear decomposition algorithm for second-order linear calibration. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 75–86.
32. Chen ZP, Li Y, Yu RQ. Pseudo alternating least squares algorithm for trilinear decomposition. *J. Chemom.* 2001; **15**: 149–167.
33. Jiang JH, Wu HL, Li Y, Yu RQ. Alternating coupled vectors resolution (ACOVER) method for trilinear analysis of threeway data. *J. Chemom.* 1999; **13**: 557–578.
34. Jiang JH, Wu HL, Li Y, Yu RQ. Three-way data resolution by alternating slice-wise diagonalization (ASD) method. *J. Chemom.* 2000; **14**: 15–36.
35. Li Y, Jiang JH, Wu HL, Chen ZP, Yu RQ. Alternating coupled matrices resolution method for three-way arrays analysis. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 33–43.
36. Faber NM, Bro R, Hopke PK. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics Intell. Lab. Syst.* 2003; **65**: 119–137.
37. Liu X, Sidiropoulos N. Cramér-Rao Lower Bound for Low-Rank Decomposition of Multidimensional Arrays. *IEEE Trans. Signal Processing* 2001; **49**: 2074.
38. Baunsgaard D, Andersson CA, Arnadal Allan, Munck L. Multi-way chemometrics for mathematical separation of fluorescent colorants and colour precursors from spectrofluorimetry of beet sugar and beet sugar thick juice as validated by HPLC analysis. *Food Chem.* 2000; **70**: 113–121.
39. JiJi RD, Andersson GG, Booksh KS. Application of PARAFAC for calibration with excitation-emission matrix fluorescence spectra of three classes of environmental pollutants. *J. Chemom.* 2000; **14**: 170–185.
40. JiJi RD, Cooper GA, Booksh KS. Excitation-emission matrix fluorescence based determination of carbamate

pesticides and polycyclic aromatic hydrocarbons. *Anal. Chim. Acta* 1999; **397**: 61–72.
41. Bro R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemometrics Intell. Lab. Syst.* 1999; **46**: 33–147.
42. Moberg L, Robertsson G, Karlberg B. Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis. *Talanta* 2001; **54**: 161–170.
43. Pedersen DK, Munck L, Engelsen SB. Screening for dioxin contamination in fish oil by PARAFAC and N-PLSR analysis of fluorescence landscapes. *J. Chemom.* 2002; **16**: 451–460.
44. Ingle JD, Crouch SR. *Spectrochemical Analysis*. Prentice Hall: New Jersey, 1972.
45. Vega-Montoto L, Wentzell PD. Maximum Likelihood Parallel Factor Analysis (MLPARAFAC). *J. Chemom.* 2003; **17**: 237–253.
46. Vega-Montoto L, Gu H, Wentzell PD. Mathematical improvements to maximum likelihood parallel factor analysis: theory and simulations. *J. Chemom.* (in press).
47. Leger M, Vega-Montoto L, Wentzell PD. Methods for systematic investigation of measurement error covariance matrices. *Chemometrics Intell. Lab. Syst.* 2005; **77**: 181–205.
48. Wentzell PD, Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemometrics Intell. Lab. Syst.* 1999; **45**: 65.
49. Brown CD, Vega-Montoto L, Wentzell PD. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Appl. Spectrosc.* 2000; **54**: 1055.
50. Mardia K, Kent JT, Bibby JM. *Multivariate Analysis*. Academic Press: 1979. New York.
51. Harshman RA, Lundy ME. The PARAFAC model for three-way factor analysis and multidimensional scaling. In *Research Methods for Multimode Data Analysis*. Praeger: New York; 1984, 122.
52. Andersson CA, Bro R. The N-way toolbox for MATLAB. *Chemometrics Intell. Lab. Syst.* 2000; **52**: 1.
53. Bro R, Kiers HAL. A new efficient method for determining the number of component in PARAFAC models. *J. Chemom.* 2003; **17**: 274.

