INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

MAXIMUM LIKELIHOOD METHODS FOR THREE-WAY ANALYSIS IN CHEMISTRY

BY

LORENZO JORGE VEGA MONTOTO

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
AT
DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA

APRIL 2005

DALHOUSIE UNIVERSITY

To comply with the Canadian Privacy Act the National Library of Canada has requested that the following pages be removed from this copy of the thesis:

Preliminary Pages

Examiners Signature Page (pii)

Dalhousie Library Copyright Agreement (piii)

Appendices

Copyright Releases (if applicable)

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D-NMR	Two-Dimensional Nuclear Magnetic Resonance
Ace	Acenaphthylene
ALS	Alternating Least Squares
CANDECOMP	Canonical Decomposition
CANDELINC	Canonical Decomposition with Linear Constraints
CCR	Classical Curve Resolution
CR	Continuum Regression
DAD	Diode Array Detector
DECRA	Direct Exponential Curve Resolution Algorithm
DTLD	Direct Trilinear Decomposition
ECM	Error Correlation Matrix
EEM	Excitation-Emission Matrix
GRAM	Generalized Rank Annihilation Method
iid	Independent and Identically Distributed
LM-PARAFAC	Levenberg-Marquartd Parallel Factor Analysis
LS	Least Squares
MCR	Multivariate Curve Resolution
MILES	Maximum Likelihood via Iterative Least Squares Estimation
ML	Maximum Likelihood
MLLRR	Maximum Likelihood Latent Root Regression
MLPARAFAC	Maximum Likelihood Parallel Factor Analysis
MLPCA	Maximum Likelihood Principal Component Analysis
MLPCR	Maximum Likelihood Principal Component Regression
MLR	Multiple Linear Regression
MS	Mass Spectrometry
MSC	Multiplicative Signal Correction
Nap	Naphthalene
NIR	Near Infrared

PARAFAC	Parallel Factor Analysis
PBM	Postponed Basis Matrix Multiplication
PCR	Principal Component Regression
Phe	Phenanthrene
PLS	Partial Least Squares
PMF	Positive Matrix Factorization
PR	Performance Ratio
PSGE	Pulse Gradient spin Echo
RAFA	Rank Annihilation Factor Analysis
RMSEE	Root Mean Square Error of the Estimate
RR	Ridge Regression
RRMSEE	Relative Root Mean Square Error of the Estimate
S/N	Signal to Noise Ratio
SBE	Successive Bayesian Estimation
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
TLD	Trilinear Decomposition
TLS	Total Least Squares
UV-Vis	Ultraviolet-Visible
WCR	Weighted Curve Resolution
W-PARAFAC	Weighted Parallel Factor Analysis

Abstract

Analytical instrumentation that produces multivariate data is now commonplace in chemical laboratories. Such data includes first-order tensors (*e.g.* a spectrum), second-order tensors (*e.g.* chromatography with multichannel detection), and third- and higher order tensors (*e.g.* fluorescence excitation-emission-lifetime measurements). From a data analysis perspective, the "cubes" of data that form third-order tensors, or three-way data, offer unique advantages not observed for lower order measurements. In particular, data that exhibit a trilinear structure can be decomposed in such a way that unique underlying factors are extracted, without the rotational ambiguity that exists when bilinear data are used. Common data analysis tools employed to carry out this type of decomposition include the well-known Parallel Factor Analysis (PARAFAC) and Direct Trilinear Decomposition (DTLD) algorithms. The application of these tools to trilinear data has tremendous potential to extract fundamental information such as spectra, concentration profiles, rate constants, and equilibrium constants from complex mixtures with little or no prior information. However, this potential is mitigated by the fact that these methods do not optimally accommodate the complex error structures commonly found in three-way data.

In this work, the development and application of Maximum Likelihood Parallel Factor Analysis (MLPARAFAC) is described. This approach is designed to incorporate prior measurement error information, including information about heteroscedascity and correlation of errors, into the decomposition procedure. Although MLPARAFAC is an extension of maximum likelihood methods for two-way data, the application to three-way data greatly expands the types of error interactions that can be observed, the size of matrices produced, and complexity of the algorithms involved. The principles behind the generalized MLPARAFAC algorithm, as well as several simplifications based on different measurement error structures and pre-compression of the data, are described. These algorithms are applied to simulated data and three fluorescence data sets to demonstrate their statistical validity, computational efficiency and estimation accuracy. Further implementation of MLPARAFAC in conjunction with the Direct Exponential Curve Resolution Algorithm (DECRA) is also examined using simulated data and spectral data from two widely studied reaction systems.

Acknowledgments

I am grateful to a number of people that have made important contributions to the flourishing of both my personal and professional skills during my graduate studies at Dalhousie University. Foremost, I would like to express my sincere gratitude to my *master* Dr. Peter D. Wentzell for giving me the opportunity to become his *disciple*. His constant advice, support, thoroughness and exemplary leadership would be sufficient to highlight his excellent qualities as a supervisor and teacher. However, it was his guidance and perseveration in trying to mold me into a more organized individual in order to fully span my potential that makes him an invaluable influence in my professional formation. I will take his many pieces of advice with me everywhere. I also thank Dr. Hong Gu for lengthy discussions that allowed me to apprehend the mathematical knowledge employed in this work and Dr. Rasmus Bro at KVL in Copenhagen for providing me with his constant support and insights with respect to three-way methods.

I had the opportunity to share my time at Dalhousie with a number of amazing individuals that, I am sure, are going to excel in their future life. Former and current members of the Wentzell's Club such as Chris Brown, Marc Leger, Praneeth D. Edirisinghev, Tobias Karakach, Hannes Hochreiner and Robert Flight have made every day of this journey an adventure into the deep subtleties of the human soul. My stay in Halifax was also blessed by meeting people such as Mina, Laura, Carl, Ivonne, Carolina, Tijana, and Family Fraser-Yax among many other great friends that we will always treasure.

Financial support for this work was provided in part by the Natural Science and Engineering Research Council and the Dow Chemical Company. I am deeply grateful to Dalhousie University and the Department of Chemistry for giving me the opportunity to come here and providing financial support through scholarships and other helpful means. I want to thank the Centro de Ingeniería y Manejo Ambiental de Bahías y Costas (CIMAB) and the late director Manuel Alepuz for giving me the opportunity to come to Canada to carry out graduate studies.

Esta parte será escrita en español porque a pesar que mucho de sus protagonistas hablan ingles, estoy seguro que preferirán ser reconocidos en su lengua natal.

Quisiera empezar con mis amigos, porque en mi opinión por excelente e influyente que haya sido la familia de cualquiera de nosotros, lo que somos en gran medida tiene que ver con ese contacto íntimo y social que solo se logra a través de años llenos de cosas buenas y malas, pero siempre compartidas. Primero quisiera reconocer el impacto que tuvieron mis amigos, los que yo quisiera llamar los Leninistas, no por creencias políticas sino por circunstancias: Rubén, Ayón, Ariel, Liliam, Lisandro, Liván entre muchos otros, gracias por los mejores años de mi vida de estudiante. Después quisiera reconocer el impacto de mi generación universitaria: Yaima, Joe, Nelaine, Chocky, Erick, Roberto, Yuri y Vlado por esos tiempos duramente humanos, pero llenos de vida. Canadá estuvo llena de nuevas caras, caras frescas y con mucho que ofrecer gente como Aurora, Keila, Jorge, Yan, Sonia, Castillo, la familia Motta-Allen y otros ya mencionados hicieron este pedazo de polo sentirse un poco más tropical. Otros amigos como Fernando, Oscarito, Otto, William, Rafael, Julio Cesar, esos que no caben dentro de estos grupos establecidos por razones temporales, pero que siempre tuvieron un impacto positivo en lo que hoy soy.

El papel de mi familia es tan importante como ubicuo, no puedo pensar en algo que sea capaz de proponerme o hacer que no este de algún modo motivado o influido por algún miembro de ese núcleo compacto que llamo familia, para todos Uds., este doctorado no es solo mío sino de todos NOSOTROS. A mis padres he de agradecerles no solo haberme traído al mundo y darme la oportunidad de concentrarme en mis estudios en una de las épocas más complejas de la historia de mi país, sino les agradezco el constante apoyo y creencia en mi potencial para ser una persona mejor tanto emocional como intelectualmente y sobre todo por su amor incondicional. A mi papá, Lorenzo, por darme el ejemplo de que leer, cuestionarse cosas y aprender siempre fueron tareas arduas, pero extremadamente disfrutables, a mi mamá, Sara, por inculcarme que todo es posible si uno da lo mejor de sí en el empeño. A mis hermanos Edita, Raúl y Ramón por siempre tener su apoyo y ayuda en cuanto me propusiera, aun cuando pareciera algo totalmente impensable. A mis sobrinos Raulito, Carlitos y Ramoncito, que ya no son tan niños, y que siempre los he visto como una extensión más de mi ser y que me han hecho replantearme mi vida en términos en los cuales pueda ser un mejor ejemplo para Uds. A la Familia Delgado Romero, no solo por los lazos que me unen con Perla, sino porque Uds. con su

amor no solo me han bienvenido a su hogar sino que me han hecho ser uno más de esa familia.

A mi esposa Perla, *amora* muchas gracias no solo por tu amor y tu entrega sin limites, los cuales ya te valdrían para merecer este doctorado mucho más que yo, sino por mucho más, por esas pequeñas cosas que no se pueden resumir en palabras, por esos momentos que duran menos que un instante pero que pueden hacer un día de derrotas lucir más luminoso y por sobre todas las cosas por ser tú y permitirme ser yo para juntos poder crear algo más grande que nosotros mismos.

"Las ciencias confirman lo que el espíritu posee:
la analogía de todas las fuerzas de la naturaleza,
la semejanza de todos los seres vivos,
la igualdad de la composición de todos los elementos del Universo,
la soberanía del hombre"

José Martí
Revista Universal de México
21/9/1875

Chapter 1
Introduction

1.1 Overview

The last few decades have seen a remarkable increase in the amount and complexity of data generated by modern instruments for chemical analysis. Different types of chromatographic and spectroscopic instrumentation are now commonplace in the analytical laboratory and coupling of such instruments is now routine. A few examples of the "hyphenated methods" that generate such copious amounts of data are time-decay and emission-excitation fluorescence, chromatography-spectroscopy combinations, tandem mass spectrometry (MS/MS) and two-dimensional nuclear magnetic resonance (2D-NMR). These methods can provide data in several dimensions (*e.g.* wavelength and time) at once. The necessity to analyze several samples, either with individual techniques or with combinations, adds yet another order to the data obtained. Contrary to univariate analytical methods that require full selectivity for proper effectiveness functionality, multivariate analytical methods are much more flexible, requiring only fractional selectivity in a multivariate context. This can dramatically reduce sample preparation procedures, saving the analyst time and money. The efficient nature of these analytical procedures, in conjunction with the ability to obtain general and accurate information regarding chemical and physical properties of the sample will make them perfect candidates for a wide variety of industrial, environmental, medical, and research applications.

Unfortunately, the measurements obtained from these instruments may be rather complex, increasing the necessity for data analytical methods that can handle these types of data. In 1971 a group of pioneering scientists coined the term "*chemometrics*" to describe the growing use of mathematical, statistical and computer-based methods in the field of chemistry [1-2]. Despite the fact that many of the so-called chemometrics techniques obviously existed prior to the 1970s, the philosophy behind the techniques as

applied to the discipline of chemistry has evolved considerably over the past decade. This has been a consequence not only related to the increasing necessity for more powerful mathematical methods to treat chemical data, but also due to the increasing availability of inexpensive computational resources. Besides the partial selectivity that allows the analyst to quantify a compound in the presence of interferents, some additional advantages of employing chemometric methods in conjunction with multivariate measurements include multiple component estimation, noise reduction, and outlier detection.

In the beginning, multivariate analytical measurements were naturally arranged into vectors and matrices, but analytical methods have now developed to the point where some techniques will yield measurements that are better arranged in multi-way arrays (*i.e.* cubes or hyper-cubes) due to the intrinsic structure of these data. The application of chemometric methods specially designed to handle this type of data will produce additional benefits and has led to terms such as "second-order advantage" and "mathematical chromatography", both of which are intimately related to the ability to decompose this type of data into the underlying contributions from individual components. Nonetheless, these benefits will only be attainable if the proper model is estimated, and the estimation process will depend heavily on the nature of the errors affecting the data. The characteristics of measurement errors and their proper treatment in the analysis of multi-way data is the central theme of this thesis.

Given the complexity of modern instrumentation, it is evident that measurement errors can arise from a wide variety of sources, and have a correspondingly complex range of properties and characteristics. Usually, the description of instrumental noise is based on its dominant source, which is associated with certain distributional and frequency characteristics. For instance, the source of Johnson or thermal noise is the thermal agitation of electrons or other charge carriers in resistive elements in the instrument. It is typically considered as a type of fundamental noise, since it does not arise from instrument or component deficiencies, and is the same at every frequency and can never be totally neglected. It is ubiquitous in resistive elements whether they are carrying current or not. Another type of fundamental noise is shot noise, or quantum

noise, that arises from the random statistical nature of discrete events, whose rate is subject to statistical fluctuations well-described by the Poisson distribution. The overall effect of shot noise is usually much lower than that of Johnson noise, although this may not be the case in measurements based on a low number of events, such as molecular fluorescence. Pink or $1/f$ noise can arise from a variety of sources and is named after the inverse proportionality between the magnitude of the noise fluctuations and the frequency of the observed signal (f). This type of noise is considered non-fundamental and generally arises from longer-term (lower-frequency) variation of instrumental components. Some other terms used to refer to this type of noise are flicker noise and drift noise. In addition to those already mentioned, there are a variety of other noise sources including such things as detector noise, read-out noise, quantization noise, and noise from environmental sources (interference noise).

From a more fundamental point of view, the noise can be classified using some of its characteristics in various ways, the most common of which include: its source, its distribution, and its characteristics in the frequency and time domains. Unfortunately, classifications based on these categories are not all mutually exclusive. Instead, a more concise but less precise term will be coined to classify noise. It will be defined in terms of whether or not it is independent and identically distributed with a normal distribution (*iid*-normal). The term "*iid* normal noise" summarizes a lot of information. The concept of independence with regard to measurement errors implies that the error observed at any one element in the signal vector is independent or unrelated to the error observed at any other (different) element in the signal vector. Independence in the measurement errors is equivalent to saying that measurement errors are uncorrelated. The term *identically distributed* implies a homogeneity in the error variance across all elements in the signal vector; *i.e.*, the standard deviation at every elements in the signal vector is the same. The terms *homoscedastic* and *heteroscedastic* are also often used to indicate whether measurement errors are identically distributed, or not. The "normal" condition simply refers to the normal distribution often assumed for the noise observed at one channel in the signal vector over many repeat measurements. Therefore, measurement errors are classified to be *iid*-normal if all of the aforementioned conditions are fulfilled and non-*iid*-normal if any of these conditions are violated. Assumptions regarding normality are

seldom significantly violated (recall the Central Limit Theorem), and therefore, in the future, references to *iid*-normal errors will often be shortened to simply "*iid* errors". Mathematically, violations of the *iid* condition can be detected and corrected through the use of the error covariance matrix.

Unfortunately, the vast majority of chemometric methods employed are typically extensions of existing univariate methods in terms of their assumptions about the error structure (*i.e.* the assumption that it is *iid*-normal). Although some consideration of heteroscedascity in univariate calibration has been made through weighted regression methods [3, 4], measurements are generally assumed to be independent. Because single channel measurements are used, the existence of correlation among errors at different channels is impossible, and correlation among samples is generally considered negligible based on the preparation of standards. Until recently, the migration to multichannel instrumentation and, therefore to multivariate data analysis has occurred, without the corresponding advance in handling of measurement errors.

The initial attempts to accommodate non-*iid* error structures in multivariate data analysis addressed the problems of heteroscedasticity while assuming that measurement errors remained independent. One of the first approaches was to extend the principles of weighted univariate regression to multivariate regression [5]. This strategy can be successful for certain calibration problems that can be treated with multiple linear regression (MLR), but does not address broader methodologies of calibration and decomposition. Paatero and Tapper [6] showed that heteroscedastic noise could be treated in bilinear decomposition through a simple scaling of the data, as long as the error structure could be represented as a rank one matrix. For cases that did not meet this criteria, Paatero developed a method called Positive Matrix Factorization (PMF2) for bilinear data [7]. This was later extended to three-way data (PMF3) [8]. On the other hand, the impact of correlated noise has attracted much less interest among analytical chemists [9-12], although its consequences can be quite severe [13]. Nevertheless, the formulation of Maximum Likelihood Principal Component Analysis (MLPCA) and other derived techniques, such as Maximum Likelihood Principal Component Regression (MLPCR) and Maximum Likelihood Latent Root Regression (MLLRR), by Wentzell and

collaborators [9, 14] have shed some light on the positive effects of including error information in the form of the error covariance matrix in multivariate data analysis to obtain optimal or near-optimal models from a theoretical point of view. This methodology is closely related to a number of other methods currently in use, including Total Least Squares (TLS) [10] and Maximum Likelihood via Iterative Least Squares Estimations (MILES) [12]. These different approaches are essentially the same in terms of the statistical model that they use to incorporate measurement error information, but differ substantially in their numerical implementations.

The transition from two-way data to three-way and multi-way data is accompanied by a concomitant increase in the complexity of the model and the noise structure. The objective of the present study is to extend maximum likelihood principles developed for MLPCA to multi-way analysis by formulating optimal algorithms to estimate the Parallel Factor Analysis (PARAFAC) model. Following a concise introductory chapter explaining the relationship between different physical, structural and statistical multi-way models, Chapter 2 introduces the general formulation of the MLPARAFAC algorithm to estimate the PARAFAC model in cases where heteroscedasticity and/or correlation occur, as well as two mathematically exact simplifications and some statistical testing of the methods. In Chapter 3, a more extensive exploration of the possible simplifications is carried out and a number of different simplifications for a variety of error structures are formulated. Chapter 4 introduces some tools to characterize the error covariance matrix for three-way arrays and describes the application and comparison of the different formulations of MLPARAFAC to three experimental data sets. Chapter 5 explores the application of MLPARAFAC to a particular problem in the analysis of kinetic data, through the use of the recently introduced Direct Exponential Curve Resolution Algorithm (DECRA). The thesis closes with a summary of the results for this research and possible avenues for future investigation.

Note that the main body of this thesis is composed of a published paper (Chapter 2) or manuscripts that have been submitted for publication (Chapters 3-5).

1.2 Instrumentation and physical models

The mathematical tools used to analyze a set of experimental measurements are closely related to the order of the data tensor obtained from the analytical device. In general, instruments can be classified by the order of data they produce; for instance, zero-order, first-order, second-order and so on. The relationship between the form of the data and the order of the instrumentation is presented in Figure 1.1. As a general rule, the use of an n-order instrument analyzing a single sample produces an n-order tensor. The use of the same instrument analyzing a data set formed by multiple samples yields an $(n+1)$–order tensor.

Zero-order instruments are analytical instruments that produce a single measurement per sample analyzed, for instance, pH, viscosity or absorbance at a single wavelength. However, most of the time the measurement obtained is not ultimately the property sought, but it is theoretically or empirically related to a more attainable property such as acidity, temperature or concentration, requiring a process called calibration. A well-known example is the application of spectroscopy in chemistry. The specimen of interest is a homogeneous solution of chromophores contained in a cuvette of known path length, L (usually 1 cm). Integration of the ratio dI/I leads to Beer's Law, shown in equation 1.1:

$$A = -\log_{10}(I/I_o) = \varepsilon(\lambda)cL \tag{1.1}$$

where A is the absorbance, I/I_o is the transmittance of a beam of light of wavelength λ, and c is the concentration of chromophore. The fundamental property of the chromophore is $\varepsilon(\lambda)$, the molar absorptivity. Unfortunately, this quantity which characterizes a particular analyte can also change with other parameters such as temperature, pH, and solvent. Therefore, calibration needs to be used to estimate this parameter, which in turn defines the structural model relating absorbance obtained and the concentration sought. The calibration of this type of device can be carried out using univariate calibration. This type of calibration is well understood and easy to adopt, since it works in a very simple statistical framework and the equipment used is inexpensive. While the simplicity of

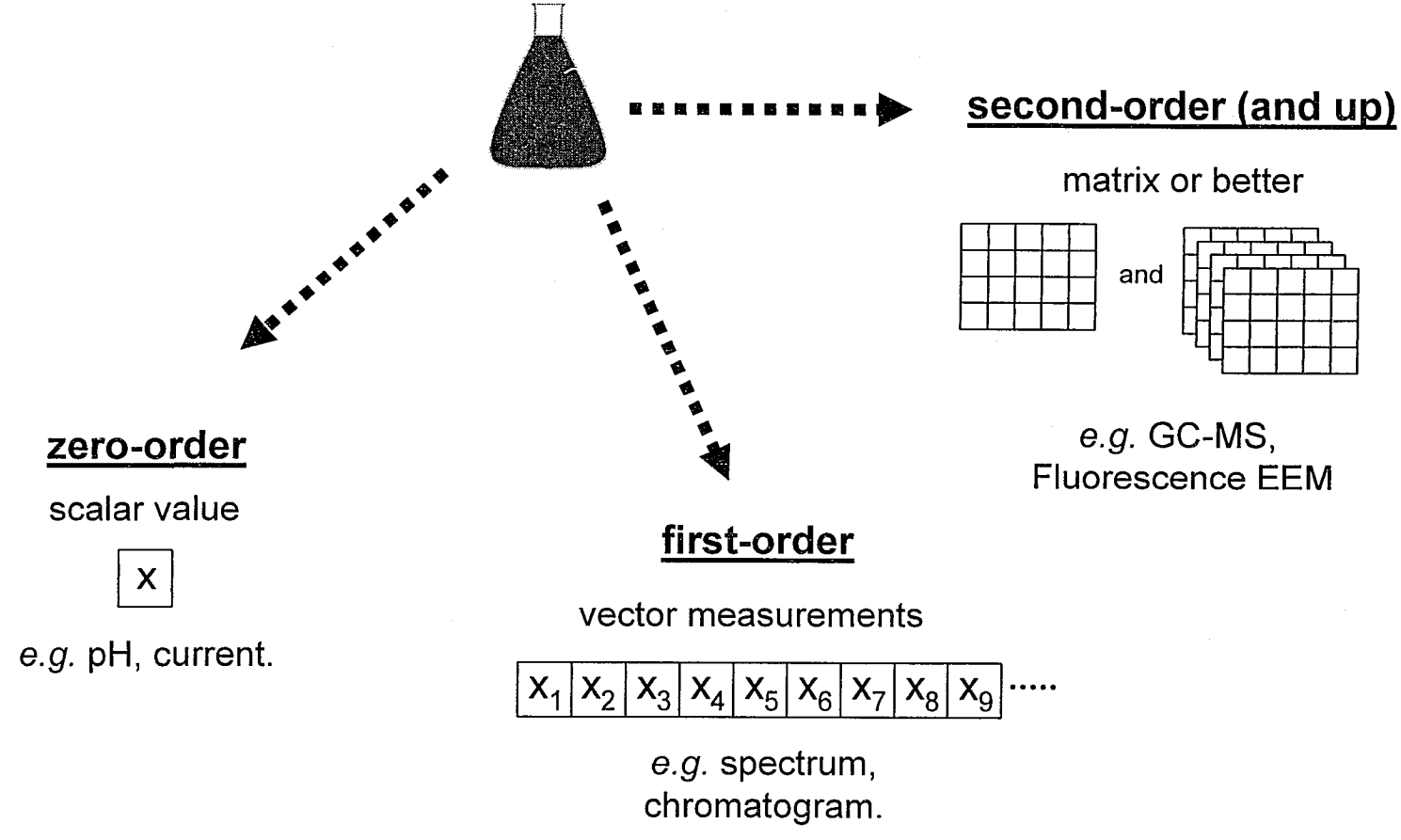

Figure 1.1. Relationship between the order of data and the instrumental characteristics.

univariate calibration is an attractive attribute, there are several fundamentally limiting properties of the approach. Univariate calibration methods naturally require full selectivity for the analyte of interest. Interferences can therefore only be handled in the rather ideal case in which the amount of the interference is constant in all calibration and prediction samples. This severe limitation mathematically precludes doing calibration in the presence of interferences, and simultaneous multicomponent analysis. In addition to the disadvantage of full selectivity, zero-order instruments cause difficulties in detecting outliers present in the analytical data. These intrinsic deficiencies can be overcome by migrating from zero order instruments to first order instrumentation.

With multiple P chromophores that need to be quantified, the experiment can involve additional independent variables. Therefore, first-order instruments are analytical devices that yield a vector of first-order data per sample as shown in Figure 1.1. The spectrometer is a common example of a first-order instrument, where the primary independent variable is the energy or wavelength of the photons absorbed or emitted. Innovations such as diode array detectors (DAD) for UV-visible absorption instruments, charge-coupled devices for fluorescence instruments, and Fourier transform techniques for nuclear magnetic resonance and infrared spectroscopy have made first-order data more readily available. The combination of two first-order techniques can produce second-order data, but the most common way to obtain second-order data is when the first-order instrument is used to analyze multiple samples. An example of the mathematical representation of the physical model behind second-order data is given in equation 1.2, which is an expression of Beer's Law for measurements at multiple wavelengths for a single component.

$$A_{ik} = c_i \varepsilon_k L \tag{1.2}$$

Here c_i is the concentration of the chromophore in sample i, ε_k is the molar absorptivity of the chromophore at wavelength k, and L is the path length of the cuvette. Theoretically, the physical model can be seen as an explicit causal (or hard) model that relates the instrumental measurements to the level of the analyte of interest, and often to the level of interfering components via an extension of Beer's law when multiple components are

involved. For multiple samples, the concentration of chromophore i can be represented by a vector **c**, which is an I x 1 vector containing the concentration of the chromophore for each sample. If P different spectroscopically active components are present in the mixtures, then equation 1.2 can be represented by:

$$\begin{aligned} \mathbf{A} &= (\mathbf{c}_1\boldsymbol{\varepsilon}_1^{\mathrm{T}} + \mathbf{c}_2\boldsymbol{\varepsilon}_2^{\mathrm{T}} + \ldots + \mathbf{c}_P\boldsymbol{\varepsilon}_P)L \\ &= \mathbf{C}\mathbf{S}^{\mathrm{T}} \end{aligned} \tag{1.3}$$

where $\boldsymbol{\varepsilon}$ is a K x 1 vector of molar absorptivities for each of K wavelength channels, **C** (I x P) and **S** (K x P) are the concentration and spectral profiles (molar absorptivities multiplied by the path length), respectively, and **A** is an I x K (samples x wavelengths) matrix of absorbances, with each row corresponding to the spectrum of a different sample. The matrix of spectra arises from the simple outer product of the concentration vector, and the pure-component spectrum for the component. It should be noted that, although this example uses spectroscopy, equation 1.3 applies equally well in its general form to other first-order instruments. A visual representation of this structure is shown in Figure 1.2, where the concentration vectors are defined by HPLC elution profiles for a two-component mixture.

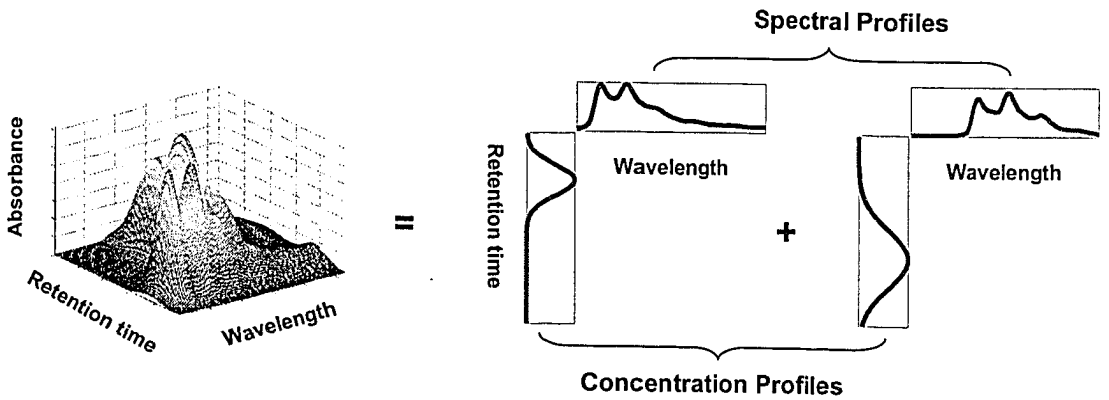

Figure 1.2. Physical model representation for second order data using a HPLC-DAD example.

Instruments that generate second order arrays of data or higher order tensors per sample are now more commonplace in the analytical laboratory. Time-decay and emission-excitation fluorescence, and other combinations of chromatographic and

spectroscopic techniques are a few examples of the many hyphenated methods that generate such data. In addition, the use of multiple samples with these techniques makes the data one order higher. A simple example to introduce the physical model is data from steady-state fluorescence. In this technique, the absorption of a photon puts a chromophore into a higher energy excited state, which can then rapidly decay, emitting a photon as fluorescence. In fluorescence spectroscopy, the sample is illuminated with light of wavelength λ^{ex}, and the consequent emission of light is measured at wavelength λ^{em}. For one fluorophore, the process is mathematically expressed as:

$$\mu_{ijk} = \alpha_i \pi_j c_k \tag{1.4}$$

where α_i is the coefficient of absorption of the fluorophore (incorporating molar absorptivity, quantum yield and source intensity) at excitation wavelength λ_i^{ex}, π_j is the relative emission (incorporating the emission profile and the detector response profile) at detection wavelength λ_j^{em} and c_k is the concentration of fluorophore in question in sample k. Equation 1.4 represents the mathematical relationship describing a single element of the three-way array, $\underline{\mathbf{U}}$, which is an I x J x K cube of data for I excitation wavelength channels, J emission wavelength channels, and K samples. This array can be viewed as K matrices representing excitation-emission spectra that are "slices" in the cube of data. Equation 1.5 represents the physical model for such a system with P components:

$$\begin{aligned} \mathbf{U}^k &= \boldsymbol{\alpha}_1 c_1^k \boldsymbol{\pi}_1^{\mathrm{T}} + \boldsymbol{\alpha}_2 c_2^k \boldsymbol{\pi}_2^{\mathrm{T}} + \cdots + \boldsymbol{\alpha}_p c_p^k \boldsymbol{\pi}_p^{\mathrm{T}} \\ &= \mathbf{A}\mathbf{D}^k\mathbf{B}^{\mathrm{T}} \end{aligned} \tag{1.5}$$

In this specific case, **A** and **B** are matrices which contain the excitation spectra ($[\alpha_1\ \alpha_2 \cdots\ \alpha_P]$) and the emission spectra ($[\pi_1\ \pi_2 \cdots\ \pi_P]$), respectively, of the pure components along the columns. $\mathbf{D}^k$ represents a diagonal matrix formed by the concentration of all P components for the k^{th} sample placed on the main diagonal of **D**. This type of data is diagrammatically depicted in Figure 1.3. Simple application of this physical model requires that sample absorbance be small and that excitation energy not be transferred between chromophores. The violation of these conditions will prevent the use of this simple physical model, since more complex models will be necessary.

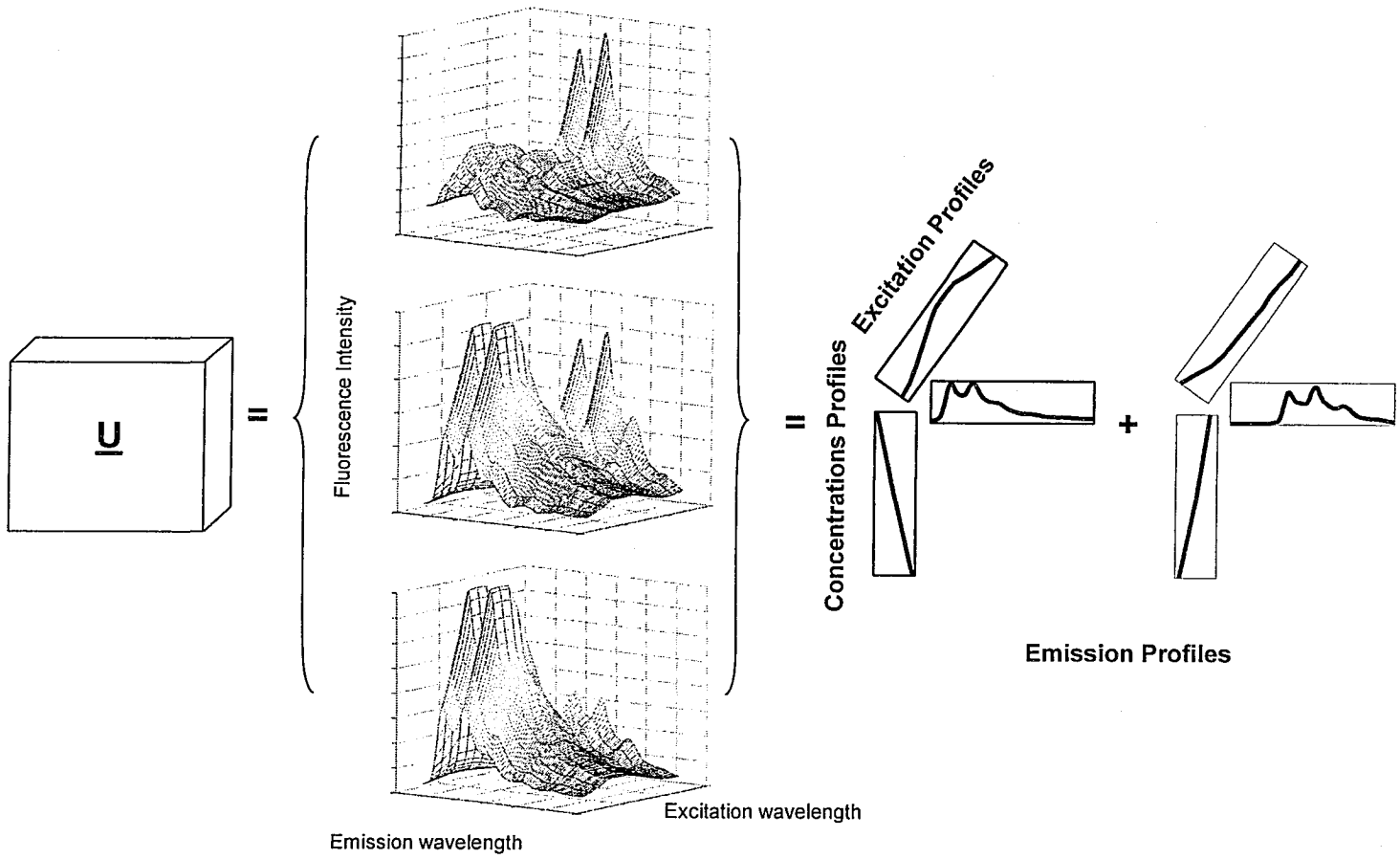

Figure 1.3. Physical model representation for third order data using an EEM example.

1.3 Structural models

Structural models can be divided into two categories: models that can often be regarded as a good approximate model of the true underlying phenomena, and models where there is little or no theory to describe how the data are related to the underlying phenomena. The latter group can be exemplified with process or sensory data while the former corresponds to spectral data of the type described in the preceding sections. In this thesis, all of the models treated belong to the first group.

1.3.1 Second-order data

Second-order data can be represented by bilinear models. A major difficulty with general bilinear models is that the underlying physical factors are usually impossible to separate by using only the data at hand, since the factors are subject to rotational and scaling ambiguities. This means that many sets of paired **C**- and **S**–type matrices can equivalently reproduce the original data set **X**. Equation 1.6 shows how the product of $\mathbf{RR}^{-1}$, which is equal to the identity matrix, can be placed between profiles **C** and **S**, leaving **X** unchanged and modifying the concentration and spectral profiles to **CR** and $\mathbf{R}^{-1}\mathbf{S}^{T}$, respectively.

$$\begin{aligned} \mathbf{X} &= \mathbf{C}\mathbf{S}^{T} \\ &= \mathbf{C}\mathbf{R}\mathbf{R}^{-1}\mathbf{S}^{T} = \mathbf{T}\mathbf{V}^{T} \end{aligned} \tag{1.6}$$

This situation reflects the ambiguity, since **R** can be any non-singular square matrix; therefore, an infinite number of equivalent solutions exists.

The existence of this intrinsic ambiguity opens two different avenues for the data analysis. Two different strategies can be employed, depending on whether the final aim of the analysis process is the prediction of future samples or the estimation of the first-order profiles. The prediction of future samples is carried out using any of the many possible variations of inverse calibration, such as principal component regression (PCR), partial least squares regression (PLS), ridge regression (RR), or continuum regression (CR), providing that a representative set of the future samples exists and the property sought has been determined by a reference method. The estimation of the individual profiles composing this type of data is carried out using a wide variety of methods

described as multivariate curve resolution (MCR) algorithms [15]. To obtain these profiles, MCR methods employ different types of constraints, ranging from empirical ones, such as non-negativity and unimodality [16] to theoretically justified ones, such as kinetic hard models [17]. It is worth noting that the method referred to as classical least squares calibration, also known as direct calibration, is the best-case scenario for MCR, in which all of the information about one of the orders is known in advance.

Geometrically speaking, inverse calibration tries to estimate the subspace which contains the property sought and the magnitude of the property in this space, while MCR methods estimate the subspace and try to estimate the best rotational matrix **R**, to align the vectors defining the subspace with the vectors representing the profiles in each mode. From here we can see that both avenues have a common core, which is the estimation of the subspace which contains the data. Many different methods have been created to do this, but a vast majority of these techniques have as a common core a method called principal component analysis (PCA) [18], which is able to estimate the subspace spanned by the underlying factors forming the data. Therefore, a short introduction to PCA will be given in the next section. This also provides a good starting point for the introduction to the different trilinear methods that are the focus of this work.

1.3.1.1 Principal components analysis

Principal components analysis (PCA) is the most widely used variable reduction method in chemistry. In addition to forming the basis for PCR, PCA is widely used in other applications, such as pattern recognition and curve resolution, and it is related to other regression methods such as PLS, RR and CR. The first known chemical application of PCA appeared in 1878 [19], and even at that time it was used in a regression problem. The extensive use of this method was paralleled by the evolution of computing technology which made computation times more reasonable, so that even large arrays are now easily handled. From a vectorial point of view, PCA transforms the data from a redundant coordinate system into a more natural coordinate system uniquely reflecting the sources of variability in the data.

The bilinear structural model will be introduced using the same scheme of the bilinear data represented by equation 1.3. Instead of using the matrices **C** and **S**,

representing the physical underlying profiles, a more general representation based on rotated matrices **T** and **V**, as used in equation 1.6, will be employed.

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} \tag{1.7}$$

Uniqueness is an important issue for structural models. It can be said that a structural model is unique when no additional constraints are necessary to identify the model. For a unique structural model, the parameters cannot be changed without changing the fit of the model. Therefore, in order to eliminate the ambiguities characterizing the bilinear model, PCA introduces orthogonality constraints in its scores and loadings (*i.e.* the dot product of different columns within **T** or $\mathbf{V}^{\mathrm{T}}$ should be zero). Principal components analysis goes a step further by constraining the first principal component to be the vector describing the most observed variance. From a mathematical point of view the first loading is defined as the normalized vector that maximizes the variance of ($\mathbf{X}\,\mathbf{v}_1$,) or in other words maximizes:

$$\mathbf{v}_1^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}_1 = \mathbf{t}_1^{\mathrm{T}}\mathbf{t}_1 \tag{1.8}$$

The next loading, $\mathbf{v}_2$, is defined as the vector maximizing the same quantity, *i.e.* $\mathbf{v}_2^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}_2 = \mathbf{t}_2^{\mathrm{T}}\mathbf{t}_2$ under the constraint that $\mathbf{t}_1$ and $\mathbf{t}_2$ are orthogonal, *i.e.* $\mathbf{t}_1^{\mathrm{T}}\mathbf{t}_2 = 0$. The procedure continues this way until all dimensions have been accounted for, always under the constraint that new scores are orthogonal to previous ones. It is important to note that the combination of the structure of this model and the imposed constraint makes the PCA model the best subspace and low rank approximation solution for bilinear data.

1.3.1.2 Numerical Implementation

Computationally, PCA can be carried out by a variety of numerical methods, such as the power method used for matrix diagonalization. However, singular value decomposition (SVD) is now the most commonly used method to perform PCA in chemometrics, since it is efficient, reliable and can accommodate rectangular matrices of the sort that appear frequently in chemical applications. Using SVD, the data matrix **X** (*I* x *J*) is decomposed into the product of three different matrices as is shown in equation 1.9.

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{1.9}$$

Here, the columns of **U** (I x I) represent the eigenvectors of the covariance matrix ($\mathbf{XX}^T$). These eigenvectors are orthonormal (orthogonal to each other and of unit length). The main diagonal of **S** (I x J) contains the singular values, which are the square roots of the eigenvalues of the covariance matrices ($\mathbf{XX}^T$) and ($\mathbf{X}^T\mathbf{X}$). The elements of the main diagonal of **S** are ordered in a decreasing fashion ($s_{ii} \geq s_{jj}, i < j$), representing the square root of the variance accounted for by each PC. Off diagonal elements are zero ($s_{ij} = 0$, $i \neq j$). $\mathbf{V}^T$ is a (J x J) matrix formed by the eigenvectors of the covariance matrix ($\mathbf{X}^T\mathbf{X}$). The row vectors of $\mathbf{V}^T$ are orthonormal and identical to the loadings described for PCA. The score matrix T is the product of matrices **U** and **S** (**T** = **US**). Truncation is easily achieved by discarding the uninformative rows and columns of the matrices.

$$\tilde{\mathbf{X}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^T \tag{1.10}$$

To summarize, PCA is a method to perform a regression of **X** onto a lower dimensional estimate of itself, thereby reducing the high dimension of the original space to a relatively simple P-dimensional subspace. It also provides a bilinear decomposition avoiding the rotational and scaling ambiguities by imposing two important constraints. These two constraints work by extracting loadings, which describe most of the variance and are mutually orthogonal among each other. In this way the PCA model can accomplish the estimation of a model that at the same time is the best low rank and best subspace approximation of bilinear data.

1.3.2 Third order data

1.3.2.1 Parallel Factor Analysis

The physical model shown in equation 1.5 is equivalent to a structural model called Parallel Factor Analysis (PARAFAC), independently introduced by Harshman [20] and by Carroll and Chang [21] in 1970. Contrary to second-order data, that can always be represented by a bilinear model providing the best subspace and low rank approximation solution, for third order data these two seemingly similar decompositions are obtained by two structurally different three-way models called Tucker [22] and PARAFAC models, providing the best subspace and low rank approximation solution, respectively [23]. Both

models can be generically represented by the same equation, but some restrictions are applied to distinguish one from the other.

Before explaining the different structural representations, it is important to note that this model is symmetric in the sense that if the array is reordered so that, for example, the first and third modes are exchanged, then the models of the two data sets will be identical, except that the first and third mode loadings have been interchanged. The structural model can be represented in different but equivalent ways depending on the way the three-way data are arranged. In the following chapters, some cases will be presented in which different representations are needed in order to estimate the parameters defining the structural model.

The matricial/unfolded representation is the most common form used to represent the PARAFAC model, since it can be easily expressed using standard linear-algebra identities and also provides an excellent starting point for describing certain estimation methods. The $I \times J \times K$ cube of data can be unfolded, yielding a matricial representation of the data. Only the equation representing the data array when it is unfolded by retaining the first dimension will be shown, since, due to the symmetry of the model, the other unfolded representations are equivalent after permutation:

$$\mathbf{X}_a = [\mathbf{X}^1, \mathbf{X}^2, ..., \mathbf{X}^K] = \mathbf{A}\mathbf{G}_a(\mathbf{C} \otimes \mathbf{B})^T \tag{1.11}$$

Equivalently, the elements forming the data can be arranged as a sequence of elements providing the vectorial representation:

$$vec(\underline{\mathbf{X}}) = (\mathbf{A} \otimes \mathbf{C} \otimes \mathbf{B})vec(\underline{\mathbf{G}}) \tag{1.12}$$

where **A**, **B** and **C** are the matrices whose columns represent the loadings for the three modes. It is common three-way practice not to distinguish between scores and loadings as these are treated equally from a numerical perspective. The dimensions of these loading matrices mark the first difference between the Tucker and the PARAFAC models. In the former, the second dimension of the loadings can be different, while in the latter the second dimension is the same for all the loadings. For instance, if a three-way array of data $\underline{\mathbf{X}}$ is given as an $I \times J \times K$ block of data, a Tucker3 model decomposition with a rank

D in the first mode, E in the second mode, and F in the third mode can be formulated. On the other hand, a PARAFAC model will be defined as a rank P model since all the loadings have the same number of columns. The second difference is related to the structure of $\underline{\mathbf{G}}$. The matrix $\mathbf{G}_a$ is the unfolded representation of the core array $\underline{\mathbf{G}}$ arranged as a matrix. This matrix is similar to the $\mathbf{S}$ matrix found in the PCA case, since the elements of the core define how individual loading vectors in the different modes interact and these elements also carry the magnitude of these interactions. Mathematically, in PARAFAC, $\underline{\mathbf{G}}$ will be composed of zeros everywhere but in the superdiagonal (g_{111}, g_{222}, …, g_{PPP}). For the Tucker model, non-zero values are allowed for the elements besides the superdiagonal elements in $\underline{\mathbf{G}}$.

The structure of $\underline{\mathbf{G}}$ for both models will have a profound impact on the uniqueness of these models. The Tucker3 model has rotational freedom, and is hence not structurally unique. This can be seen by replacing the model in equation 1.11 with a model where the first mode loadings have been rotated by a non-singular square matrix, $\mathbf{S}$. Then, by counter-rotating $\mathbf{G}_a$ with the inverse of this matrix, the model is obtained.

$$\mathbf{X}_a = \mathbf{A}\mathbf{S}\mathbf{S}^{-1}\mathbf{G}_a(\mathbf{C} \otimes \mathbf{B})^T \tag{1.13}$$

As we can see, this model is completely equivalent to the original model, hence rotation is possible. Even with the orthogonality constraints imposed, the model does not provide an identifiable solution, since rotations by any orthogonal matrix of any of the loading matrices will provide new orthogonal loading matrices. In fact, it has been shown that the structural model is so redundant that several parameters, often more than half of the elements, in the core, $\mathbf{G}$, can be forced to zero without changing the fit of the model [24]. This clearly shows that the Tucker models are unnecessarily complex and explains why they give "ambiguous" results. As in two-way analysis the rotational freedom has prompted the need for rotations of solutions in order to increase interpretability [25, 26].

Before analyzing the uniqueness of the PARAFAC model, it is convenient to introduce an alternative representation, since this will make the proof simpler and will clarify how the PARAFAC model is related to the principle of parallel proportional profiles. It is important to say that Tucker3 model cannot be represented by this expression since some important interactions cannot be explicitly included. The

PARAFAC model can be written in terms of frontal slices. Let $\mathbf{X}_k$ be the I x J matrix obtained from $\underline{\mathbf{X}}$ by fixing the third mode at the value k. Therefore, the model is expressed in terms of its frontal slices as:

$$\mathbf{X}^k = \mathbf{A}\mathbf{D}^k\mathbf{B}^T \tag{1.14}$$

Here matrices **A** and **B** represent the loadings for mode A and B, respectively and $\mathbf{D}^k$ is the diagonal matrix, which contains the kth row of matrix **C** on its diagonal.

Equation 1.14 has an equivalent representation to the one shown above when a similar transformation is carried out and the model is reformulated as:

$$\mathbf{X}^k = \mathbf{A}\mathbf{P}\mathbf{P}^{-1}\mathbf{D}^k\mathbf{S}\mathbf{S}^{-1}\mathbf{B}^T \tag{1.15}$$

After the transformation, if the trilinear model holds, **AP**, $\mathbf{B}(\mathbf{S}^{-1})^T$ and $\mathbf{P}^{-1}\mathbf{D}^k\mathbf{S}$ can be considered the loadings instead of considering **A**, **B** and **C**. However, in order for the trilinear model to still hold, the product matrix $\mathbf{P}^{-1}\mathbf{D}^k\mathbf{S}$ must remain diagonal to represent the third mode. This fact constrains the matrices **P** and **S** (and hence $\mathbf{P}^{-1}$ and $\mathbf{S}^{-1}$) to be a very special type of matrix called permutation and/or scaling matrices. Therefore, solutions must be unique up to trivial differences in factor order and relative scaling across modes. Hence, contrary to the bilinear case, where a rotational ambiguity exists, in trilinear and multilinear data in general, this ambiguity does not exist, which allows for the unique estimation of the profiles. This intrinsic property of this type of data makes this technique very suitable for chemistry, since the estimated solutions can be obtained directly without any major assumption and compared to other profiles included in databases, allowing the identification of the compounds present in the samples. Another advantageous consequence of the unique estimation is the so-called "second order advantage", which allows the quantitation of an analyte in the presence of interferences with only one calibration sample, since only a scaling indeterminacy has to be solved.

Many studies have been carried out to understand the uniqueness of the PARAFAC model estimates. Harshman [27] and Leurgans *et al.* [28] showed that unique solutions can be expected if the loading vectors are linearly independent in two of the modes and there are no more than two linearly dependent loading vectors in the third mode. Kruskal [29 ,30] found less restricted conditions for uniqueness by using the *k*-

rank of the loading matrices, which is a term introduced by Harshman and Lundy [31]. The proof says that if any combination of k_A columns of **A** have full column-rank, and this does not hold for k_A+1, then the *k*-rank of **A** is k_A. The *k*-rank is thus related, but not equal, to the rank of the matrix, as the *k*-rank can never exceed the rank. Kruskal formulated the following equation as a proof:

$$k_A + k_B + k_C = 2P+2 \tag{1.16}$$

If under this set of conditions this proof holds, then the PARAFAC solution is unique. Here k_A is the *k*-rank of **A**, k_B is the *k*-rank of **B**, k_C is the *k*-rank of **C** and *P* is the number of PARAFAC components sought. None of the above conditions are strong enough to cover all situations where uniqueness can be expected, but they do give sufficient conditions for uniqueness. Note that, regardless of the above rule, a one-component solution is always unique. This even holds for a two-way decomposition.

1.3.2.2 Numerical Implementation

There are many published algorithms for fitting the PARAFAC model. Some of the most important methods will be briefly explained in this section. The first algorithms used to estimate the PARAFAC model were least squares iterative algorithms [20, 21] based on the principle of alternating least squares (ALS) [32] and it has been the most frequently used algorithm to date. Numerically, the principle behind ALS is to divide the highly nonlinear optimization problem into a sequence of conditional linear sub-problems and solve these in a least squares sense via simple established numerical methods. Equation 1.11 is an excellent starting point to understand the way ALS is implemented. The PARAFAC model can also be expressed as in equations 1.18 and 1.19 for the three-way data matricized in the second and third modes. Let $\mathbf{X}_a$ be the I x JK matricized array and define $\mathbf{X}_b$ (J x IK) and $\mathbf{X}_c$ (K x IJ) similarly as the arrays matricized in the second and third modes, respectively.

$$\mathbf{X}_a = \mathbf{A}\mathbf{I}_a(\mathbf{C} \otimes \mathbf{B})^T \tag{1.17}$$

$$\mathbf{X}_b = \mathbf{B}\mathbf{I}_b(\mathbf{C} \otimes \mathbf{A})^T \tag{1.18}$$

$$\mathbf{X}_c = \mathbf{C}\mathbf{I}_c(\mathbf{B} \otimes \mathbf{A})^T \tag{1.19}$$

Using expression 1.17, **A** can be estimated by assuming **B** and **C** by a simple least squares step as shown in equation T3 in Table 1.1. The full ALS algorithm is shown in Table 1.1. As can be seen from the table, the ALS algorithm consists of three least squares estimates, each providing a better estimate of one set of loadings, and the overall algorithm will therefore improve the least squares fit of the model to the data. It is worth noting that, even though a number of algorithms exist to estimate the PARAFAC model, only the ALS algorithm is extendable to higher order data. Despite the excellent convergence properties and well-defined optimization problem characterizing the ALS procedure, the iterative nature of the method is a major inconvenience.

This inconvenience can be overcome by using the direct trilinear decomposition method (DTLD) [33] which solves an eigenvalue problem. Historically, the evolution of DTLD is linked to a method called Rank Annihilation Factor Analysis (RAFA)[34]. RAFA tries to estimate the concentration of an analyte in an unknown matrix solely using the unknown sample and a pure standard. Mathematically, the idea behind RAFA was based on reducing the rank of the calibration sample by subtracting the contribution from the analyte of interest. That is, if the signal from the analyte of interest is subtracted from the sample data, then the rank of this matrix will decrease by one, as the contribution of the analyte of interest to the rank is one in the case of ordinary bilinear rank-one data like chromatographic or fluorescence data. Later, Lorber [35, 36] found that the algorithm could be automated by expressing the process as a generalized eigenvalue problem. In 1986, Sanchez and Kowalski generalized the method into the Generalized Rank Annihilation Method (GRAM) [37]. In GRAM several components could be present or absent in both calibration and standard samples. The method starts by finding two sets of orthogonal bases that represent the subspace in each mode using two slices of the three-way array. Subsequently, transformation matrices are found that transform these bases into estimated pure components. In DTLD, the GRAM method is extended to data with more than two slices by generating two pseudo-slices as differently weighted averages of all the slices. The loadings in two modes are estimated by using the synthetic data set and

Table 1.1. PARAFAC algorithm.

1. Given an *IxJxK* cube of data $\underline{\mathbf{X}}$, the algorithm is initialized using random values of the correct dimensions or using estimates obtained by TLD.

$$[\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}] = tld(\underline{\mathbf{X}},P)$$

Using the estimates of $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ the initial value for the objective function can be calculated using equation T1.

$$S_1^2 = trace\left((\mathbf{X}_a - \hat{\mathbf{X}}_a)(\mathbf{X}_a - \hat{\mathbf{X}}_a)^{\mathrm{T}}\right) \tag{T1}$$

2. Unfold $\underline{\mathbf{X}}$ retaining the first order and estimate $\hat{\mathbf{A}}$ conditional on $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$.

$$\hat{\mathbf{Z}}_a = \mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}} \tag{T2}$$

$$\hat{\mathbf{A}} = \mathbf{X}_a\hat{\mathbf{Z}}_a^{\mathrm{T}}(\hat{\mathbf{Z}}_a\hat{\mathbf{Z}}_a^{\mathrm{T}})^{-1} \tag{T3}$$

3. Unfold $\underline{\mathbf{X}}$ retaining the second order and estimate of $\hat{\mathbf{B}}$ conditional on $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$.

$$\hat{\mathbf{Z}}_b = \mathbf{I}_b(\mathbf{C}\otimes\mathbf{A})^{\mathrm{T}} \tag{T4}$$

$$\hat{\mathbf{B}} = \mathbf{X}_b\hat{\mathbf{Z}}_b^{\mathrm{T}}(\hat{\mathbf{Z}}_b\hat{\mathbf{Z}}_b^{\mathrm{T}})^{-1} \tag{T5}$$

4. Unfold $\underline{\mathbf{X}}$ retaining the third order and estimate $\hat{\mathbf{C}}$ conditional on $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

$$\hat{\mathbf{Z}}_c = \mathbf{I}_c(\mathbf{B}\otimes\mathbf{A})^{\mathrm{T}} \tag{T6}$$

$$\hat{\mathbf{C}} = \mathbf{X}_c\hat{\mathbf{Z}}_c^{\mathrm{T}}(\hat{\mathbf{Z}}_c\hat{\mathbf{Z}}_c^{\mathrm{T}})^{-1} \tag{T7}$$

Using $\hat{\mathbf{C}}$ and the estimates of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ the objective function can be calculated using equation T8.

$$S_2^2 = trace\left((\mathbf{X}_c - \hat{\mathbf{X}}_c)(\mathbf{X}_c - \hat{\mathbf{X}}_c)^{\mathrm{T}}\right) \tag{T8}$$

5. Calculate the convergence parameter λ.

$$\lambda = (S_1^2 - S_2^2)/S_2^2 \tag{T9}$$

If λ is less than the convergence limit (typically 10^{-8} in this work), terminate. Otherwise make $S_1^2 = S_2^2$ and return to step 2.

GRAM. The loading in the last mode can be calculated using equations 1.17, 1.18 or 1.19, depending on which modes were used to obtain the pseudo-slices.

It is important to note that several other algorithms have been proposed in the literature. Many of them claim to be insensitive to over-factorization and faster than the standard methods [38-43]. However, it has been demonstrated that they are suboptimal from a statistical and structural point of view, since they lack an objective function and also need more stringent conditions to achieve uniqueness [44]. On the other hand, Positive Matrix Factorization applied to three-way data (PMF3) [8] is statistically correct, but its Gauss-Newton optimization is computationally involved, since the memory requirement of the algorithm increases quickly with the size of the problem, preventing the algorithm from being applicable to large data sets in practice.

1.4 Statistical models

Thus far, the focus of this discussion has been the analytical measurements obtained in multivariate/multi-way analysis and the structural model that best represents those measurements. However, the analytical measurements are always disturbed by uncontrollable variation, which obscures the measurement of the signal of interest (*e.g.* the true signal described by the structural model). In general, this noise may be caused by everything from the sampling process to the instrumental acquisition. The analytical signal will consist of two inseparable parts: a pure or undistorted true signal and a corrupting part called noise. This signal is measured as a function of some other ordinal variable such as wavelength, or time. For instance, it is helpful if we imagine a discretely sampled signal vector, $\mathbf{x}$ (e.g., a spectrum or chromatogram), from which we could break the signal in two parts: the pure signal component, $\mathbf{x}^{\circ}$ and the noise, $\mathbf{e}$, as shown below in equation 1.20, and depicted in Figure 1.4.

$$\mathbf{x} = \mathbf{x}^{\circ} + \mathbf{e} \tag{1.20}$$

The estimation of the parameter defining the previously introduced structural model for the pure data, generically symbolized by $\mathbf{x}^{\circ}$, can be carried out by applying a variety of numerical algorithms to the experimental data obtained from instrumental

sources. However, the optimality of these methods will be guaranteed only if the measurement errors fulfill certain conditions. The term, **e**, in equation 1.20 is a random

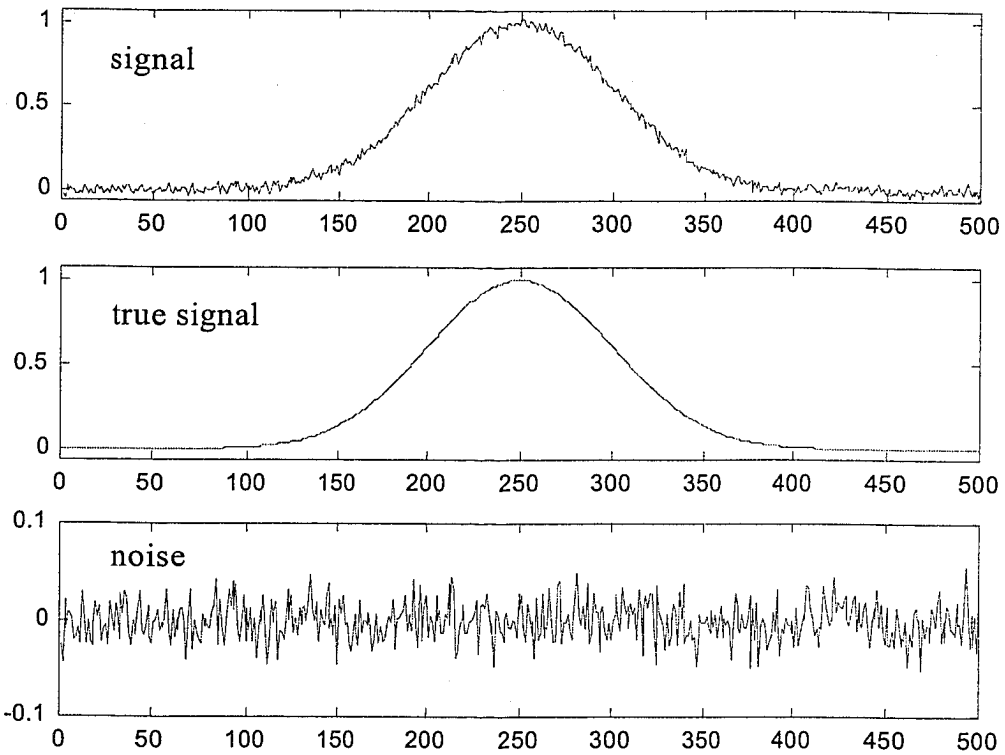

Figure 1.4. An observed signal, which can be envisioned as the contribution of two parts: the true signal and the noise.

variable and by definition cannot be predicted, only characterized. There are different methodologies and figures of merit used to characterize the error. Nonetheless, the focus of this work will be the error covariance matrix, since it is not only the most simple and complete way to describe this term, but it is also the one usually utilized to carry this information into the estimation process.

1.4.1 Error covariance matrix

The error covariance matrices are composed of two closely related statistical quantities, the variance and covariance terms. For the first element in the noise vector, e_1 ($\mathbf{e} = [e_1 \quad e_2 \quad \cdots \quad e_n]$), the measurement error variance is defined by:

$$\sigma_1^2 = E(e_1^2) \tag{1.21}$$

where $E(\ )$ indicates the expectation value of the quantity in brackets. In practice this, theoretical value cannot be obtained and the variance is estimated in the usual way:

$$s_1^2 = \frac{\sum_{r=1}^{R} (e_{1,r})^2}{R-1} \tag{1.22}$$

where R is the number of replicate measurements used in the estimate. This number is a quantitative estimate of the magnitude of the error variance at element one, but it says nothing about the relation of the errors at element one to errors at another element. This relationship can be quantified by calculating the error covariance, given for elements 1 and 2:

$$\sigma_{12} = E(e_1 e_2) \tag{1.23}$$

As for the variance, the covariance between elements 1 and 2 is estimated using:

$$s_{12} = \frac{\sum_{r=1}^{R} (e_{1,r})(e_{2,r})}{R-1} \tag{1.24}$$

where the summation product now includes the errors at different channels (1 and 2). The expectation of the error covariance term is positive when the errors at channels i and j are correlated, negative when the errors are anticorrelated, and zero when the errors are independent of one another. The calculation of variances, and error covariances for every channel in a signal vector, then, allows one to map the structure of the variations in the measurement errors, and how they are correlated between channels. This structure is conveniently summarized in an error covariance matrix, Ψ, a mapping of the variance and covariance of the measurement errors, which has the general form for an n x n matrix as depicted in equation 1.25.

$$\Psi = \begin{bmatrix} s_1^2 & s_{12} & s_{13} & \cdots & s_{1n} \\ s_{21} & s_2^2 & s_{23} & \cdots & s_{2n} \\ s_{31} & s_{32} & s_3^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & \cdots & s_n^2 \end{bmatrix} \tag{1.25}$$

Since $s_{ij} = s_{ji}$, error covariance matrices are necessarily symmetric.

For *iid* noise, the error covariance terms should approach zero in the expectation, and if the noise is identically distributed, then all of the diagonal terms (variances) should be approximately the same. This can be equivalently expressed using all of the elements of the n-element vector of errors, e,

$$\mathbf{\Psi} = E\left(\mathbf{e}\mathbf{e}^{\mathrm{T}}\right) = \sigma^2 \cdot \mathbf{I}_{\mathrm{n}} \tag{1.26}$$

This structure arises because, in the *iid* case, the expectations of the individual variances, and covariances are

$$E\left(e_i^2\right) = E\left(e_i e_i\right) = \sigma^2 \tag{1.27}$$

$$E\left(e_i e_j\right) = \sigma_{ij} = 0 \tag{1.28}$$

Therefore, under *iid* noise conditions, the error covariance matrix should be diagonal. Deviations from the *iid* condition have easily recognizable influences on the error covariance matrix. The loss of the independence condition corresponds to error covariance terms being significantly different from zero, and Ψ deviates from the diagonal form of equation 1.26. Heteroscedasticity (unequal variance for different elements) is characterized by unequal diagonal elements in Ψ. Figure 1.5 shows some examples of error covariance matrices estimated from 100 replicates. Figures 1.5a and b correspond to error covariance matrices of measurements affected by *iid*, and non-*iid* noise respectively.

Thus far, the structure of the error covariance matrix for two extreme scenarios has been introduced. The simplest case characterizes scenarios where the noise is *iid* and

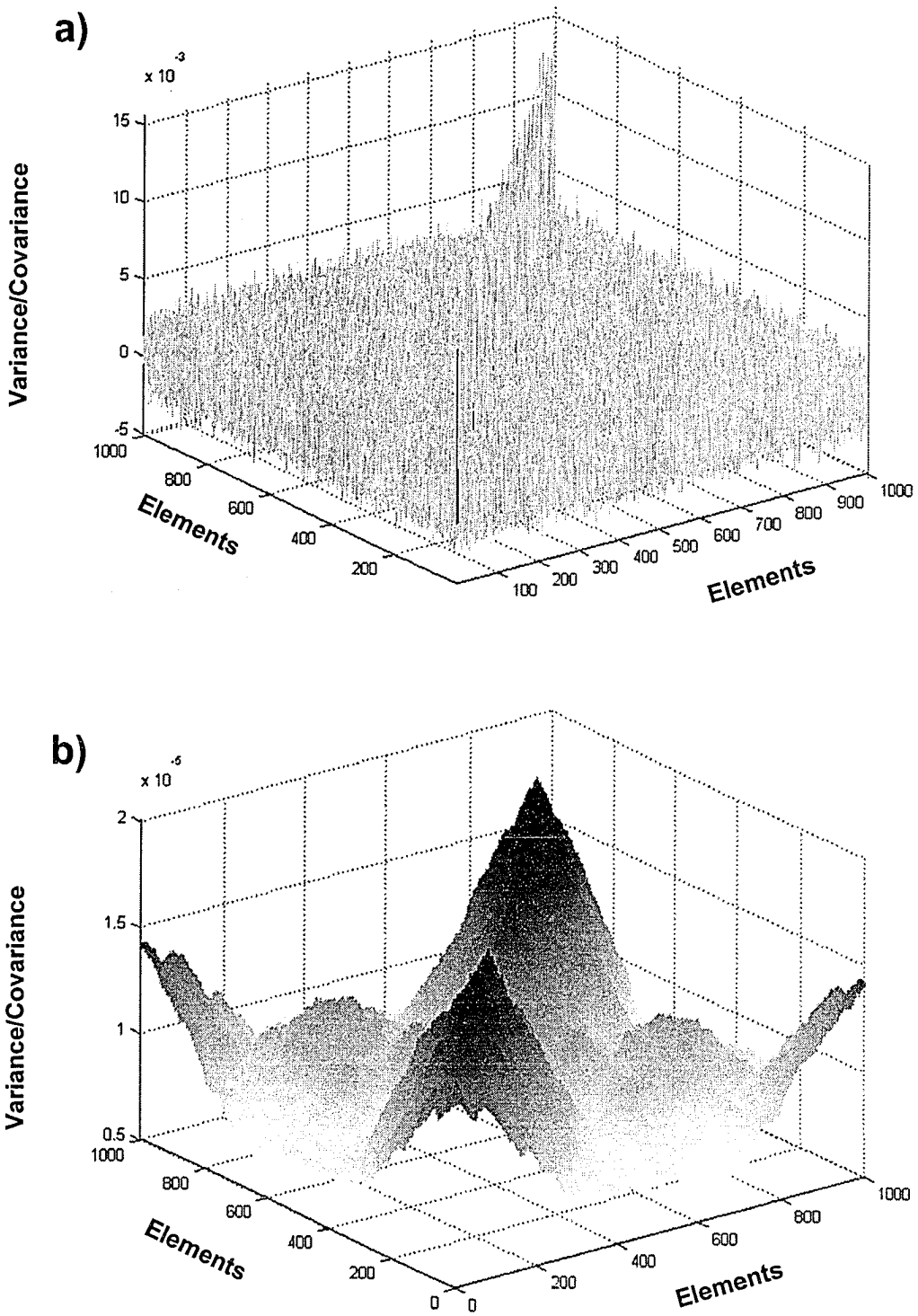

Figure 1.5. Error covariance matrices for measurement error, which follow the a) *iid* condition and the b) non – *iid* condition.

is numerically represented by only one element, σ^2. On the other hand, in some cases the error structure is so pervasive and unequal among the elements that a full error covariance matrix Ψ (n x n) in needed. A number of more "grey" scenarios exist in between these extremes. For instance, correlations may only exist among measurements within one order of the data matrix. This is usually the case, for example, when measurements from absorption spectrometers are employed, since errors in adjacent channels are correlated by source flicker noise or cell positioning errors, but there are no errors from sample to sample. Similarly, for thee-way data obtained by pairing emission-excitation profiles for different sample compositions, we might expect measurement error correlation to exist over the entire excitation-emission matrix (EEM). Other cases, simpler and more complex, can be envisioned, producing all kinds of simplifications to the full error covariance matrix due to the inherent symmetry. Therefore, the information forming the full error covariance matrix can be equivalently conveyed by smaller matrices carrying the same information in some cases. A more extensive discussion about the error structure and the error covariance matrix can be found in reference 46.

1.4.2 Measurement error structure and model estimation

The numerical methods previously introduced to estimate structural models such as PCA and PARAFAC can be applied to data with any error structure. However, the optimal estimates of the true parameters, which are in essence, the maximum likelihood parameter estimates, will be only obtained if the measurement errors corrupting the data are *iid* and normal. As mentioned above, the *iid* error assumption is largely a remnant of the age of univariate calibration. The appearance of multichannel instruments leads to measurement errors that can be correlated in several different ways in a data matrix **X**. There may be correlations among channels in the calibration spectra (columns in the matrix), for instance. This is often the prevailing scenario in spectroscopic calibration, and can arise from a variety of conditions including sensor spatial correlations in the instrument, source flicker, and numerous signal-processing techniques. Another example can be samples that are not run in random order, in which case low-frequency variations in the instrument performance (e.g., temperature drift, source degradation) can become embedded in the measurement error structure. The presence of non-*iid* noise will

seriously hamper the estimation process, since most of the methods used assume these characteristics. The graphical example depicted in Figure 1.6, can be used to illustrate the problem and the proposed solution.

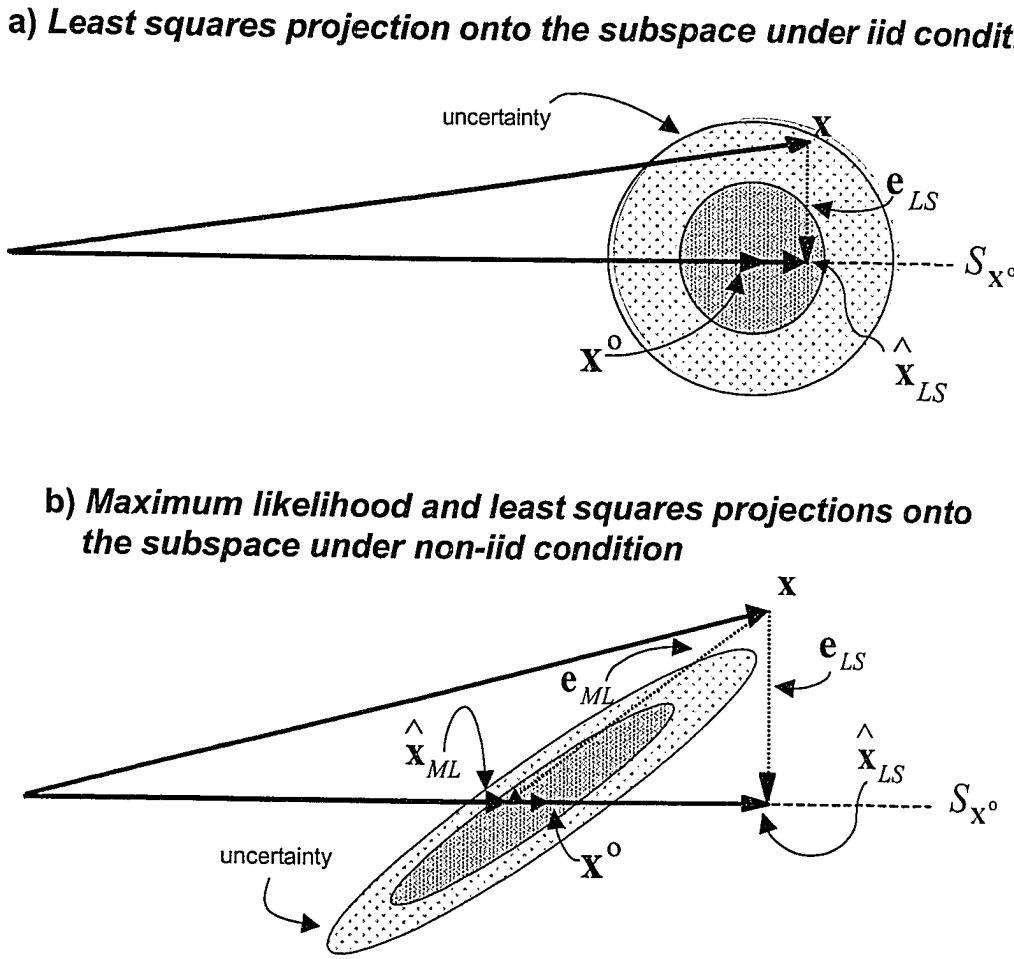

Figure 1.6. Maximum likelihood estimation approaches for different noise structures: a) least squares projection under *iid* conditions, and b) maximum likelihood projection and least squares projection when the *iid* conditions are violated.

Figure 1.6a shows the simple scenario of a noise-corrupted bivariate vector, $\mathbf{x}$, related to the unattainable true vector, $\mathbf{x}^{\circ}$. The subspace $S_{\mathbf{x}^{\circ}}$ where the true vector resides is assumed known. The best representation of the vector $\mathbf{x}$ in the subspace $S_{\mathbf{x}^{\circ}}$ is obtained by the orthogonal projection of $\mathbf{x}$ onto the subspace $S_{\mathbf{x}^{\circ}}$. This orthogonal (least-

squares) projection minimizes the length of the error between $\mathbf{x}$ and $\hat{\mathbf{x}}$ ($\hat{\mathbf{x}}$ is the projection of $\mathbf{x}$ onto $S_{\mathbf{x}^\circ}$). On the other hand, Figure 1.6b shows a case where the errors are non-*iid*. In this case, the sphere shown in the previous figure is stretched along the coordinate axes in some fashion. By considering the shape of this ellipse, it can be said that the data are corrupted by some heteroscedaticity and correlation, since the principal axes of the ellipse have different radii and are not aligned with the coordinate axes. If a standard method is used, the best estimate will be obtained by orthogonally projecting the vector onto the $S_{\mathbf{x}^\circ}$ subspace. As seen here, this vector is very distant from the true vector $\mathbf{x}^\circ$, therefore it is an unlikely estimate. Rather, if the error covariance matrix is utilized to provide a directional guide for the projection of the vector $\mathbf{x}$ onto the subspace $S_{\mathbf{x}^\circ}$, this oblique projection yields an estimate that is very close to the true vector $\mathbf{x}^\circ$.

In order to translate this geometrical example mathematically, it is important to realize that a measurement error vector can be described by a multivariate probability distribution. In practice, multivariate probability density functions are difficult to obtain but, as previously mentioned, the normal condition is seldom violated and, therefore, a multivariate normal distribution is usually assumed. For the vector of measurement errors, $\mathbf{e}$ from the previous example, this is shown as:

$$\begin{aligned} PDF(\mathbf{x}) &= \frac{1}{2\pi|\boldsymbol{\Psi}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\boldsymbol{\Psi}^{-1}(\mathbf{x}-\hat{\mathbf{x}})\right] \\ &= \frac{1}{2\pi|\boldsymbol{\Psi}|^{1/2}} \exp\left[-\frac{1}{2}\mathbf{e}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\mathbf{e}\right] \end{aligned} \tag{1.29}$$

In maximum likelihood estimation, the error ellipsoid, $\boldsymbol{\Psi}$, is associated with some "true" point which lies on the hyperplane described by a trial model. Since the actual error-free measurement is not known, a best guess is required, and this is the maximum likelihood estimate for the point. For the general case, the maximum likelihood estimate is obtained by finding the point in the subspace where the measured value maximizes the multivariate probability density function for that measurement. Thus, the maximum likelihood estimate of $\mathbf{x}$ is the one for which the observed measurement is "most

probable". The joint probability density function is called the likelihood function. The objective of this method is to maximize its value with respect to $\hat{\mathbf{x}}$ (although this method usually maximizes with respect to a parametric model of $\hat{\mathbf{x}}$). Alternatively, we could minimize the constant part of the logarithm of this function likelihood represented in equation 1.29:

$$S^2 = (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}} \boldsymbol{\Psi}^{-1} (\mathbf{x} - \hat{\mathbf{x}}) \tag{1.30}$$

This is the goodness-of-fit objective function. In this equation $\mathbf{x}$ and $\hat{\mathbf{x}}$ represent the observed and the maximum likelihood estimates of the measurement vector. In fact, this is a sum of squared residuals for all measurements weighted by the appropriate error covariance matrix. For *iid* measurement errors, this function can be reduced to the least-squares objective function, represented in equation 1.31, and is usually mimimized by methods such as PCA and PARAFAC:

$$S^2 = \frac{1}{\sigma^2} (\mathbf{x} - \hat{\mathbf{x}})^{\mathrm{T}} (\mathbf{x} - \hat{\mathbf{x}}) \tag{1.31}$$

It is important to note that the use of a vector to formulate the statistical model in a generic way is not introduced only for the sake of simplicity, but also because it provides a more general way in which this problem can be treated, since the errors for a given analytical experiment can be correlated along any order. In many cases, this general scenario, which can be prohibitive for practical cases, can be alleviated by equivalent but more compact representations, as in the case shown in equation 1.31 for the *iid* case. Additional simplifications are possible by using equivalent representations of the structural models in conjunction with alternative optimal representation of the error covariance matrix. For example, in kinetic studies, the course of the reaction is followed spectroscopically, giving rise to errors that are correlated in both the time and wavelength modes. The other mode may be composed of samples, but the structure of this correlated noise changes from sample to sample independently. This type of situation is not uncommon when spectroscopic techniques such as NIR spectroscopy are used due to path length variations. Mathematically, the trilinear errors-in-variable model best suited to describe these data can be obtained by minimizing equation 1.32:

$$S^2 = \sum_{i=1}^{I} ({}^i\mathbf{x}_a - {}^i\tilde{\mathbf{a}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})\,{}^i\boldsymbol{\Psi}_a^{-1}({}^i\mathbf{x}_a - {}^i\tilde{\mathbf{a}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}} \tag{1.32}$$

For this case, we will consider that the three-way data $\underline{\mathbf{X}}$, will be unfolded preserving the samples of different compositions in mode A, while modes B and C will be combined in one composite mode, formed by the spectral information and the time information for each sample. However, if the sample contribution to the error structure can be alleviated somehow, the objective function in this case can be expressed as shown in equation 1.33.

$$S^2 = trace[(\mathbf{X}_a - \mathbf{A}\mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a - \mathbf{A}\mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}}] \tag{1.33}$$

The philosophy introduced above has been successfully used by MLPCA to treat bilinear data in a number of analytical scenarios [9, 14, 45, 47, 48]. The implementation details of MLPCA will not be given here as they can be found elsewhere [17,49]. For the case of three-way data, a generic method called Maximum Likelihood via Iterative Least Squares Estimation (MILES) was introduced by Bro *et al.* [12]. 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. The method is based on a majorization strategy in which the original objective function is substituted by a simpler and equivalent objective function in each step of the estimation process. Unfortunately, the simplicity of this numerical implementation is hindered by the amount of computation time needed. Since the method runs the full least squares optimization in each step, the time needed to obtain an estimate is sometimes excessive. Another important drawback of this approach is that the physical problem becomes obscured by the efficient but unfamiliar numerical methodology.

The present work introduces the theoretical foundations for a maximum likelihood implementation of the PARAFAC model. The method will be called Maximum Likelihood Parallel Factor Analysis (MLPARAFAC) and is an errors-in-variables modeling method in that it accounts for measurement errors in the estimation of model parameters. It is an optimal modeling method in a maximum likelihood sense for functional models with no errors in the model equations. The present method is a natural extension to PARAFAC of the MLPCA method introduced by Wentzell *et al.* [9]. The mathematical aspects of the algorithm will be described in detail to allow the principles to

be readily applied. The algorithm can accommodate heteroscedastic and correlated noise in one or more dimensions and has excellent convergence characteristics since its core is based on an ALS framework.

1.5 References

1. Wold S. *Kem. Tidskr.* 1972; **3**: 34.
2. Sheperd P. *J. Chemom.* 1987; **1**: 3.
3. Lawton W.H. and Sylvestre E.A. *Technometrics* 1971; **13**: 617.
4. Laeven J.M. and Smith C. *J. Chromatogr.* 1985; **176**: 77.
5. Martens H. and Næs T. *Multivariate Calibration*; John Wiley & Sons; New York, 1989.
6. Paatero P. and Tapper U. *Chemom. Intell. Lab. Syst.* 1993; **18**: 183.
7. Paatero P. and Tapper U. *Environmetrics* 1994; **5**: 111.
8. Paatero P. *Chemom. Intell. Lab. Syst.* 1997; **38**: 223.
9. Wentzell P.D., Andrews D.T., Hamilton D.C., Faber K. and Kowalski B.R. *J. Chemom.* 1997; **11**: 339
10. Van Huffel S. and Vanderwalle J. *The Total Least Squares Problem: Computational Aspect and Analysis*; SIAM; Philadelphia, 1991.
11. Martens H, Høy M., Wise B.M., Bro R. and Brockhoff P.M.B *J. Chemom.* 2003, **17**: 153.
12. Bro R., Sidiropoulos N.D. and Smilde A.K. *J. Chemom.* 2002; **16**: 387
13. Arnold M.A., Burmeister J.J. and Small G.W. *Anal. Chem.* 1998; **70**: 1773.
14. Wentzell P.D., Andrews D.T. and Kowalski B.R. *Anal. Chem.* 1997; **69**: 2299.
15. de Juan A. and Tauler R. *Anal. Chim. Acta* 2003; **500**: 195.
16. Tauler R., Smilde A.K., and Kowalski B.R. *J. Chemom.* 1995; **9**: 31.
17. Bezemer E. and Rutan S.C. *Anal. Chem.* 2001; **73**: 4403.
18. Kalivas J.H. *J. Chemom.* 1999; **13**: 111.
19. Adcock R.J. *The Analyst* 1878; **5**: 53.

20. Harshman R.A. *UCLA Working Papers in Phonetics* 1970; **16**: 1.

21. Carroll J.D. and Chang *J. Psychometrika* 1970; **35**: 283.

22. Kroonenberg P.M. and de Leeuw J. *Psychometrika* 1980; **45**: 69.

23. Sidiropoulos N.D. and Bro R. *J. Chemom.* 2000; **14**: 229.

24. Kiers H.A.L., ten Berge J.M.F. and Rocci R. *Psychometrika* 1997; **62**: 349.

25. Brouwer P., Kroonenberg P.M. *J. Classification* 1991; **8**: 93.

26. Kiers H.A.L. *Statistica Applicata* 1992; **4**: 659.

27. Harshman R.A. *UCLA Working Papers in Phonetics* 1972; **22**: 111.

28. Leurgans S., Ross R.T. and Abel R.B. *J. Matrix Anal. Appl.* 1993; **14**: 1064.

29. Kruskal J.B. *Linear algebra and its applications* 1977; **18**: 95.

30. Kruskal J.B., Rank, decomposition, and uniqueness for 3-way and N-way arrays, In: Multiway data analysis, (Eds. R. Coppi, S. Bolasco), Elsevier Science Pub. (North-Holland), 1989, 7.

31. Harshman R.A. and Lundy M.E., Data preprocessing and the extended PARAFAC model, In: Research methods for multimode data analysis, (Eds. H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald) Praeger, New York, 1984, 216.

32. de Leew J. Block relaxation algorithms in statistics, In: Information systems and data analysis, (Eds. H.H. Bock, W. Lenski, M.M Ritcher) Spinger, New York, 1994, 308.

33. Sanchez E. and Kowalski B.R. *J. Chemom.* 1990; **4**: 29.

34. Ho C., Christian G.D. and Davidson E.R. *Anal. Chem.* 1978; **50**: 1108.

35. Lorber A. *Anal. Chim. Acta* 1984; **164**: 293.

36. Lorber A. *Anal. Chem.* 1985; **57**: 2395.

37. Sanchez E. and Kowalski B.R. *Anal. Chem.* 1986; **58**: 496.

38. Wu H.L., Shibukawa M. and Oguma K. *Anal. Sci.* 1997; **13**: 53.

39. Chen Z.P., Wu H.L., Jiang J.H., Li Y. and Yu R.Q. *Chemom. Intell. Lab. Syst.* 2000; **52**: 75.

40. Chen Z.P, Li Y. and Yu R.Q. *J. Chemom.* 2001; **15**: 149.

41. Jiang J.H., Wu H.L., Li Y. and Yu R.Q. *J. Chemom.* 1999; **13**: 557.

42. Jiang J.H., Wu H.L., Li Y. and Yu R.Q. *J. Chemom.* 2000; **14**: 15.

43. Li Y., Jiang J.H., Wu H.L., Chen Z.P. and Yu R.Q. *Chemom. Intell. Lab. Syst.* 2000; **52**: 33.

44. Faber N.K.M., Bro R. Hopke P.K. *Chemom. Intell. Lab. Syst.* 2003; **65**: 119.

45. Schreyer S., Bidinosti M. and Wentzell P.D. *Appl. Spectrosc.* 2002; **56**: 789.

46. Leger M., Vega-Montoto L. and Wentzell P.D. *Chemom. Intell. Lab. Syst.* 2004; **In press**.

47. Wentzell P.D. and Lohnes M.T. *Chemom. Intell. Lab. Syst.* 1999; **45**: 65.

48. Andrews D.T. and Wentzell P.D. *Anal. Chim. Acta* 1997; **350**: 341

Chapter 2

Maximum Likelihood Parallel Factor Analysis (MLPARAFAC)[1]

2.1 Abstract

Algorithms for carrying out maximum likelihood parallel factor analysis (MLPARAFAC) for three-way data are described. These algorithms are based on the principle of alternating least squares, but differ from conventional PARAFAC algorithms in that they incorporate measurement error information into the trilinear decomposition. This information is represented in the form of an error covariance matrix. Four algorithms are discussed for dealing with different error structures in the three-way array. The simplest of these treats measurements with non-uniform measurement noise which is uncorrelated. The most general algorithm can analyze data with any type of noise correlation structure. The other two algorithms are simplifications of the general algorithm which can be applied with greater efficiency to cases where the noise is correlated only along one mode of the three-way array. Simulation studies carried out under a variety of measurement error conditions were used for statistical validation of the maximum likelihood properties of the algorithms. The MLPARAFAC methods are also shown to produce more accurate results than PARAFAC under a variety of conditions.

---

[1] Submitted to *Journal of Chemometrics* (Published, April 2003)

2.2 Introduction

With advancing technology of analytical instrumentation, data in the form of tensors of second order and higher have become more commonplace. Examples of such techniques include fluorescence excitation-emission spectroscopy and chromatography with multichannel detectors. In 1980, Hirschfeld [1] provided a very complete table of all the feasible combinations of techniques capable of providing second order data at that time and estimated that about 60% of the techniques are bilinear under certain conditions. Extension to trilinear data is easily accomplished when several samples are analyzed by these methods. This list has continued to expand in terms of the number of techniques and possible analytical orders as this instrumentation becomes commonplace in chemistry laboratories. Ever since Appellof and Davidson [2] provided the first application of trilinear decomposition to chemistry using both simulated and real LC/emission/excitation fluorescence data, the number of applications has expanded to many branches of chemistry, ranging from basic research to environmental and food chemistry.

Trilinear data (and multilinear tensors in general) share common properties with bilinear data that make the latter structure central to modern chemometrics. Both types of data can model deterministic relationships among variables, especially in cases where a high degree of collinearity exists. These types of models allow multivariate and multiorder data to be represented by a smaller number of variables. Using this smaller set of variables, the data can be described within experimental error as a P-dimensional hyperplane. In this case, P is called the chemical rank or "true" rank of the data set in order to distinguish it from the mathematical rank. In general, the chemical rank is typically related to the number of underlying chemical factors or chemical components present in the mixture. However, contrary to what happens in bilinear models, where the smaller set of variables are abstract solutions of the underlying physical factors which are not unique due to rotational ambiguities, the trilinear and higher multilinear models can produce unique and well-identified solutions (up to trivial differences in factor order and relative scaling across modes) [3]. In addition, the uniqueness of the solution gives rise

to the "second order advantage" which allows the quantitation of an analyte in the presence of interferences with only one calibration sample.

A variety of algorithms have been developed to estimate the multilinear model, including parallel factor analysis [4] (PARAFAC), direct trilinear decomposition [5] (DTLD), and positive matrix factorization [6] (PMF3). These algorithms are based on different numerical approaches, namely alternating least squares (ALS), eigenproblem formulation and a Gauss–Newton approach, respectively. Each has its own advantages and disadvantages that make it suitable in a specific situation. However, PARAFAC (ALS) is currently the most widely used algorithm, mainly due to its good convergence properties. ALS, which was introduced by Yates [7] in 1933, works by simply dividing the parameters into several sets. Each set of parameters is estimated in a least squares sense conditionally on the remaining parameters. The estimation of the parameters is repeated iteratively until a certain stop criterion is reached. In this way, a very complex nonlinear problem becomes a sequence of simpler least squares steps in which the parameter sets are improved in each step. As all estimates of parameters are least squares estimates, the procedure can only improve the fit or keep it the same if converged. It follows from this that the objective function decreases monotonically, and, since it is also bounded from below (the objective function cannot be less than zero), convergence is always reached. This does not imply that the global minimum is guaranteed, since a problem like this is characterized by several local minima. Global convergence can be assessed when repetitions using different starting points yield similar sets of parameters. In addition to the reliable convergence characteristics of ALS, it is also used because it yields maximum likelihood estimation under certain noise characteristics.

Methods such as PARAFAC give maximum likelihood estimates of the model parameters when the noise is independently and identically distributed with a normal distribution (*iid* normal). Noise can be broadly defined as an undesirable variation in a measured signal which obscures the measurement of interest, the true signal. Based on the specific advantages of multilinear data, this definition will be narrowed to undesirable variation attributable to non-chemical sources (*e.g.* instrumental sources). Noise can have many different origins, having a very complex range of properties and characteristics. Unfortunately, these properties and characteristics are not mutually

exclusive making the number of possibilities of noise structures very large. The term *iid* has been coined in the chemometric literature to make a precise and concise description of the fundamental properties needed to characterize the instrumental noise in the "ideal" case. It conveys information about independence (*i.e.* the error observed at any one channel is uncorrelated with the error observed at any other channel) and the homogeneity of distributions (*i.e.* identically distributed implies the error variance and distribution are the same for all measurements). Conventional least squares approaches to trilinear decomposition are maximum likelihood methods only under *iid* conditions. These naïve assumptions about the noise structure corrupting the multilinear data can lead to poor models, since all of the methods rely on a least squares procedure. PARAFAC and DTLD are the most affected since both independence and homoscedasticity (identical distributions) need to be satisfied to yield the maximum likelihood solution. PMF3 can overcome the need for homoscedastic noise to yield the maximum likelihood solution because it applies a weighting scheme that solves this impediment. When minor variations from the assumption of *iid* normal noise are observed, some scaling techniques can be used with PARAFAC in order to alleviate the deviations from the *iid* condition, but this will only yield a maximum likelihood solution when the noise is uncorrelated and the heteroscedasticity follows a certain structure. A more general approach to tackle this problem, W-PARAFAC, was introduced in 1997 by Kiers [8], who used a weighted objective function to remedy the problem of heteroscedastic noise. The algorithm is based in a majorization procedure instead of an ALS algorithm. W-PARAFAC and PMF3 both overcome the heteroscedasticity of the noise using a weighted objective function, but the issue of the noise correlation is still a problem for both methods, since they cannot accommodate error covariance terms in the procedure.

The presence of covariance among measurement errors is an ubiquitous and pernicious effect produced by several sources ranging from the temporal correlation of pump noise in chromatography to the spatial correlation of array detectors in spectroscopy. Another important source of correlation in the measurement errors is signal processing, particularly electronic or digital smoothing filters. Because of all of these effects, correlated measurement errors are likely to be the rule rather than the exception

for multivariate data sets, implying that standard methods of analysis (both two-way and multiway) that make assumptions of *iid* normal noise are suboptimal. The only optimal means to account for the correlation in measurement errors is using a maximum likelihood approach to estimate model parameters that are most likely to give rise to the observed measurements. For bilinear data, this problem has been addressed through the development of maximum likelihood principal component analysis (MLPCA) [9], which has been shown to provide improved results where the effects of noise correlation are significant.

Correlation among measurement errors in three-way data is complicated by the unfolding/matricization process usually used in ALS algorithms. Elements with correlated measurement errors which may appear adjacent to one another in a "slice" of the three-way array may become spatially separated from one another when the cube is unfolded in certain ways. Because of this, conceptualization and simplification of error covariance structures for three-way data is more difficult, and this has impeded the development of maximum likelihood methods for three-way data. Until recently, this problem was avoided by the standard estimation algorithms. Recently, a method called MILES [10], which is based on a majorization-ALS algorithm, was introduced to address the problem of correlated measurement errors for multilinear data. The extent to which this method yields maximum likelihood estimates is unclear since no validation of the results was done in this context and the theoretical foundation of the method is obscured by the complexity of the algorithm.

This paper introduces the theoretical foundations for Maximum Likelihood Parallel Factor Analysis (MLPARAFAC). MLPARAFAC is an errors-in-variables modeling method in that it accounts for measurement errors in the estimation of model parameters. It is an optimal modeling method in a maximum likelihood sense for functional models with no errors in the model equations. The present method is a natural extension to PARAFAC of the MLPCA method introduced by Wentzell *et al.* [9]. The mathematical aspects of the algorithm are described in detail to allow the principles to be readily applied. The algorithm can accommodate heteroscedastic and correlated noise in one or more dimensions and has excellent convergence characteristics because its core is

based on an alternating least squares procedure. Although, all the cases used in this paper will be three-way data this algorithm is extensible to N-way data.

2.2.1 Notation

In this paper, scalars are indicated by italics and vectors by bold lower-case characters. Bold upper-case letters are used for two-way matrices and underlined bold upper-case letters for three-way data. The letters A, B ,C and *I*, *J*, *K* are reserved for indicating the first, second and third mode of three-way data and the dimensions of those modes respectively. Also the letter *P* is reserved to represent the number of factors used in the model. The terms mode, way and order are used indistinctively, as well as the terms factors and components. When three way arrays are unfolded to matrices, the following notation will be used. If $\underline{\mathbf{X}}$ (*I* x *J* x *K*) is unfolded while retaining the first order to produce a (*I* x *JK*) matrix, this will be designated $\mathbf{X}_a$. In the same way, matrices $\mathbf{X}_b$ (*J* x *IK*) and $\mathbf{X}_c$ (*K* x *IJ*) will be used to represent unfolded matrices which retain the second and the third orders, respectively. In general, other matrices with subscripts *a*, *b* and *c* represent unfolding while retaining the first, second and third modes. The symbol "⊗" will be used primarily to indicate the Kronecker product, but will also be used to represent the tensor product in certain cases which will be clearly distinguished.

2.3 Theory

PARAFAC is an acronym used to refer to two different, though closely related concepts. It is used to describe the model that the trilinear structure of the data follows, and it is also used to refer to one of the various algorithms used to estimate the parameters of the aforementioned model. PARAFAC was originally introduced by Harshman [4] and simultaneously and independently by Carroll and Chang [11], who referred to it as Canonical Decomposition (CANDECOMP). The model can be seen as an extension of bilinear PCA to higher orders. The PARAFAC model for a three-way array is defined by three loading matrices, **A**, **B** and **C**, with elements a_{ip}, b_{jp} and c_{kp}. It can be written as a tensor product, as shown in equation 2.1:

$$\underline{\mathbf{X}} = \sum_{p=1}^{P} \mathbf{a}_p \otimes \mathbf{b}_p \otimes \mathbf{c}_p + \underline{\mathbf{E}} \qquad (2.1)$$

where $\mathbf{a}_p$, $\mathbf{b}_p$ and $\mathbf{c}_p$ are the pth columns of the loading matrices **A**, **B** and **C**, respectively. The model can also be expressed in scalar form as shown in equation 2.2:

$$x_{ijk} = \sum_{p=1}^{P} a_{ip} b_{jp} c_{kp} + e_{ijk} \tag{2.2}$$

Here, x_{ijk} is an element of the three-way array $\underline{\mathbf{X}}$, and e_{ijk} is an element of the corresponding residual matrix, $\underline{\mathbf{E}}$, where the indices refer to modes A, B, and C, respectively.

Since most of the mathematical/statistical tools and concepts used in chemometrics rely on the foundations of linear algebra, a matrix representation of a three-way array is very useful. The process of converting a cube or higher order arrangement of data into a matrix is called unfolding or matricization and it can be done in at least as many ways as the array has orders. Equation 2.3 represents the unfolded data when the first order is retained.

$$\mathbf{X}_a = \mathbf{A}\mathbf{Z}_a + \mathbf{E}_a \tag{2.3}$$

The $\mathbf{X}_a$ matrix is obtained from the matrix multiplication of loading matrix **A** and a matrix $\mathbf{Z}_a$ which is formed from loading matrices **B** and **C**. The $\mathbf{Z}_a$ matrix can be obtained as a Khatri-Rao product [12] of matrices **B** and **C** or as a Kronecker product [13] of matrices **B** and **C** premultiplied by the unfolded superdiagonal "identity" matrix of order P ($\mathbf{I}_a$). These alternative representations are shown in equations 2.4 and 2.5, respectively.

$$\mathbf{Z}_a = (\mathbf{C}^T \,|\otimes|\, \mathbf{B}^T) \tag{2.4}$$

$$\mathbf{Z}_a = \mathbf{I}_a(\mathbf{C}^T \otimes \mathbf{B}^T) \tag{2.5}$$

Analogous equations can be used to represent $\underline{\mathbf{X}}$ as the matrices obtained when the second and third orders are retained ($\mathbf{X}_b$ and $\mathbf{X}_c$).

Assuming **B** and **C** are known (or estimated) and *iid* noise conditions, then an estimate of **A** can be obtained solving the conditional least squares problem to minimize the sum of the squares of the residuals in $\underline{\mathbf{E}}$. The solution to this problem is given by equation 2.6.

$$\hat{\mathbf{A}} = \mathbf{X}_a \hat{\mathbf{Z}}_a^T (\hat{\mathbf{Z}}_a \hat{\mathbf{Z}}_a^T)^{-1} \tag{2.6}$$

This least squares estimate of **A** can in turn be used to obtain estimates of **B** and **C** (given $\hat{\mathbf{C}}$ and $\hat{\mathbf{B}}$, respectively) by employing similar equations involving $\mathbf{X}_b$ and $\mathbf{X}_c$. This leads naturally to the iterative ALS procedure which can be used to estimate all of the loadings in a stepwise procedure.

2.3.1 Non-uniform Measurement Errors

Unfortunately, in cases where *iid* noise conditions are violated, the conventional ALS algorithm will produce suboptimal estimates of the loadings. In those cases where measurement errors remain independent but the condition of homoscedasticity is violated (*i.e.* each measurement can have a different variance), a more general objective function can be minimized to satisfy the maximum likelihood criterion. Consider the three-way array of measurements $\underline{\mathbf{X}}$ and an associated array $\underline{\mathbf{\Sigma}}$, which contains the variances of the measurements of the corresponding elements in $\underline{\mathbf{X}}$. For a given trial solution $\hat{\underline{\mathbf{X}}}$ (based on estimates of $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and $\hat{\mathbf{C}}$ such that $\hat{\mathbf{X}} = \hat{\mathbf{A}}\hat{\mathbf{Z}}_a$), equation 2.7 gives the likelihood function in terms of the matrices unfolded in the A mode.

$$L = \prod_{i=1}^{I} \frac{1}{(2\pi)^{jk/2}\left|{}^{i}\mathbf{\Psi}_a\right|^{1/2}} \exp\left[-\frac{1}{2}({}^{i}\mathbf{x}_a - {}^{i}\hat{\mathbf{x}}_a)\,{}^{i}\mathbf{\Psi}_a^{-1}({}^{i}\mathbf{x}_a - {}^{i}\hat{\mathbf{x}}_a)^{\mathrm{T}}\right] \tag{2.7}$$

In this equation, ${}^{i}\mathbf{x}_a$ represents the *i*th row of the unfolded matrix $\mathbf{X}_a$ and ${}^{i}\hat{\mathbf{x}}_a$ represent the corresponding vector of estimates of $\hat{\mathbf{X}}_a$. The matrix ${}^{i}\mathbf{\Psi}_a$ is the measurement error covariance matrix for the *i*th row of $\mathbf{X}_a$, which in the case of uncorrelated errors will be a diagonal matrix (*JK* x *JK*) containing the variance of the measurement errors of ${}^{i}\mathbf{x}_a$; that is, it is the diagonalized form of the *i*th row of $\mathbf{\Sigma}_a$. The error covariance matrix is defined according to:

$${}^{i}\mathbf{\Psi}_a = E[({}^{i}\mathbf{x}_a - {}^{i}\mathbf{x}_a^{\circ})^{\mathrm{T}} \cdot ({}^{i}\mathbf{x}_a - {}^{i}\mathbf{x}_a^{\circ})] \tag{2.8}$$

where "*E*" designates an expectation value and ${}^{i}\mathbf{x}_a^{\circ}$ represents *i*th row of $\mathbf{X}_a^{\circ}$, which is the true or expectation value of $\underline{\mathbf{X}}^{\circ}$ unfolded in the A mode. Since $\underline{\mathbf{X}}^{\circ}$ is not normally known, it is normally estimated on the basis of mean values, or else ${}^{i}\mathbf{\Psi}_a$ is estimated on the basis of prior information (*e.g.* an assumption of proportional errors).

Obtaining the maximum likelihood estimates of $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$, and $\hat{\mathbf{C}}$ means maximizing the likelihood function in equation 2.8 with respect to these loading parameters. This is equivalent to minimizing the logarithm of the likelihood function, which, when constant terms are ignored, results in the objective function in equation 2.9.

$$\begin{aligned} S^2 &= \sum_{i=1}^{I} ({}^i\mathbf{x}_a - {}^i\hat{\mathbf{x}}_a)\,{}^i\boldsymbol{\Psi}_a^{-1}({}^i\mathbf{x}_a - {}^i\hat{\mathbf{x}}_a)^{\mathrm{T}} = \sum_{i=1}^{I} S_i^2 \\ &= \sum_{i=1}^{I} ({}^i\mathbf{x}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\,{}^i\mathbf{x}_a^{\mathrm{T}} - {}^i\mathbf{x}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\,{}^i\hat{\mathbf{x}}_a^{\mathrm{T}} - {}^i\hat{\mathbf{x}}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\,{}^i\mathbf{x}_a^{\mathrm{T}} + {}^i\hat{\mathbf{x}}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\,{}^i\hat{\mathbf{x}}_a^{\mathrm{T}}) \end{aligned} \tag{2.9}$$

To minimize the objective function, S^2, with respect to the loadings $\hat{\mathbf{A}}$ given $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$, we first recognize that each term, S_i^2, in the summation is an independent function of the ith row of $\hat{\mathbf{A}}$, designated as ${}^i\hat{\mathbf{a}}$, and the given matrix $\hat{\mathbf{Z}}_a$, that is ${}^i\hat{\mathbf{x}}_a = {}^i\hat{\mathbf{a}}\hat{\mathbf{Z}}_a$. This means that S^2 can be minimized by minimizing the individual terms, allowing each row of $\mathbf{A}$ to be estimated independently as shown in equations 2.10 to 12.

$$S^2 = {}^i\mathbf{x}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\,{}^i\mathbf{x}_a^{\mathrm{T}} - {}^i\mathbf{x}_a\,{}^i\boldsymbol{\Psi}_a^{-1}(\hat{\mathbf{a}}\mathbf{Z}_a)^{\mathrm{T}} - \hat{\mathbf{a}}\mathbf{Z}_a\,{}^i\boldsymbol{\Psi}_a^{-1\,i}\,\mathbf{x}_a^{\mathrm{T}} + \hat{\mathbf{a}}\mathbf{Z}_a\,{}^i\boldsymbol{\Psi}_a^{-1}(\hat{\mathbf{a}}\mathbf{Z}_a)^{\mathrm{T}} \tag{2.10}$$

$$\frac{\partial S^2}{\partial \hat{\mathbf{a}}} = 0 - {}^i\mathbf{x}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}} - \mathbf{Z}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\,{}^i\mathbf{x}_a^{\mathrm{T}} + 2\mathbf{Z}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}}\,\hat{\mathbf{a}}^{\mathrm{T}} \tag{2.11}$$

$$\hat{\mathbf{a}} = {}^i\mathbf{x}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}}(\mathbf{Z}_a\,{}^i\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}})^{-1} \tag{2.12}$$

It should be emphasized that, in these equations, ${}^i\hat{\mathbf{a}}$ is used to designate a row of $\hat{\mathbf{A}}$ and does not represent a loading vector of $\hat{\mathbf{A}}$. From equation 2.12, estimates of ${}^i\hat{\mathbf{a}}$ can be combined to give $\hat{\mathbf{A}}$. In cases where the error covariance matrix is the same for all the rows of $\mathbf{X}_a$, equation 2.12 can be generalized to the matrix form represented in equation 2.13.

$$\hat{\mathbf{A}} = \mathbf{X}_a\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}}(\mathbf{Z}_a\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}})^{-1} \tag{2.13}$$

This equation can also be reduced easily to equation 2.6 in cases where the noise is the same (homoscedastic) for all the channels.

Since the requirement for this development was independence of measurement errors, the error covariance matrices for all the orders are diagonals. Unfolding $\underline{\mathbf{X}}$ in the

other two directions leads to similar equations for $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$, allowing an equivalent maximum likelihood estimation of $\hat{\underline{\mathbf{X}}}$ in all the spaces, subject to the constraint that two of the spaces remain fixed. This occurs because the objective function of $\underline{\mathbf{X}}$ unfolded in all the orders reduces to the same summation but in a different order. To obtain the unrestricted maximum likelihood estimation of $\hat{\underline{\mathbf{X}}}$, it is necessary to optimize the objective function with respect to all three sets of loading vectors. An alternative to such a direct optimization is an iterative approach using ALS.

The algorithm for the Maximum Likelihood PARAFAC in cases of heteroscedastic noise is given in Table 2.1. The algorithm alternately uses the maximum likelihood estimates of two modes, say $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$, to update the estimates in the mode left out, say $\hat{\mathbf{A}}$. This procedure is carried out iteratively, using the previously estimated mode and one of the other two modes, say $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$, to estimate the other, say $\hat{\mathbf{B}}$. This procedure has been found to be simple, fast and reliable. Although, global convergence is not guaranteed, it does not seem to be susceptible to local minima as is the case with gradient methods. In addition, this method is very attractive since its core is based on an ALS framework, which ensures an improvement of the solution in each step. The algorithm is easily applied in cases where there are missing values by incorporating large variances for the missing measurements. Convergence time depends on the dimensionality of the data, the degree of similarity of the components forming the system, the accuracy of the initial estimates and the structure of the errors. The two most important factors increasing the convergence time are the dimension of the model and the degree of similarity, especially the former, which makes each step longer and increases the necessity for more iteration. Some strategies have been reported to improve the

Table 2.1. Standard MLPARAFAC algorithm (uncorrelated errors).

1. Given an I x J x K cube of data $\underline{\mathbf{X}}$ and a corresponding I x J x K cube $\underline{\boldsymbol{\Sigma}}$ of measurement error variances, the algorithm is initialized using random values of the correct dimensions or using estimates obtained by TLD.

$$[\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}] = tld(\underline{\mathbf{X}},P) \tag{T.1}$$

2. Unfold $\underline{\mathbf{X}}$ and $\underline{\boldsymbol{\Sigma}}$ retaining the first order and calculate the maximum likelihood estimation of $\hat{\mathbf{A}}$ conditional on $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$.

$$\mathbf{X}_a = unfold(\underline{\mathbf{X}},a);\ \boldsymbol{\Sigma}_a = unfold(\underline{\boldsymbol{\Sigma}},a);\ {}^{i}\boldsymbol{\Psi}_a = diag({}^{i}\boldsymbol{\Sigma}_a) \tag{T.2}$$

$${}^{i}\hat{\mathbf{a}}^{\mathrm{T}} = {}^{i}\mathbf{x}_a\,{}^{i}\boldsymbol{\Psi}_a^{-1}\hat{\mathbf{Z}}_a^{\mathrm{T}}(\hat{\mathbf{Z}}_a\,{}^{i}\boldsymbol{\Psi}_a^{-1}\hat{\mathbf{Z}}_a^{\mathrm{T}})^{-1} \tag{T.3}$$

Here ${}^{i}\hat{\mathbf{a}}^{\mathrm{T}}$ is a row vector of $\hat{\mathbf{A}}$. Using this estimate and the estimates of $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ the objective function can be calculated using equation T.4.

$$S_a^2 = \sum_{i=1}^{I}({}^{i}\mathbf{x}_a - {}^{i}\hat{\mathbf{x}}_a)\,{}^{i}\boldsymbol{\Psi}_a^{-1}({}^{i}\mathbf{x}_a - {}^{i}\hat{\mathbf{x}}_a)^{\mathrm{T}} \tag{T.4}$$

3. Unfold $\underline{\mathbf{X}}$ and $\underline{\boldsymbol{\Sigma}}$ retaining the second order and calculate the maximum likelihood estimation of $\hat{\mathbf{B}}$ conditional on $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$.

$$\mathbf{X}_b = unfold(\underline{\mathbf{X}},b);\ \boldsymbol{\Sigma}_b = unfold(\underline{\boldsymbol{\Sigma}},b);\ {}^{j}\boldsymbol{\Psi}_b = diag({}^{j}\boldsymbol{\Sigma}_b) \tag{T.5}$$

$${}^{j}\hat{\mathbf{b}}^{\mathrm{T}} = {}^{j}\mathbf{x}_b\,{}^{j}\boldsymbol{\Psi}_b^{-1}\hat{\mathbf{Z}}_b^{\mathrm{T}}(\hat{\mathbf{Z}}_b\,{}^{j}\boldsymbol{\Psi}_b^{-1}\hat{\mathbf{Z}}_b^{\mathrm{T}})^{-1} \tag{T.6}$$

Here ${}^{j}\hat{\mathbf{b}}^{\mathrm{T}}$ is a row vector of $\hat{\mathbf{B}}$. Using this estimate and the estimates of $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ the objective function can be calculated using equation T.7.

$$S_b^2 = \sum_{j=1}^{J}({}^{j}\mathbf{x}_b - {}^{j}\hat{\mathbf{x}}_b)\,{}^{j}\boldsymbol{\Psi}_b^{-1}({}^{j}\mathbf{x}_b - {}^{j}\hat{\mathbf{x}}_a)^{\mathrm{T}} \tag{T.7}$$

4. Unfold $\underline{\mathbf{X}}$ and $\underline{\boldsymbol{\Sigma}}$ retaining the third order and calculate the maximum likelihood estimation of $\hat{\mathbf{C}}$ conditional on $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

$$\mathbf{X}_c = unfold(\underline{\mathbf{X}},c);\ \boldsymbol{\Sigma}_c = unfold(\underline{\boldsymbol{\Sigma}},c);\ {}^{k}\boldsymbol{\Psi}_c = diag({}^{k}\boldsymbol{\Sigma}_c) \tag{T.8}$$

$${}^{k}\hat{\mathbf{c}}^{\mathrm{T}} = {}^{k}\mathbf{x}_c\,{}^{k}\boldsymbol{\Psi}_c^{-1}\hat{\mathbf{Z}}_c^{\mathrm{T}}(\hat{\mathbf{Z}}_c\,{}^{k}\boldsymbol{\Psi}_c^{-1}\hat{\mathbf{Z}}_c^{\mathrm{T}})^{-1} \tag{T.9}$$

Here ${}^{k}\hat{\mathbf{c}}^{\mathrm{T}}$ is a row vector of $\hat{\mathbf{C}}$. Using this estimate and the estimates of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ the objective function can be calculated using equation T.10.

$$S_c^2 = \sum_{k=1}^{K}({}^{k}\mathbf{x}_c - {}^{k}\hat{\mathbf{x}}_c)\,{}^{k}\boldsymbol{\Psi}_c^{-1}({}^{k}\mathbf{x}_c - {}^{k}\hat{\mathbf{x}}_c)^{\mathrm{T}} \tag{T.10}$$

5. Calculate the convergence parameters λ_1 and λ_2.

$$\lambda_1 = (S_b^2 - S_a^2)/S_a^2;\ \lambda_2 = (S_c^2 - S_a^2)/S_a^2 \tag{T.11}$$

If λ_1 and λ_2 are less than the convergence limit (typically 10^{-8} in this work), terminate. Otherwise return to step 2.

efficiency of the algorithm [14], but these will not be incorporated here. Comparative data on convergence time will be reported in a future paper.

It is worth noting that the algorithm presented in Table 2.1 imposes restrictions on the presence of offsets in any mode. Normally, this would be equivalent to saying that the data have been properly mean centered [15], but in the case of non-uniform measurement errors, mean centering is not equivalent to eliminating offsets. The case of offsets will be treated from a more optimal, though incomplete perspective in Section 2.3.4.

Although, the problem of heteroscedastic noise has been addressed in the literature using weighted PARAFAC algorithms, the description presented here marks the first time (to our knowledge) that a formal theoretical treatment of this problem from a maximum likelihood perspective has been given. It also represents a good starting point to generalize this algorithm to more complicated scenarios, such as systems affected by correlated noise and heteroscedastic and correlated noise in two or more dimensions.

2.3.2 Correlated Measurement Errors

The incorporation of uncorrelated, heteroscedastic measurement errors into the ALS framework as described in the preceding section is relatively straightforward. On the surface it may appear that extension to correlated measurement errors is a trivial matter, since the likelihood function expressed by equation 2.7 should be equally applicable for error covariance matrices that are not diagonal. However, there is a critical difference that relates to the way in which the information in the error covariance matrices is transformed when the three-way array is unfolded. In the case of uncorrelated measurement errors, the diagonal error covariance matrices in each mode contain all of the information about the uncertainty in the measurements, although the order in which this information appears varies with the modes. In the case of correlated measurements, some of this information will be lost in one or more modes, making it impossible to maintain consistency in the ALS estimates obtained when using the same strategy as for independent errors.

To illustrate this point consider the relatively simple case where the errors are correlated in one order only. For example, we may have a case where multiple samples

of different composition are separated by chromatography with multichannel detection and there is significant correlation in the time domain due to pump noise. Alternatively, we could imagine fluorescence excitation-emission measurements for a series of samples, which are correlated in the emission domain due to source fluctuations, but uncorrelated in the excitation domain because it is scanned at longer time intervals as the second order. For convenience, we will say that the measurements along the rows, which make up mode B are correlated, but there is no correlation among these rows in the three-way array. This situation is conceptually illustrated with a small 4x3x2 array in Figure 2.1.

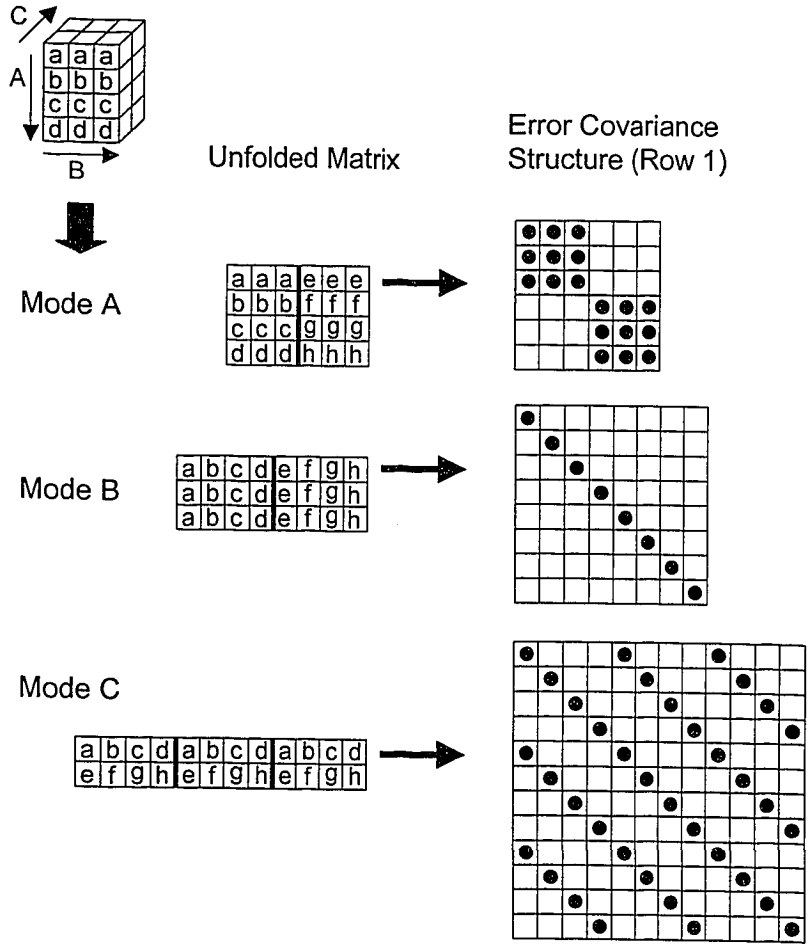

Figure 2.1. Illustration of the unfolding of a three-way array and its effect on the structure of the error covariance matrix for the first row of the unfolded matrix. Elements with correlated measurement errors are labelled with the same letter.

The elements of the array that are labeled with the same letters are considered to be correlated in this example. Considering unfolding in the A mode first, the figure shows the structure of the error covariance matrix for the first row of $\mathbf{X}_a$, which is block

diagonal due to the presence of two sets of correlated measurement errors. The remaining 3 rows will have the same error covariance structure, resulting in 72 non-zero elements in total describing error covariance. On the other hand, the error covariance matrix for the first row of $\mathbf{X}_b$ has a diagonal form since the correlated measurements appear in the columns. Considering all three rows of $\mathbf{X}_b$, this results in only 24 non-zero elements describing the error structure. Information on the covariance has been lost in this representation. Finally, the error covariance matrix structure for the first row of $\mathbf{X}_c$ is band diagonal. The two error covariance matrices resulting from this unfolded matrix will have a total of 72 non-zero values describing the error covariance and contain the same information as the A mode, only in a different representation. However, because the error covariance matrices for $\mathbf{X}_b$ contain incomplete information, the sequence of steps in the ALS algorithm described in the previous section cannot be completed using this approach.

As correlation among the orders becomes more complex, the inability to represent this information becomes more obvious. This is clear if one realizes that a complete description of all correlations in the general case would require $(IJK)^2$ elements, but the total number of elements in the row covariance matrices for, say $\mathbf{X}_a$, is only $I(JK)^2$. In order to circumvent this problem a more general solution for correlated errors will be obtained redefining the problem and modeling the measurements as a single point in an IJK-dimensional space. To do this, $\underline{\mathbf{X}}$ (or alternatively any unfolded representation) is vectorized by applying the '*vec*' operator and the equations are adapted as necessary. The generalization of equations 2.12 and 2.10 are:

$$vec(\hat{\mathbf{A}}^{\mathrm{T}}) = (\mathbf{V}_a^{\mathrm{T}}\boldsymbol{\Omega}_a^{-1}\mathbf{V}_a)^{-1}\mathbf{V}_a^{\mathrm{T}}\boldsymbol{\Omega}_a^{-1}vec(\mathbf{X}_a^{\mathrm{T}}) \tag{2.14}$$

$$S^2 = vec(\Delta\mathbf{X}_a^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{\Omega}_a^{-1}vec(\Delta\mathbf{X}_a^{\mathrm{T}}) \tag{2.15}$$

where,

$$\mathbf{V}_a = \mathbf{I}_I \otimes \hat{\mathbf{Z}}_a^{\mathrm{T}} \tag{2.16}$$

$$\boldsymbol{\Omega}_a = E[(vec((\mathbf{X}_a - \mathbf{X}_a^{\circ})^{\mathrm{T}})) \cdot (vec((\mathbf{X}_a - \mathbf{X}_a^{\circ})^{\mathrm{T}}))^{\mathrm{T}}] \tag{2.17}$$

$$\Delta \mathbf{X}_a = (\mathbf{X}_a - \hat{\mathbf{X}}_a) \tag{2.18}$$

In these equations, the '*vec*' operator reshapes a matrix into a column vector by taking the elements in sequence column-wise [13]. The symbol '⊗' as used here identifies the Kronecker product such that each element of $\mathbf{I}_I$ is multiplied by $\hat{\mathbf{Z}}_a^{\mathrm{T}}$ therefore $\mathbf{V}_a$ is an *IP* x *IJK* matrix with $\mathbf{Z}_a^{\mathrm{T}}$ repeating along the diagonal. The matrix $\mathbf{\Omega}_a$ is the full error covariance matrix for $vec(\mathbf{X}_a^{\mathrm{T}})$, providing information about the error covariance among all the measurements. Similar equations can be obtained by making the appropriate substitutions for the second and third mode in a trilinear case, or to the other dimensions in a multilinear case. Based on this, an alternating regression algorithm similar to the one in the preceding section can be formulated as shown in Table 2.2.

In a manner analogous to the ALS algorithm for heteroscedastic errors, the generalized algorithm presented in Table 2.2 uses the maximum likelihood estimates in two spaces to estimate the solution in the other space. In order to exchange the solutions, the error covariance matrix for $vec(\mathbf{X}_a^{\mathrm{T}})$, given by $\mathbf{\Omega}_a$, needs to be modified to give the error covariance matrix for $vec(\mathbf{X}_b^{\mathrm{T}})$ and $vec(\mathbf{X}_c^{\mathrm{T}})$, given by $\mathbf{\Omega}_b$ and $\mathbf{\Omega}_c$ respectively. This can be done on an element-by-element basis; but since these matrices contain the same elements in a different order, it is simpler to apply a special type of matrix called permutation matrix to carry out the rearrangement. The permutation matrix is an orthonormal matrix that changes the arrangement of the elements. Conveniently, the same permutation matrix that is used to convert error covariance matrices can also be used to convert between the vectorized forms of $\mathbf{X}_a$, $\mathbf{X}_b$, and $\mathbf{X}_c$. Equations 2.19 through 2.22 show how this is done.

$$vec(\mathbf{X}_b^{\mathrm{T}}) = \mathbf{P}_b vec(\mathbf{X}_a^{\mathrm{T}}) \tag{2.19}$$

$$vec(\mathbf{X}_c^{\mathrm{T}}) = \mathbf{P}_c vec(\mathbf{X}_a^{\mathrm{T}}) \tag{2.20}$$

$$\mathbf{\Omega}_b = \mathbf{P}_b \mathbf{\Omega}_a \mathbf{P}_b^{\mathrm{T}} \tag{2.21}$$

$$\mathbf{\Omega}_c = \mathbf{P}_c \mathbf{\Omega}_a \mathbf{P}_c^{\mathrm{T}} \tag{2.22}$$

The construction of the permutation matrices $\mathbf{P}_b$ and $\mathbf{P}_c$, which consist only of ones and zeros, is conceptually straightforward but algorithmically involved, so the details of this will not be presented here.

The algorithm presented in Table 2.2 represents a completely general treatment for the case where correlation can exist among all of the measurement errors. Although it is presented for the trilinear case, extension to higher orders is trivial. The algorithm also has very good convergence characteristics and gives results that are identical to those obtained by the algorithm in Section 2.2.1 in the presence of uncorrelated noise. In practice, implementation of the algorithm is limited to some extent by the size and stability of the matrices and the convergence time. These three factors are not completely independent from one another. For example, as $\underline{\mathbf{X}}$ becomes large, the associated error covariance matrices tend to become ill-conditioned, causing convergence problems. A variety of approaches, such as compression [14], line search extrapolation [16], and simplifications based on the error structure [17] may be adapted to the present algorithm to avoid these problems. The first two modifications will not be treated in this paper since the first is beyond the scope of the present work and the second is primarily an algorithmic modification to the ALS algorithm. However, the third approach has important practical implications and for this reason will be the focus of the next section.

2.3.3 Simplification: Correlation Along One Order Only

For many chemical applications, error covariance affects only one order or at least the covariance in other orders can be neglected. This can, in certain cases, result in substantial simplification of the generalized algorithm. For the purpose of illustration, only the case where correlations exist along the rows (*i.e.* in the second order, as illustrated in Figure 2.1) will be considered, since correlations along other orders can be rendered equivalent through permutation of the original array or appropriate adjustment

Table 2.2. General MLPARAFAC algorithm (correlated measurement errors).

1. Given an I x J x K cube of data $\underline{\mathbf{X}}$, a corresponding IJK x IJK matrix $\mathbf{\Omega}_a$ of error covariances for $vec(\mathbf{X}_a)$ and two permutation matrices $\mathbf{P}_b$ and $\mathbf{P}_c$ to permute from $vec(\mathbf{X}_a)$ to $vec(\mathbf{X}_b)$ and $vec(\mathbf{X}_c)$, respectively, the algorithm is initialized using random values of the correct dimensions or using estimates obtained by TLD.

$$[\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}] = tld(\underline{\mathbf{X}},P) \tag{T.12}$$

2. Unfold and vectorize $\underline{\mathbf{X}}$ retaining the first order and calculate the maximum likelihood estimation of $\hat{\mathbf{A}}$ conditional on $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$.

$$vec(\mathbf{X}_a^{\mathrm{T}}) = vec(unfold(\underline{\mathbf{X}},a)^{\mathrm{T}});\ \hat{\mathbf{V}}_a = \mathbf{I}_I \otimes \hat{\mathbf{Z}}_a^{\mathrm{T}};\ \mathbf{\Omega}_a \tag{T.13}$$

$$vec(\hat{\mathbf{A}}^{\mathrm{T}}) = (\hat{\mathbf{V}}_a^{\mathrm{T}}\mathbf{\Omega}_a^{-1}\hat{\mathbf{V}}_a)^{-1}\hat{\mathbf{V}}_a^{\mathrm{T}}\mathbf{\Omega}_a^{-1}vec(\mathbf{X}_a^{\mathrm{T}}) \tag{T.14}$$

Here $vec(\hat{\mathbf{A}}^{\mathrm{T}})$ is the vectorized row form of $\hat{\mathbf{A}}$. Using this estimate and the estimates of $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ the objective function can be calculated using equation T.15.

$$\Delta\mathbf{X}_a = (\mathbf{X}_a - \hat{\mathbf{X}}_a);\ S_a^2 = vec(\Delta\mathbf{X}_a)^{\mathrm{T}}\mathbf{\Omega}_a^{-1}vec(\Delta\mathbf{X}_a) \tag{T.15}$$

3. Vectorize $\underline{\mathbf{X}}$ retaining the second order and calculate the maximum likelihood estimation of $\hat{\mathbf{B}}$ conditional on $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$.

$$vec(\mathbf{X}_b^{\mathrm{T}}) = \mathbf{P}_b vec(\mathbf{X}_a^{\mathrm{T}});\ \hat{\mathbf{V}}_b = \mathbf{I}_J \otimes \hat{\mathbf{Z}}_b^{\mathrm{T}};\ \mathbf{\Omega}_b = \mathbf{P}_b\mathbf{\Omega}_a\mathbf{P}_b^{\mathrm{T}} \tag{T.16}$$

$$vec(\hat{\mathbf{B}}^{\mathrm{T}}) = (\hat{\mathbf{V}}_b^{\mathrm{T}}\mathbf{\Omega}_b^{-1}\hat{\mathbf{V}}_b)^{-1}\hat{\mathbf{V}}_b^{\mathrm{T}}\mathbf{\Omega}_b^{-1}vec(\mathbf{X}_b^{\mathrm{T}}) \tag{T.17}$$

Here $vec(\hat{\mathbf{B}}^{\mathrm{T}})$ is the vectorized row form of $\hat{\mathbf{B}}$. Using this estimate and the estimates of $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ the objective function can be calculated using equation T.18.

$$\Delta\mathbf{X}_b = (\mathbf{X}_b - \hat{\mathbf{X}}_b);\ S_b^2 = vec(\Delta\mathbf{X}_b)^{\mathrm{T}}\mathbf{\Omega}_b^{-1}vec(\Delta\mathbf{X}_b) \tag{T.18}$$

4. Vectorize $\underline{\mathbf{X}}$ retaining the third order and calculate the maximum likelihood estimation of $\hat{\mathbf{C}}$ conditional on $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

$$vec(\mathbf{X}_c^{\mathrm{T}}) = \mathbf{P}_c vec(\mathbf{X}_a^{\mathrm{T}});\ \hat{\mathbf{V}}_c = \mathbf{I}_K \otimes \hat{\mathbf{Z}}_c^{\mathrm{T}};\ \mathbf{\Omega}_c = \mathbf{P}_c\mathbf{\Omega}_a\mathbf{P}_c^{\mathrm{T}} \tag{T.19}$$

$$vec(\hat{\mathbf{C}}^{\mathrm{T}}) = (\hat{\mathbf{V}}_c^{\mathrm{T}}\mathbf{\Omega}_c^{-1}\hat{\mathbf{V}}_c)^{-1}\hat{\mathbf{V}}_c^{\mathrm{T}}\mathbf{\Omega}_c^{-1}vec(\mathbf{X}_c^{\mathrm{T}}) \tag{T.20}$$

Here $vec(\hat{\mathbf{C}}^{\mathrm{T}})$ is the vectorized row form of $\hat{\mathbf{C}}$. Using this estimate and the estimates of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ the objective function can be calculated using equation T.21.

$$\Delta\mathbf{X}_c = (\mathbf{X}_c - \hat{\mathbf{X}}_c);\ S_c^2 = vec(\Delta\mathbf{X}_c)^{\mathrm{T}}\mathbf{\Omega}_c^{-1}vec(\Delta\mathbf{X}_c) \tag{T.21}$$

5. Calculate the convergence parameters λ_1 and λ_2.

$$\lambda_1 = (S_b^2 - S_a^2)/S_a^2;\ \lambda_2 = (S_c^2 - S_a^2)/S_a^2 \tag{T.22}$$

If λ_1 and λ_2 are less than the convergence limit (typically 10^{-8} in this work), terminate. Otherwise return to step 2.

of equations which will be presented. For this case, three common cases can be distinguished: (1) the error covariance is different among all of the rows forming the array; (2) the error covariance is different among rows forming different slices but identical among the rows of the same slice; and (3) the error covariance is identical among the rows of all the slices. This section will focus in the second and third cases, since the first case can only be treated using the general algorithm. To begin, however, it is helpful to examine the second case, which is more general and can be extended to the third case in a straightforward manner.

Imagine a trilinear data set such as the examples presented in Section 2.2.2, where the error correlation can be expected to affect only one order, which we will assume to be the second order as noted above. In addition, in certain cases where this assumption applies, it may be possible to make the additional assumption that the error covariance matrix is the same for each row in the same vertical slice of data. Considering that the correlation occurs along the rows of $\mathbf{X}_a$ and is the same in each row, all the covariance information is contained in a single *JK* x *JK* covariance matrix $\mathbf{\Psi}_a$ defined by:

$$\mathbf{\Psi}_a = E[(\mathbf{x}_a - \mathbf{x}_a^o)^T(\mathbf{x}_a - \mathbf{x}_a^o)] \tag{2.23}$$

Here, $\mathbf{x}_a$ and $\mathbf{x}_a^o$ can represent any row of $\mathbf{X}_a$ and $\mathbf{X}_a^o$, the unfolded forms of the measured data array and the error-free data array, respectively. Of course, $\mathbf{X}_a^o$ is not generally known, so in the absence of *a priori* knowledge of the error covariance matrix, $\mathbf{\Psi}_a$ might typically be estimated by obtaining replicates of the measurements for each row and using the means in place of $\mathbf{x}_a^o$, then pooling all of the results, as indicated in equation 2.24.

$$\mathbf{\Psi}_a \approx \frac{1}{I}\sum_{i=1}^{I}\frac{1}{(N-1)}\sum_{n=1}^{N}({}^{in}\mathbf{x}_a - {}^{i}\bar{\mathbf{x}}_a)^T({}^{in}\mathbf{x}_a - {}^{i}\bar{\mathbf{x}}_a) \tag{2.24}$$

In this equation, ${}^{in}\mathbf{x}_a$ is the *n*th replicate measurement of the *i*th row of $\mathbf{X}_a$ and ${}^{i}\bar{\mathbf{x}}_a$ is the mean of the *N* replicates for that row. (Note that these replicates would likely be obtained through separate experiments for each of the *K* slices.) Other strategies are also possible, but these will not be discussed in detail here. The full covariance matrix, $\mathbf{\Omega}_a$, will now be block diagonal, consisting of *I* identical diagonal units of dimension *JK* x *JK*.

This situation offers a number of advantages to the algorithm. From a storage capacity point, the improvement is related to the reduction of the number of non-zero elements from a maximum of $(IJK)^2$ in the general case to $(JK)^2$, since $\mathbf{\Omega}_a$, that has the form represented in equation 2.26, can also be represented as the sparse Kronecker product shown in equation 2.26.

$$\mathbf{\Omega}_a = \begin{bmatrix} \mathbf{\Psi}_a & & & & \\ & \mathbf{\Psi}_a & & & \\ & & . & & \\ & & & . & \\ & & & & \mathbf{\Psi}_a \end{bmatrix} \tag{2.25}$$

$$\mathbf{\Omega}_a = \mathbf{I}_I \otimes \mathbf{\Psi}_a \tag{2.26}$$

Additionally, this improves the numerical stability of the algorithm since the Kronecker form allows $\mathbf{\Omega}_a$ to be inverted by inversion of the individual covariance matrix $\mathbf{\Psi}_a$ as shown in equation 2.27.

$$\mathbf{\Omega}_a^{-1} = \mathbf{I}_I \otimes \mathbf{\Psi}_a^{-1} \tag{2.27}$$

The companion error covariance matrices for the other orders can be obtained using the permutation matrices via equations 2.28 and 2.29.

$$\mathbf{\Omega}_b^{-1} = \mathbf{P}_b \mathbf{\Omega}_a^{-1} \mathbf{P}_b^{\mathrm{T}} \tag{2.28}$$

$$\mathbf{\Omega}_c^{-1} = \mathbf{P}_c \mathbf{\Omega}_a^{-1} \mathbf{P}_c^{\mathrm{T}} \tag{2.29}$$

Based on these equations and in the identical block diagonal form of $\mathbf{\Omega}_a$, it is easy to demonstrate that the maximum likelihood solution for the **A** loadings is obtained using equation 2.30.

$$\hat{\mathbf{A}} = \mathbf{X}_a \mathbf{\Psi}_a^{-1} \mathbf{Z}_a^{\mathrm{T}} (\mathbf{Z}_a \mathbf{\Psi}_a^{-1} \mathbf{Z}_a^{\mathrm{T}})^{-1} \tag{2.30}$$

Although, the equation for order A under this assumptions is analogous in form to equation 2.13 for the heteroscedastic case, the rest of the equations needed to implement the ALS algorithm are different. In order to obtain these equations, it should first be realized that $\mathbf{\Omega}_c^{-1}$ can be represented as shown in equation 2.31, as is apparent from Figure 2.1, while $\mathbf{\Omega}_b^{-1}$ cannot be similarly simplified under these circumstances.

$$\boldsymbol{\Omega}_c^{-1} = \begin{bmatrix} {}^1\boldsymbol{\Psi}_c^{-1} & & & \\ & {}^2\boldsymbol{\Psi}_c^{-1} & & \\ & & \ddots & \\ & & & {}^K\boldsymbol{\Psi}_c^{-1} \end{bmatrix} \tag{2.31}$$

This leads to equations 2.32 and 2.33 for the maximum likelihood estimation of **B** and **C**, respectively under this assumption.

$$vec(\hat{\mathbf{B}}^{\mathrm{T}}) = (\mathbf{V}_b^{\mathrm{T}}\boldsymbol{\Omega}_b^{-1}\mathbf{V}_b)^{-1}\mathbf{V}_b^{\mathrm{T}}\boldsymbol{\Omega}_b^{-1}vec(\mathbf{X}_b^{\mathrm{T}}) \tag{2.32}$$

$${}^k\hat{\mathbf{c}}^{\mathrm{T}} = \mathbf{x}_c\,{}^k\boldsymbol{\Psi}_c^{-1}\mathbf{Z}_c^{\mathrm{T}}(\mathbf{Z}_c\,{}^k\boldsymbol{\Psi}_c^{-1}\mathbf{Z}_c^{\mathrm{T}})^{-1} \tag{2.33}$$

Besides the storage improvements achieved, speed enhancements are also realized since **A** can now be estimated projecting the data at once onto a smaller set of matrices. In order to estimate the **C** loading, a row-by-row procedure has to be implemented since the error covariance matrices change from slice to slice. The estimation of **B** has to be done using the full error covariance matrix as in the general case since the error covariance terms needed cannot be summarized in a more efficient manner. This algorithm for this simplified case is presented in Table 2.3.

A further simplification is possible when the error covariance matrix is the same for each row in all the slices, a situation which is not uncommon, at least to a first approximation. In this case equations 2.27 and 2.30 can be used to estimate **A** as before, and analogous equations can be used to estimate **C** by making the appropriate substitutions, since all of the $\boldsymbol{\Psi}_c$'s are now the same. The calculations are further simplified by realizing that $\boldsymbol{\Omega}_b^{-1}$, under these noise characteristics, can be expressed as in equation 2.34, since the permutation matrix in this case is similar to the commutation matrix used in reference17, reducing the estimation of **B** to equation 2.35.

$$\boldsymbol{\Omega}_b^{-1} = \boldsymbol{\Psi}_b^{-1} \otimes \mathbf{I}_J \tag{2.34}$$

$$\hat{\mathbf{B}} = \mathbf{X}_b\mathbf{Z}_b^{\mathrm{T}}(\mathbf{Z}_b\mathbf{Z}_b^{\mathrm{T}})^{-1} \tag{2.35}$$

Table 2.4 gives the algorithm under these assumptions.

Table 2.3. Simplified MLPARAFAC algorithm (Simplification 1 - same error covariance matrix for each row in a slice, but different between slices).

1. Given an I x J x K cube of data $\underline{\mathbf{X}}$, a corresponding IJK x IJK matrix $\mathbf{\Omega}_b$ of error covariances for $vec(\mathbf{X}_b)$ and two permutation matrices $\mathbf{P}_b$ and $\mathbf{P}_c$ to migrate from $vec(\mathbf{X}_a)$ to $vec(\mathbf{X}_b)$ and $vec(\mathbf{X}_c)$ respectively. The algorithm is initialized using estimates obtained by TLD.

$$[\hat{\mathbf{A}},\hat{\mathbf{B}},\hat{\mathbf{C}}] = tld(\underline{\mathbf{X}},P) \tag{T.23}$$

2. Unfold $\underline{\mathbf{X}}$ retaining the first order and calculate the maximum likelihood estimation of $\hat{\mathbf{A}}$ conditional on $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$. Since $\mathbf{\Omega}_a$ is block diagonal as shown in T.24, $\hat{\mathbf{A}}$ can be calculated at once.

$$\mathbf{X}_a = unfold(\underline{\mathbf{X}},a);\ \mathbf{\Omega}_a = \mathbf{P}_b^{\mathrm{T}}\mathbf{\Omega}_b\mathbf{P}_b;\ \mathbf{\Omega}_a = \mathbf{I}_J \otimes \mathbf{\Psi}_a \tag{T.24}$$

$$\hat{\mathbf{A}} = \mathbf{X}_a\mathbf{\Psi}_a^{-1}\hat{\mathbf{Z}}_a^{\mathrm{T}}(\hat{\mathbf{Z}}_a\mathbf{\Psi}_a^{-1}\hat{\mathbf{Z}}_a^{\mathrm{T}})^{-1} \tag{T.25}$$

Using $\hat{\mathbf{A}}$ and the estimates of $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ the objective function can be calculated using equation T.26.

$$S_a^2 = tr((\mathbf{X}_a - \hat{\mathbf{X}}_a)\mathbf{\Psi}_a^{-1}(\mathbf{X}_a - \hat{\mathbf{X}}_a)^{\mathrm{T}}) \tag{T.26}$$

3. Unfold $\underline{\mathbf{X}}$ retaining the second order and calculate the maximum likelihood estimation of $\hat{\mathbf{B}}$ conditional on $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ using $\mathbf{\Omega}_b$.

$$vec(\mathbf{X}_b^{\mathrm{T}}) = vec(unfold(\mathbf{X}_b))^{\mathrm{T}};\ \hat{\mathbf{V}}_b = \mathbf{I}_J \otimes \hat{\mathbf{Z}}_b^{\mathrm{T}};\ \mathbf{\Omega}_b \tag{T.27}$$

$$vec(\hat{\mathbf{B}}^{\mathrm{T}}) = (\hat{\mathbf{V}}_b^{\mathrm{T}}\mathbf{\Omega}_b^{-1}\hat{\mathbf{V}}_b)^{-1}\hat{\mathbf{V}}_b^{\mathrm{T}}\mathbf{\Omega}_b^{-1}vec(\mathbf{X}_b^{\mathrm{T}}) \tag{T.28}$$

Here $vec(\hat{\mathbf{B}}^{\mathrm{T}})$ is the vectorized row form of $\hat{\mathbf{B}}$. Using this estimate and the estimates of $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ the objective function can be calculated using equation T.29.

$$\Delta\mathbf{X}_b = (\mathbf{X}_b - \hat{\mathbf{X}}_b);\ S_b^2 = vec(\Delta\mathbf{X}_b)^{\mathrm{T}}\mathbf{\Omega}_b^{-1}vec(\Delta\mathbf{X}_b) \tag{T.29}$$

4. Unfold $\underline{\mathbf{X}}$ retaining the third order and calculate the maximum likelihood estimation of $\hat{\mathbf{C}}$ conditional on $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$. ${}^k\mathbf{\Psi}_c$ is constructed taking the corresponding block of $\mathbf{\Omega}_c$ since it is block diagonal.

$$\mathbf{X}_c = unfold(\underline{\mathbf{X}},c);\ \mathbf{\Omega}_c = \mathbf{P}_c\mathbf{P}_b^{\mathrm{T}}\mathbf{\Omega}_b\mathbf{P}_b\mathbf{P}_c^{\mathrm{T}};\ {}^k\mathbf{\Psi}_c \tag{T.30}$$

$${}^k\hat{\mathbf{c}}^{\mathrm{T}} = {}^k\mathbf{x}_c\,{}^k\mathbf{\Psi}_c^{-1}\hat{\mathbf{Z}}_c^{\mathrm{T}}(\hat{\mathbf{Z}}_c\,{}^k\mathbf{\Psi}_c^{-1}\hat{\mathbf{Z}}_c^{\mathrm{T}})^{-1} \tag{T.31}$$

Here ${}^i\hat{\mathbf{c}}^{\mathrm{T}}$ is a row vector of $\hat{\mathbf{C}}$. Using this estimate and the estimates of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ the objective function can be calculated using equation T32.

$$S_c^2 = \sum_{k=1}^{K}({}^k\mathbf{x}_c - {}^k\hat{\mathbf{x}}_c)\,{}^k\mathbf{\Psi}_c^{-1}({}^k\mathbf{x}_c - {}^k\hat{\mathbf{x}}_c)^{\mathrm{T}} \tag{T.32}$$

5. Calculate the convergence parameters λ_1 and λ_2.

$$\lambda_1 = (S_b^2 - S_a^2)/S_a^2;\ \lambda_2 = (S_c^2 - S_a^2)/S_a^2 \tag{T.33}$$

If λ_1 and λ_2 are less than the convergence limit (typically 10^{-8} in this work), terminate. Otherwise return to step 2.

Table 2.4. Simplified MLPARAFAC algorithm (Simplification 2 - same error covariance matrix for each row in each slice).

1. Given an I x J x K cube of data $\underline{\mathbf{X}}$, and the error covariance matrices $\mathbf{\Psi}_a$ and $\mathbf{\Psi}_c$ for the A and B orders respectively. The algorithm is initialized using random values of the correct dimensions or using estimates obtained by TLD.

$$[\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{C}}] = tld(\underline{\mathbf{X}}, P) \tag{T.34}$$

2. Unfold $\underline{\mathbf{X}}$ retaining the first order and calculate the maximum likelihood estimation of $\hat{\mathbf{A}}$ conditional on $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$.

$$\mathbf{X}_a = unfold(\underline{\mathbf{X}}, a);\ \mathbf{\Psi}_a \tag{T.35}$$

$$\hat{\mathbf{A}} = \mathbf{X}_a \mathbf{\Psi}_a^{-1} \hat{\mathbf{Z}}_a^{\mathrm{T}} (\hat{\mathbf{Z}}_a \mathbf{\Psi}_a^{-1} \hat{\mathbf{Z}}_a^{\mathrm{T}})^{-1} \tag{T.36}$$

Using $\hat{\mathbf{A}}$ and the estimates of $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ the objective function can be calculated using equation T.37.

$$S_a^2 = tr((\mathbf{X}_a - \hat{\mathbf{X}}_a)\mathbf{\Psi}_a^{-1}(\mathbf{X}_a - \hat{\mathbf{X}}_a)^{\mathrm{T}}) \tag{T.37}$$

3. Unfold $\underline{\mathbf{X}}$ retaining the second order and calculate the maximum likelihood estimation of $\hat{\mathbf{B}}$ conditional on $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$.

$$\mathbf{X}_b = unfold(\underline{\mathbf{X}}, b) \tag{T.38}$$

$$\hat{\mathbf{B}} = \mathbf{X}_b \hat{\mathbf{Z}}_b^{\mathrm{T}} (\hat{\mathbf{Z}}_b \hat{\mathbf{Z}}_b^{\mathrm{T}})^{-1} \tag{T.39}$$

Using $\hat{\mathbf{B}}$ and the estimates of $\hat{\mathbf{C}}$ and $\hat{\mathbf{A}}$ the objective function can be calculated using equation T.40.

$$S_b^2 = tr((\mathbf{X}_b - \hat{\mathbf{X}}_b)(\mathbf{X}_b - \hat{\mathbf{X}}_b)^{\mathrm{T}}) \tag{T.40}$$

4. Unfold $\underline{\mathbf{X}}$ retaining the third order and calculate the maximum likelihood estimation of $\hat{\mathbf{C}}$ conditional on $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$.

$$\mathbf{X}_c = unfold(\underline{\mathbf{X}}, c);\ \mathbf{\Psi}_c \tag{T.41}$$

$$\hat{\mathbf{C}} = \mathbf{X}_c \mathbf{\Psi}_c^{-1} \hat{\mathbf{Z}}_c^{\mathrm{T}} (\hat{\mathbf{Z}}_c \mathbf{\Psi}_c^{-1} \hat{\mathbf{Z}}_c^{\mathrm{T}})^{-1} \tag{T.42}$$

Using $\hat{\mathbf{C}}$ and the estimates of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ the objective function can be calculated using equation T.43.

$$S_c^2 = tr((\mathbf{X}_c - \hat{\mathbf{X}}_c)\mathbf{\Psi}_c^{-1}(\mathbf{X}_c - \hat{\mathbf{X}}_c)^{\mathrm{T}}) \tag{T.43}$$

5. Calculate the convergence parameters λ_1 and λ_2.

$$\lambda_1 = (S_b^2 - S_a^2)/S_a^2;\ \lambda_2 = (S_c^2 - S_a^2)/S_a^2 \tag{T.44}$$

If λ_1 and λ_2 are less than the convergence limit (typically 10^{-8} in this work), terminate. Otherwise return to step 2.

2.3.4 MLPARAFAC with Offsets

So far, it has been assumed that the multilinear data is not affected by offsets in any mode. Unfortunately, it is not uncommon in chemical system to have offsets affecting different orders. The sources of offsets range from instrumental artifacts, such as a spectral background for all samples or variations in cell position, to factors related to sample preparation. One general model to describe trilinear data affected by different kinds of offsets is represented by equation 2.36.

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \sum_{p=1}^{P} a_{ip} b_{jp} c_{kp} \tag{2.36}$$

Here, μ is the grand mean of $\underline{\mathbf{X}}$ and α_i, β_j and γ_k represent the offsets for mode A, B and C, respectively. It has been reported [15] that, in cases where an overall offset exists; it can be removed by eliminating the offset associated with any mode. Therefore, the grand mean can be incorporated into any or all the offset terms affecting the individual modes. When the measurements in $\underline{\mathbf{X}}$ are corrupted by *iid* noise, proper mean-centering to remove the offset is a convenient approach since this pre-processing step does not destroy the multilinear characteristics of the data. It is important to note, however, that mean-centering will alter the structure of the loadings so that they may no longer be readily identified with real factors, counteracting one of the main benefits of trilinear decomposition.

From a mathematical point of view, the mean-centering is equivalent to adding trilinear factors that are formed by the product of a vector of offsets and two other loading vectors set to ones as shown in equation 2.37.

$$\underline{\mathbf{X}} = (\boldsymbol{\alpha} \otimes \mathbf{1}_J \otimes \mathbf{1}_K) + (\mathbf{1}_I \otimes \boldsymbol{\beta} \otimes \mathbf{1}_K) + (\mathbf{1}_I \otimes \mathbf{1}_J \otimes \boldsymbol{\gamma}) + \sum_{p=1}^{P} \mathbf{a}_p \otimes \mathbf{b}_p \otimes \mathbf{c}_p \tag{2.37}$$

Note that equation 2.37 is a general formulation and in a given application, the offset affecting any of the modes could be set to zero, *i.e.* α, β or γ could be a zero vector. In addition, it could even be constrained to be a general offset affecting all the measurements equally and then, loadings representing each mode would be equal to a vector of ones and everything multiplied by a scalar representing the offset. However, the presence of non-uniform and/or correlated error distribution makes mean centering no

longer optimal from a maximum likelihood point of view. This can be understood considering that mean centering in any mode is the projection of $\underline{\mathbf{X}}$ unfolded in this mode onto the null space spanned by the vector of ones corresponding to this mode. Therefore, this projection will only be optimal under *iid* conditions. In order to mean center optimally, the procedure should be incorporated into the ALS algorithm. Contrary to what happens in MLPCA, where the loadings are constrained to be orthogonal, PARAFAC does not impose any constraints on the estimation of the loadings, making the inclusion of offsets in the ALS algorithm a more straightforward task. Additionally, an important benefit is that the offsets may often be uniquely determined because of the uniqueness of the PARAFAC model.

A relatively simple approach to handling offsets can be used when the offsets follow the structure represented by equation 2.37. It is clear from this equation that the offsets can be incorporated by using from one to three more factors (in the trilinear case) than the number of factors expected in the absence of offsets. The number of additional factors which should be added depends on how many modes exhibit the offsets in equation 2.37. This means of dealing with offsets is easily incorporated into the MLPARAFAC algorithm, and will yield maximum likelihood estimates of the loadings in accordance with the model, but is not the best approach. This is because the loadings in the two modes other than the one in which the offsets occur are allowed to "float"; that is, they are not constrained to unity (or, more generally, a constant value). While these loadings may be nearly constant and will constitute a maximum likelihood solution to the expanded-factor model, all of the loadings in this case should experience a greater variance than would be expected with the true model. The situation is analogous to fitting simple bivariate straight line data with an intercept of zero to linear models. The data could be fit using only a slope term (intercept forced to zero), or with a slope and intercept term. Both approaches will yield a maximum likelihood solution, but the latter strategy (which has a closer fit but fewer degrees of freedom) will produce a larger variance in the slope, so it is the less preferred method given *a priori* knowledge of a zero intercept. Likewise, if we have prior knowledge of a structure such as that in equation 2.37, it is better to incorporate this into the modeling process.

Equation 2.37 is only one of many possible constrained structures that can exist in trilinear models, and it is clear that the question of offsets is part of a more general issue of constrained factors. The nature of these constraints is very application dependent and relies on prior information. While such constraints can be incorporated into the MLPARAFAC algorithm, a general discussion of strategies is premature and beyond the scope of the current paper. However, one example will be presented in Section 2.4.5 to demonstrate the performance of MLPARAFAC in the presence of offsets.

2.3.5 Estimation of Error Covariance Matrices

The error covariance matrix is of critical importance in maximum likelihood methods such as MLPCA and MLPARAFAC. Consequently, questions often arise related to procedures used to estimate the error covariance matrix, the quality of these estimates, and the implications of this on the subsequent analysis. While the emphasis of this work is on the development of the algorithm, it is legitimate to raise these concerns, so they will be addressed here, although only briefly.

Perhaps the most obvious way to estimate the error covariance matrix is through the use of replicates, as indicated in the discussion related to equation 2.24. In practice, such an approach may be limited by experimental design considerations or realistic constraints on the number of experiments that can be conducted. Covariance estimates, like variance estimates, are notoriously imprecise unless a large number of replicates is employed. This is often impractical, although pooling can sometimes be used. The question then becomes whether it is better to employ traditional methods (which assume an *iid*-normal error structure) or maximum likelihood methods with a noisy error covariance matrix. Maximum likelihood methods will generally be favored in situations where the number of replicates is large and/or the level of heteroscedastic/correlated noise is high. The precise point at which the use of maximum likelihood methods becomes advantageous will depend on the particular application and a detailed examination of this is beyond the scope of the present work.

An alternative to the often unpopular approach of measuring replicates is to characterize the error covariance structure for a particular instrument or application based

on empirical evidence or theoretical considerations. In the same way that certain instruments are known to exhibit proportional noise, it may be possible to obtain a functional form for the error covariance in certain types of applications. This is already done to some degree when multiplicative signal correction (MSC) is applied to near-infrared data dominated by scatter noise. Furthermore, in some circumstances, it may be possible to describe covariance arising from techniques such as filtering or transformation on a purely theoretical basis. By using such approaches, more reliable error covariance matrices can be obtained that are not subject to the statistics of a small number of replicates.

For the work presented here, which is intended to validate the algorithm rather than to demonstrate its practical application, the theoretical error covariance matrix based on noise simulation was used. This removed any uncertainties associated with the error covariance in the statistical validation.

2.4 Experimental

2.4.1 Data Sets

Since the objective of this work is to describe the theoretical basis of the MLPARAFAC algorithm and to validate its capabilities, all of the data sets employed in this work were simulated so that the rank and error structure could be known with confidence. Future studies will examine the performance of the algorithm for real experimental systems. Although a wide range of simulations were carried out, the results from only six data sets are presented here to support the main conclusions. In all cases, the data sets were relatively small, since the studies generally involved statistical validation requiring numerous runs.

Data Set 1 was a rank-three data set of dimensions 8x7x4 used to test the degrees of freedom with conventional PARAFAC algorithm under conditions of *iid* normal noise and compare it with the new algorithms. The loadings for mode A were represented by an 8x3 matrix drawn from a uniform distribution of random numbers from zero to three ($U(0,3)$). Similarly, **B** was a 7x3 matrix from $U(0,2)$ and **C** was a 4x3 matrix from

$U(0,5)$. The error free data were generated using equation 2.3, yielding the 8x28 matrix of error-free data, unfolded to maintain the A mode. The matrix of measurement errors was an 8x28 matrix of normally distributed random numbers ($\mu = 0$, $\sigma = 0.1$, or $N(0,0.1)$), which was added to the error-free data to generate the unfolded form of Data Set 1. This matrix was then folded into a three-way array and passed to the PARAFAC algorithms.

Data Set 2 was a rank-three data set of dimensions 6x7x3 and was used to test the algorithm under conditions of heteroscedastic but uncorrelated noise. The error-free data was generated in the same fashion as Data Set 1, with the same ranges of loadings but using the corresponding dimensions. The matrix of measurement errors was created by the Hadamard (element-by-element) multiplication of a 6x21 matrix of normally distributed random numbers drawn from $N(0,1)$ and a 6x21 matrix of random numbers, $\mathbf{Q}_a$, drawn from $U(0,0.1)$, representing the matrix of standard deviation for each measurement in $\mathbf{X}_a$. The noise matrix and the error-free data set were added and the resultant matrix was folded.

The error-free part of Data Set 3, which was used to test the general algorithm for correlation in multiple orders, is identical to Data Set 1. The noise matrix was created to introduce non-uniform and correlated noise at the same time. Initially, an 8x28 matrix of normally distributed random numbers drawn from $N(0,1)$ was generated and multiplied in an element-by-element fashion by one-tenth of the value of the error-free measurements. The resultant matrix was treated with a 15 point moving average filter along each row in order to produce error covariance. At the boundaries of the error matrix the filter was wrapped around the to the opposite side in order to eliminate edge effects. Since the error matrix was unfolded to maintain mode A, this approach produced correlation among the measurements in the two other modes. Although this approach is not particularly realistic, it represents a general case for which the covariance structure could be easily predicted. Again, the error-free data set was added to the noise matrix in order to generate the data set.

Data Sets 4 and 5 were 5x8x4 matrices, again formed by three components in the same manner as already described for error-free data. These data sets were used to test

simplifications to the general algorithm related to the error covariance structure. In both cases, the error-free data were the same and only the measurement error matrices differed. The noise matrix for Data Set 4 was generated to simulate a system where the errors are correlated along only one order (the B mode) and the error covariance matrix is identical for each vector in this mode. To do this, four 5x8 matrices of normally distributed random numbers drawn from $N(0,0.1)$ were generated and all of them were individually treated with a 5 point moving average filter along the rows. The filtered error matrix was added to the error-free matrix and used in the simulations. For Data Set 5, the correlated errors were also only in one order and all the vectors in a given 'slice' (mode C fixed) had the same error covariance structure, but this structure varies from slice to slice. The measurement error matrices for this data set were generated in the same manner as for Data Set 4, but the standard deviation of the normal distribution and the filter width were varied between slices ($\sigma = 0.15, 0.15, 0.2, 0.1, 0.05$; $w = 3, 5, 7, 3$).

Data Set 6, which was used to test the performance in the presence of offsets, was constructed from a 7x8x4 rank three matrix with the same distribution of loading values and the same noise correlation structure as Data Set 3 – heteroscedastic and correlated in two orders. In this case, however, a single vector offset was added to the second order; that is, a 1x8 vector of values drawn from $U(0,2)$ was added to each row of the three way array.

2.4.2 Computational Aspects

All the calculations performed in this work were carried out on a Sun Ultra 60 workstation with 2 x 300 MHz processors and 512 MB of RAM and a 700MHz Pentium-III PC with 128 MB of RAM. All programs were written in-house using Matlab 6.0 (The MathWorks Inc., Natick, MA).

2.5 Results and Discussion

2.5.1 Statistical Validation

In order to validate the various proposed algorithms, it was necessary to verify that they yield the maximum likelihood solution. This can be accomplished by exploiting

the statistical characteristics of S^2 values for the correct model. Operationally, this is done by analyzing replicate data sets, each with the same matrix of error-free data and the same error structure, but with different realizations of the measurement error each time. If the distribution of S^2 values for these replicates follows a χ^2 distribution with the appropriate degrees of freedom, it can then be concluded that the algorithm is finding the maximum likelihood solution. Probability plots are used in this work to make this comparison. Initially, the replicate data sets (normally 100 replicates) are analyzed and the S^2 values are stored. Then, the S^2 values are sorted from the smallest to the largest and assigned a cumulative probability according to their position in the list; this is called the observed probability. For instance, the third element in the list would be assigned an observed probability of $2/n$, where n is the number of replicates. The expected probability is then calculated using the χ^2 distribution. The cumulative probability density function for χ^2 can be calculated using the incomplete gamma function included in Matlab as shown in equation 2.38:

$$P(S^2 \mid \nu) = \Gamma_{inc}\left(\frac{S^2}{2}, \frac{\nu}{2}\right) \tag{2.38}$$

where ν is the number of degrees of freedom. If the two distributions are the same, a plot of the observed probabilities *vs.* the expected probabilities should yield a straight line with a slope of unity. If the model is insufficient to account for the systematic variance, either because the form of the model is incorrect or the parameters are suboptimal, then the points of the plot will lie above the ideal line. If the model accounts for an excessive amount of variance, *i.e.* the estimated rank is too high and measurement variance is modeled, the points will lie below the ideal line. It should be pointed out that the only way to employ this approach is to use simulated data where the true noise characteristics are known. Because error estimates for virtually all experimental measurements will have some (often substantial) degree of uncertainty, the resulting S^2 values will not follow a χ^2 distribution. (For this reason, it can be argued that the present methods are not truly, "maximum likelihood", since they should also estimate the error covariance, but this is not practical in most situations.)

The issue of degrees of freedom for trilinear data is far from being trivial. Bro has suggested that degrees of freedom do not exist *a priori* [18], but have to be determined from the specific data. This situation arises from the fact that the rank of a trilinear data set cannot be calculated based on the same approach used in bilinear data. For instance, the maximum rank of a 3x3x3 array is five [19] contrary to what happens in bilinear data, where the maximum rank of a 3x3 matrix is always three. Unfortunately, there is no simple rule for calculating the maximum rank of arrays in general, except for the bilinear case and some simple trilinear arrays. However, Durell *et al* [20]. reported two equations to calculate the degrees of freedom in trilinear and quadrilinear models, as given in equation 2.39 and 2.40.

$$\nu(\underline{\mathbf{X}})_{3\text{-way}} = IJK - P(I+J+K-2) \tag{2.39}$$

$$\nu(\underline{\mathbf{X}})_{4\text{-way}} = IJKL - P(I+J+K+L-3) \tag{2.40}$$

The theoretical foundation of these equations is not completely clear, but it has been suggested in the literature that they might be used for exploratory (qualitative) purpose. In other words, they should not be used as the statistically correct number of degrees of freedom. In the present work, the approach was to use equation 2.39 as estimator of the statistically meaningful number of degrees of freedom for a trilinear case. In order to assess the merit of this approach, trilinear data corrupted with *iid* normal noise, such as Data Set 1, were submitted to the standard PARAFAC algorithm, which is well-known to yield the maximum likelihood solution under these noise characteristics. The replication procedure described above was performed using 100 and 1000 replicates and the probability plot, shown in Figure 2.2, was constructed. It is observed that the plot follows the theoretical slope very closely for 1000 replicates, indicating that equation 2.39 provides a credible number of degrees of freedom, at least for the purposes of this study. For 100 replicates, the agreement is not as good due to the smaller sample size, but these results are included as a point of reference for other studies that involve only 100 replicates. It is worth noting that, even though the results are not shown, analysis of all of the trilinear data structures used in this work was carried out under *iid* conditions using PARAFAC to confirm the estimated degrees of freedom.

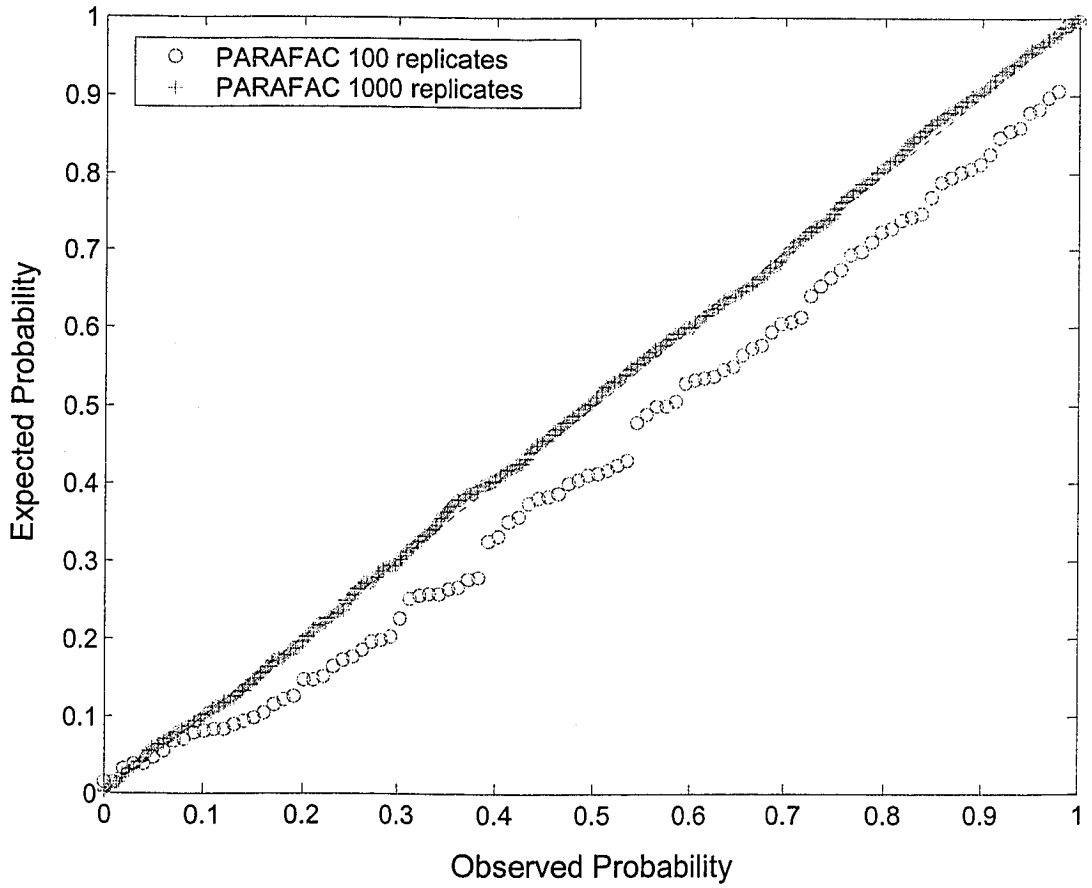

Figure 2.2. Probability plot for PARAFAC results under conditions of *iid* normal errors for 100 (O) and 1000 (+) replicates. The solid line with unity slope indicates ideal behaviour for maximum likelihood estimation.

2.5.2 Non-uniform (Uncorrelated) Measurement Errors: Data Set 2

In order to test the validity of the algorithm depicted in Table 2.1, Data Set 2, which was corrupted with heteroscedastic error, was employed. Since the main objective of this study is the statistical validation of the algorithm, the theoretical error covariance matrix obtained from the simulation parameters was employed. The theoretical error covariance matrix for each row is calculated using equation 2.41:

$$ {}^{i}\mathbf{\Psi}_{a} = diag({}^{i}\mathbf{q}_{a})^{2} \tag{2.41} $$

where *diag*( ) represents the diagonalization operator that transforms the vector argument into a diagonal matrix. The result is a diagonal matrix with the squared elements of ${}^{i}\mathbf{q}_{a}$ (the *i*th row of $\mathbf{Q}_{a}$, the matrix of standard deviations unfolded in the A mode) along the

diagonal. Error covariance matrices for the other orders were obtained using the same equation applied to $\mathbf{Q}_b$ and $\mathbf{Q}_c$ accordingly.

Figure 2.3 shows the results obtained for the analysis of Data Set 2 using PARAFAC and the version of MLPARAFAC designed to accommodate heteroscedastic noise.

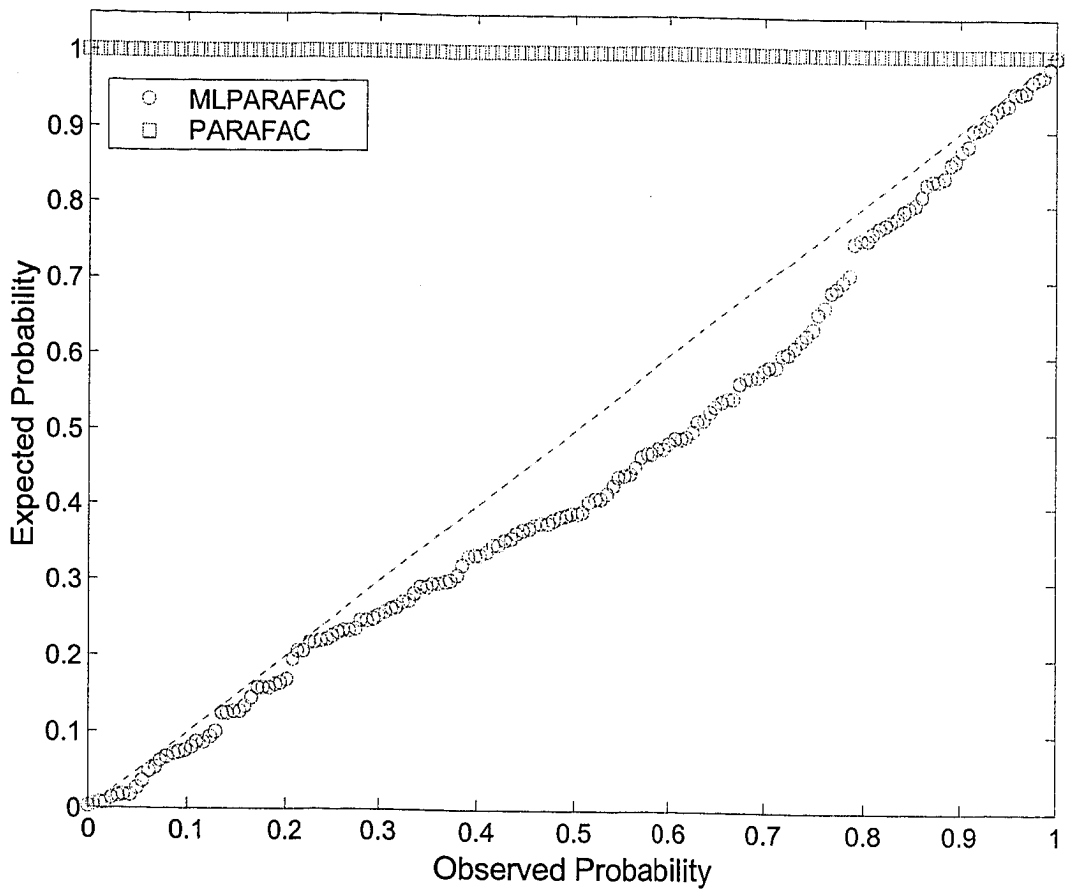

Figure 2.3. Probability plot for the analysis 100 replicates of Data Set 2 (non-uniform, uncorrelated errors) by MLPARAFAC (○) and PARAFAC (□).

The S^2 values in both cases were calculated in the same manner, *i.e.* using equation 2.9 with either the PARAFAC or MLPARAFAC estimates of $\hat{\mathbf{X}}_a$. It is clear from the figure that the estimates obtained using MLPARAFAC follow the expected behavior for maximum likelihood estimation, with only minor deviations attributable to the statistical limitations of the study. On the other hand, the models obtained by PARAFAC do not adequately account for the systematic variance in the data set, producing suboptimal solutions that deviate radically from the line representing expected

χ^2 distribution in the probability plot. Although this data set was not designed to test the more general algorithm depicted in Table 2.2, it was also analyzed using that algorithm to test its generality. The general algorithm produced exactly the same set of solutions, indicating that the two algorithms are equivalent under these noise characteristics.

2.5.3 Non-uniform and Correlated Measurement Errors: Data Set 3

In the preceding section, it was noted that the general MLPARAFAC algorithm for correlated errors was able to handle the case of uncorrelated errors as well. Data Set 3 was designed to test the general algorithm in the presence of errors which were correlated and heteroscedastic. Again, the theoretical error covariance matrix was used. For this specific data set, the covariance matrix in the A mode is given by:

$$\mathbf{\Omega}_a = \begin{bmatrix} {}^1\mathbf{\Psi}_a & & & \\ & {}^2\mathbf{\Psi}_a & & \\ & & \ddots & \\ & & & {}^I\mathbf{\Psi}_a \end{bmatrix} \tag{2.42}$$

where ${}^i\mathbf{\Psi}_a$ represents the error covariance matrix of the i row of $\mathbf{X}_a$ and was calculated using equation 2.43.

$${}^i\mathbf{\Psi}_a = \mathbf{F}^{\mathrm{T}}(diag((0.1)\cdot {}^i\mathbf{x}_a^{\circ}))^2\mathbf{F} \tag{2.43}$$

Here, $\mathbf{F}$ is the 28x28 filter matrix designed to carry out the 15-point moving average smooth on the noise ($\mathbf{e}_{corr} = \mathbf{e}_{iid}\,\mathbf{F}$), and second term is a diagonal matrix of the variance of the noise in the ith row of noise matrix prior to smoothing, equal to 10% of the error-free measurement squared. The companion error covariance matrices, $\mathbf{\Omega}_b$ and $\mathbf{\Omega}_c$, were calculated using their respective permutation matrices as shown in equation 2.21 and 2.22.

Figure 2.4 shows the probability plots obtained using Data Set 3. Results for the general MLPARAFAC algorithm, which can accommodate any covariance structure, are shown for both 100 and 1000 replicates. Both of these show good agreement with the expected slope of unity, indicating that a maximum likelihood solution has been obtained.

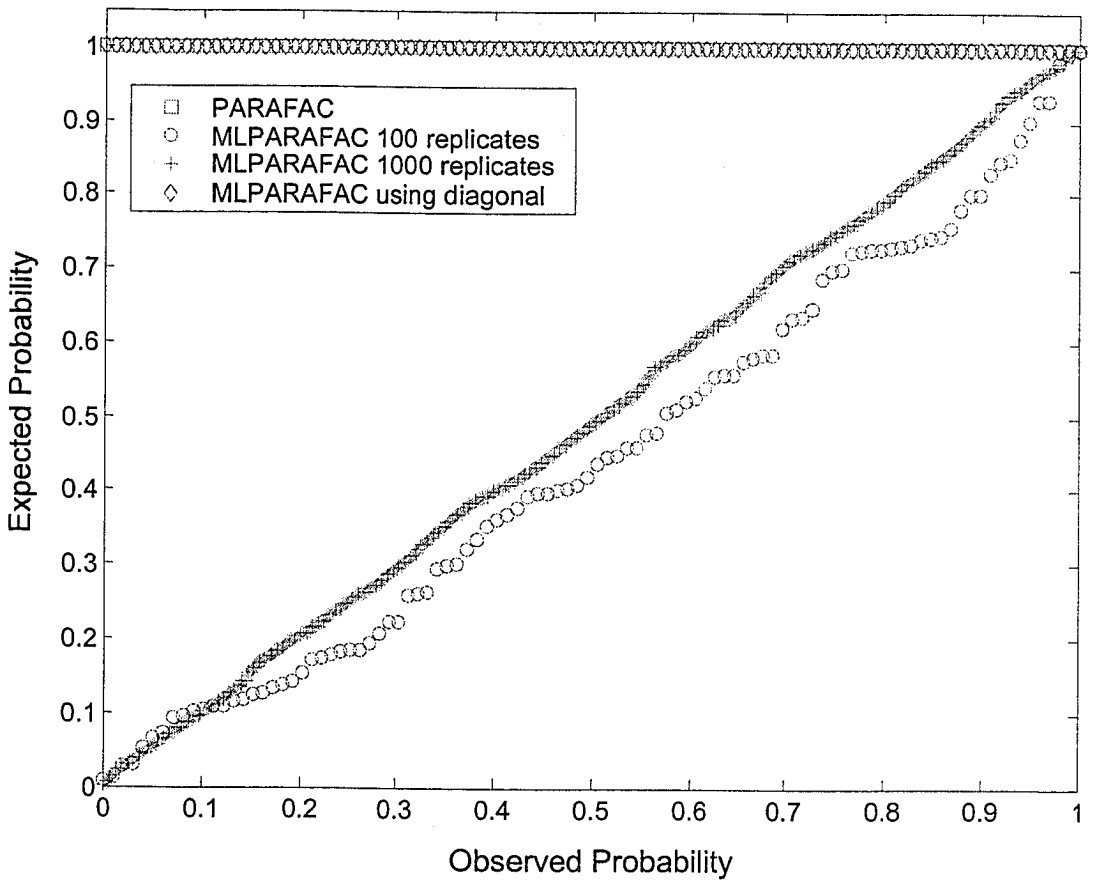

Figure 2.4. Probability plot for the analysis of Data Set 3 (correlated measurement errors in two modes) using the general MLPARAFAC algorithm with 100 (○) and 1000 (+) replicates, the standard MLPARAFAC algorithm for uncorrelated errors with 100 replicates (◇) and PARAFAC with 100 replicates (□).

In contrast, it is clear that the PARAFAC model has substantial systematic error, since it generates a maximum expected probability of unity across all observed probabilities. In order to test whether the superior performance of the general MLPARAFAC algorithm was due to its inclusion of the error covariance structure or simply because it accounts for heteroscedasticity, results were also generated using the version of MLPARAFAC designed to accommodate only heteroscedasticity. For this analysis, only the diagonal elements of the full error covariance matrix ($\boldsymbol{\Omega}$) were employed. Like the standard PARAFAC algorithm, these models result in systematic errors, indicating that modeling the covariance structure is a critical factor.

2.5.4 Simplified Error Covariance Structures: Data Sets 4 and 5

While the general MLPARAFAC algorithm should be able to deal with any error covariance structure, in many cases it may be possible to use the simplified algorithms presented in Tables 2.3 and 2.4. These algorithms were tested using Data Sets 4 and 5. Data Set 4, which has a simple error covariance structure consisting of correlation in one mode only and identical error covariance matrices for all the vectors in this mode, was used to test the corresponding algorithm in Table 2.4, which will be referred to as Simplification 2. The probability plots for this study are shown in Figure 2.5, together with the results of the generalized algorithm and conventional PARAFAC.

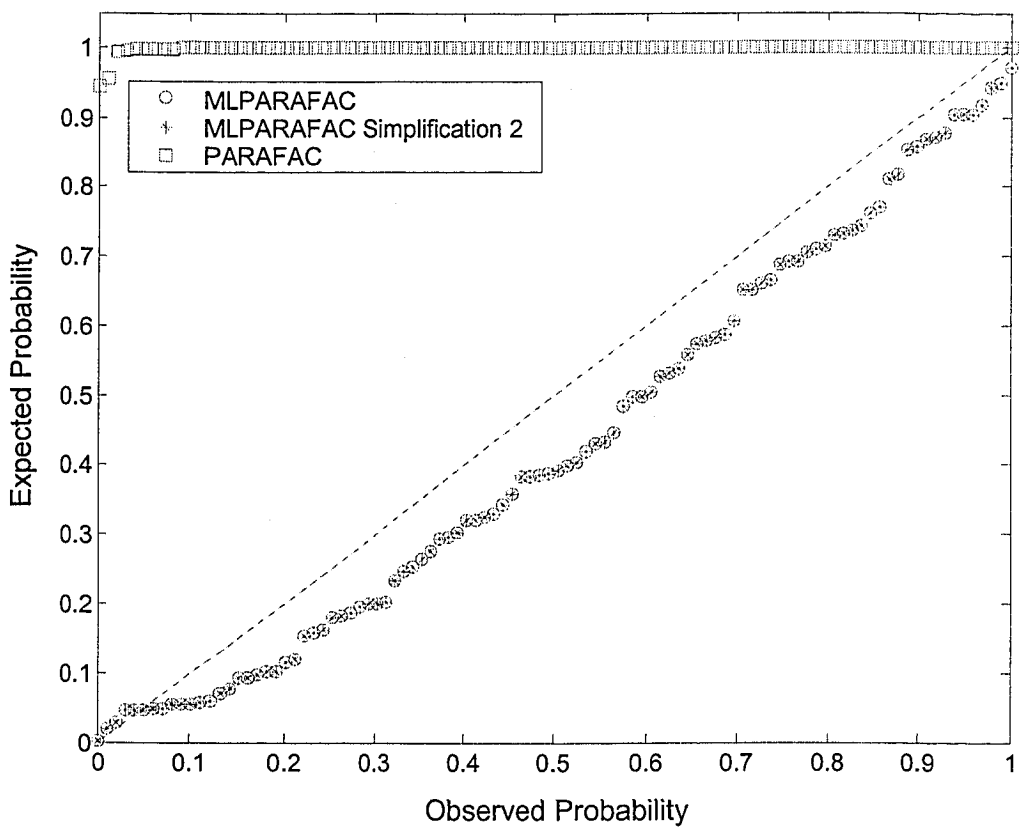

Figure 2.5. Probability plot for the analysis of 100 replicates of Data Set 4 (identical row correlations) using the general MLPARAFAC algorithm (○), Simplification 2 of the general MLPARAFAC algorithm (✳) and PARAFAC (□).

Note that the results of the general algorithm and Simplification 2 are identical, confirming that the latter is a special case of the former, and that both appear to produce the maximum likelihood results. As before, the performance of PARAFAC is suboptimal. Simplification 1, which appears in Table 2.3, is designed to handle the case

where: (i) error correlation exists in one mode only, and (ii) the error covariance structure differs from vector to vector along one of the remaining modes, but is the same along the other remaining mode. Data Set 5, which was simulated to test this algorithm, was created such that errors were correlated along the rows (mode B) and the error covariance matrix was identical for rows within the same slice (mode A), but different across different slices (mode C). The results from analysis of 100 replicates are summarized in Figure 2.6.

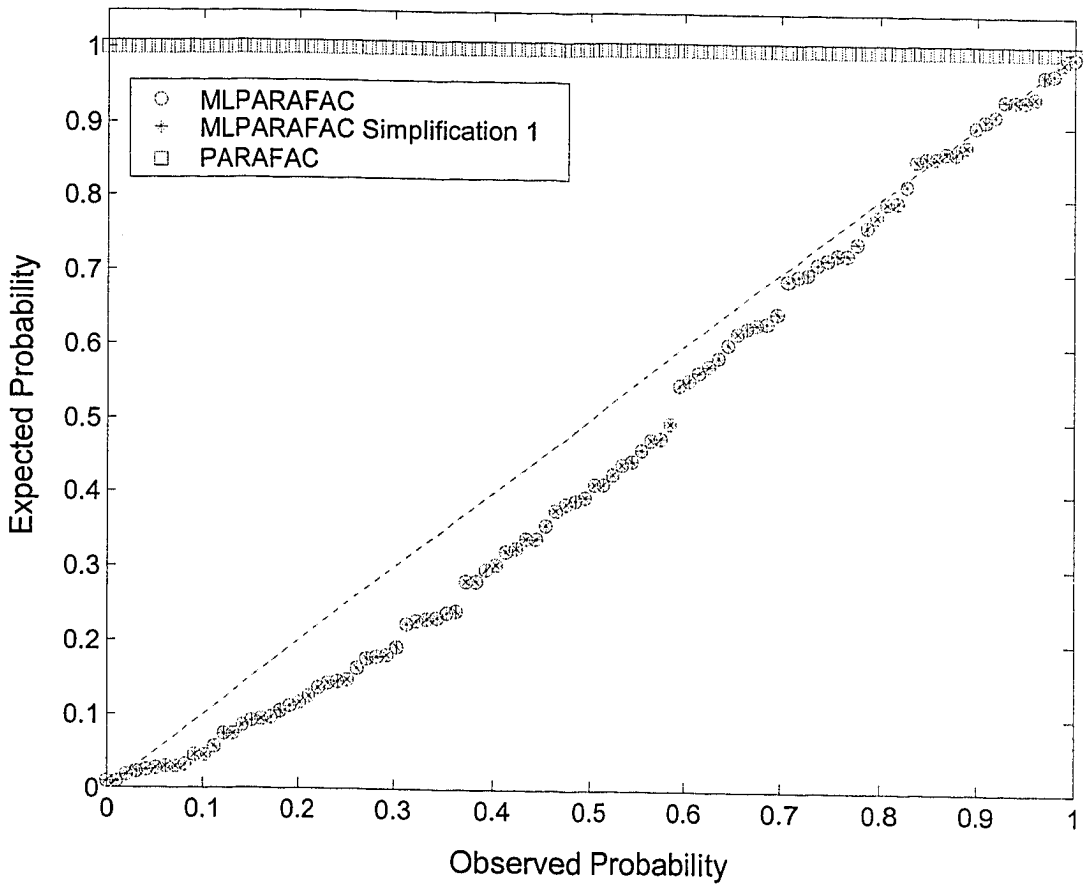

Figure 2.6. Probability plot for the analysis of 100 replicates of Data Set 5 (different row correlations along mode A, same row correlations along mode C) using the general MLPARAFAC algorithm (○), Simplification 1 of the MLPARAFAC algorithm (✳) and PARAFAC (□).

As with Simplification 2, the figure shows the identical results for Simplification 1 and the generalized algorithm, both of which produce maximum likelihood estimates, and poor results for PARAFAC.

2.5.5 MLPARAFAC with Offsets: Data Set 6

As noted in Section 2.2.4, the inclusion of certain kinds of offsets in the trilinear structure can be modeled by using an expanded rank model. This can be demonstrated with Data Set 6, which has offsets added to one order (*i.e.* α and γ are zero in equation 2.37, but β is not). Therefore, expansion of the PARAFAC model to rank four should accommodate the offsets. This is demonstrated with the probability plots in Figure 2.7, which compares the results of MLPARAFAC (general algorithm) with conventional

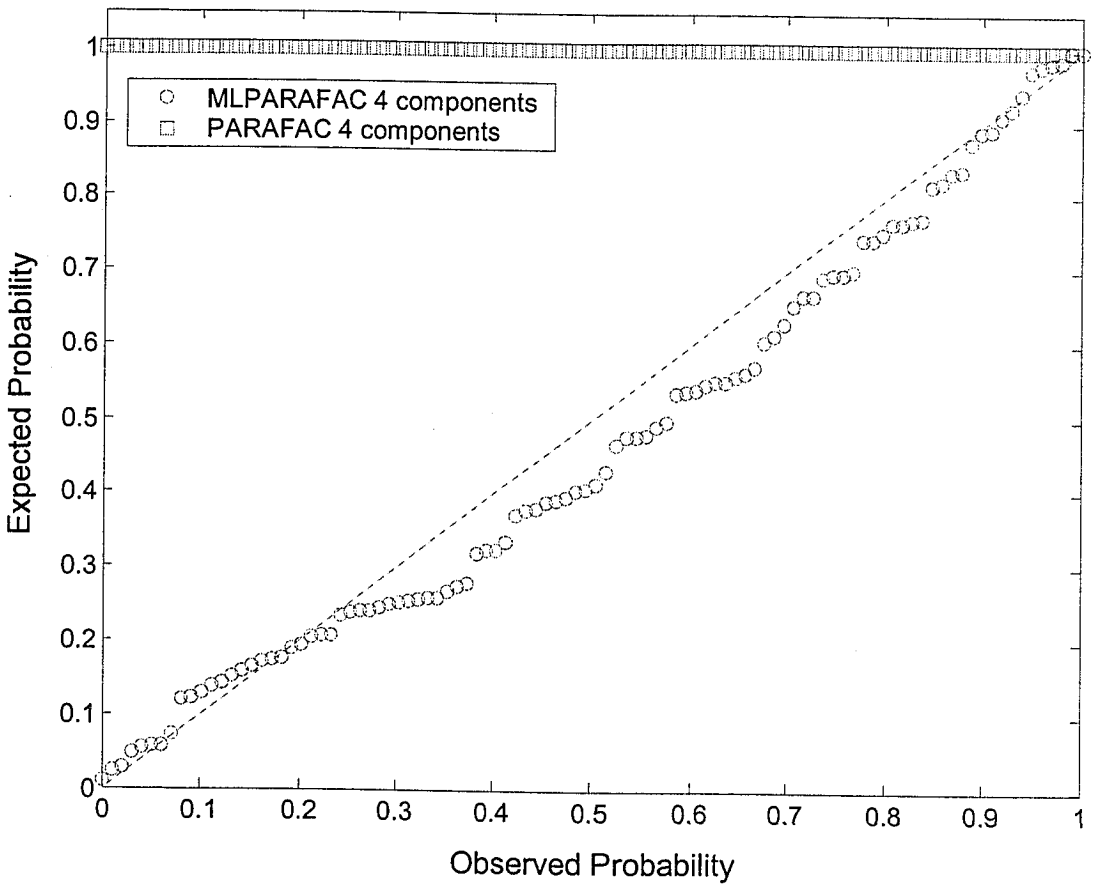

Figure 2.7. Probability plot for the analysis of 100 replicates of Data Set 6 (correlation along modes B and C plus offset on modeB) using the general MLPARAFAC algorithm (○) and PARAFAC (□).

PARAFAC, both with rank four models. It is clear that MLPARAFAC produces the maximum likelihood solution while PARAFAC does not. Furthermore, this approach to

handling offsets is superior to mean-centering in that the integrity of the loading vectors is retained.

As noted in Section 2.2.4, the maximum likelihood solution extracted in this manner does not represent the "best" solution in this application because information about constraints on the loading vectors in the A and C modes of the offset factor (*i.e.* that they are fixed) is not incorporated into the ALS algorithm. While it is possible to do this, the inclusion of constrained factors adds algorithmic complications and introduces questions regarding degrees of freedom, so this issue will not be dealt with in this paper.

2.5.6 Model Quality

The preceding sections dealt with the statistical validation of the maximum likelihood estimation process, but nothing has been said about the quality of the estimates obtained using these new algorithms. Although the implication has been that the MLPARAFAC solutions are better, two reasonable questions that arise are: (1) Are the MLPARAFAC estimates closer to the true underlying factors than the PARAFAC estimates?, and (2) Do the MLPARAFAC estimates offer a significant advantage over the estimates obtained by PARAFAC?. The first question can be answered easily using simulated data. The second question is more general in essence and it can only be addressed on a case-by-case basis since the advantages gained by MLPARAFAC will strongly depend on the type and magnitude of error corrupting the data and the correct use of a number of parameters related to the estimation of the model. Some of the parameters determining the success of MLPARAFAC over PARAFAC are the number of components, accuracy of the estimation of the error covariance matrix, and the use of the correct algorithm based on the error structure present.

The first issue, the closeness of estimates to the true factors, will be addressed using vector angles as a figure of merit. This figure of merit is the angular difference between the true loading vectors and the estimated loading vectors in each mode. For example, the vector angle between two loading vectors in mode A is given by:

$$\theta_p^a = \cos^{-1}\left(\frac{\hat{\mathbf{a}}_p^{\mathrm{T}}\mathbf{a}_p}{\|\hat{\mathbf{a}}_p\|\,\|\mathbf{a}_p\|}\right) \tag{2.44}$$

where $\mathbf{a}_p$ and $\hat{\mathbf{a}}_p$ are the true and estimated values for the pth loading vector of **A**. Analogous equations can be used for the other orders. Smaller angles mean a greater similarity, so by comparing the vector angles obtained by MLPARAFAC with those of PARAFAC, the agreement with the true vector can be assessed. An alternative measure is the correlation coefficient of the vectors, which is simply the term in parentheses, but since this approaches unity with small differences, it is less sensitive.

To evaluate the accuracy of the loadings extracted by MLPARAFAC and PARAFAC under different conditions, loadings extracted from 100 replicates of Data Sets 2 through 6 by both MLPARAFAC and PARAFAC were used to calculate vector angles for each of the loadings. These angles were then averaged over the 100 replicates to give 9 mean angles and their standard deviations (3 modes x 3 factors) for each method. These results are summarized in compressed form in Table 2.5, which, in the interest in saving space, shows only the results for the first loading vector in each mode. The uncertainty given is the population standard deviation.

Table 2.5. Comparison of vector angle accuracies for PARAFAC and MLPARAFAC. Results are based on 100 replicates and uncertainties are given as standard deviations.

Data Set	Mean angular deviation (°) PARAFAC			Mean angular deviation (°) MLPARAFAC		
	A	**B**	**C**	**A**	**B**	**C**
2	0.27± 0.15	0.33± 0.13	0.21± 0.18	0.17± 0.09	0.19± 0.08	0.14± 0.11
3	0.90± 0.36	0.61± 0.34	0.58± 0.37	0.08± 0.02	0.14± 0.05	0.09± 0.04
4	0.17± 0.07	0.27± 0.14	0.21± 0.16	0.07± 0.04	0.19± 0.08	0.10± 0.09
5	0.25± 0.12	0.43± 0.23	0.32± 0.25	0.10± 0.05	0.23± 0.16	0.16± 0.16
6	1.77± 1.30	3.04± 1.16	1.52± 1.03	0.24± 0.12	0.47± 0.14	0.31± 0.19

The results in Table 2.5 support the general view that MLPARAFAC produces more accurate estimates of the loading vectors than PARAFAC. Both the mean vector angles and their uncertainties are smaller in all cases for MLPARAFAC, although the degree to which this is true varies with the data set. For Data Set 2, the differences

between the two methods is relatively small. This might be expected, however, since this data set contains heteroscedastic errors only with no correlated errors, and the degree of heteroscedasticity, arising from proportional errors, is not very large. Nevertheless, differences are statistically significant (note that the standard deviation of the mean will be the value reported in the table divided by 10). The differences are much more dramatic for Data Set 3, which has correlated errors in two modes, and illustrates the importance of modeling error covariance. To further emphasize this point, the analysis of Data Set 3 by MLPARAFAC assuming only heteroscedastic errors (*i.e.* using only the diagonal) produced corresponding vector angles of 0.92 ± 0.37, 0.59 ± 0.34, and 0.57 ± 0.35, which are not significantly different from the PARAFAC results. Data Sets 4 and 5, which exhibit a smaller degree of error covariance than Data Set 3, also show less dramatic differences between MLPARAFAC and PARAFAC, but the angular differences are still about a factor of two and are statistically very significant. The analysis of these two data sets employed the simplified algorithms, but it should be noted that the general algorithm produced identical results, as expected. In Data Set 6, the addition of a fourth factor representing the offset decreases the quality of the estimates in general compared to Data Set 3 (the most similar data set). Because of the highly correlated error structure, this data set exhibits a difference of a factor of five or more in the mean vector angles obtained by the two methods. For comparison purposes, the corresponding vector angles for the rank three MLPARAFAC model are 1.52 ± 0.55, 3.79 ± 0.17 and 1.44 ± 0.42, indicating that the inclusion of the fourth factor to model the offset is essential.

These results clearly demonstrate that improved estimates of loadings can be obtained from the trilinear model when information about the measurement error structure is available and is incorporated into the modeling process in the correct way. As already noted, the extent to which these improvements will be significant for a given application depends on nature of the application and the characteristics of the noise. Furthermore, the results presented here were obtained assuming an absolute knowledge of the measurement error covariance matrix, but in practice this is typically estimated on the basis of replicate measurements and hence may be less reliable. The benefits of including measurement error information must therefore be weighed against the detrimental effects of including poor quality information. The development of the

algorithms presented here has demonstrated the potential for improvements that could be achieved and facilitates application to more practical situations in which an experimental assessment of their benefits can be made.

2.6 Conclusions

Four algorithms for carrying out MLPARAFAC based on an ALS framework have been described in this work. The simplest of these is designed to work with cases where the measurement errors are non-uniform (heteroscedastic) but uncorrelated. The most general form of the algorithm can treat data with any type of error covariance structure. Two simplifications of the general algorithm were also presented which more efficiently handle more restricted error covariance structures. All of the algorithms were shown to produce maximum likelihood estimates through a comparison of the distribution of the objective function with the χ^2 distribution. It was also shown that the quality of the estimated loading vectors for MLPARAFAC was significantly better than for the PARAFAC models in cases where the error covariance matrix is known.

Although the principles of MLPARAFAC have been established here, a number of more practical aspects related to its implementation remain to be examined. These include issues related to the computational efficiency and stability of the algorithms for large arrays, the estimation of error covariance matrices for three-way data, and the implementation of constraints on the loadings within the algorithms. These subjects will be the focus of future investigations.

Acknowledgements:

The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Dow Chemical Company (Midland, MI).

2.7 References

1. Hirschfeld T. The hy-phen-ated methods. *Anal. Chem.* 1980; **52**: 297A.

2. Apellof CJ and Davidson ER. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Anal. Chem.* 1981; **53**: 2053.

3. Leurgans S and Ross RT. Multilinear models: applications in spectroscopy. *Statist. Sci.* 1992; **7**: 289.

4. Harshman RA. Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis. *WPP* 1970; **16**: 1.

5. Sanchez E and Kowalski BR. Tensorial resolution: a direct trilinear decomposition. *J. Chemom.* 1990; **4**: 29.

6. Paatero P. A weighted non-negative least squares algorithm for three-way "PARAFAC" factor analysis, *Chemom. Intell. Lab. Syst.* 1997; **38**: 223.

7. Yates F. The analysis of replicated experiments when the field results are incomplete. *Empire J. Exp. Agric.* 1933; **1**: 129.

8. Kiers HAL. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika.* 1997; **62**: 251.

9. Wentzell PD, Andrews DT, Hamilton DC, Faber K and Kowalski BR. Maximum likelihood principal component analysis. *J. Chemom.* 1997; **11**: 339

10. Bro R, Sidiropoulos ND and Smilde AK. Maximum likelihood fitting using simple least squares algorithms. *J. Chemom.* 2002; **16**: 387.

11. Carroll JD and Chang J. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika.* 1970; **35**: 283.

12. Rao CR and Mitra S. *Generalized Inverse of Matrices and its Applications.* Wiley: New York ,1971.

13. Magnus JR and Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley: Chichester, 1988.

14. Andersson CA and Bro R. Improving the speed of multi-way algorithms. Part I: Tucker3, *Chemom. Intell. Lab. Syst.* 1998; **42**: 93.

15. Bro R. and Smilde A.K. Centering and scaling in component analysis. *J. Chemom.* 2003; **17**: 16.

16. Ross RT and Leurgans S. Component resolution using multilinear models, *Methods Enzymol.* 1995; **246**: 679.

17. Wentzell PD and Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemom. Intell. Lab. Syst.* 1999; **45**: 65.

18. Bro R. Multi-way analysis in the food industry: models, algorithms and applications. PhD Thesis, Univerteit van Amsterdam. 1998.

19. Kruskal JB. Rank, decomposition, and uniqueness for 3-way and N-way arrays, In: Multiway data analysis. Elsevier: North-Holland,1989, 7, (Eds. R. Coppi, S. Bolasco).

20. Durell SR, Lee C, Ross RT and Gross EL. Factor analysis of the near-ultraviolet absorption spectrum of plastocyanyn using bilinear, trilinear and quadrilinear models. *Arch. Biochem. Biophys.* 1990; **278**: 148.

Chapter 3

Mathematical Improvements to Maximum Likelihood Parallel Factor Analysis: Theory and Simulations[2]

3.1 Abstract

A number of simplified algorithms for carrying out Maximum Likelihood Parallel Factor Analysis (MLPARAFAC) for three-way data affected by different error structures are described. The MLPARAFAC method was introduced to establish the theoretical basis to treat heteroscedastic and/or correlated noise affecting trilinear data. Unfortunately, the large size of the error covariance matrix employed in the general formulation of this algorithm prevents its application to solve standard three-way problems. The algorithms developed here are based on the principle of alternating least squares, but differ from the generalized MLPARAFAC algorithm in that they do not use equivalent alternatives of the objective function to estimate the loadings for the different modes. Instead, these simplified algorithms tackle the loss of symmetry of the PARAFAC model by using only one representation of the objective function to estimate the loadings of all of the modes. In addition, a compression step is introduced to allow the use of the generalized algorithm. Simulation studies carried out under a variety of measurement error conditions were used for statistical validation of the maximum likelihood properties of the algorithms and to assess the quality of the results and computation time. The simplified MLPARAFAC methods are also shown to produce more accurate results than PARAFAC under a variety of conditions.

---

[2] Submitted to *Journal of Chemometrics* (Submitted March 2005)

3.2 Introduction

Over the past three decades, the use of multivariate [1-2] and multi-way [3-5] methods have driven a change in the analytical laboratory from a univariate and chemically selective paradigm into a multivariate/multi-way and mathematically selective philosophy. Nonetheless, it was not until the 1990's that some researchers [6-11] started to consider in a consistent manner the nature of the noise corrupting these measurements in the context of multivariate analysis. The assumption of *iid* -normal (independent and identically distributed noise with a normal distribution) upon which univariate least squares methods [12] relied to provide optimal estimates was recognized as a limitation in the presence of other types of noise cases. The nature of the noise affecting multivariate measurements is strongly related to the nature of the experiment and the type of instrument employed [13], as well as different cosmetic manipulations [14-15] that make the noise deviate from the *iid* condition. Instrumental factors, such as spatial correlation in the detector sensors, detector response variation, source intensity instability, temperature fluctuations and physical variation in the sample and in the positioning of the sample within the instrument are a few examples of the causes of the existence of correlated noise.

In 1994, Paatero and Tapper [6] resurrected the idea of introducing some kind of weight information related to the uncertainty of the variables when the method Positive Matrix Factorization (PMF) was introduced. Unfortunately, this weighting information was only related to the variance of these variables, correcting for the violation of the identical distribution of the noise, but their method still assumed that the errors were independent from channel to channel. A more complete alternative was available a few years later when Wentzell *et al.* [7] formulated Maximum Likelihood Principal Component Analysis (MLPCA) which considered cases where the *iid* condition was completely violated due to the presence of heteroscedasticity and correlated noise. A principal innovation of this method was the use of the error covariance matrix (ECM), which is a more general way of describing the magnitude of the errors and the relationships among them. A few other closely related methods [8-10] have also been introduced to handle bilinear data in a maximum likelihood fashion, sometimes adding other constraints or information.

The application of this philosophy to multi-way data lagged behind the bilinear case until Bro *et al.* [11] introduced a generic method called Maximum Likelihood via Iterative Least Squares Estimation (MILES), which worked as a iterative preprocessing tool to condition the data from a maximum likelihood perspective in order that least squares methods such as PCA and PARAFAC could optimally handle the estimation process. The method is based on a majorization strategy in which the original objective function is substituted by a simpler and equivalent objective function in each step of the estimation process. Unfortunately, the simplicity of this numerical implementation is hindered by the amount of computation time needed. Since the method runs the full least squares optimization in each step, the time needed to get an estimate is sometimes excessive. Another important drawback of this approach is that the physical problem becomes obscured by the efficient but unfamiliar numerical methodology.

Recently, a method called Maximum Likelihood Parallel Factor Analysis (MLPARAFAC) was introduced to the chemometrics literature [16]. The main difference with respect to MILES is that MLPARAFAC is a method based solely on an alternating least squares (ALS) optimization. The implementation is straightforward and runs faster since the noise information is introduced in each iteration rather than in each optimization step as it is in MILES. The method was designed to estimate the parameters of the well-known PARAFAC model from a maximum likelihood perspective in cases where different violations of the assumed *iid*-normal error condition exist. Four algorithms for carrying out MLPARAFAC based on an ALS framework were described in this work. The simplest of these was designed to work with cases where the measurement errors are non-uniform (heteroscedastic) but uncorrelated. The most general form of the algorithm can treat data with any type of error covariance structure. Two simplifications of the general algorithm were also presented which more efficiently handle more restricted error covariance structures. All of the algorithms were shown to produce maximum likelihood estimates through a comparison of the distribution of the objective function with the χ^2 distribution. It was also shown that the quality of the estimated loading vectors for MLPARAFAC is significantly better than for the PARAFAC models in cases where the error covariance matrix is known.

Although the original paper on MLPARAFAC outlined the theory for dealing with correlated error, demonstrated its validity through simulation, and introduced some exact simplifications based on mathematical properties of the matrices used in the estimation process, it was found that many important situations remained uncovered and they are the subject of this paper. This work will be divided in two parts: the first part will introduce, test and apply the methodology to simulated data, while a companion paper will treat the application of MLPARAFAC to three experimental data sets. This paper will initially analyze the two simplifications introduced in the earlier work, since more interesting and useful simplifications can be found when those algorithmic alternatives are approached from a geometrical and computational point of view. This will lead us to the extension of one of these alternatives to more general cases where the noise structure along one order is less restricted and to cases were the error structure is correlated along two orders. Cases where the noise structure is correlated along more than two orders will need to be treated using the general algorithm, but since this is usually impractical from a computational point of view when the raw data are used, a compression approach will be introduced. After the algorithmic issues have been covered, a thorough analysis will be provided in order to go from these mathematically clear and well-defined cases to the more "grey" real cases. Also, some simulations will show the effects in the estimate when some cases with a considerable departure from the assumed structure are used with the corresponding simplification. Figure 3.1 shows the scenarios treated in this work.

3.2.1 Notation

In this paper, scalars are indicated by italics and vectors by bold lower-case characters. Bold upper-case letters are used for two-way matrices and underlined bold upper-case letters for three-way data. The letters **A**, **B** ,**C** and I, J, K are reserved for indicating the first, second and third mode of three-way data and the dimensions of those modes respectively. Also, the letter P is reserved to represent the number of factors used

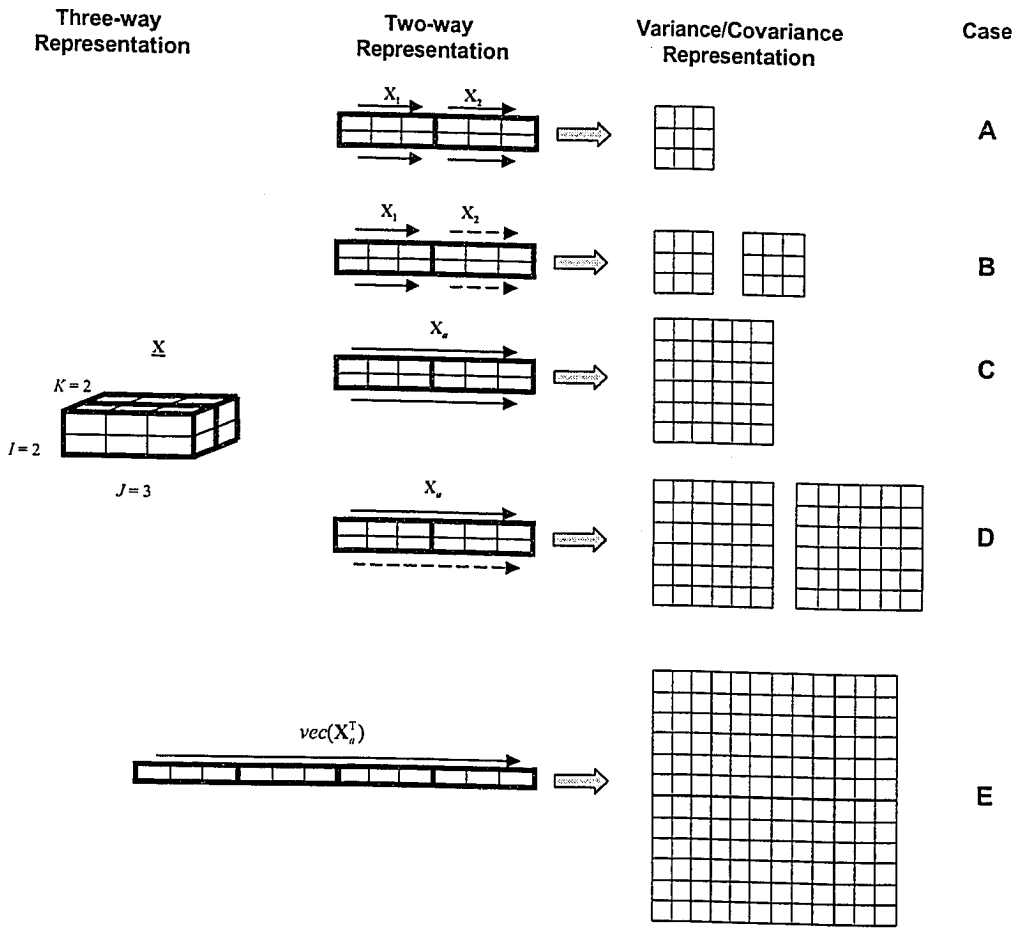

Figure 3.1. Illustration of the possible scenarios in which correlated errors might pervade a three-way array and the corresponding representations of the structure of the error covariance matrix to describe all the sources of variation. Arrows indicate which elements of the unfolded or vectorized three-way array have correlated errors. Different arrows represent different error structures.

in the model. The terms mode, way and order are used interchangeably, as well as the terms factors and components. When three-way arrays are unfolded to matrices, the following notation will be used. If $\underline{\mathbf{X}}$ (I x J x K) is unfolded while retaining the first order to produce a (I x JK) matrix, this will be designated $\mathbf{X}_a$. In the same way, matrices $\mathbf{X}_b$ (J x IK) and $\mathbf{X}_c$ (K x IJ) will be used to represent unfolded matrices which retain the second and the third orders, respectively. In general, other matrices with subscripts a, b and c represent unfolding while retaining the first, second and third modes. The use of subscripts i, j, k and p accompanying matrices and vectors refers to the use of the i, j, k and p – th slice or row of the corresponding data array or matrix. An important exception

to this notational rule is when subscripts i, j, k and p accompany matrix $\mathbf{I}$ in which case it refers to the identity matrix of order represented by the subscript. The use of superscript "-T" accompanying square matrices indicates that the inverse of the transpose of the corresponding matrix is calculated. The symbol "$\otimes$" will be used primarily to indicate the Kronecker product, but will also be used to represent the tensor product in certain cases which will be clearly distinguished. The symbol "$|\otimes|$" will be used to indicate the Khatri-Rao product, which is a compact version of a column-wise Kronecker product.

3.3 Theory

In the original paper introducing MLPARAFAC [16], it was noted that for many chemical applications, error covariance affects only one order, or at least the covariance in other orders can be neglected. This can, in certain cases, result in substantial simplification of the generalized algorithm. For the purpose of illustration, only the case where correlations exist along the rows will be considered, since correlations along other orders can be rendered equivalent through permutation of the original array or appropriate adjustment of equations introduced. For this case, three common situations can be distinguished: (1) the error covariance is different among all of the rows forming the array; (2) the error covariance is different among rows forming different slices but identical among the rows of the same slice; and (3) the error covariance is identical among the rows of all the slices. Simplifications for cases (2) and (3) were formulated based on mathematical identities and the more general scenario represented by case (1) was considered unsolvable by any simplification. Deeper scrutiny of these simplifications led the authors to realize the existence of more powerful and general simplifications for these cases. The next subsections will revisit these two simplifications from a geometrical and algorithmic point of view. One of these simplifications will be further extended to the case where error covariance is different among all of the rows forming the array and to the case where correlation is present along two modes.

3.3.1 Correlation along One Order

3.3.1.1 Case 1A

Imagine a trilinear data set such that the error correlation can be expected to affect only one order, which we will assume to be the second order. In addition, in certain cases where this assumption applies, it may be possible to make the additional assumption that the error covariance matrix is the same for each row in all the slices of data. Given that the observed data, $\underline{\mathbf{X}}$, can be considered the sum of the true data, $\underline{\mathbf{X}}^{\circ}$, and a array of measurement errors, $\underline{\mathbf{E}}$, this can be mathematically represented using any of the following three equations:

$$\begin{aligned} \underline{\mathbf{X}} &= \underline{\mathbf{X}}^{\circ} + \underline{\mathbf{E}} \\ \mathbf{X}_k &= \mathbf{X}_k^{\circ} + \mathbf{E} \\ \mathbf{x}_{i:k} &= \mathbf{x}_{i:k}^{\circ} + \mathbf{e} \end{aligned} \tag{3.1}$$

The trilinear data can be equivalently represented as a three-way array of elements, a slice by slice representation or as a vector representation, respectively. As mentioned above, all these representations are equivalent but only the last representation allows a clear representation of the characteristics of the noise, which follows a normal distribution around zero and with variance /covariance matrix Σ, $\mathbf{e} \sim N(\mathbf{0}, \Sigma)$. Since the errors are correlated, Σ cannot be expressed as a multiple of the identity matrix. This case is conceptually similar to the case treated by Brown *et al.* [15] in which bilinear data corrupted by drift noise were accommodated by applying an optimally designed filter. Therefore, we can consider our problem as a similar preprocessing problem in which each frontal slice k of data is multiplied by a filter as shown in Equation 3.2:

$$^{\mathrm{F}}\mathbf{X}_k = \mathbf{X}_k\mathbf{F} \tag{3.2}$$

$\mathbf{F}$ is an optimal filter matrix that will be applied to the data, and thus to the individual error vectors in each slice as shown in Equation 3.3:

$$^{\mathrm{F}}\mathbf{X}_k = (\mathbf{X}_k^{\circ} + \mathbf{E})\mathbf{F} = \mathbf{X}_k^{\circ}\mathbf{F} + \mathbf{E}\mathbf{F} \tag{3.3}$$

The error covariance matrix after filtering can be expressed as:

$$^{\mathrm{F}}\Sigma = E(\mathbf{F}^{\mathrm{T}}\mathbf{e}^{\mathrm{T}}\mathbf{e}\mathbf{F}) \tag{3.4}$$

Since the filter matrices are constant, they can be extracted from the expectation operator $E(\bullet)$ to obtain:

$$^{F}\mathbf{\Sigma} = \mathbf{F}^{T}E(\mathbf{e}^{T}\mathbf{e})\mathbf{F} = \mathbf{F}^{T}\mathbf{\Sigma}\mathbf{F} \tag{3.5}$$

F is an optimally constructed filter in the sense that it will rotate and scale the data yielding a new noise data, **EF**, which follows a normal distribution around zero with variance/covariance matrix equal to a multiple of the identity matrix, $^{F}\mathbf{\Sigma} = \sigma^{2}\mathbf{I}$. Therefore, dropping the proportionality constant (which can be viewed simply as a scaling factor) and substituting this equality into Equation 3.5 yields:

$$\mathbf{I} = \mathbf{F}^{T}\mathbf{\Sigma}\mathbf{F} \tag{3.6}$$

The filter matrix **F** which solves Equation 3.6 can be readily obtained considering the estimation process as an extended eigenproblem in which matrix Σ is initially rotated to yield a diagonal matrix that then goes through a scaling process producing the identity matrix. This linear transformation can only be executed when **F** is defined as the product of the eigenvectors of **F**, **U**, multiplied by the inverse of the diagonal matrix **S** formed by the square root of corresponding eigenvalues of **F** as shown in Equation 3.7.

$$\begin{aligned} \mathbf{F} &= \mathbf{U} * \mathbf{S}^{-1} \\ \mathbf{\Sigma} &= \mathbf{U}\mathbf{S}^{2}\mathbf{U}^{T} = \mathbf{U}\mathbf{S}\mathbf{S}\mathbf{U}^{T} \end{aligned} \tag{3.7}$$

It is worth noting that, even though the term filter has been used thus far, this optimal filter will not have the typical form of a least-squares polynomial filter such as the usual symmetric/antisymmetric band diagonal Savitzky-Golay filters [17]. In fact, it will not technically be a filter since no noise reduction is carried out. It can be better understood as a "modulator" which transforms the original signal corrupted by non-*iid* noise to a signal corrupted by *iid*-noise. This transformation not only affects the noise but also affects the imbedded true signal that is the aim of the estimation process. Fortunately, this transformation will not affect the trilinear structure, since all the slices are going to be rotated and scaled equivalently, as is evident from Equation 3.3. Additionally, uniqueness, which is one of the most appealing characteristics of trilinear data, will be preserved since the inverse transformation exists and can be easily applied to the

estimated loadings describing the order along which the noise is correlated. This is mathematically represented by Equation 3.8:

$$\begin{aligned} {}^{F}\hat{\mathbf{X}}_a &= \hat{\mathbf{A}}\,\mathbf{I}_a(\hat{\mathbf{C}}^{T}\otimes{}^{F}\hat{\mathbf{B}}^{T}) \\ \hat{\mathbf{B}} &= \mathbf{F}^{-1}\left({}^{F}\hat{\mathbf{B}}\right) \end{aligned} \tag{3.8}$$

The advantages of this approach with respect to the previous approach formulated in Table IV of reference 16 to treat this type of data optimally is two-fold. First, it will not be necessary to calculate the inverse of Ψ_a ($\Psi_a = \mathbf{I}_K \otimes \Sigma$) in order to estimate the parameters of the model since the error structure information is reflected in the data and not in the projection of the data. This is rather convenient, since Σ can be rank deficient for a variety of reasons. Second, the estimation procedure will be carried out using the standard PARAFAC algorithm, which is more stable and less computationally involved than the algorithm in Table IV of reference 16.

3.3.1.2 Case 1B

In addition to the simplest case treated above, Figure 3.1 represents a few other cases where the complexity of the error structure increases gradually up to the most complex case where the errors affecting all the elements of the multi-way data are related. Case B represented in Figure 3.1 takes the simplest error structure one step further to the case where noise is still correlated along one dimension but the structure and/or magnitude of it changes from slice to slice. The first reasonable approach to treat such a case might be to use the previously described strategy, utilizing in each case a filter matrix derived from the error covariance matrix obtained for each individual slice as shown in Equation 3.9:

$${}^{F}\mathbf{X}_k = \mathbf{X}_k\mathbf{F}_k \tag{3.9}$$

Equation 3.10 shows that the reasoning holds from a noise treatment perspective, since the local filtering will produce a diagonal matrix because the filter matrices are going to rotate and scale the original error covariance matrix for each slice in order to fulfill the *iid* condition.

$${}^{F}\mathbf{\Sigma}_k = \mathbf{F}_k^{T}E(\mathbf{e}_k^{T}\mathbf{e}_k)\mathbf{F}_k = \mathbf{F}_k^{T}\mathbf{\Sigma}_k\mathbf{F}_k \tag{3.10}$$

However, when this strategy is thoroughly explored via equation 3.11, it is clear that the "cleaning effect" produced over the noise has a negative collateral effect over the part of the data related to the chemical information since the trilinearity is destroyed by applying different rotation to the data in each slice.

$$^{F}\mathbf{X}_k = (\mathbf{X}_k^{\circ} + \mathbf{E}_k)\mathbf{F}_k = \mathbf{X}_k^{\circ}\mathbf{F}_k + \mathbf{E}_k\mathbf{F}_k \tag{3.11}$$

A solution based on a mathematical simplification of the full error covariance matrix was introduced in reference 16. The approach used to obtain this simplification was based on the idea of finding a simpler representation of the error covariance matrix to express the normal equations to estimate the loading for each mode. A relatively concise and computationally efficient formulation was found for the estimation of the loading for the modes A and C, but the equation for the estimation of B was still a function of the full error covariance matrix for this particular mode, as can be seen in equation 3.12:

$$vec(\hat{\mathbf{B}}^{T}) = (\mathbf{V}_b^{T}\boldsymbol{\Omega}_b^{-1}\mathbf{V}_b)^{-1}\mathbf{V}_b^{T}\boldsymbol{\Omega}_b^{-1}vec(\mathbf{X}_b^{T}) \tag{3.12}$$

Equivalently to the notation in reference 16, $\mathbf{V}_b$ is a *JP* x *IJK* matrix with $\mathbf{Z}_b^{T} = (\mathbf{C} \otimes \mathbf{A})^{T}$ repeating along the diagonal. The matrix $\boldsymbol{\Omega}_b$ is the full error covariance matrix for $vec(\mathbf{X}_b^{T})$, providing information about the error covariance among all the measurements. The presence of $\boldsymbol{\Omega}_b$ in this equation makes this simplification practically useless since its dimensions in a practical application will make the storage and manipulation for this equation prohibitive.

The lack of success of this approach can be attributed to the well-established strategy in standard PARAFAC in which the different estimation sub-steps are formulated using 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 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 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, in this paper, a new approach is 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 the sub-steps as minimization problems of the objective function written to preserve mode A alone. It is worth noting in advance that this alternative is laborious, since the equations will no longer be simple bilinear representations in which the mode to be determined is represented independently and the other two modes are represented as a composite mode, but as a more complex set of equations in which the modes are going to be interrelated most of the time.

We will start by showing the estimation of the normal equation for case B represented in Figure 3.1. For this particular case, we will be able to see how the equation obtained for mode A is exactly the same as the equation shown in reference 16 as proof that this strategy is equivalent to the standard strategy used in the past. Also, our goal will be accomplished by formulating a tractable equation for mode B, making equation 3.12 unnecessary. Even though the estimation of the loading for mode C was not particularly complex, the new strategy will provide a set of equations that is less demanding from a storage point of view. We start by defining the objective function as equation 3.13:

$$f = \sum_{k=1}^{K} trace[(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^T)\mathbf{\Psi}_k^{-1}(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^T)^T] \tag{3.13}$$

In this equation, $\mathbf{X}_k$ represents the k-th slice of the three-way array $\underline{\mathbf{X}}$, $\mathbf{A}$ and $\mathbf{B}$ are matrices of dimensions I x P and J x P representing the loading vectors for mode A and B respectively, $\mathbf{D}_k$ is a P x P diagonal matrix with the k-th row of the K x P matrix $\mathbf{C}$ along the diagonal and $\mathbf{\Psi}_k^{-1}$ is the inverse of the error covariance matrix that describes the noise affecting all the rows of the k-th slice of the three-way array $\underline{\mathbf{X}}$. The implementation of an alternating least squares algorithm for the estimation of mode A loadings assumes B and C are known and then equation 3.13 is minimized with respect to each element forming $\mathbf{A}$. Before proceeding with the derivation, it will be convenient to express equation 3.13 as the quadratic form shown in equation 3.14 where $\mathbf{M}_k = \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^T$:

$$f = \sum_{k=1}^{K} trace(\mathbf{M}_k\, \mathbf{\Psi}_k^{-1}\mathbf{M}_k^T) \tag{3.14}$$

equation 3.15 shows the derivation:

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{A}_{ip}} &= \frac{\partial f}{\partial \mathbf{M}_k}\left(\frac{\partial \mathbf{M}_k}{\partial \mathbf{A}_{ip}}\right)^{\mathrm{T}} \\
&= \sum_{k=1}^{K} trace\left(2\mathbf{M}_k \boldsymbol{\Psi}_k^{-1} \frac{\partial \mathbf{M}_k^{\mathrm{T}}}{\partial \mathbf{A}_{ip}}\right) \\
&= \sum_{k=1}^{K} trace\left(2\mathbf{M}_k \boldsymbol{\Psi}_k^{-1} \frac{\partial (\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}}{\partial \mathbf{A}_{ip}}\right) \\
&= \sum_{k=1}^{K} trace\left(2\mathbf{M}_k \boldsymbol{\Psi}_k^{-1}(-\mathbf{E}_{ip}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}\right) \\
&= \sum_{k=1}^{K} trace\left(2(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_k^{-1}(-\mathbf{E}_{ip}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}\right) \\
&= \sum_{k=1}^{K} trace\left(-2\mathbf{X}_k \boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}} + 2\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}}\right) \\
&= -2\sum_{k=1}^{K} trace\left(\mathbf{X}_k \boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}}\right) + 2\sum_{k=1}^{K} trace(\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}})
\end{aligned} \tag{3.15}$$

Equation 3.15 represents the first derivative of the objective function with respect to the elements of **A**. The matrix $\mathbf{E}_{ip}$ is an elementary I x P matrix with all of its elements equal to zero with the exception to the element located in the position i x p, which is equal to 1. This equation will be equal to zero for the optimum value of $\mathbf{A}_{ip}$ given **B** and **C**. In order to calculate this value of $\mathbf{A}_{ip}$ equation 3.15 is transformed as follows:

$$\begin{aligned}
&\sum_{k=1}^{K} trace\left(\mathbf{X}_k \boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}}\right) = \sum_{k=1}^{K} trace(\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}}) \\
&\sum_{k=1}^{K} vec(\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{X}_k^{\mathrm{T}})^{\mathrm{T}} vec(\mathbf{E}_{pi}) = \sum_{k=1}^{K} vec(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} vec(\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\mathbf{E}_{ip}^{\mathrm{T}}) \\
&\sum_{k=1}^{K} vec(\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{X}_k^{\mathrm{T}})^{\mathrm{T}} vec(\mathbf{E}_{pi}) = \sum_{k=1}^{K} vec(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} (\mathbf{I}_I \otimes \mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k) vec(\mathbf{E}_{pi}) \\
&vec\left(\sum_{k=1}^{K} (\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{X}_k^{\mathrm{T}})\right)^{\mathrm{T}} vec(\mathbf{E}_{pi}) = vec(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} \sum_{k=1}^{K} (\mathbf{I}_I \otimes \mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k) vec(\mathbf{E}_{pi})
\end{aligned} \tag{3.16}$$

Equation 3.16 is one of the *IP* equations necessary to estimate the loadings of **A**. The rest of the equations are obtained as the right and left parts of this equation are multiplied by the different vectorized $\mathbf{E}_{pi}$ matrices. Since this term is completely independent in both

sides of the equation, the process can be carried out in a straightforward manner using a matrix $\mathbf{E}$ formed as $[vec(\mathbf{E}_{11})\ vec(\mathbf{E}_{21})\ \ldots\ vec(\mathbf{E}_{IP})]$. A closer examination of this matrix reveals that $\mathbf{E}$ is the identity matrix of order IP, making the multiplication theoretically sound but numerically unnecessary and providing equation 3.17 to estimate the loadings of mode A:

$$vec(\mathbf{A}^{\mathrm{T}}) = \left(\sum_{k=1}^{K}(\mathbf{I}_I \otimes \mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k)\right)^{-1} vec\left(\sum_{k=1}^{K}(\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{X}_k^{\mathrm{T}})\right) \tag{3.17}$$

Taking into consideration the properties of the *vec* operator and the Kronecker product, equation 3.17 can be transformed to:

$$\mathbf{A} = \sum_{k=1}^{K}\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\left(\sum_{k=1}^{K}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{D}_k\right)^{-1} \tag{3.18}$$

For this scenario, equation 3.18 is a more compact and computationally efficient representation of the equivalent equation 30 in reference 16 and reproduced here as equation 3.19.

$$\mathbf{A} = \mathbf{X}_a\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}}(\mathbf{Z}_a\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}})^{-1} \tag{3.19}$$

The summations over k found in equation 3.18 can be eliminated by using the unfolded representation of $\underline{\mathbf{X}}$ retaining mode A ($\mathbf{X}_a$) and by expressing Ψ_a as the block diagonal error covariance matrix with the individual error covariance matrices for each slice along the diagonal and expressing the projection space by $\mathbf{Z}_a = \mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}}$.

For convenience, the mathematical procedure to derive the estimation equations for the rest of the loadings in this and the rest of the different scenarios are provided in Appendix I. The results of these derivations are given below.

$$vec(\mathbf{B}) = \left(\sum_{k=1}^{K}(\mathbf{D}_k\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{D}_k \otimes \boldsymbol{\Psi}_k^{-1})\right)^{-1} vec\left(\sum_{k=1}^{K}(\mathbf{D}_k\mathbf{A}^{\mathrm{T}}\mathbf{X}_k\boldsymbol{\Psi}_k^{-1})\right) \tag{3.20}$$

$$\mathbf{c}_k = \left[\left(\mathbf{B}'\boldsymbol{\Psi}_k^{-1}\mathbf{B}\otimes\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1} vec\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\right)\right]^{\mathrm{T}}\mathbf{E} \tag{3.21}$$

It is interesting to note how both equations are composed of the two key parts of a standard weighted least squares estimator: a projection matrix spanning the space where the best approximation of the noiseless signal is located and a vector representing a

weighted projection of the data onto the space where the signal is located. The awkward form of these two components is a consequence of the manner in which these equations were obtained, as anticipated at the beginning of this section. Two important details have to be mentioned for the expression to obtain loadings for mode C: vector $\mathbf{c}_k^T$ represents the k-th row of the K x P matrix $\mathbf{C}$ and the matrix $\mathbf{E}$ in this case is a matrix formed as $[vec(\mathbf{E}_{11})\ vec(\mathbf{E}_{22})\ \ldots\ vec(\mathbf{E}_{PP})]$ and is used to choose the necessary elements for the estimation of $\mathbf{c}_k$, since the optimization was originally designed to have this vector along the main diagonal of $\mathbf{D}_k$.

3.3.2 Correlation along Two Orders

3.3.2.1 Case 1C

Figure 3.1C represents cases where the error structure becomes more complex by affecting elements of the data set located in two different modes. Such types of scenarios are not unusual, for example in kinetic studies where the course of the reaction is followed spectroscopically giving rise to errors that are correlated in both the time and wavelength modes, while the other mode may be composed of samples with different compositions of the reactants that are run independently of one another. For this case, we will consider that the three-way data $\underline{\mathbf{X}}$, will be unfolded preserving the samples of different compositions in mode A, while modes B and C will be combined in one composite mode formed by the spectral information and the time information for each sample. The objective function in this case can be expressed as shown in equation 3.22.

$$\begin{aligned} f &= trace[(\mathbf{X}_a - \mathbf{A}\mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^T)\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a - \mathbf{A}\mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^T)^T] \\ &= trace[(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^T)\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^T)^T] \end{aligned} \tag{3.22}$$

As mentioned in the notation section, the variables with the "a" subscript such as $\mathbf{X}_a$ and $\mathbf{I}_a$ represent the three-way arrays $\underline{\mathbf{X}}$ and $\underline{\mathbf{I}}$ unfolded preserving mode A independently. Array $\underline{\mathbf{I}}$ is P x P x P with all the elements equal zero but those on the superdiagonal, which are equal to unity. A small modification was made in the second expression in equation 3.22 to make it more compact by expressing $\tilde{\mathbf{A}}$ as the product of $\mathbf{A}$ by $\mathbf{I}_a$. Equation 3.22 will be used only to obtain the loadings for modes B and C, since the loadings for mode A can be obtained by equation 3.19. It is important to anticipate that

the expressions obtained are not going to have the visual clarity to be interpreted as equation 3.19 due to the manner in which they were obtained. The expression for the estimation of the loadings **B** for this noise characteristic is shown in equation 3.23:

$$vec(\mathbf{B}^{\mathrm{T}}) = \left(\sum_{m=1}^{K} \sum_{n=1}^{K} (\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes \mathbf{L}_{mn}) \right)^{-1} \left(\sum_{m=1}^{K} \sum_{n=1}^{K} (\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes \mathbf{R}_{mn}) \right) vec(\mathbf{I}_J) \tag{3.23}$$

Equations 3.24 and 3.25 show the expressions to calculate matrices **R** and **L**, respectively.

$$\mathbf{L} = (\mathbf{C} \otimes \mathbf{I}_P) \tilde{\mathbf{A}}^{\mathrm{T}} \tilde{\mathbf{A}} (\mathbf{C} \otimes \mathbf{I}_P)^{\mathrm{T}} \tag{3.24}$$

$$\mathbf{R} = (\mathbf{C} \otimes \mathbf{I}_P) \tilde{\mathbf{A}}^{\mathrm{T}} \mathbf{X}_a \tag{3.25}$$

It should be noted that in order to obtain equation 3.23, a number of manipulations of the different matrices involved in the estimation process are performed as shown in the Appendix. The most remarkable manipulation the reader must be aware of in order to understand equation 3.23 is the partitioning of the *JK* x *JK* inverse error covariance matrix $\boldsymbol{\Psi}^{-1}$, the *KP* x *KP* matrix **L** and the *KP* x *KJ* matrix **R** into three *K* x *K* super-matrices composed of the corresponding *J* x *J*, *P* x *P* and *P* x *J* matrices. A graphical representation is presented in equation A.19 in the Appendix. It is clear from the equation that subscripts *m* and *n* indicate the use of different partitioned pieces of $\boldsymbol{\Psi}^{-1}$, **L** and **R**. Although equation 3.23 does not resemble the traditional representation of a weighted least squares estimator, a closer look will actually indicate, as before, that it is formed by the key pieces of this type of estimator: a projection matrix spanning the space where the vector to be estimated resides (the term within the inverse operator) and a weighted image of the signal in the same space (the term following the inverse operator). The equation to estimate the loadings for mode C are equivalently obtained and will have a similar structure as can be seen in equation 3.26:

$$vec(\mathbf{C}^{\mathrm{T}}) = \left(\sum_{m=1}^{P} \sum_{n=1}^{J} (\mathbf{T}_{mn} \otimes \mathbf{S}_{nm}^{\mathrm{T}}) \right)^{-1} \left(\sum_{m=1}^{J} \sum_{n=1}^{J} (\mathbf{L}_{mn}^{\mathrm{T}} \otimes \mathbf{R}_{nm}) \right) vec(\mathbf{I}_K) \tag{3.26}$$

Equations 3.27 to 30 show the necessary expressions to calculate the matrices involved in equation 3.26. As before, matrices $\mathbf{K}_{JK}$ and $\mathbf{K}_{KJ}$ are *JK* x *KK* commutation matrices.

$$\mathbf{T} = (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{I}_K) \mathbf{K}_{JK} \boldsymbol{\Psi}_a^{-1} \mathbf{K}_{KJ} \tag{3.27}$$

$$\mathbf{S} = (\mathbf{B} \otimes \mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}} \tag{3.28}$$

$$\mathbf{L} = \boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ} \tag{3.29}$$

$$\mathbf{R} = (\mathbf{B} \otimes \mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a \tag{3.30}$$

Again, the subscripts *m* and *n* indicate the use of different partitioned pieces of the full matrices previously shown. In all cases, we have tried to produce the most compact representation for the expression used to calculate the estimates, but it is possible that further simplifications have been unnoticed by the authors. Also, some of these expressions will be computationally implemented in a more efficient way than the one used here, which was preferred for its notational simplicity.

3.3.2.2 Case 1D

Figure 3.1D represents chemical scenarios that are very similar to the previous case. The complexity of the system is taken a step further by considering that the noise propagates in a correlated fashion along two modes but the structure of this correlated noise changes from sample to sample independently. This type of situation is not uncommon when spectroscopic techniques susceptible to path length variations such as NIR spectroscopy are used. Mathematically, the trilinear errors-in-variable model best suited to describe these data can be obtained by minimizing equation 3.31:

$$f = \sum_{i=1}^{I} ({}^{i}\mathbf{x}_a - {}^{i}\tilde{\mathbf{a}}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}})\,{}^{i}\boldsymbol{\Psi}_a^{-1}({}^{i}\mathbf{x}_a - {}^{i}\mathbf{a}(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}})^{\mathrm{T}} \tag{3.31}$$

This objective function yields expressions for the estimates that are very similar to the previous case, but in this particular case the estimates are obtained in a row by row fashion for mode A, and as a summation over the *I* objects in mode A for modes B and C, as can be seen in equations 3.32 to 34:

$${}^{i}\mathbf{a} = {}^{i}\mathbf{x}_a\,{}^{i}\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}}(\mathbf{Z}_a\,{}^{i}\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}})^{-1} \tag{3.32}$$

$$vec(\mathbf{B}^{\mathrm{T}}) = \left(\sum_{i=1}^{I}\sum_{m=1}^{K}\sum_{n=1}^{K}({}^{i}\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes {}^{i}\mathbf{L}_{mn})\right)^{-1}\left(\sum_{i=1}^{I}\sum_{m=1}^{K}\sum_{n=1}^{K}({}^{i}\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes {}^{i}\mathbf{R}_{mn})\right)vec(\mathbf{I}_J) \tag{3.33}$$

$$vec(\mathbf{C}^{\mathrm{T}}) = \left(\sum_{i=1}^{I} \sum_{m=1}^{P} \sum_{n=1}^{J} ({}^{i}\mathbf{T}_{mn} \otimes {}^{i}\mathbf{S}_{nm}^{\mathrm{T}}) \right)^{-1} \left(\sum_{i=1}^{I} \sum_{m=1}^{J} \sum_{n=1}^{J} ({}^{i}\mathbf{L}_{mn}^{\mathrm{T}} \otimes {}^{i}\mathbf{R}_{nm}) \right) vec(\mathbf{I}_K) \quad (3.34)$$

The estimations of the loadings for mode B and C will use the same equations shown before, but in all cases the matrix $\tilde{\mathbf{A}}$ will be replaced by the corresponding row vector ${}^{i}\tilde{\mathbf{a}}$ and the I x JK matrix $\mathbf{X}_a$ will be replaced by the row vector ${}^{i}\mathbf{x}_a$. It is important to emphasize that a set of I error covariance matrices of dimensions JK x JK will be used by this method making this alternative very expensive from a storage and computational point of view.

Thus far in section 3.2, a number of different simplified scenarios have been examined, ranging from the simplest, where the error covariance matrix can be fully represented by a J x J matrix, to the most complex case, where it is necessary to consider I different JK x JK error covariance matrices. From the estimation equations, it is evident that the computational effort and the storage space increase as the complexity of the error structure characterizing the noise affecting the data grows. Therefore, the main advantage of using a simpler alternative will be the reduction of time needed to estimate the loadings for each mode. On the other hand, some scenarios will show the merit of using the more complex alternatives in order to provide the maximum likelihood estimation for each mode. The situation in which practioners will have to compromise to estimate the best possible errors-in-variables model using the minimum amount of time will depend on the characteristics of the data at hand and will be difficult to assess on an *a priori* basis. In the experimental section of this paper, a number of simulated data sets are used to validate the statistical properties of these algorithms and also to show the advantages of using one algorithm over the other in terms of time, computational power and quality of the results.

3.3.3 Correlation along Three Orders

In the previous sections, the expressions for a number of simplified algorithms were derived for a variety of scenarios characterized by error covariance matrices of different complexity. However, there are going to be cases where none of these simplifications will provide the best solution, making it necessary to use the full algorithm presented in reference 16. As noted in that work, the full algorithm is not a

viable alternative except when the dimensions of each order are unrealistically small. This is also the case with some of the simplifications discussed here (*e.g.* case 1D) for which the amount of storage space is prohibitive from a practical point of view. In these cases, some compression methods, taking advantage of different intrinsic levels of structure present within the data, will be introduced to tackle the situation. This section provides the theoretical basis of this approach and describes the implementation in the context of the model using the full error covariance matrix, although it is important to note that this can also be applied to some of the simplified models previously discussed.

3.3.3.1 Compression

Compression is a natural concept for two-way and multi-way data since both types of data can model deterministic relationships among variables, especially in cases where a high degree of collinearity and multilinearity exist. These types of data can be represented by a smaller number of variables. Using this smaller set of variables, the data can be described within experimental error as a P-dimensional hyperplane. In this case, P is called the chemical rank or pseudorank of the data set in order to distinguish it from the mathematical rank. In general, the chemical rank is typically related to the number of underlying chemical factors or chemical components present in the mixture. For multi-way data, the theoretical basis of the idea was initially introduced by Carroll *et al.* [19] in 1980, stating the optimality theorem of the Canonical Decomposition with Linear Constraints (CANDELINC) model, which ensures that the compressed array preserves the original variation maximally when a set of orthogonal bases, usually Tucker3 factors, are used to project the original array onto the space spanned by them. In 1981, Appellof and Davidson [20] provided the first application of trilinear decomposition to chemistry using both simulated and real LC/emission/excitation measurements by compressing the original data. They used the scores provided by the principal component decomposition of the unfolded data in each mode as compression bases. Later, Alsberg and Kvalheim published a number of papers [21-22] proposing a method called Postponed Basis Matrix Multiplication (PBM) using B-spline basis sets for the compression of high dimensional arrays. A comparative study done by Kiers and Harshman [23] proved that PBM is equivalent to the more general approach based on the CANDELINC model. They also

stressed that there is no need for special algorithms in the CANDELINC approach, showing it was only necessary to compress the array using a select set of optimal bases, to use any existing multi-way algorithm on the compressed array, and to decompress the result by post-multiplying the solution with the bases. The latest additions to the arsenal of compression basis sets have been a variety of wavelet families of basis sets not only used as compression devices but also as smoothing and denoising alternatives [24]. It is worth noting that these positive side effects commonly attributed to the compression using wavelets are not completely an intrinsic characteristic of the basis set, but a consequence of the projection step involved in the compression procedure.

From a structural point of view, the possibility of using different basis sets such as Tucker3 factors, PCA factors, B-splines and wavelets is a consequence of the different levels of underlying structure present in the chemical part of any multi-way data. The type of data encountered in chemistry is normally collinear (well suited for B-splines and wavelets), bilinear (ideally treated by PCA) and, in many cases, trilinear (where Tucker3 basis sets are the perfect option). This idea will be clearly demonstrated from a mathematical point of view throughout the theoretical development of an example shown next.

An I x J x K array $\underline{\mathbf{X}}$ is given, such that matrices **U** (I x D), **V** (J x E) and **Z** (K x F), representing orthogonal basis for the systematic variation in the first, second and third mode respectively, are considered known. Dimensions D, E and F are the pseudo-rank (*i.e.* the rank of the subspace spanning the systematic variation when noise is not present [25]) for each mode. It is important to clarify that matrices **U**, **V** and **Z** as well as ranks D, E and F must be estimated beforehand, but in this case, for the sake of illustration, will be considered known. The standard estimation of the PARAFAC model can be expressed via equation 3.36:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathbf{X}_a - \mathbf{A}\mathbf{I}_a(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{3.36}$$

The CANDELINC optimality theorem expresses the existence of three matrices Δ, Θ and Φ of orders (D x P), (E x P) and (F x P) that are related to **A**, **B** and **C** through a bilinear relationship with **U**, **V** and **Z** as shown in equation 3.37:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\Delta \\ \mathbf{B} &= \mathbf{V}\Theta \\ \mathbf{C} &= \mathbf{Z}\Phi \end{aligned} \tag{3.37}$$

From a geometric point of view, this is equivalent to saying that each mode is linearly constrained to sub-spaces **U**, **V**, and **Z**. Therefore, if the minimization problem represented by equation 3.36 is to be solved subject to the constraints expressed by equation 3.37, it is only necessary to estimate the much smaller matrices Δ, Θ and Φ using the smaller array $\underline{\mathbf{Y}}$ of order D x E x F obtained after the projection. Mathematically, this is carried out by projecting $\underline{\mathbf{X}}$ onto the space spanned by **U**, **V** and **Z** as shown in equation 3.38:

$$\hat{\mathbf{X}}_a = \mathbf{U}\mathbf{U}^{\mathrm{T}}\mathbf{X}_a(\mathbf{Z}\mathbf{Z}^{\mathrm{T}} \otimes \mathbf{V}\mathbf{V}^{\mathrm{T}}) \tag{3.38}$$

Using equation 3.38, array $\underline{\mathbf{Y}}$ can be defined as:

$$\mathbf{Y}_a = \mathbf{U}^{\mathrm{T}}\mathbf{X}_a(\mathbf{Z} \otimes \mathbf{V}) \tag{3.39}$$

Equation 3.39 coincides with the expression used to calculate the core matrix for the Tucker3 model [26] when matrices **U**, **V** and **Z** represent the respective modes for this model. This is a clear mathematical proof to demonstrate the earlier statement indicating Tucker3 loadings as the perfect basis set for compression of multi-way data. As mentioned before, array **Y** can be used to estimate Δ, Θ and Φ and, using the expressions depicted in equation 3.37, the loadings in the original space can be calculated as the standard estimation problem depicted in equation 3.36, which is reduced to the one represented in equation 3.40:

$$\min_{\Delta,\Theta,\Phi} \left\| \mathbf{Y}_a - \Delta \mathbf{I}_a (\Phi \otimes \Theta)^{\mathrm{T}} \right\|_{\mathrm{F}}^2 \tag{3.40}$$

Thus far, it has been demonstrated why Tucker3 provides the best basis set for compression. In addition to the method of choice for the compression basis set, another key piece of information is the dimensions for corresponding basis set. In general, compression will provide an approximate solution, although it has been reported in the literature [27] that, in situations where only one mode is high dimensional, an exact compression can be obtained by compressing this mode with a basis set of dimension equal to the product of dimensions of the other smaller two orders. In reality, exact

compression can be considered the exception instead of the rule. No formal theory exists to choose the number of components for each Tucker3 loading. A rule of thumb is to use at least five more components than the number of components expected for the system, since the main objective is to speed up the algorithm and therefore only information related to the chemical structure is needed.

Up to this point, the theory and most important equations for the compression and estimation of multi-way data to be treated with the standard PARAFAC and other multi-way models such as PARAFAC2 and PARATUCK2 have been introduced. However, when this philosophy is to be extended for cases where a maximum likelihood method such as MLPAPAFAC is to be used, a few other equations must be introduced. These new equations will lead us to issues related to the selection, calculation and number of basis sets needed for this approach.

Even though compression can be applied to any of the simplified scenarios, we will treat here the case where the full error covariance matrix must be used. The expression used to compress the full error covariance matrix is a direct extension of the projection expression shown in equation 3.39 in a vectorized form:

$$\boldsymbol{\Xi}_a = (\mathbf{U} \otimes \mathbf{Z} \otimes \mathbf{V})^{\mathrm{T}} \boldsymbol{\Omega}_a (\mathbf{U} \otimes \mathbf{Z} \otimes \mathbf{V}) \tag{3.39}$$

Equation 3.39 will convert the original *IJK* x *IJK* full error covariance matrix describing the noise structure present in the original array $\underline{\mathbf{X}}$ in a compressed *DEF* x *DEF* full error covariance matrix describing the noise in the compressed array $\underline{\mathbf{Y}}$. Although equation 3.39 represents the theoretical expression to compress the error covariance matrix, it does not solve the size problem associated with it. In order to solve this problem, the compression step must be carried out on the original data and the compressed arrays used to calculate the compressed error covariance matrix. These alternatives are equivalent, as can be seen in equation 3.40, where equation 3.39 is used as starting point in a backward transformation.

$$
\begin{aligned}
\boldsymbol{\Xi}_a &= (\mathbf{U}\otimes\mathbf{Z}\otimes\mathbf{V})^{\mathrm{T}}\boldsymbol{\Omega}_a(\mathbf{U}\otimes\mathbf{Z}\otimes\mathbf{V}) \\
&= (\mathbf{U}\otimes\mathbf{Z}\otimes\mathbf{V})^{\mathrm{T}}E\left(vec(\mathbf{E}_a^{\mathrm{T}})vec(\mathbf{E}_a^{\mathrm{T}})^{\mathrm{T}}\right)(\mathbf{U}\otimes\mathbf{Z}\otimes\mathbf{V}) \\
&= E\left((\mathbf{U}\otimes\mathbf{Z}\otimes\mathbf{V})^{\mathrm{T}}vec(\mathbf{E}_a^{\mathrm{T}})vec(\mathbf{E}_a^{\mathrm{T}})^{\mathrm{T}}(\mathbf{U}\otimes\mathbf{Z}\otimes\mathbf{V})\right) \\
&= E\left(vec\left\{\left[\mathbf{U}^{\mathrm{T}}\mathbf{E}_a(\mathbf{Z}\otimes\mathbf{V})\right]^{\mathrm{T}}\right\}vec\left\{\left[\mathbf{U}^{\mathrm{T}}\mathbf{E}_a(\mathbf{Z}\otimes\mathbf{V}\right]^{\mathrm{T}}\right\}^{\mathrm{T}}\right) \\
&= E\left(vec(\mathbf{N}_a^{\mathrm{T}})vec(\mathbf{N}_a^{\mathrm{T}})^{\mathrm{T}}\right)
\end{aligned}
\tag{3.40}
$$

It is important to differentiate in equation 3.40, the expression $E(\bullet)$, which represents the expectation value of the expression in parenthesis from expression, $\mathbf{E}_a$ which represents the noise array $\underline{\mathbf{E}}$ unfolded as an I x JK matrix. Expression 3.40 represents the symmetric outer product of the multiplication of the unfolded error array and the compression basis set in vector form. This can be transformed to the following matrix expression to be further explored:

$$
\begin{aligned}
\mathbf{N}_a &= \mathbf{U}^{\mathrm{T}}\mathbf{E}_a(\mathbf{Z}\otimes\mathbf{V}) \\
&= \mathbf{U}^{\mathrm{T}}\left(\mathbf{X}_a - \mathbf{X}_a^{\circ}\right)(\mathbf{Z}\otimes\mathbf{V}) \\
&= \mathbf{U}^{\mathrm{T}}\mathbf{X}_a(\mathbf{Z}\otimes\mathbf{V}) - \mathbf{U}^{\mathrm{T}}\mathbf{X}_a^{\circ}(\mathbf{Z}\otimes\mathbf{V})
\end{aligned}
\tag{3.41}
$$

Here, $\mathbf{X}_a$ and $\mathbf{X}_a^{\circ}$ are the unfolded forms of the measured data array and the error-free data array, respectively. Equation 3.41 shows that a successful estimation of the noise in the compressed space can be obtained if the compression basis sets are chosen to optimally compress the chemical part represented by $\mathbf{X}_a^{\circ}$. Two detrimental effects can be foreseen if the chosen basis set does not span the space of $\mathbf{X}_a^{\circ}$ properly. The first is related to the loss of meaningful chemical information during the projection step and it is common to PARAFAC and MLPARAFAC. The second is a direct consequence of the first one and related to the inclusion of chemical variability in the error covariance matrix as if it were noise. Clearly, the second detrimental effect will only affect MLPARAFAC since PARAFAC does not use any noise information. In order to prevent these effects when compression is used with MLPARAFAC, it is necessary to retain as much variation as possible. This alternative is not advisable for PARAFAC, since including a large amount of variation can increase the uncertainty of the estimates, but in the case of MLPARAFAC there is no danger of this, since this meaningless variation (noise) will be down-weighted via the error covariance matrix during the estimation process.

It is well known that, in practice, $\mathbf{X}_a^{\circ}$ is not generally available, hence in the absence of *a priori* knowledge, the error-free data array is replaced by its best unbiased estimate (considering the normal assumption), which is the average array $\overline{\mathbf{X}}_a$ calculated by obtaining replicates of the measurements. For practical applications, equation 3.41 becomes equation 3.42:

$$\begin{aligned}\mathbf{N}_a &= \mathbf{U}^{\mathrm{T}}\mathbf{X}_a(\mathbf{Z}\otimes\mathbf{V}) - \mathbf{U}^{\mathrm{T}}\overline{\mathbf{X}}_a(\mathbf{Z}\otimes\mathbf{V}) \\ &= \mathbf{Y}_a - \overline{\mathbf{Y}}_a\end{aligned} \tag{3.42}$$

Equation 3.42 also unveils another important practical issue regarding the selection of the compression basis sets, indicating that the optimal basis set will be obtained as a Tucker3 decomposition of the mean array $\underline{\mathbf{X}}$. The compressed error covariance matrix Ξ_a will be calculated using a set of R replicates as shown in equation 3.43:

$$\Xi_a \approx \frac{1}{(R-1)}\sum_{r=1}^{R}\left[vec\left\{(\mathbf{Y}_a^r - \overline{\mathbf{Y}}_a)^{\mathrm{T}}\right\}vec\left\{(\mathbf{Y}_a^r - \overline{\mathbf{Y}}_a)^{\mathrm{T}}\right\}^{\mathrm{T}}\right] \tag{3.43}$$

Based on the theoretical expressions derived in this section, a sequence of steps to prepare the data for the most general MLPARAFAC algorithm is shown in Table 3.1.

It is important to note that, although this strategy was explained for the compression of all three orders, it can also be applied to the compression of one or two orders in a very straightforward manner. For example, if only mode A is compressed, equation 3.39 will become equation 3.44, since in that case **Z** and **V** will be the identity matrix of orders J and K respectively and $(\mathbf{Z}\otimes\mathbf{V}) = \mathbf{I}_{JK}$.

$$\mathbf{Y}_a = \mathbf{U}^{\mathrm{T}}\mathbf{X}_a \tag{3.44}$$

This result is equivalent and symmetric for all the orders. Therefore, if an order different from A is to be compressed, the data will be unfolded, keeping the desired order unmodified, and multiplied by the optimal base describing this order. Equivalently, if more than one order needs to be compressed, this methodology can be individually repeated for both orders, including a folding and unfolding intermediate step between the multiplication by each basis set.

In the experimental part of this paper, a number of simulated data sets will be used to test the performance of the compression approach under different conditions, such as the level of noise and the amount of structure in the chemical data. Also, a comparative study between Tucker3 and PCA loadings will be carried out to confirm the theoretical results.

Table 3.1. Algorithm for the MLPARAFAC algorithm using compression.

1. Given R replicates of an I x J x K cube of data $\underline{\mathbf{X}}$. The algorithm starts by calculating a Tucker3 model for the average cube of data, $\overline{\underline{\mathbf{X}}}$.

$$[\mathbf{U},\mathbf{V},\mathbf{Z},\overline{\underline{\mathbf{Y}}}] = tucker3(\overline{\underline{\mathbf{X}}},P) \tag{T.1}$$

2. For each replicate, unfold $\underline{\mathbf{X}}^r$, retain the first order and regress $\mathbf{X}_a^r$ onto the subspace spanned by $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{Y}$ in order to calculate $\mathbf{Y}_a^r$ for each replicate as shown in equation T.2:

$$\mathbf{Y}_a^r = \mathbf{U}^{\mathrm{T}}\mathbf{X}_a^r(\mathbf{Z}\otimes\mathbf{V}) \tag{T.2}$$

Using all the $\mathbf{Y}_a^r$, estimate the error covariance matrix in the compressed subspace, represented by Ξ_a in equation T.3.

$$\Xi_a \approx \frac{1}{(R-1)}\sum_{r=1}^{R}\left[vec\left\{(\mathbf{Y}_a^r-\overline{\mathbf{Y}}_a)^{\mathrm{T}}\right\}vec\left\{(\mathbf{Y}_a^r-\overline{\mathbf{Y}}_a)^{\mathrm{T}}\right\}^{\mathrm{T}}\right] \tag{T.3}$$

3. Submit Ξ_a and each $\mathbf{Y}_a^r$ to the MLPARAFAC algorithm previously introduced until convergence is achieved.

$$[\Delta^r,\Theta^r,\Phi^r] = MLPARAFAC(\underline{\mathbf{Y}}^r,\Xi_a,P) \tag{T.4}$$

4. Using the following relationships, the uncompressed MLPARAFAC loadings can be obtained.

$$\begin{aligned}\mathbf{A}^r &= \mathbf{U}\Delta^r\\ \mathbf{B}^r &= \mathbf{V}\Theta^r\\ \mathbf{C}^r &= \mathbf{Z}\Phi^r\end{aligned} \tag{T.5}$$

3.4 Experimental

3.4.1 Data Sets

Since the objective of this work is to introduce the theoretical basis and test the statistical properties and performance of a number of simplified alternatives of the MLPARAFAC algorithm, all of the data sets employed in this work were simulated so that the rank and error structure could be known with confidence. Experimental results will be presented in a companion paper to examine the performance of the algorithm for real experimental systems. Although a wide range of simulations were carried out, the results from only six data sets are presented here to support the main conclusions. In all cases, the data sets were relatively small, since the studies generally involved statistical validation requiring numerous runs.

Data Sets 1 to 5 share the same noise-free structure. This structure was a rank-three data set of dimensions 12x15x6 used to test the statistical characteristics and the performance of the different algorithms introduced. The loadings for mode A were represented by a 12x3 matrix drawn from a uniform distribution of random numbers from zero to three ($U(0,3)$). Similarly, **B** was a 15x3 matrix from $U(0,2)$ and **C** was a 6x3 matrix from $U(0,5)$. The error-free data were generated using the well known PARAFAC model, yielding the 12x90 matrix of error-free data, unfolded to maintain the A mode. Each data set is used to generate 100 replicates obtained by adding this noise-free structure to different realizations of the following error structures.

The matrix of measurement errors for Data Set 1 was a 12x90 matrix with a very simple structure. The simplest noise structure studied in this paper (Case 1A) was imposed on this data set. Initially, six different 12x15 matrices of normally distributed random numbers drawn from $N(0,0.1)$ were generated. These matrices were individually treated with a 7 point moving average filter along each row in order to produce error covariance. At the boundaries of the error matrix, the filter was wrapped around the opposite side in order to eliminate edge effects. Since these error matrices were individually treated with the same filter, this approach produced correlation among the measurements in one mode, and it is identical for all the slices (Case 1A). Although this

approach is not particularly realistic, it represents a general case for which the covariance structure could be easily predicted. Finally, the error-free data were added to the noise matrix in order to generate the data set.

The matrix of measurement errors for Data Set 2 was a created in a very similar fashion to the matrix of measurement errors for Data Set 1. The only difference is that each of the six different 12x15 matrices of normally distributed random numbers drawn from $N(0,0.1)$ were individually multiplied by a different filter matrix. The filter matrices were constructed from moving average filters (wrapped around the opposite side in order to eliminate edge effects) of dimensions 3, 5, 7, 7, 9, 5. Since these error matrices were individually treated with the same filter, this approach produced correlation among the measurements in one mode, and different from slice to slice (Case 1B).

The noise matrix of Data Set 3 was created to introduce correlated noise in two orders. Initially, a 12x90 matrix of normally distributed random numbers drawn from $N(0,0.1)$ was generated. This matrix was treated with a 67 point moving average filter along each row in order to produce error covariance. Since the error matrix was unfolded to maintain mode A, this approach produced the same row correlation among the measurements in the other two other modes (Case 1C).

The noise matrix of Data Set 4 was constructed in a similar way to the the noise matrix of Data Set 3. However, different size moving average filters were used along each row in order to produce error covariance among the measurements in the two other modes but with a different structure for each row. Twelve different moving average filter matrices with sizes in the range between 53 and 77 points were used (Case 1D).

The matrix of measurement errors for Data Set 5 was created to have the most complex noise structure studied in this paper (Case 1E). Initially, a 12x90 matrix of normally distributed random numbers drawn from $N(0,0.1)$ was generated. This matrix was vectorized by stacking the transposed rows on top of each other producing a 1080x1 vector that was multiplied by a 1080x1080 filter matrix. This filter matrix was formed by accommodating eigth 135x135 filter matrices of 127 points moving average filter along the diagonal to produce error covariance. As before, the boundaries of the filter

matrices were wrapped around the opposite side in order to eliminate edge effects. Considering that dimension of the filter matrices was 135 x 135 this approach produced correlation among the measurements in three modes. Again, the error-free data were added to the noise matrix in order to generate the data set.

The matrix of measurement errors for Data Set 6 was also created to represent the most complex in case 1E but with a more heterogeneous structure. The noise structure was constructed in a similar fashion to Data Set 5, but in this case nine different 120x120 filter matrices of 101, 133, 109, 131, 119, 121, 127 and 97 point moving average filter along the diagonal to produce error covariance. As before, the boundaries of the filter matrices were wrapped around the opposite side in order to eliminate edge effects. Considering that dimension of the filter matrices was 135 x 135 and each individual filter matrix was created with a different number of points, this approach produced correlation with a very heterogeneous structure among the measurements in three modes.

Data Sets 7 to 10 were rank-three data sets of dimensions 32x128x8 and were used to test the compression approach for different conditions of noise and data structure. In a generic way, the data sets are generated to contain the same broad spectral characteristics commonly observed in fluorescence excitation/emission matrices. The pure components for modes A and B were generated by adding Gaussian peaks of random means and standard deviations. The position of the center of each peak is a random number drawn between one and the largest channel number. The width of each peak is also drawn from a uniform distribution with a range between 10 and 40 ($U(10, 40)$). The spectra were normalized to unit length in all cases. The information about the intensity for each component is carried in mode C, in which the pure component concentrations are represented by an 8x3 matrix drawn from a uniform distribution of random numbers from zero to thirty ($U(0,30)$). Two different issues affecting the compression were investigated with these data sets: the amount of chemical information contained in the data and the level of noise affecting the data. Data Sets 7 and 8 were constructed with unimodal components for modes A and B. Data Sets 9 and 10 were constructed using components obtained by adding five Gaussian peaks for each component. All the data sets were constructed using the same error structure. In all cases,

an error structure equivalent to Data Set 1 was used to compare the results obtained after compression with results obtained without any compression using an algorithm which is optimal but not computationally involved. For this case a 61 point moving average filter was used for each row. The error structure is the same in each case but the signal-to-noise ratio (SNR) is varied to test the performance of compression with respect to the noise. Data Sets 7 and 9 have a SNR = 1000 and Data Sets 8 and 10 have a SNR = 250. The SNR values reported here represent the best case scenario, since they are calculated as the ratio between the maximum peak for the most concentrated sample and the value for the noise defined as three times the standard deviation. Therefore, there will be parts of these data sets with poorer SNR. All the data sets utilize 25 replicates calculated by adding the respective noise-free data and a different realization of the noise structure described for each data set.

3.4.2 Computational Aspects

All calculations performed in this work were carried out on a Sun Ultra 60 workstation with 2 x 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 [28].

3.5 Results and Discussion

In this section the estimation equations for each method will be validated using Data Sets 1 to 5 in order to cover different possible scenarios. In addition to the validation discussion, some general conclusions will be drawn about the merits of using the different algorithms based on the quality of the results and computational efforts invested to get them. Data Sets 7 to 10 will be used to compare the quality of the results for the compression approach with standard PARAFAC using different scenarios (noise level, amount of structural information and different basis sets) which have a very simple error structure in order to use the simplification developed for Case 1A as a benchmark value.

3.5.1 Statistical Validation

In order to validate the various proposed algorithms, it was necessary to verify that they yield the maximum likelihood solution. This can be accomplished by exploiting the statistical characteristics of S^2 values for the correct model. This methodology has been explained elsewhere [8,16] but it will be briefly reproduced here for the sake of completeness. Operationally, this is done by analyzing replicate data sets, each with the same matrix of error-free data and the same error structure, but with different realizations of the measurement error each time. If the distribution of S^2 values for these replicates follows a χ^2 distribution with the appropriate degrees of freedom [16,29], it can then be concluded that the algorithm is finding the maximum likelihood solution. Probability plots are used in this work to make this comparison. Initially, the replicate data sets (normally 100 replicates) are analyzed and the S^2 values are stored. Then, the S^2 values are sorted from the smallest to the largest and assigned a cumulative probability according to their position in the list; this is called the observed probability. For instance, the third element in the list would be assigned an observed probability of $2/n$, where n is the number of replicates. The expected probability is then calculated using the χ^2 distribution. The cumulative probability density function for χ^2 can be calculated using the incomplete gamma function as shown in equation 3.45:

$$P(S^2 \mid \nu) = \Gamma_{inc}\left(\frac{S^2}{2}, \frac{\nu}{2}\right) \tag{3.45}$$

where ν is the number of degrees of freedom [16]. If the two distributions are the same, a plot of the observed probabilities *vs.* the expected probabilities should yield a straight line with a slope of unity. If the model is insufficient to account for the systematic variance, either because the form of the model is incorrect or the parameters are suboptimal, then the points of the plot will lie above the ideal line. If the model accounts for an excessive amount of variance, (*i.e.* the estimated rank is too high and measurement variance is modeled), the points will lie below the ideal line.

Figure 3.2 shows the probability plots obtained when all of the algorithms introduced in this work, in addition the general MLPARAFAC algorithm (without

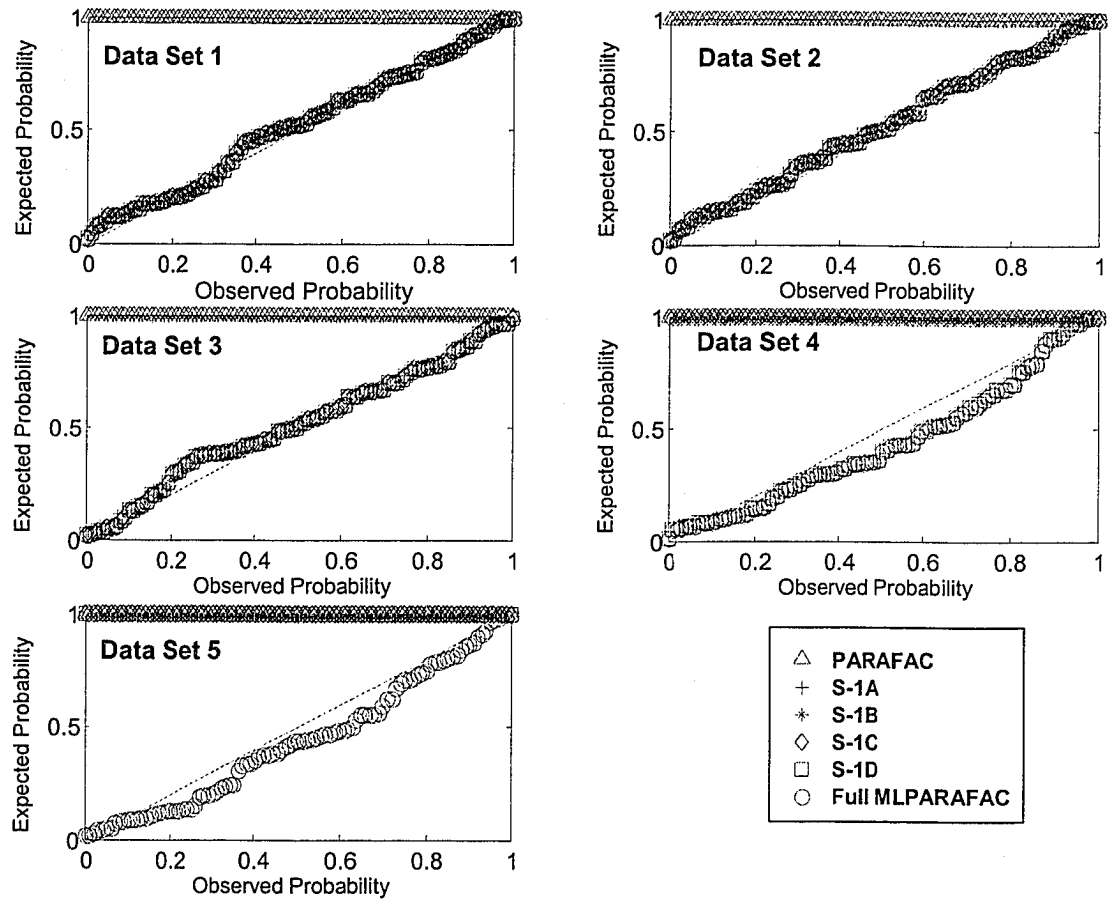

Figure 3.2. Probability plots obtained for 100 replicates of different simulated data sets using different algorithms such as PARAFAC (△); simplifications 1A (+), 1B (✳), 1C (◇), 1D (□); and full MLPARAFAC (○). The solid line with unity slope indicates ideal behaviour for maximum likelihood estimation.

compression) and PARAFAC are used to estimate Data Sets 1 to 5. The general MLPARAFAC was included as a benchmark since it can accommodate any covariance structure. Figure 3.2 shows a perfect trend, starting with all the methods but PARAFAC providing optimal models and ending with only the general MLPARAFAC algorithm providing an optimal model. As the complexity of the error structure increases, the methods designed to handle simpler error structures join the PARAFAC method, indicating the suboptimality of their estimates. Even though this trend was theoretically expected since each data set was constructed mimicking the error structure and therefore the objective function used to derive the estimation equation for each method, the results show from a numerical point of view the correctness of the estimation expressions for each case and how all of these methods are different simplified instances of a general

class of method. It is important to emphasize that this particular methodology is very sensitive to suboptimal solutions; therefore, it should not be used to compare the quality of different solutions.

3.5.2 Model Quality and Performance

The preceding sections dealt with the statistical validation of the maximum likelihood estimation process, but nothing has been said about the quality of the estimates obtained using these new algorithms. Although it has been previously demonstrated [16] that MLPARAFAC estimates are closer to the true underlying factors than the PARAFAC estimates, two reasonable questions are still not answered: (1) How do the MLPARAFAC estimates from different simplifications behave as the complexity of the error structure increases?, and (2) What is the computational price paid for the increment on complexity?. Both questions will be answered using simulated data. The computational workload and the quality of the data will be assessed using the average time needed for convergence and loading vector angles, respectively. Both magnitudes will be calculated using 100 replicates. In order to put this comparison into context, the value for each method relative to the value for the PARAFAC model will be used.

As mentioned above, the quality of the estimates will be measured as the closeness of estimates to the true factors using vector angles as a figure of merit. This figure of merit is the angular difference between the true loading vectors and the estimated loading vectors in each mode. For example, the vector angle between two loading vectors in mode A is given by:

$$\theta_p^a = \cos^{-1}\left(\frac{\hat{\mathbf{a}}_p^{\mathrm{T}}\mathbf{a}_p}{\|\hat{\mathbf{a}}_p\|\|\mathbf{a}_p\|}\right) \tag{3.46}$$

where $\mathbf{a}_p$ and $\hat{\mathbf{a}}_p$ are the true and estimated values for the pth loading vector of $\mathbf{A}$. Analogous equations can be used for the other orders. Smaller angles mean a greater similarity, so by comparing the vector angles obtained by the different simplifications of MLPARAFAC with those of PARAFAC, the agreement with the true vector can be assessed. An alternative measure is the correlation coefficient of the vectors, which is simply the term in parentheses, but since this approaches unity with small differences, it

is less sensitive. The quality of the estimates as well as the computation time will be compared in a relative fashion with respect to the corresponding values for the PARAFAC model. Equations 3.47 and 3.48 represent the expressions to calculate the relative average angle (RA) and the relative computation time (RCT):

$$RQ = \frac{\bar{\theta}_X}{\bar{\theta}_{PAR}} \tag{3.47}$$

$$RCT = \frac{\bar{t}_X}{\bar{t}_{PAR}} \tag{3.48}$$

Two completely different scenarios will be explored in order to have a broad view of the problem, since the degree to which these results will be extendable to a given application depends on the nature of the application and the characteristics of the noise. Data Sets 1 and 5 will be used since they represent very different scenarios in which clear comparisons can be made and conclusions drawn. The validation results showed that all of the simplifications provided optimal estimates for Data Set 1; therefore, this is a good scenario to test the computational advantages of using simpler algorithms over more complex algorithms when the data merit the simplification. Data Set 6 has a more complex error structure and these simplifications are also used to treat it.

Figures 3.3 and 3.4 show the results for the comparison in terms of quality and performance, respectively, when different simplifications are used. As expected, all the methods but PARAFAC provided the same results for Data Set 1 in terms of quality in Figure 3.3, since the error structure used was the simplest case. However, when the time employed to reach the convergence is taken into account (Figure 3.4), it is possible to appreciate the advantages of using simplified algorithms when the data at hand merit the use of a simplification.

In Figure 3.3, the relative average angle for Data Set 5 exhibits a nice trend, showing an improvement of the quality of the results as the complexity of the algorithms used increases. PARAFAC and general MLPARAFAC are located at the two extremes, corresponding to the methods providing the worst and best estimates. Again, a positive correlation between the complexity of the algorithm and the time needed to obtain the

best possible solution is observed, indicating in this case that a better solution will require the use of more computational effort. These results were expected, since the common wisdom tends to assume that the application of more complex algorithms (which in turn translates into error covariance matrices that are bigger and richer in information) will provide estimates of a better quality. Even though the error structures in real applications tend to be simple in general [13], the authors believe that in this particular case, the perfect monotonic trend in quality was mainly the result of an oversimplified (*i.e.* very symmetric) error structure. To avoid a misleading conclusion in this regard, this issue was further explored using a Data Set 6, which has a similar but less symmetric error structure.

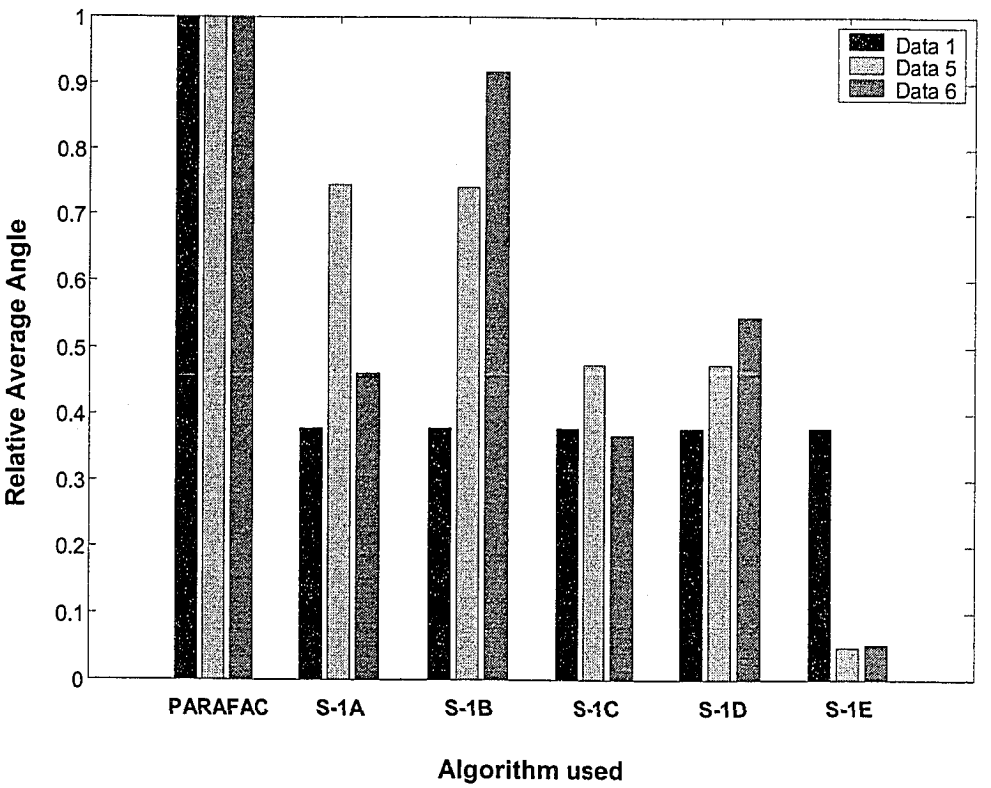

Figure 3.3. Comparison of the improvements in the quality of the estimates obtained for different MLPARAFAC algorithms for three characteristic data sets. The quality is measured using the relative average vector angle with respect to PARAFAC and the results are based on 100 replicates.

The results for Data Set 6 are quite surprising. For Data Set 5, a trend showing a monotonic improvement in the quality of the results with the complexity of the algorithm

used was observed, but the simulations for Data Set 6 show a very different scenario. PARAFAC and general MLPARAFAC were the only methods that coincide with the expected trend results. The remainder of the simplifications did not provide a clear trend in quality. For instance, the simplifications assuming that the errors are correlated along one order and are the same everywhere (Case 1B) gave estimates that are as good as the ones provided by the methodology assuming errors with the same structure affecting two orders (Case 1D).

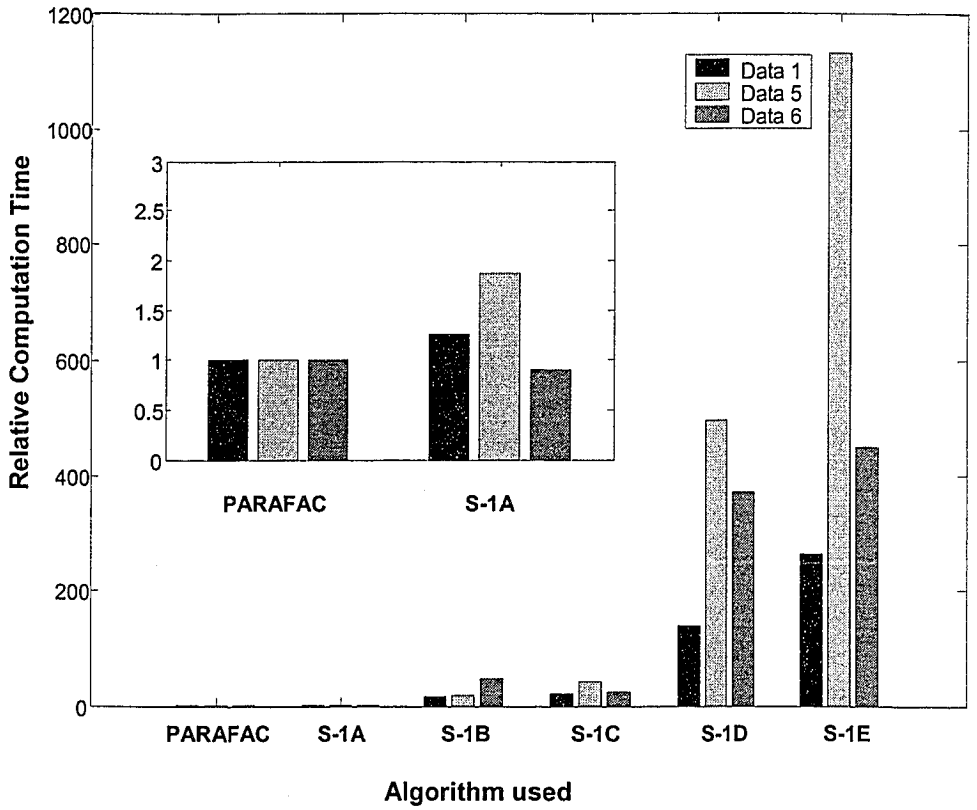

Figure 3.4. Comparison of the time utilized by different MLPARAFAC algorithms for three characteristic data sets. The performance is measured using relative time with respect to PARAFAC and the results are based on 100 replicates.

Another striking inconsistency evident from Figure 3.3 is that the quality of the results for the simplifications representing case 1B and 1D were worse than the quality of the results for Cases 1A and 1C, respectively. The mathematical theory behind these expressions makes Cases 1A and 1C subsets of the more general implementations representing Cases 1B and 1D, respectively, when the models are properly used. Therefore, all these inconsistent results clearly illustrate the importance of a thorough

characterization of the error structure, since the applications of an incorrect model can significantly degrade the quality of the result. It is important to note that the comparison of these methodologies from a computational point of view is meaningless for Data Set 6, since all of them produced a variety of sub-optimal models.

In reality, data commonly found in chemistry will have a behavior closer to the scenario illustrated by the simulations using Data Set 5. Probably, the error structure will not be exactly equivalent to the error covariance matrix used to derive the expression for a particular simplification, but it will not depart to the extent that Data Set 6 did to make the simplifications useless. However, it is important to fully characterize the error structure in order to apply the most suitable algorithm given the data set at hand in order to avoid erratic results such as the ones shown for Data Set 6. Unfortunately, due to the length and scope of this paper, only exact mathematical simplifications were shown, but in a companion paper to this work, a number of important guidelines will be introduced and used with different experimental data sets in order to cover more grey scenarios.

Finally, it is important to emphasize that, although only the results for three data sets were shown, many different data sets with the same characteristics of Data Sets 1, 5 and 6 were used to ensure the generality of the conclusions drawn.

3.5.3 Compression results

In the results shown in the previous section, general MLPARAFAC always provided the best solution, provoking the question: why not use general MLPARAFAC for every case? There are two reasons for this. The first is that general MLPARAFAC usually takes more time to produce the estimates, as already demonstrated. The second reason is that the previous results used general MLPARAFAC for a very small data set. For a more typical size data set, general MLPARAFAC cannot be applied directly due to storage and memory limitations. In order to overcome these limitations, a compression strategy was formulated. This section will show that, even though compressed MLPARAFAC will not give exactly the same results as general MLPARAFAC, the solutions will generally be superior to the PARAFAC solution. Figure 3.5 shows the comparative results for different cases (Data Sets 7, 8, 9 and 10) and compression basis

sets with respect to PARAFAC and general MLPARAFAC. In general, these results clearly demonstrate that improved estimates of loadings with respect to PARAFAC can be obtained from the general algorithm when information about the measurement error structure is compressed and is incorporated into the modeling process in the correct way. As already noted, the extent to which these improvements will be significant for a given application depends on the nature of the chemical data and the level and structure of the noise affecting the measurements. As can be seen from Figure 3.5, when the amount of information related to the chemical data increases, a larger number of factors are needed to yield better estimates using the compressed data. For Data Sets 7 and 8, which are constructed by unimodal components, six factors are enough to produce good results

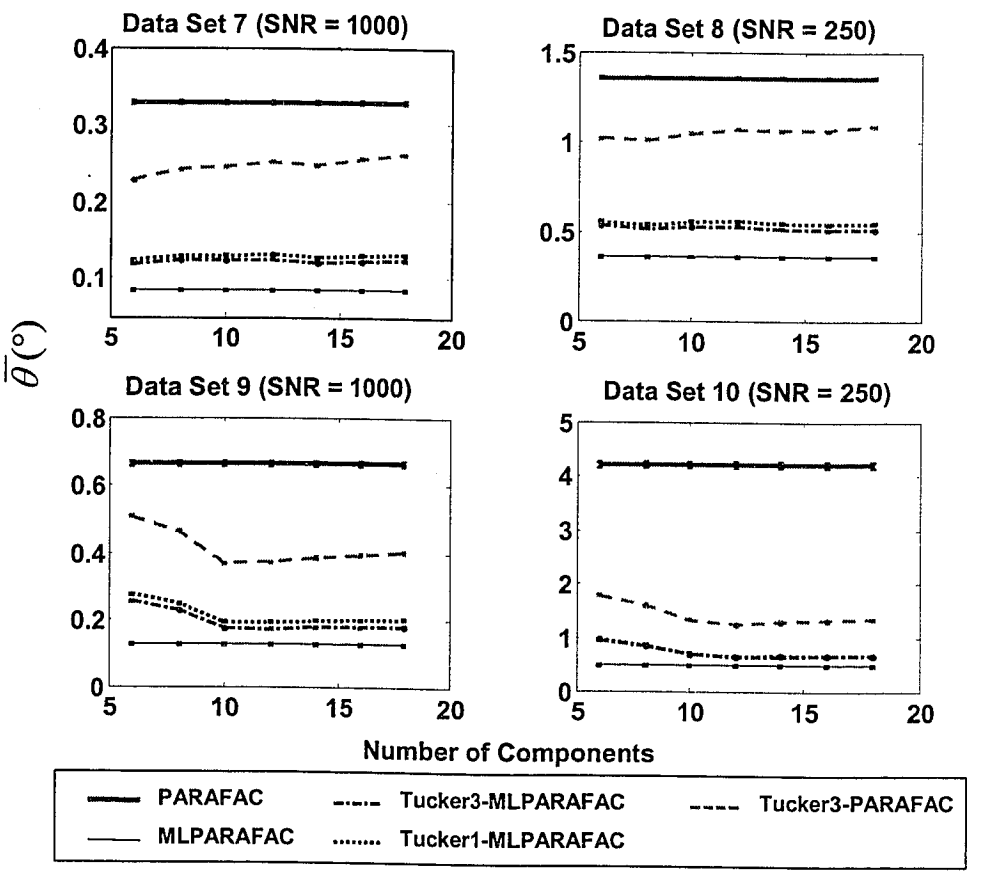

Figure 3.5. Comparison of the quality, in terms of average vector angle, of the estimates obtained for four different data sets when PARAFAC and general MLPARAFAC are employed on the original data and on compressed data. Tucker1 (PCA) and Tucker3 loadings were used as compression basis sets.

while for Data Sets 9 and 10, ten components are necessary to produce similar results. It is important to note, that Tucker3 and Tucker1 (PCA) basis sets produces very similar results in all cases, at least to the extent of these simulations. The different noise levels produce an equivalent worsening of all methods, indicating that this does not play an important role in the compression strategy. In addition to the PARAFAC, general MLPARAFAC, Tucker1-MLPARAFAC and Tucker3-MLPARAFAC, Tucker3-PARAFAC is also included to dissect the improved results with respect to PARAFAC in its two most important contributions: the effects of compression and the use of the error information in the estimation process. It can be observed in all cases that although the compression step by itself produced some improvement in the results, the use of compression and weighting yield much better estimates. It is also important to comment about the worsening of the estimates as the number of components increases, shown as a trend in all cases when the compressed data are treated with standard PARAFAC. This situation does not occur when MLPARAFAC is used due to its capacity to down-weight noisy regions as anticipated in the theory section

In reality, the difference between the quality of the estimates of compressed MLPARAFAC and PARAFAC will not be as large as the differences encountered in the simulation studies, since the results presented here were obtained assuming an absolute knowledge of the measurement error covariance matrix, while in practice this is typically estimated on the basis of replicate measurements and hence may be less reliable. Therefore, the benefits of including measurement error information must be weighed against the detrimental effects of including poor quality information. In many cases, it will be more advisable to use one of the previous simplifications because, in those situations, the advantages gained by pooling error covariance estimates may outweigh the benefits of using the full error covariance matrix.

3.6 Conclusions

In this work, the standard practice of expressing the estimation process by minimizing the different formulations of the same objective function was discarded since it does not take into account the loss of symmetry caused by the introduction of error information. A new approach, in which the same objective function is used to estimate

the loadings for all the modes, was introduced due to the benefits of locating the noise information in one or two modes as a simple representation and using it equivalently to obtain the estimation equations for each mode.

Four algorithms for carrying out simplified variations of general MLPARAFAC when the data at hand are corrupted by correlated noise affecting one or two orders have been described in this work by using the new approach. Also, a compression step was included prior to the use of general MLPARAFAC for cases where the noise structure is affecting three modes and the volume of data precludes the use of general MLPARAFAC on the raw data.

All of the algorithms were shown to produce maximum likelihood estimates through a comparison of the distribution of the objective function with the χ^2 distribution. It was also shown that the use of simplified algorithms when the data at hand merit the simplification is beneficial from a computational point of view. When the error structure was properly used, the quality of the estimates was the same for all the methods designed to handle this error structure. Two simulated scenarios where the error structure assumed departs from the actual error structure were studied to illustrate the importance of a thorough characterization of the error structure.

The merits of using compressed MLPARAFAC over PARAFAC were studied in different scenarios. Also, no significant differences were found between Tucker3 and Tucker1 basis sets, at least for the data used in the simulation studies.

Although the principles of general MLPARAFAC and a number of simplifications have been established here, a number of more practical aspects related to its application on experimental data remain to be examined. These include issues related to the characterization of the error structure and the application of the different simplifications. These subjects will be the focus of a companion paper.

3.7 Appendix

Case 1B:

Mode B

This scenario is represented by the following objective function:

$$f = \sum_{k=1}^{K} trace[(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_k^{-1}(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}] \tag{A.1}$$

Defining: $\mathbf{M}_k = \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}$, equation A.1 can be modified to yield:

$$f = \sum_{k=1}^{K} trace(\mathbf{M}_k\,\boldsymbol{\Psi}_k^{-1}\mathbf{M}_k^{\mathrm{T}}) \tag{A.2}$$

Using standard relations for derivatives of matrices and vectors [20], this gives:

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{B}_{jp}} &= \frac{\partial f}{\partial \mathbf{M}_k}\left(\frac{\partial \mathbf{M}_k}{\partial \mathbf{B}_{jp}}\right)^{\mathrm{T}} \\
&= \sum_{k=1}^{K} trace\left(2\mathbf{M}_k\boldsymbol{\Psi}_k^{-1}\frac{\partial \mathbf{M}_k^{\mathrm{T}}}{\partial \mathbf{B}_{jp}}\right) \\
&= \sum_{k=1}^{K} trace\left(2\mathbf{M}_k\boldsymbol{\Psi}_k^{-1}\frac{\partial(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})}{\partial \mathbf{B}_{jp}}\right) \\
&= \sum_{k=1}^{K} trace\left(2\mathbf{M}_k\boldsymbol{\Psi}_k^{-1}(-\mathbf{A}\mathbf{D}_k\mathbf{E}_{jp}^{\mathrm{T}})^{\mathrm{T}}\right) \\
&= \sum_{k=1}^{K} trace\left(2(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_k^{-1}(-\mathbf{A}\mathbf{D}_k\mathbf{E}_{jp}^{\mathrm{T}})^{\mathrm{T}}\right) \\
&= \sum_{k=1}^{K} trace\left(-2\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{E}_{jp}\mathbf{D}_k\mathbf{A}^{\mathrm{T}} + 2\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{E}_{jp}\mathbf{D}_k\mathbf{A}^{\mathrm{T}}\right) \\
&= -2\sum_{k=1}^{K} trace\left(\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{E}_{jp}\mathbf{D}_k\mathbf{A}^{\mathrm{T}}\right) + 2\sum_{k=1}^{K} trace(\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{E}_{jp}\mathbf{D}_k\mathbf{A}^{\mathrm{T}})
\end{aligned} \tag{A.3}$$

Setting this derivative equal to zero to find the minimum leads to:

$$\sum_{k=1}^{K} trace\left(\mathbf{X}_k \boldsymbol{\Psi}_k^{-1} \mathbf{E}_{jp}^{\mathrm{T}} \mathbf{D}_k \mathbf{A}^{\mathrm{T}}\right) = \sum_{k=1}^{K} trace(\mathbf{A}\mathbf{D}_k \mathbf{B}^{\mathrm{T}} \boldsymbol{\Psi}_k^{-1} \mathbf{E}_{jp} \mathbf{D}_k \mathbf{A}^{\mathrm{T}})$$
$$\sum_{k=1}^{K} trace\left(\mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{X}_k \boldsymbol{\Psi}_k^{-1} \mathbf{E}_{jp}\right) = \sum_{k=1}^{K} trace(\mathbf{B}^{\mathrm{T}} \boldsymbol{\Psi}_k^{-1} \mathbf{E}_{jp} \mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{A}\mathbf{D}_k) \quad \text{(A.4)}$$

Expressing the traces as the product of two vectors [20] yields:

$$\sum_{k=1}^{K} vec(\boldsymbol{\Psi}_k^{-1} \mathbf{X}_k^{\mathrm{T}} \mathbf{A}\mathbf{D}_k)^{\mathrm{T}} vec(\mathbf{E}_{jp}) = \sum_{k=1}^{K} vec(\mathbf{B})^{\mathrm{T}} vec(\boldsymbol{\Psi}_k^{-1} \mathbf{E}_{jp} \mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{A}\mathbf{D}_k)$$
$$vec\left(\sum_{k=1}^{K} (\mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{X}_k \boldsymbol{\Psi}_k^{-1})\right)^{\mathrm{T}} vec(\mathbf{E}_{jp}) = vec(\mathbf{B})^{\mathrm{T}} \sum_{k=1}^{K} (\boldsymbol{\Psi}_k^{-1} \otimes \mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{A}\mathbf{D}_k) vec(\mathbf{E}_{jp}) \quad \text{(A.5)}$$

Equation A.5 is one of the *JP* equations necessary to estimate the loadings of **B**. The rest of the equations are obtained as the right and left parts of this equation are multiplied by the different vectorized $\mathbf{E}_{jp}$ matrices. Since this term is completely independent in both sides of the equation, the process can be carried out in a straightforward manner using a matrix **E** formed as $[vec(\mathbf{E}_{11})\ vec(\mathbf{E}_{21})\ \ldots\ vec(\mathbf{E}_{JP})]$. A closer look of this matrix shows that **E** is the identity matrix of order *JP*, making the multiplication theoretically sound but numerically unnecessary and providing equation A.6 to estimate the loading of B:

$$vec(\mathbf{B})^{\mathrm{T}} = \left(\sum_{k=1}^{K} (\boldsymbol{\Psi}_k^{-1} \otimes \mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{B}\mathbf{D}_k)\right)^{-1} vec\left(\sum_{k=1}^{K} (\mathbf{D}_k \mathbf{A}^{\mathrm{T}} \mathbf{X}_k \boldsymbol{\Psi}_k^{-1})\right) \quad \text{(A.6)}$$

Mode C

Similarly, this objective function is used to represent the following scenario. It is important to realize that it can be expressed as the summation over the *K* slices:

$$
\begin{aligned}
f &= \sum_{k=1}^{K} trace[(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_k^{-1}(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}] \\
&= \sum_{k=1}^{K} f_k
\end{aligned}
\tag{A.7}
$$

Defining: $\mathbf{M}_k = \mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}$, equation A.7 can be modified to yield:

$$
f_k = trace(\mathbf{M}_k\, \boldsymbol{\Psi}_k^{-1}\mathbf{M}_k^{\mathrm{T}}) \tag{A.8}
$$

Using standard relations for derivatives of matrices and vectors [20], this gives:

$$
\begin{aligned}
\frac{\partial f_k}{\partial \mathbf{C}_{kp}} &= \frac{\partial f_k}{\partial \mathbf{M}_k}\left(\frac{\partial \mathbf{M}_k}{\partial \mathbf{C}_{kp}}\right)^{\mathrm{T}} \\
&= trace\left(2\mathbf{M}_k\boldsymbol{\Psi}_k^{-1}\frac{\partial \mathbf{M}_k^{\mathrm{T}}}{\partial \mathbf{C}_{kp}}\right) \\
&= trace\left(2\mathbf{M}_k\boldsymbol{\Psi}_k^{-1}\frac{\partial (\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}}{\partial \mathbf{C}_{kp}}\right) \\
&= trace\left(2\mathbf{M}_k\boldsymbol{\Psi}_k^{-1}(-\mathbf{A}\mathbf{E}_{pp}\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}\right) \\
&= trace\left(2(\mathbf{X}_k - \mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_k^{-1}(-\mathbf{A}\mathbf{E}_{pp}\mathbf{B}^{\mathrm{T}})^{\mathrm{T}}\right) \\
&= trace\left(-2\cdot\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}} + 2\cdot\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}}\right) \\
&= -2\cdot trace\left(\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}}\right) + 2\cdot trace(\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}})
\end{aligned}
\tag{A.9}
$$

Setting this derivative equal to zero to find the minimum leads to:

$$
\begin{aligned}
&trace\left(\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}}\right) = trace(\mathbf{A}\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}}) \\
&trace(\mathbf{A}^{\mathrm{T}}\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}) = trace(\mathbf{D}_k\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}}\mathbf{A}) \\
&vec(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{X}_k^{\mathrm{T}}\mathbf{A})^{\mathrm{T}}\, vec(\mathbf{E}_{pp}) = vec(\mathbf{D}_k)^{\mathrm{T}}\, vec(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\mathbf{E}_{pp}\mathbf{A}^{\mathrm{T}}\mathbf{A}) \\
&vec(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{X}_k^{\mathrm{T}})^{\mathrm{T}}\, vec(\mathbf{E}_{pp}) = vec(\mathbf{D}_k)^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\mathbf{A}\otimes\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B})vec(\mathbf{E}_{pp})
\end{aligned}
\tag{A.10}
$$

The last expression in equation A.10 is one of the *PP* equations necessary to estimate the loadings of k row of matrix **C**. The rest of the equations are obtained as the right and left

parts of this equation are multiplied by the different vectorized $\mathbf{E}_{pp}$ matrices. Since this term is completely independent in both sides of the equation, the process can be carried out in a straightforward manner, using a matrix $\mathbf{E}$ formed as $[vec(\mathbf{E}_{11})\ vec(\mathbf{E}_{22})\ \ldots\ vec(\mathbf{E}_{PP})]$. Contrary to what happened in the estimation of mode B, matrix $\mathbf{E}$ is used to pick the relevant elements in both members, since we are only interested in the estimation of the elements located in the diagonal of $\mathbf{D}_k$. Therefore, equation A.11 is used to estimate the loading of C in a row by row fashion:

$$\mathbf{c}_k = \left[\left(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Psi}_k^{-1}\mathbf{B}\otimes\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1} vec\left(\mathbf{A}^{\mathrm{T}}\mathbf{X}_k\boldsymbol{\Psi}_k^{-1}\mathbf{B}\right)\right]^{\mathrm{T}}\mathbf{E} \tag{A.11}$$

Case 1C:

Mode B

This scenario is well represented by the following objective function:

$$f = trace[(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}}] \tag{A.12}$$

In order to make the equation more tractable the following modifications were applied: $\mathbf{M} = (\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{I}_a$ to yield:

$$f = trace(\mathbf{M}^{\mathrm{T}}\boldsymbol{\Psi}_a^{-1}\mathbf{M}) \tag{A.13}$$

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{B}_{jp}} &= \frac{\partial f}{\partial \mathbf{M}}\left(\frac{\partial \mathbf{M}}{\partial \mathbf{B}_{jp}}\right)^{\mathrm{T}} \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}\mathbf{M}\frac{\partial \mathbf{M}}{\partial \mathbf{B}_{jp}}\right) \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}\mathbf{M}\frac{\partial\left[(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}}\right]}{\partial \mathbf{B}_{jp}}\right) \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}\mathbf{M}(-\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{E}_{jp})^{\mathrm{T}})\right) \\
&= trace\left(2\cdot\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a^{\mathrm{T}} - (\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}})(-\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{E}_{jp})^{\mathrm{T}})\right) \\
&= -2\cdot trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{X}_a^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{E}_{jp})^{\mathrm{T}}\right) + 2\cdot trace(\boldsymbol{\Psi}_a^{-1}(\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{E}_{jp})^{\mathrm{T}})
\end{aligned} \tag{A.14}$$

Setting this derivative equal to zero to find the minimum leads to:

$$
\begin{aligned}
&trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{X}_a^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{E}_{jp})^{\mathrm{T}}\right)=trace(\boldsymbol{\Psi}_a^{-1}(\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{E}_{jp})^{\mathrm{T}})\\
&trace\left(\boldsymbol{\Psi}_a^{-1}(\mathbf{C}\otimes\mathbf{E}_{jp})\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=trace(\boldsymbol{\Psi}_a^{-1}(\mathbf{C}\otimes\mathbf{E}_{jp})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})\\
&trace\left(\boldsymbol{\Psi}_a^{-1}(\mathbf{I}_K\otimes\mathbf{E}_{jp})(\mathbf{C}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=\\
&trace(\boldsymbol{\Psi}_a^{-1}(\mathbf{I}_K\otimes\mathbf{E}_{jp})(\mathbf{C}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}^{\mathrm{T}}\otimes\mathbf{I}_P)(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}}))
\end{aligned}
\tag{A.15}
$$

Equation A.15 becomes equation A.18 using the matrices $\mathbf{L}$ and $\mathbf{R}$ as defined in equation A.16 and A.17 respectively.

$$\mathbf{R}=(\mathbf{C}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a \tag{A.16}$$

$$\mathbf{L}=(\mathbf{C}\otimes\mathbf{I}_p)\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{I}_p)^{\mathrm{T}} \tag{A.17}$$

$$trace\left(\boldsymbol{\Psi}_a^{-1}(\mathbf{I}_K\otimes\mathbf{E}_{jp})\mathbf{R}\right)=trace(\boldsymbol{\Psi}_a^{-1}(\mathbf{I}_K\otimes\mathbf{E}_{jp})\mathbf{L}(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}})) \tag{A.18}$$

Equation A.18 can be expressed as equation A.20 when the matrices forming both members of the previous equation are partioned as shown in equation A.19. Matrices ${}^{mn}\boldsymbol{\Psi}_a^{-1}$, ${}^{nm}\mathbf{R}$ and ${}^{nm}\mathbf{L}$ have orders J x J; P x J and P x P, respectively.

$$
\begin{aligned}
&trace\left(\begin{bmatrix} {}^{11}\boldsymbol{\Psi}_a^{-1} & \cdots & {}^{1K}\boldsymbol{\Psi}_a^{-1}\\ \vdots & \ddots & \vdots \\ {}^{K1}\boldsymbol{\Psi}_a^{-1} & \cdots & {}^{KK}\boldsymbol{\Psi}_a^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{E}_{jp} & & \\ & \mathbf{E}_{jp} & \\ & & \mathbf{E}_{jp}\end{bmatrix}\begin{bmatrix} {}^{11}\mathbf{R} & \cdots & {}^{1K}\mathbf{R}\\ \vdots & \ddots & \vdots \\ {}^{K1}\mathbf{R} & \cdots & {}^{KK}\mathbf{R}\end{bmatrix}\right)=\\
&trace\left(\begin{bmatrix} {}^{11}\boldsymbol{\Psi}_a^{-1} & \cdots & {}^{1K}\boldsymbol{\Psi}_a^{-1}\\ \vdots & \ddots & \vdots \\ {}^{K1}\boldsymbol{\Psi}_a^{-1} & \cdots & {}^{KK}\boldsymbol{\Psi}_a^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{E}_{jp} & & \\ & \mathbf{E}_{jp} & \\ & & \mathbf{E}_{jp}\end{bmatrix}\begin{bmatrix} {}^{11}\mathbf{L} & \cdots & {}^{1K}\mathbf{L}\\ \vdots & \ddots & \vdots \\ {}^{K1}\mathbf{L} & \cdots & {}^{KK}\mathbf{L}\end{bmatrix}(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}})\right)
\end{aligned}
\tag{A.19}
$$

$$
\begin{aligned}
&trace\left(\sum_{m=1}^{K}\sum_{n=1}^{K}{}^{mn}\boldsymbol{\Psi}_a^{-1}\mathbf{E}_{jp}{}^{nm}\mathbf{R}\right)=trace\left(\left(\sum_{n=1}^{K}\sum_{m=1}^{K}{}^{mn}\boldsymbol{\Psi}_a^{-1}\mathbf{E}_{jp}{}^{nm}\mathbf{L}\right)\mathbf{B}^{\mathrm{T}}\right)\\
&trace\left(\sum_{m=1}^{K}\sum_{n=1}^{K}\mathbf{E}_{jp}{}^{nm}\mathbf{R}^{mn}\boldsymbol{\Psi}_a^{-1}\right)=trace\left(\mathbf{B}^{\mathrm{T}}\left(\sum_{n=1}^{K}\sum_{m=1}^{K}{}^{mn}\boldsymbol{\Psi}_a^{-1}\mathbf{E}_{jp}{}^{nm}\mathbf{L}\right)\right)\\
&vec(\mathbf{E}_{pj})^{\mathrm{T}}vec\left(\sum_{n=1}^{K}\sum_{m=1}^{K}{}^{nm}\mathbf{R}^{mn}\boldsymbol{\Psi}_a^{-\mathrm{T}}\right)=vec(\mathbf{B})^{\mathrm{T}}vec\left(\sum_{n=1}^{K}\sum_{m=1}^{K}{}^{mn}\boldsymbol{\Psi}_a^{-1}\mathbf{E}_{jp}{}^{nm}\mathbf{L}\right)\\
&vec(\mathbf{E}_{pj})^{\mathrm{T}}\left(\sum_{n=1}^{K}\sum_{m=1}^{K}\left({}^{mn}\boldsymbol{\Psi}_a^{-\mathrm{T}}\otimes{}^{nm}\mathbf{R}\right)\right)vec(\mathbf{I}_J)=vec(\mathbf{B})^{\mathrm{T}}\left(\sum_{n=1}^{K}\sum_{m=1}^{K}({}^{nm}\mathbf{L}^{\mathrm{T}}\otimes{}^{mn}\boldsymbol{\Psi}_a^{-1})\right)vec(\mathbf{E}_{jp})\\
&vec(\mathbf{E}_{pj})^{\mathrm{T}}\left(\sum_{n=1}^{K}\sum_{m=1}^{K}\left({}^{mn}\boldsymbol{\Psi}_a^{-\mathrm{T}}\otimes{}^{nm}\mathbf{R}\right)\right)vec(\mathbf{I}_J)=vec(\mathbf{E}_{jp})^{\mathrm{T}}\left(\sum_{n=1}^{K}\sum_{m=1}^{K}({}^{nm}\mathbf{L}\otimes{}^{mn}\boldsymbol{\Psi}_a^{-\mathrm{T}})\right)vec(\mathbf{B})
\end{aligned}
\tag{A.20}
$$

Equation A.20 is one of the *JP* equations necessary to estimate the loadings of **B**. The rest of the equations are obtained as the right and left parts of this equation are multiplied by the different vectorized $\mathbf{E}_{pj}$ and $\mathbf{E}_{jp}$ matrices, respectively. Since these terms are completely independent on both sides of the equation, the process can be carried out in a straightforward manner using matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ formed as [*vec*($\mathbf{E}_{11}$) *vec*($\mathbf{E}_{21}$) ... *vec*($\mathbf{E}_{PJ}$)] and [*vec*($\mathbf{E}_{11}$) *vec*($\mathbf{E}_{21}$) ... *vec*($\mathbf{E}_{JP}$)], respectively. A closer look at these matrices shows that $\mathbf{E}_1$ is the identity matrix of order *JP* while $\mathbf{E}_2$ is equal to the commutation matrix $\mathbf{K}_{PJ}$. When the equation is rearranged to estimate the loading of B equation A.21 is obtained:

$$vec(\mathbf{B}^{\mathrm{T}}) = \left(\sum_{m=1}^{K}\sum_{n=1}^{K}(\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes \mathbf{L}_{mn})\right)^{-1}\left(\sum_{m=1}^{K}\sum_{n=1}^{K}(\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes \mathbf{R}_{mn})\right)vec(\mathbf{I}_J) \qquad \text{(A.21)}$$

Mode C

This scenario is well represented by the following objective function:

$$f = trace[(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}}] \qquad \text{(A.22)}$$

In order to make the equation more tractable the following modifications were applied: $\mathbf{M} = (\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}}$ and $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{I}_a$ to yield:

$$f = trace(\mathbf{M}^{\mathrm{T}}\boldsymbol{\Psi}_a^{-1}\mathbf{M}) \qquad \text{(A.23)}$$

$$\begin{aligned}
\frac{\partial f}{\partial \mathbf{C}_{kp}} &= \frac{\partial f}{\partial \mathbf{M}}\left(\frac{\partial \mathbf{M}}{\partial \mathbf{C}_{kp}}\right)^{\mathrm{T}} \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}\mathbf{M}\frac{\partial \mathbf{M}^{\mathrm{T}}}{\partial \mathbf{C}_{kp}}\right) \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}\mathbf{M}\frac{\partial\left[(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})\right]}{\partial \mathbf{C}_{kp}}\right) \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}\mathbf{M}(-\tilde{\mathbf{A}}(\mathbf{E}_{kp}\otimes\mathbf{B})^{\mathrm{T}})\right) \\
&= trace\left(2\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a^{\mathrm{T}} - (\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}})(-\tilde{\mathbf{A}}(\mathbf{E}_{kp}\otimes\mathbf{B})^{\mathrm{T}})\right) \\
&= -2\cdot trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{X}_a^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{E}_{kp}\otimes\mathbf{B})^{\mathrm{T}}\right) + 2\cdot trace(\boldsymbol{\Psi}_a^{-1}(\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{E}_{kp}\otimes\mathbf{B})^{\mathrm{T}})
\end{aligned} \qquad \text{(A.24)}$$

Setting this derivative equal to zero to find the minimum leads to:

$$
\begin{aligned}
&trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{X}_a^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{E}_{kp}\otimes\mathbf{B})^{\mathrm{T}}\right)=trace\left(\boldsymbol{\Psi}_a^{-1}(\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{E}_{kp}\otimes\mathbf{B})^{\mathrm{T}}\right)\\
&trace\left(\boldsymbol{\Psi}_a^{-1}(\mathbf{E}_{kp}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=trace\left((\mathbf{C}\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{E}_{pk}\otimes\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}\right)\\
&trace\left(\boldsymbol{\Psi}_a^{-1}(\mathbf{E}_{kp}\otimes\mathbf{I}_J)(\mathbf{I}_P\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=\\
&trace\left((\mathbf{C}\otimes\mathbf{I}_J)(\mathbf{I}_P\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{E}_{pk}\otimes\mathbf{I}_P)(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}\right)\\
&trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ}(\mathbf{I}_J\otimes\mathbf{E}_{kp})\mathbf{K}_{JP}(\mathbf{I}_P\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=\\
&trace((\mathbf{I}_J\otimes\mathbf{C})\mathbf{K}_{JP}(\mathbf{I}_P\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}\mathbf{K}_{PP}(\mathbf{I}_P\otimes\mathbf{E}_{pk})\mathbf{K}_{PK}(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ})\\
&trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ}(\mathbf{I}_J\otimes\mathbf{E}_{kp})\mathbf{K}_{JP}(\mathbf{I}_P\otimes\mathbf{B})\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=\\
&trace\left((\mathbf{I}_J\otimes\mathbf{C})(\mathbf{B}\otimes\mathbf{I}_P)\mathbf{K}_{PP}\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}\mathbf{K}_{PP}(\mathbf{I}_P\otimes\mathbf{E}_{pk})\mathbf{K}_{PK}(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ}\right)\\
&trace\left(\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ}(\mathbf{I}_J\otimes\mathbf{E}_{kp})(\mathbf{B}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a\right)=\\
&trace\left((\mathbf{I}_J\otimes\mathbf{C})(\mathbf{B}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}}(\mathbf{I}_P\otimes\mathbf{E}_{pk})\mathbf{K}_{PK}(\mathbf{I}_K\otimes\mathbf{B}^{\mathrm{T}})\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ}\right)
\end{aligned}
\tag{A.25}
$$

It is worth noting two important manipulations carried out in equation A.25. First, the commutation matrices are introduced in order to invert the order of the Kronecker products $(\mathbf{E}_{kp}\otimes\mathbf{I}_J)$ and $(\mathbf{C}\otimes\mathbf{I}_J)$. Second, due to the sparse nature of $\tilde{\mathbf{A}}$, the following equality holds:

$$\mathbf{K}_{PP}\tilde{\mathbf{A}}^{\mathrm{T}}=\tilde{\mathbf{A}}^{\mathrm{T}} \tag{A.26}$$

Equation A.25 becomes equation A.31 using the matrices **T**, **S**, **L** and **R** as defined in equations A.27 to A.30:

$$\mathbf{L}=\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ} \tag{A.27}$$

$$\mathbf{R}=(\mathbf{B}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\mathbf{X}_a \tag{A.28}$$

$$\mathbf{S}=(\mathbf{B}\otimes\mathbf{I}_P)\tilde{\mathbf{A}}^{\mathrm{T}}\tilde{\mathbf{A}} \tag{A.29}$$

$$\mathbf{T}=(\mathbf{B}^{\mathrm{T}}\otimes\mathbf{I}_K)\mathbf{K}_{JK}\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ} \tag{A.30}$$

$$trace\left(\mathbf{L}(\mathbf{I}_J\otimes\mathbf{E}_{kp})\mathbf{R}\right)=trace\left((\mathbf{I}_J\otimes\mathbf{C})\mathbf{S}(\mathbf{I}_P\otimes\mathbf{E}_{pk})\mathbf{T}\right) \tag{A.31}$$

Equation A.31 can be expressed as equation A.33 when the matrices forming both members of the previous equation are partioned as shown in equation A.32. Matrices ${}^{mn}\mathbf{L}$, ${}^{nm}\mathbf{R}$, ${}^{mn}\mathbf{S}$ and ${}^{nm}\mathbf{T}$ have dimensions K x K; P x K ; P x P and K x K respectively.

$$trace\left(\begin{bmatrix} {}^{11}\mathbf{L} & \cdots & {}^{1J}\mathbf{L} \\ \vdots & \ddots & \vdots \\ {}^{J1}\mathbf{L} & \cdots & {}^{JJ}\mathbf{L} \end{bmatrix}\begin{bmatrix} \mathbf{E}_{kp} & & \\ & \mathbf{E}_{kp} & \\ & & \mathbf{E}_{kp} \end{bmatrix}\begin{bmatrix} {}^{11}\mathbf{R} & \cdots & {}^{1J}\mathbf{R} \\ \vdots & \ddots & \vdots \\ {}^{J1}\mathbf{R} & \cdots & {}^{JJ}\mathbf{R} \end{bmatrix}\right) =$$

$$trace\left((\mathbf{I}_J \otimes \mathbf{C})\begin{bmatrix} {}^{11}\mathbf{S} & \cdots & {}^{1P}\mathbf{S} \\ \vdots & \ddots & \vdots \\ {}^{J1}\mathbf{S} & \cdots & {}^{JP}\mathbf{S} \end{bmatrix}\begin{bmatrix} \mathbf{E}_{pk} & & \\ & \mathbf{E}_{pk} & \\ & & \mathbf{E}_{pk} \end{bmatrix}\begin{bmatrix} {}^{11}\mathbf{T} & \cdots & {}^{1J}\mathbf{T} \\ \vdots & \ddots & \vdots \\ {}^{P1}\mathbf{T} & \cdots & {}^{PJ}\mathbf{T} \end{bmatrix}\right) \quad \text{(A.32)}$$

$$trace\left(\sum_{m=1}^{J}\sum_{n=1}^{J} {}^{mn}\mathbf{L}\mathbf{E}_{kp}{}^{nm}\mathbf{R}\right) = trace\left(\mathbf{C}\sum_{n=1}^{J}\sum_{m=1}^{P} {}^{mn}\mathbf{S}\mathbf{E}_{pk}{}^{nm}\mathbf{T}\right)$$

$$\sum_{m=1}^{J}\sum_{n=1}^{J} trace\left(\mathbf{E}_{kp}{}^{nm}\mathbf{R}{}^{mn}\mathbf{L}\right) = vec(\mathbf{C}^{\mathrm{T}})^{\mathrm{T}}\, vec\left(\sum_{n=1}^{J}\sum_{m=1}^{P} {}^{mn}\mathbf{S}\mathbf{E}_{pk}{}^{nm}\mathbf{T}\right) \quad \text{(A.33)}$$

$$vec(\mathbf{E}_{pk})^{\mathrm{T}}\left(\sum_{m=1}^{J}\sum_{n=1}^{J}\left({}^{mn}\mathbf{L}^{\mathrm{T}} \otimes {}^{nm}\mathbf{R}\right)\right)vec(\mathbf{I}_K) = vec(\mathbf{E}_{pk})^{\mathrm{T}}\left(\sum_{n=1}^{J}\sum_{m=1}^{P}({}^{nm}\mathbf{T} \otimes {}^{mn}\mathbf{S}^{\mathrm{T}})\right)vec(\mathbf{C}^{\mathrm{T}})$$

Equation A.33 is one of the KP equations necessary to estimate the loadings of **C**. The rest of the equations are obtained as the right and left parts of this equation are multiplied by the vectorized $\mathbf{E}_{pk}$ matrix. Since these terms are completely independent on both sides of the equation, the process can be carried out in a straightforward manner using matrices $\mathbf{E}_I$ formed as $[vec(\mathbf{E}_{11})\ vec(\mathbf{E}_{21})\ \ldots\ vec(\mathbf{E}_{KP})]$. A closer look at these matrices shows that $\mathbf{E}_I$ is the identity matrix of order PK, providing equation A.34 to estimate the loadings of C:

$$vec(\mathbf{C}^{\mathrm{T}}) = \left(\sum_{m=1}^{P}\sum_{n=1}^{J}(\mathbf{T}_{mn} \otimes \mathbf{S}_{nm}^{\mathrm{T}})\right)^{-1}\left(\sum_{m=1}^{J}\sum_{n=1}^{J}(\mathbf{L}_{mn}^{\mathrm{T}} \otimes \mathbf{R}_{nm})\right)vec(\mathbf{I}_K) \quad \text{(A.34)}$$

Case 1D:

This scenario is structurally similar to the previous case, but more complex since the error covariance matrix changes from row to row. Therefore, the estimation process cannot be carried out in one step, but rather as a sequence of I independent problems solved by minimizing equation A.35:

$$f = \sum_{i=1}^{I} ({}^{i}\mathbf{x}_a - {}^{i}\tilde{\mathbf{a}}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}}) {}^{i}\boldsymbol{\Psi}_a^{-1} ({}^{i}\mathbf{x}_a - {}^{i}\mathbf{a}(\mathbf{C}\otimes\mathbf{B})^{\mathrm{T}})^{\mathrm{T}} \tag{A.35}$$

As mentioned before, the only difference in the minimization process between equation A.35 and equation A.12 is the sequential manner in which the former is solved. This situation leads to estimation equations that are similar to the previous case, but solved in a sequential manner. Mathematically, this is carried out by solving row by row in mode A and solving over a sequence of I summations for mode B and C as shown next:

$${}^{i}\mathbf{a} = {}^{i}\mathbf{x}_a \, {}^{i}\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}}(\mathbf{Z}_a \, {}^{i}\boldsymbol{\Psi}_a^{-1}\mathbf{Z}_a^{\mathrm{T}})^{-1} \tag{A.36}$$

$$vec(\mathbf{B}^{\mathrm{T}}) = \left(\sum_{i=1}^{I}\sum_{m=1}^{K}\sum_{n=1}^{K}({}^{i}\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes {}^{i}\mathbf{L}_{mn})\right)^{-1}\left(\sum_{i=1}^{I}\sum_{m=1}^{K}\sum_{n=1}^{K}({}^{i}\boldsymbol{\Psi}_{nm}^{-\mathrm{T}} \otimes {}^{i}\mathbf{R}_{mn})\right) vec(\mathbf{I}_J) \tag{A.37}$$

where:

$${}^{i}\mathbf{R} = (\mathbf{C}\otimes\mathbf{I}_P) {}^{i}\tilde{\mathbf{a}}^{\mathrm{T}} {}^{i}\mathbf{x}_a \tag{A.38}$$

$$\mathbf{L} = (\mathbf{C}\otimes\mathbf{I}_P) {}^{i}\tilde{\mathbf{a}}^{\mathrm{T}} {}^{i}\tilde{\mathbf{a}}(\mathbf{C}\otimes\mathbf{I}_p)^{\mathrm{T}} \tag{A.39}$$

and:

$$vec(\mathbf{C}^{\mathrm{T}}) = \left(\sum_{i=1}^{I}\sum_{m=1}^{P}\sum_{n=1}^{J}({}^{i}\mathbf{T}_{mn} \otimes {}^{i}\mathbf{S}_{nm}^{\mathrm{T}})\right)^{-1}\left(\sum_{i=1}^{I}\sum_{m=1}^{J}\sum_{n=1}^{J}({}^{i}\mathbf{L}_{mn}^{\mathrm{T}} \otimes {}^{i}\mathbf{R}_{nm})\right) vec(\mathbf{I}_K) \tag{A.40}$$

$${}^{i}\mathbf{L} = {}^{i}\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ} \tag{A.41}$$

$${}^{i}\mathbf{R} = (\mathbf{B}\otimes\mathbf{I}_P) {}^{i}\tilde{\mathbf{a}}^{\mathrm{T}} {}^{i}\mathbf{x}_a \tag{A.42}$$

$${}^{i}\mathbf{S} = (\mathbf{B}\otimes\mathbf{I}_P) {}^{i}\tilde{\mathbf{a}}^{\mathrm{T}} {}^{i}\tilde{\mathbf{a}} \tag{A.43}$$

$${}^{i}\mathbf{T} = (\mathbf{B}^{\mathrm{T}}\otimes\mathbf{I}_K)\mathbf{K}_{JK} {}^{i}\boldsymbol{\Psi}_a^{-1}\mathbf{K}_{KJ} \tag{A.44}$$

Acknowledgements:

The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Dow Chemical Company (Midland, MI).

3.8 References

1. Martens H. and Næs T. *Multivariate Calibration*; John Wiley & Sons; New York, 1989.

2. de Juan A. and Tauler R. Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution. *Anal. Chim. Acta* 2003; **500**: 195

3. Bro, R. PARAFAC. Tutorial and applications. *Chemom. Intell. Lab. Syst.*1997; **38**: 149.

4. Kiers H.A.L., ten Berge J.M.F. and Bro R. PARAFAC2 - part I. A direct fitting algorithm for the PARAFAC2 model. *J. Chemom.* 1999; **13**: 275.

5. Bro R. Multi-way calibration. Multi-linear PLS. *J. Chemom.* 1996; **10**: 47.

6. Paatero P. and Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 1994; **5**: 111.

7. Wentzell P.D., Andrews D.T., Hamilton D.C., Faber K. and Kowalski B.R. Maximum Likelihood Principal Component Analysis. *J. Chemom.* 1997; **11**: 339

8. Van Huffel S. and Vandewalle J. *The Total Least Squares Problem: Computational Aspect and Analysis*; SIAM: Philadelphia 1991.

9. Kiers H.A.L. Weighted least squares fitting using ordinary least squares algorithms. *Psychometrika.* 1997; **62**: 251.

10. Martens H., Hoy, Wise B.M., Bro R. and Brockhoff P.B. Pre-whitening of data covariance-weighted pre-processing. *J. Chemom.* 2003; **17**: 153.

11. Bro R., Sidiropoulos N.D. and Smilde A.K. Maximum likelihood fitting using simple least squares algorithms. *J. Chemom.* 2002; **16**: 183.

12. Leger M., Vega-Montoto L. and Wentzell P.D. Methods for systematic investigation of measurement error covariance matrices. *Chemom. Intell. Lab. Syst.* 2004; **In Press**.

13. Weisberg S. *Applied Linear Regression* Wiley: Chichester, 1985.

14. Brown C.D. and Wentzell P.D. Hazard of digital smoothing filters as a preprocessing tool in multivariate calibration. *J. Chemom.* 1999; **13**: 133.

15. Brown C.D., Vega-Montoto L. and Wentzell P.D. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Appl. Spectrosc.* 2000; **54**: 1055.

16. Vega-Montoto L. and Wentzell P.D. Maximum Likelihood Parallel Factor Analysis (MLPARAFAC), *J. Chemom.* 2003; **17**: 237.

17. Savitzky A. and Golay M.J.E. Smoothing and differentiation of data by simplified least squares. *Anal. Chem.* 1964; **36**: 1627.

18. Magnus J.R. and Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley: Chichester, 1988.

19. Carroll J.D., Pruzansky S., Kruskal J.B. Candelinc: A general approach to multidimensional analysis of many-ways arrays with linear constraints on parameters, *Psychometrika* 1980; **45**: 3

20. Apellof C.J. and Davidson E.R. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Anal. Chem.* 1981; **53**: 2053.

21. Alsberg B.K., Kvalheim O.M. Speed improvement of multivariate algorithms by the method of postponed basis matrix multiplications. Part I. Principal component analysis, *Chemom. Intell. Lab. Syst.* 1994; **24**: 31

22. Alsberg B.K., Kvalheim O.M. Speed improvement of multivariate algorithms by the method of postponed basis matrix multiplications. Part II. Three-mode principal component analysis. *Chemom. Intell. Lab. Syst.* 1994; **24**: 43.

23. Kiers H.A.L, Harshman R.A. Relating two proposed methods for speed-up of algorithms for fitting two- and three-way principal component and related multilinear methods. *Chemom. Intell. Lab. Syst.* 1997; **39:** 31

24. Leger M.N. and Wentzell P.D. Maximun Likelihood Principal Component Regression on wavelet-compressed data. *Appl. Spectrosc.* 2004; **58**: 855.

25. Tomiši V. and Simeon V. Assessment of the effective rank of a (co)variance matrix: A nonparametric goodness-of-fit test. *J. Chemom.* 1993; **7**: 381

26. Andersson C.A. and Bro R. Improving the speed of multi-way algorithms. Part I: Tucker3. *Chemom. Intell. Lab. Syst.* 1998; **42**: 93

27. Kiers H.A.L, Krijnen W. P. An efficient algorithm for PARAFAC of three-way data with large numbers of observation units. *Psychometrika*. 1991; **56**: 147.

28. Andersson C.A. and Bro R. The N-way toolbox for MATLAB. *Chemom. Intell. Lab. Syst.* 2000; **52**: 1.

29. Durell S.R., Lee C., Ross R.T. and Gross E.L. Factor analysis of the near-ultraviolet absorption spectrum of plastocyanyn using bilinear, trilinear and quadrilinear models. *Arch. Biochem. Biophys.* 1990; **278**: 148.

Chapter 4

Mathematical Improvements to Maximum Likelihood Parallel Factor Analysis: Experimental Studies[3]

4.1 Abstract

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.

---

[3] Submitted to *Journal of Chemometrics* (Submitted March 2005)

4.2 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 involves 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 non-ideal 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, 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 measurements 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. 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 [36-41].

In general, even though the characteristics of the noise affecting fluorescence EEMs are well documented [42], 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 MILES (Maximum Likelihood via Iterative Least Squares Estimation) and MLPARAFAC (Maximum Likelihood Parallel Factor Analysis), have been introduced to the chemometrics literature [26, 43] 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 [44], a number of important simplifications of the general MLPARAFAC [43] 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 [43] 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 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 sub-steps 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 three-fold. 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.

4.3 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.* [45] 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 practioner choose the proper estimation method to accommodate the error structure in an optimal way. Leger *et al.* [45] 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.

4.3.1 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 mode 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 4.1 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.

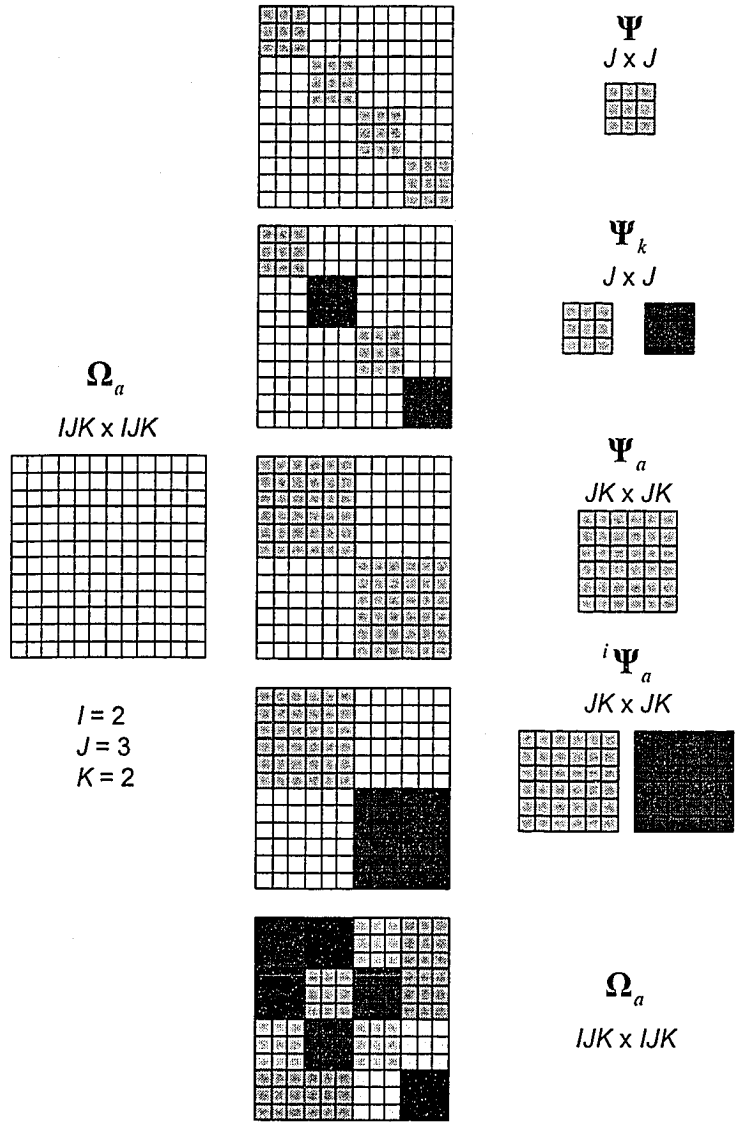

Figure 4.1. Illustration of the possible scenarios in which a full error covariance matrix can be expressed using different simplified representations of the error structure to describe all of the sources of variation.

It is clear from Figure 4.1 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 [45-47], 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, ${}^{r}\underline{\mathbf{X}}$ (I x J x K) is unfolded while retaining mode A, producing ${}^{r}\mathbf{X}_a$ (I x JK). Then, ${}^{r}\mathbf{X}_a$ is transposed and used to calculate the individual experimental error covariance matrix for each object included in mode B and C via equation 4.1.

$$\boldsymbol{\Sigma}_o = \frac{1}{(R-1)}\sum_{r=1}^{R}({}^{r}\mathbf{x}_o - \bar{\mathbf{x}}_o)^{\mathrm{T}}({}^{r}\mathbf{x}_o - \bar{\mathbf{x}}_o) \tag{4.1}$$

where ${}^{r}\mathbf{x}_o$ is the o^{th} 1 x I row vector of replicate r taken from ${}^{r}\mathbf{X}_a^{\mathrm{T}}$ and $\bar{\mathbf{x}}_o$ is the 1 x I mean vector of the replicate measurements. The superscript "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 Σ) 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, Σ_{avg}. For this example Σ_{avg} can be calculated as follows:

$$\boldsymbol{\Sigma}_{avg} = \frac{1}{JK}\sum_{o=1}^{JK}\boldsymbol{\Sigma}^{o} \tag{4.2}$$

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 out to determine the extent of the error correlation effect. Mathematically, equations 4.1 and 4.2 can be combined to give a clearer view of this calculation. This is done by considering the I x JKR matrix of residuals for all replicates of all objects, $\mathbf{E}_a$. The equation can then be written as:

$$\mathbf{\Sigma}_{avg} = \frac{1}{JK(R-1)}\mathbf{E}_a\mathbf{E}_a^{\mathrm{T}} \tag{4.3}$$

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:

$$\rho_{rs} = \frac{\sigma_{rs}}{\sigma_r\sigma_s} \tag{4.4}$$

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:

$$\mathbf{\Sigma}_{corr} = \mathbf{\Sigma} ./ \sqrt{diag(\mathbf{\Sigma}) \cdot diag(\mathbf{\Sigma})^{\mathrm{T}}} \tag{4.5}$$

where the notation "./" indicates an element-wise division (Hadamard quotient), the function "*diag*" converts the diagonal of $\mathbf{\Sigma}$ 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:

$$\mathbf{\Sigma}_a^i = \frac{1}{(R-1)} \sum_{r=1}^{R} ({}^r\mathbf{x}_a^i - \bar{\mathbf{x}}_a)^{\mathrm{T}} ({}^r\mathbf{x}_a^i - \bar{\mathbf{x}}_a) \tag{4.6}$$

where ${}^r\mathbf{x}_a^i$ is the 1 x *JK* row vector of replicate *r* and $\bar{\mathbf{x}}$ is the 1 x *JK* mean vector of the replicate measurements. The pooled error covariance matrix is calculated as shown:

$$\mathbf{\Sigma}_a^{avg} = \frac{1}{I} \sum_{i=1}^{I} \mathbf{\Sigma}_a^i \tag{4.7}$$

4.3.2 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 [48], 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 45 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 of 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 [45]. 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, $\mathbf{E}$ ($= \mathbf{X}\text{-}\mathbf{X}^{o}$), could be represented as:

$$\mathbf{E} = \begin{bmatrix} \mathbf{x}_{1\bullet} - \mathbf{x}^{o} \\ \mathbf{x}_{2\bullet} - \mathbf{x}^{o} \\ \vdots \\ \mathbf{x}_{R\bullet} - \mathbf{x}^{o} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_{1\bullet} \\ \mathbf{e}_{2\bullet} \\ \vdots \\ \mathbf{e}_{R\bullet} \end{bmatrix} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1R} \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} e_{21} \\ e_{22} \\ \vdots \\ e_{2R} \end{bmatrix} \begin{bmatrix} \sqrt{x_1^{o}} & \sqrt{x_2^{o}} & \cdots & \sqrt{x_J^{o}} \end{bmatrix} \tag{4.8}$$

$$= \mathbf{e}_1 \cdot \mathbf{1}_J^{T} + \mathbf{e}_2 \cdot \sqrt{\mathbf{x}^{o}}$$

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 x 1 vector of ones, and $\mathbf{x}^{o}$ is a row vector representing the

error-free spectrum. The R x 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:

$$\mathbf{\Sigma} = E\left(\mathbf{e}^{\mathrm{T}} \cdot \mathbf{e}\right) = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 = \sigma_1^2 \cdot \mathbf{1}_J \cdot \mathbf{1}_J^{\mathrm{T}} + \sigma_2^2 \cdot \left(\sqrt{\mathbf{x}^o}\right)^{\mathrm{T}} \cdot \sqrt{\mathbf{x}^o} \tag{4.9}$$

It is important to mention that the structural model shown in equation 4.1 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 4.9 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 $\Sigma_1, \Sigma_2, \ldots, \Sigma_O$. 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 x P matrix representing the sources of variation (*i.e.* structural factors) and $\mathbf{D}_o$ is the P x 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 $\mathbf{e}_1$ and $\mathbf{e}_2$ (contribution vectors) share the same statistical characteristics for all samples (i.e. same σ_1 and σ_2). As explained in reference 45, 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 4.10. It is recommended that the PARAFAC algorithm be run in a split-half [49] fashion to make sure of the validity of the estimates.

4.3.3 Assessment of the error structure

Figure 4.2 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 though 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 x J, JK x JK or IJK x 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

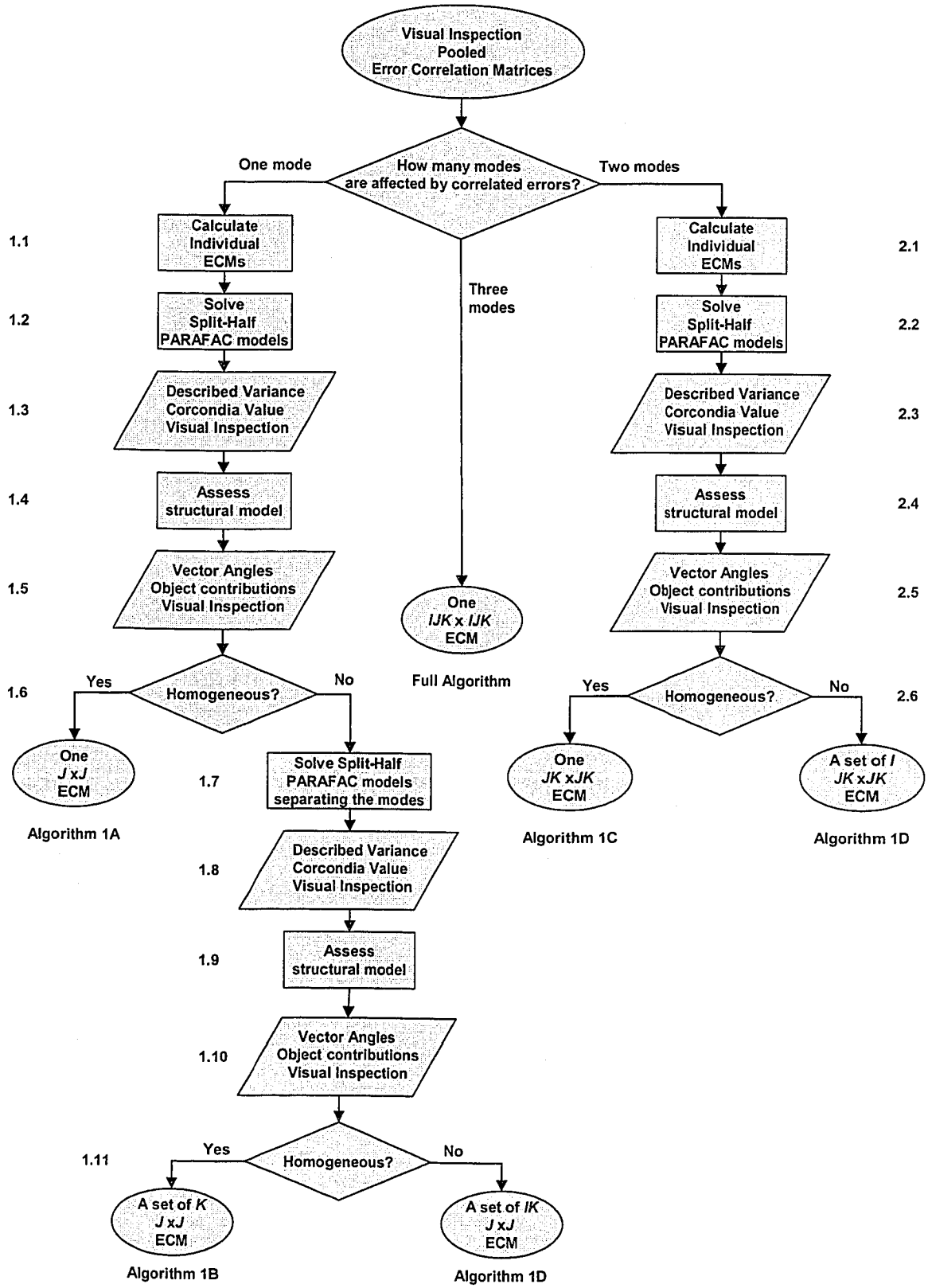

Figure 4.2. Flow chart employed to characterize the error structure.

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. [45].

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 [49] 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 [45]. 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 4.4.

4.4 Experimental

4.4.1 Sample preparation

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 to 0.34 μg/g (ace), 0.018 to 0.063 μg/g (nap), and 0.0072 to 0.027 μg/g (phe).

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 nm 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 ten replicates using the same experimental conditions. These are shown in Figure 4.3.

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.

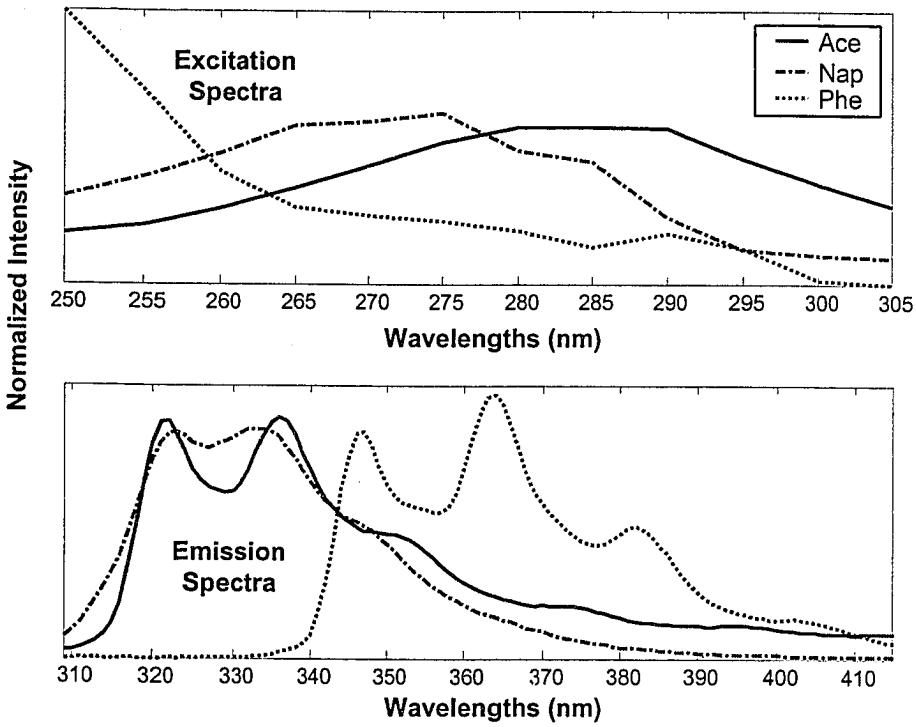

Figure 4.3. Pure excitation (top panel) and emission (bottom panel) normalized spectra of the compounds employed in this work. Each spectrum is the average of ten replicate measurements.

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 [42]. Again, the excitation range was scanned from the highest to the lowest excitation wavelength. These different designs are represented pictorially in Figure 4.4.

4.4.2 Computational Aspects

All the calculations performed in this work were carried out on a Sun Ultra 60 workstation with 2 x 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 [50].

4.5 Results and Discussion

4.5.1 Analysis of the error covariance matrices

Figures 4.5 to 4.7 show the pooled correlation matrices of each mode for Data Sets 1, 2 and 3, respectively. They are plotted 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

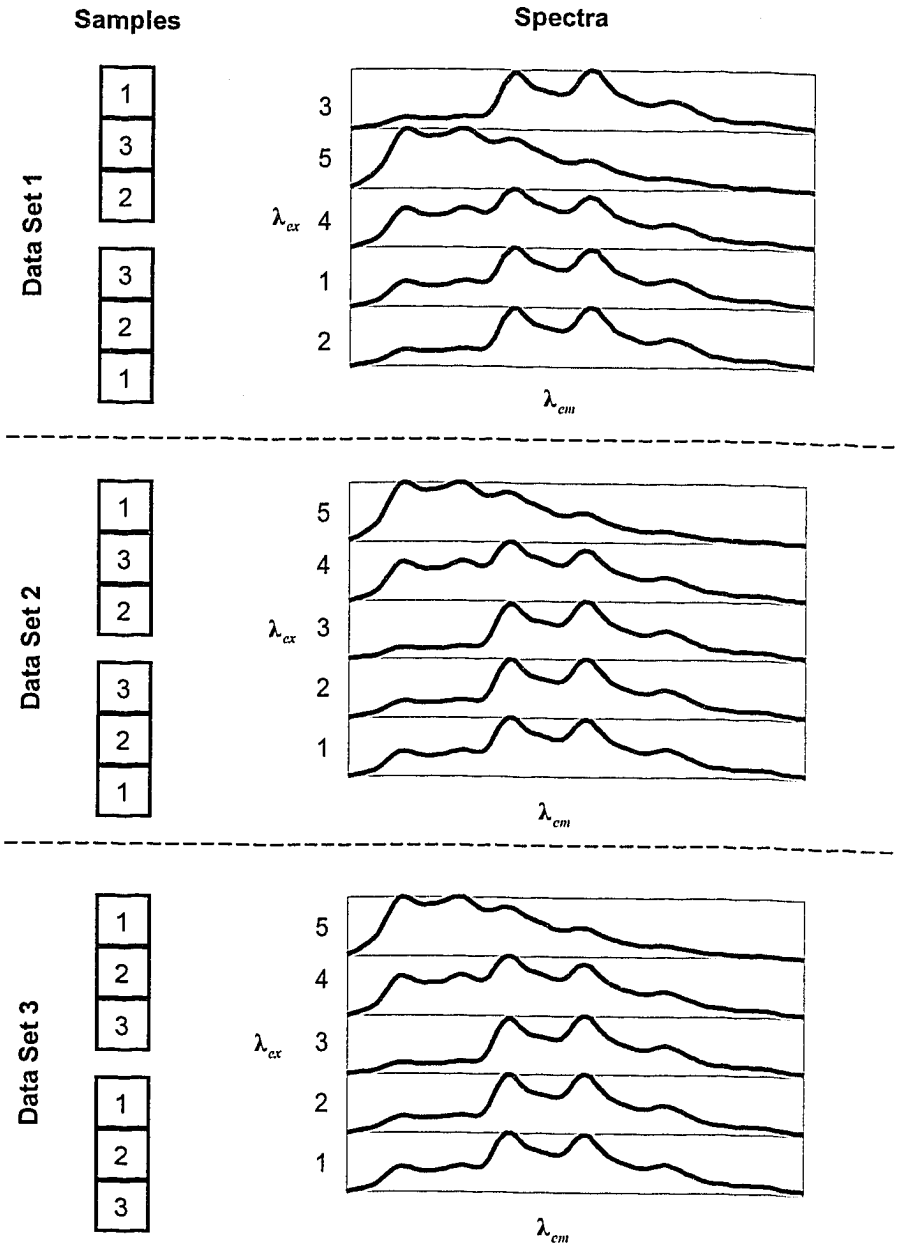

Figure 4.4. Simplified pictorial representation of the experimental designs employed to acquire Data Sets 1, 2 and 3.

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* x *JK* format for the error covariance matrix. Order permutations are not necessary for Data Set 3, since the general MLPARAFAC algorithm will be needed to provide optimal estimates, requiring a full *IJK* x *IJK* error covariance matrix and merit the use of compression in order to use the algorithm.

Although, previous results indicate that the use of *J* x *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 that the different objects pooled produce different sources of structure (results not shown) indicating again that the use a pooled *J* x *J* error covariance matrix would be sub-optimal.

The flow chart in Figure 4.2 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.

Following the right-hand side of the flowchart in Figure 4.2, individual error covariance matrices were first calculated as given in step 2.1. These were organized into four different split-half groups and each was decomposed by PARAFAC (step 2.2). Figure 4.8 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

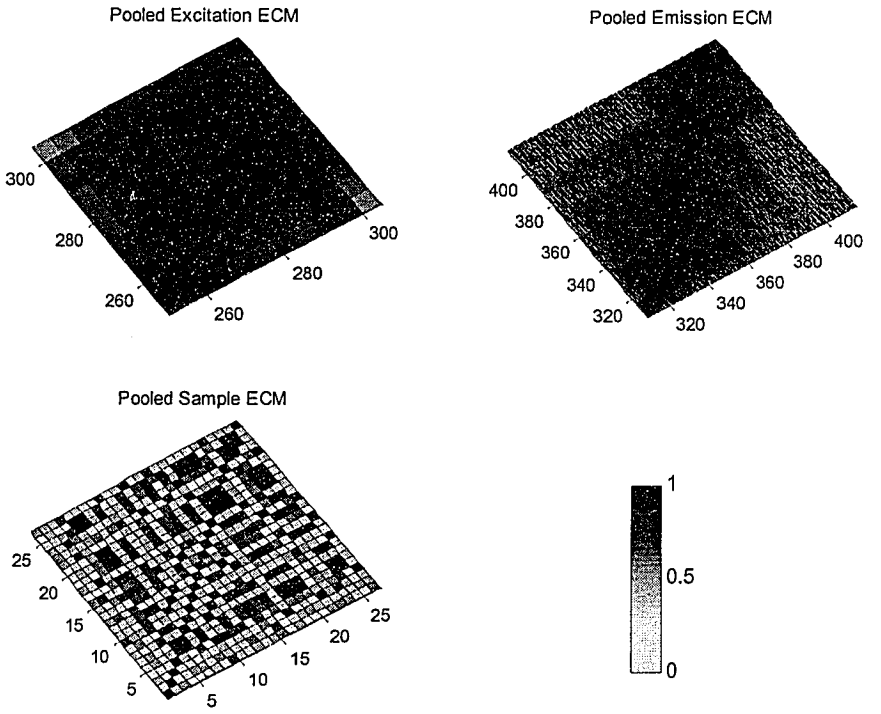

Figure 4.5. Pooled correlation matrices for each mode of Data Set 1 using intensity maps.

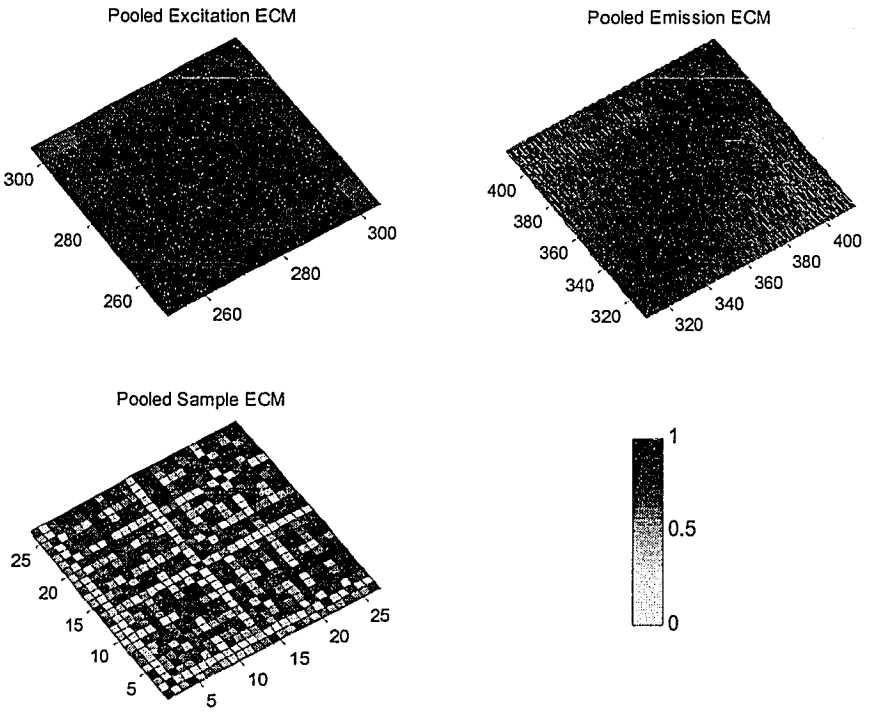

Figure 4.6. Pooled correlation matrices for each mode of Data Set 2 using intensity maps.

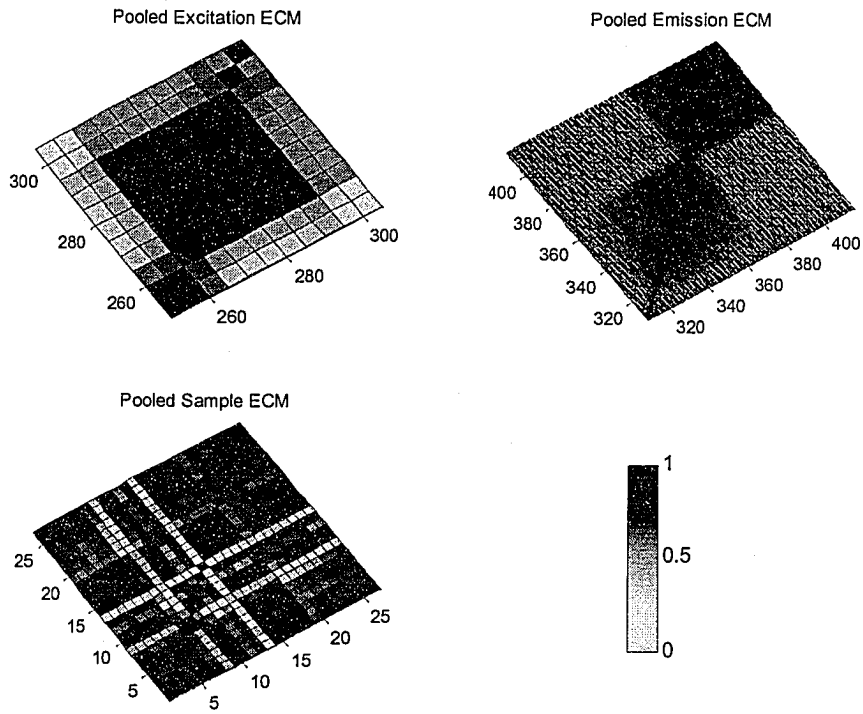

Figure 4.7. Pooled correlation matrices for each mode of Data Set 3 using intensity maps.

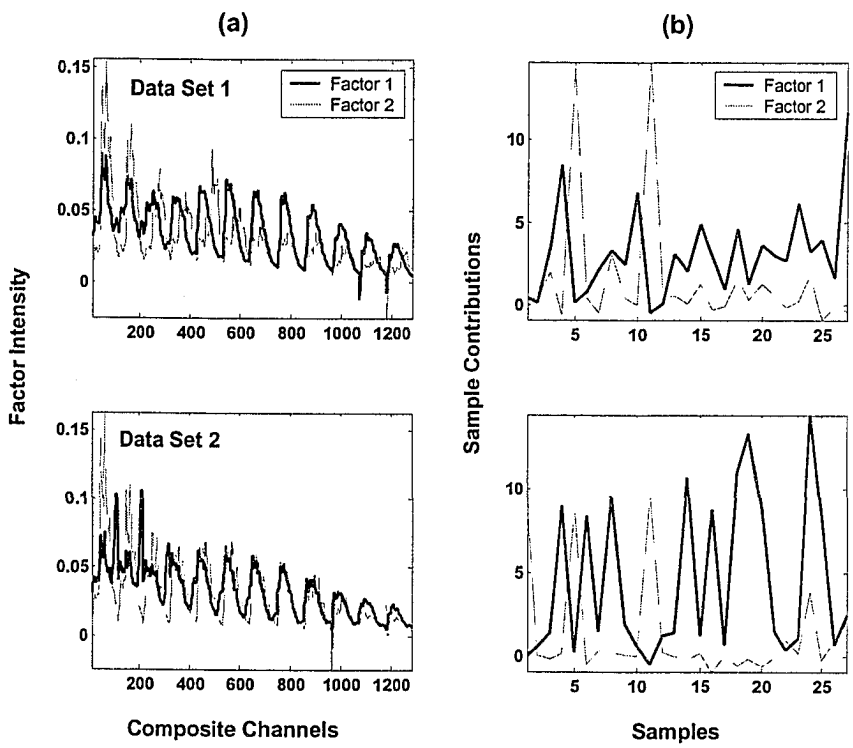

Figure 4.8. Results of two-component PARAFAC decomposition of the individual error covariance matrices for the composite mode formed by excitation and emission modes for Data Sets 1 and 2: (a) structural factors, (b) sample contributions.

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 channels. As discussed in reference 45 and shown in step 2.3 of Figure 4.2, different pieces of information, such as the variance accounted for the model, the corcondia value [51], 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. Based on these evidences, it was concluded that the error covariance matrices could be represented by two factors.

The next step in this process was to assess the homogeneity of the structural factors (step 2.4 in the flowchart) using a combination of metrics and visual analysis (step 2.5 in the flowchart). 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 to 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 4.3.

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 4.8 and 4.10. 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 [45], the departure from homogeneity due to sample contribution will not 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. Based on this, it was concluded that the error covariance matrices were sufficiently homogeneous to permit pooling (step 2.6 in the flowchart) and the use of one *JK* x *JK* error covariance is recommended.

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.

4.5.2 Estimation assessment

4.5.2.1 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:

$$RMSEE_p^r = \sqrt{\frac{\sum(\hat{\mathbf{y}}_p^r - \mathbf{y}_p^o)^2}{N_s - 1}} \tag{4.11}$$

where $\hat{\mathbf{y}}_p^r$ represents the estimated N_s x 1 vector of concentrations for component p and replicate block r, $\mathbf{y}_p^o$ is the corresponding N_s x 1vector of standard concentrations, and N_s is the number of samples. The use of (N_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 4.12:

$$\overline{RMSEE_p} = \frac{\sum_{r=1}^{R} RMSEE_p^r}{R} \tag{4.12}$$

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

$$\overline{RRMSEE}_p = \frac{\overline{RMSEE}_p}{\bar{y}_p} \tag{4.13}$$

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 4.14:

$$\theta_p^r = \cos^{-1}\left(\frac{\mathbf{f}_p^{\mathrm{T}}\mathbf{f}_p^r}{\|\mathbf{f}_p\|\|\mathbf{f}_p^r\|}\right) \tag{4.14}$$

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:

$$\bar{\theta}_p = \frac{\sum_{r=1}^{R} \theta_p^r}{R} \tag{4.15}$$

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:

$$PR_{spec} = \frac{\left(\sum_{p=1}^{P} \bar{\theta}_p^{\mathrm{X}}\right)}{\left(\sum_{p=1}^{P} \bar{\theta}_p^{\mathrm{PARAFAC}}\right)}$$

$$PR_{conc} = \frac{\left(\sum_{p=1}^{P} \overline{RRMSEE}_p^{\mathrm{X}}\right)}{\left(\sum_{p=1}^{P} \overline{RRMSEE}_p^{\mathrm{PARAFAC}}\right)} \tag{4.15}$$

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 perrformance 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.

4.5.2.2 Performance of the algorithms

Although a proper permutation arrangement and format for the error covariance matrix were suggested for each data set in Section 4.4.1, 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 4.1, 4.2 and 4.3 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 $\overline{RRMSEE}_p$ 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, 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,

Table 4.1 Results obtained by different algorithms when applied to different arrangements of Data Set 1. Row(s) in bold represent(s) best case scenario.

		$\overline{RRMSEE}$				$\bar{\theta}$ (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

Table 4.2 Results obtained by different algorithms when applied to different arrangements of Data Set 2 Row(s) in bold represent(s) best case scenario.

		$\overline{RRMSEE}$				$\bar{\theta}$ (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

Table 4.3 Results obtained by different algorithms when applied to different arrangements of Data Set 1. Row(s) in bold represent(s) best case scenario.

		$\overline{RRMSEE}$				$\bar{\theta}$ (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**

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 x JK error covariance matrix, and 1D, which use a set of JK x 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 behaviour of some odd-numbered samples in the first half of the data set (see Section 4.4.1). 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 [44], but results were not especially sensitive to them as long as a sufficient number of components was 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 4.1-3, 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.

4.6 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 [44] 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 [45], 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 emision 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.

Acknowledgements:

The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Dow Chemical Company (Midland, MI).

4.7 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. Chim. 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. *Chemom. 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. *Chemom. Intell. Lab. Syst.* 1997; **38**: 149-171
8. Xie YL, Baeza-Baeza JJ, Ramis-Ramos G. Second-order tensorial calibration for

kinetic spectrophotometric determination. *Chemom. Intell. Lab. Syst.* 1996; **32**: 215-32.

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**: 718-25.

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**: 3293-9.

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**: 449-53.

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**: 3328-36

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**: 2561-9

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**: 2552-60.

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**: 2067-70.

17. Appellof CJ, Davidson ER. Strategies for analyzing data from video fluorometric monitoring of liquid chromatographic effluents. *Anal. Chem.* 1981; **53**: 2053-6.

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**: 2242-7

20. Phillips GR, Georghiou S. Global analysis of steady-state polarized fluorescence spectra using trilinear curve resolution. *Biophys. J.* 1993; **65**: 918-26

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**: 9158-62.

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**: 317-20

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. *Chemom. 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 R.A. Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode 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, *Chemom. 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, *Chemom. 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, *Chemom. Intell. Lab. Syst.* 2000; **52**: 33– 43.

36. 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.

37. JiJi RD, Andersson GG, Booksh KS. Applocation of PARAFAC for calibration with excitation-emission matrix fluorescence spectra of three classes of environmental pollutants. *J. Chemom.* 2000; **14**: 170-185.

38. 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

39. Bro, R. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemom. Intell. Lab. Syst.* 1999; **46**: 33-147.

40. Moberg L., Robertsson G, Karlberg B. Spectrofluorimetric determination of chlorophylls and pheopigments using parallel factor analysis. *Talanta* 2001; **54**: 161-170.

41. 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.

42. Ingle JD, Crouch SR. *Spectrochemical Analysis* Prentice Hall: New Jersey, 1972.

43. Vega-Montoto L, Wentzell PD. Maximum Likelihood Parallel Factor Analysis (MLPARAFAC) *J. Chemom.* 2003; **17**: 237.

44. Vega-Montoto L, Gu H, Wentzell PD. Mathematical Improvements to maximum likelihood parallel factor analysis: theory and simulations *J. Chemom.* **Submitted**

45. Leger M, Vega-Montoto L, Wentzell PD. Methods for systematic investigation of measurement error covariance matrices. *Chemom. Intell. Lab. Syst.* 2004; **In press**

46. Wentzell PD, Lohnes MT. Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemom. Intell. Lab. Syst.* 1999; **45**: 65.

47. Brown CD, Vega-Montoto L, Wentzell PD. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Appl. Spectrosc.* 2000; **54**: 1055.

48. Mardia K, Kent JT, Bibby JM, *Multivariate Analysis* Academic Press: 1979.

49. 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

50. Andersson CA, Bro R, The N-way toolbox for MATLAB. *Chemom. Intell. Lab. Syst.* 2000; **52**: 1.

51. Bro R, Kiers HAL. A new efficient method for determining the number of component in PARAFAC models. *J. Chemom.* 2003; **17**: 274.

Chapter 5

Approaching the Direct Exponential Curve Resolution Algorithm from a Maximum Likelihood Perspective[4]

5.1 Abstract

The implementation of Maximum Likelihood Parallel Factor Analysis (MLPARAFAC) in conjunction with the Direct Exponential Curve Resolution Algorithm (DECRA) is described. DECRA takes advantage of the intrinsic exponential structure of some bilinear data sets to produce trilinear data by a simple shifting scheme, but this manipulation generates an error structure that is not optimally handled by traditional three-way chemometrics methods such as TLD and PARAFAC. In this work, the effects of these violations are studied using simulated and experimental data used in conjunction with the well-established TLD and PARAFAC. The results obtained by both methods are compared with the results obtained by MLPARAFAC, which is a method designed to optimally accomodate a variety of measurement error structures. The impact on the estimates of different parameters linked to the data sets and the DECRA method is investigated using simulated data. The results indicate that PARAFAC produces estimates of much poorer quality than TLD and MLPARAFAC. Also, it was found that the quality TLD estimates was comparable or only marginally poorer than the MLPARAFAC estimates. A number of commonly used algorithms were also compared to MLPARAFAC using two sets of published experimental data from kinetic studies. The MLPARAFAC estimates of rate constants were more precise than the other methods examined.

---

[4] Submitted to Analytica Chimica Acta (Submitted March 2005)

5.2 Introduction

The method called the Direct Exponential Curve Resolution Algorithm (DECRA), introduced to the chemometrics literature by Antalek and Windig [1], was originally formulated as a manageable alternative to treat pulse-gradient-spin-echo (PGSE) NMR data for separating highly overlapped spectra. In PGSE-NMR experiments, the strength of two sets of magnetic field gradient pulses is varied to produce a data set in which the signal of each component forming the sample decays exponentially. In turn, this leads to two dimensions: one dimension is conventional chemical shift and the other is diffusion time. The self-diffusion coefficient is a property intimately related to characteristics of an individual sample, such as size, shape, mass and charge, as well as its surrounding environment, such as solution, temperature and aggregation state. This quantity can be calculated for each component from these experiments since the decays are a function of time. From an experimental point of view, the use of this technique is extremely advantageous since it provides a non-invasive and straightforward method to obtain both physical and chemical information. However, the main challenge of this technique resides in the mathematical implementation of data analysis. Two classes of techniques are commonly used to interpret the data: single channel methods and multivariate methods. These two approaches differ from a fundamental perspective. The single channel methods, represented by SPLMOD and CONTIN [2,3], rely on a hard model where pure or nearly pure channels are sought to use during the estimation process. Unfortunately, overlapping regions are commonly present in the spectra of real mixtures, hence single channel methods find it difficult to deal with complex mixtures. Unlike single channel methods, multivariate methods analyze the total information available in the data set simultaneously. The most important characteristic of this approach is the capability to analyze overlapping regions. Mathematically, the ideal method of choice is multivariate curve resolution (MCR) [4] due to the intrinsic bilinear structure of the data. However, it is well known that an infinite number of solutions are possible with MCR even when the model is estimated using several constraints. This deficiency led Schulze and Stilbs [5] to utilize a method proposed by Kubista [6] for the estimation process. The solution for Kubista's method is equivalent to the already well-established generalized

rank annihilation method (GRAM), as shown by Sanchez and Kowalski [7]. GRAM solves the structural PARAFAC model [8] using an extended eigenvalue problem. The strict trilinear structure needed to estimate the model was theoretically sound but impractical from an experimental point of view since the proposed experiment required the use of a complicated data acquisition scheme in order to avoid peak shift and gradient level shift while a second sample was recorded.

These problems motivated Antalek and Windig to propose DECRA. The innovative idea behind DECRA's implementation is the use of the existing underlying exponential structure in the decay data, eliminating the need for a second sample. The method utilizes only a single sample described by an exponential decay. Then this sample is shifted in the "time domain" yielding a second sample which is proportional to the first. To illustrate this basic principle with a simple example, consider a vector of exponentially decaying values, **d**, as given in equation 5.1:

$$\mathbf{d} = e^{-k\mathbf{t}} \tag{5.1}$$

where k is the decay constant and **t** represents a vector of uniformly increasing time values of length N. We can now parse this vector into new vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$:

$$\begin{aligned} \mathbf{d}_1 &= [d_1\ d_2\ d_3 \ldots\ d_{N-S}]^{\mathrm{T}} \\ \mathbf{d}_2 &= [d_{S+1}\ d_{S+2}\ d_{S+3} \ldots\ d_N]^{\mathrm{T}} \end{aligned} \tag{5.2}$$

where S is a positive integer less than N. These vectors can be combined into a matrix **D**:

$$\begin{aligned} \mathbf{D} = [\mathbf{d}_1\ \mathbf{d}_2] &= [\exp(-k\mathbf{t}')\ \ \exp(-k\{\mathbf{t}'+\mathbf{t}_S\})] \\ &= [\exp(-k\mathbf{t}')\ \{\exp(-k\mathbf{t}')\exp(-k\mathbf{t}_S)\}] \\ &= \exp(-k\mathbf{t}')[1\ \ \exp(-kt_S)] \end{aligned} \tag{5.3}$$

Here, **t'** represents the vector formed by the first (N-S) elements of **t** and t_S is element S of **t**. Note that matrix **D** has a rank of unity since the second column will simply be a scalar multiple of the first. Furthermore, the decay constant can be obtained from the ratio of the elements in the two columns:

$$k = \frac{1}{t_S}\ln\left(\frac{d_{i1}}{d_{i2}}\right) \tag{5.4}$$

Thus, this simple process of creating a matrix (two-way data) by offsetting a single vector (one-way data) in time, leads to a method which can extract the decay constant. With DECRA, this relatively simple concept is extended in a number of ways. First, DECRA is based on the idea that any system that can be decomposed into a linear combination of exponentials can be subjected to an analysis similar to above. Moreover, DECRA starts with bilinear matrices (two-way data) and, through the shifting process, generates trilinear data (three-way data). The decomposition of these matrices, which is described in greater detail in Section 5.2.1, yields spectral and decay constant information.

This simple but elegant idea was easily extended to the PGSE-NMR bilinear data, yielding a two sample trilinear array suitable to be decomposed using GRAM. The use of DECRA instantly gained a number of advocates because of the possibility of obtaining unique solutions to the decomposition problem. GRAM was initially preferred over MCR and other multilinear methods [9] due to the non-iterative nature of the algorithm.

The application of this technique to other types of data such as magnetic resonance spectra and images [1, 11-14], short-wavelength near-infrared [15], UV-Vis [16], solid-state NMR and mid-infrared spectroscopy [17], motivated the exploration of other decomposition methods such as direct trilinear decomposition (DTLD) [9], Levenberg-Marquardt-PARAFAC (LM-PARAFAC) [15], weighted curve resolution (WCR) [9] and successive Bayesian estimation (SBE) [18]. With the exception of DTLD, which is an extension of GRAM for cases where more than two samples are analyzed, the other methods share a common property that makes them different from GRAM. GRAM and DTLD provide exact solutions to the decomposition problem when the data are not corrupted by noise. For cases where noise is present, the solutions are approximated but not in an optimal sense. In contrast, all the other methodologies have a well-defined objective function and optimization strategy providing the optimal values when the noise follows a normal-*iid* (independent and identically distributed) structure. Unfortunately, these noise conditions are rarely met in real applications, making the estimation suboptimal. Furthermore, the situation becomes worse when the bilinear data are shifted in order to produce the second pseudo-sample, since some correlated errors appear as a consequence of the shift. Even though this situation has been barely recognized in the literature [19-20], Pedersen *et al.* [19] tried to correct for it by

introducing a shift scheme in which non-overlapping measurements are used to produce the second pseudo-sample, but their results were not promising. Recently, a method called maximum likelihood parallel factor analysis (MLPARAFAC) was introduced [21-23] to treat cases where the standard assumptions about the noise were violated for trilinear data. The presence of a trilinear structure corrupted by non-*iid* noise makes the matrices used for DECRA quite suitable for MLPARAFAC. These three-way data sets fulfill the trilinear condition needed to apply the PARAFAC model, and are constructed by shifting bilinear data so that a non-*iid* structure will result, making MLPARAFAC the most optimal alternative. This new method will be applied to two experimental data sets and some simulated data to show the improvements in terms of accuracy and precision in the estimates.

5.3 Theoretical aspects

5.3.1 DECRA

In this paper, MLPARAFAC will be used to estimate the rate constants for two kinetic studies previously reported in the literature [9]. Accordingly, a first-order kinetic model will be used to illustrate the effects of applying the principle of DECRA to the structure of the chemical data and the noise. Suppose that a system under study involves consecutive first-order reactions as shown below:

$$\mathrm{U} \xrightarrow{k_1} \mathrm{V} \xrightarrow{k_2} \mathrm{W}$$

Let the M x N matrix $\mathbf{Y}$ be a collection of spectra obtained during the time course of the reaction, with M equidistant time points at N wavelengths for the mixture of the three species. Assuming the Lambert-Beer's Law holds, $\mathbf{Y}$ can be decomposed into the product shown in equation 5.5:

$$\mathbf{Y} = \mathbf{F}\mathbf{L}^{\mathrm{T}} + \mathbf{E} \tag{5.5}$$

where every row in $\mathbf{Y}$ denotes a spectrum recorded at certain time. $\mathbf{F}$ is an M x P matrix (P is the number of components) with the variation of the concentration profile in time for each component along each column. $\mathbf{L}$ is an N x P matrix containing the pure spectra for each component along the columns, and $\mathbf{E}$ is an M x N residual matrix that includes

model errors, experimental errors and instrumental noise. Analytically, the columns of **F** can be represented by the integrated rate laws shown in equations 5.6 to 5.8:

$$\mathbf{f}_U = F_U^{\circ} e^{-k_1 t} \tag{5.6}$$

$$\mathbf{f}_V = \frac{k_1 F_U^{\circ}}{(k_2 - k_1)}(e^{-k_1 t} - e^{-k_2 t}) = kF_U^{\circ}(e^{-k_1 t} - e^{-k_2 t}) \tag{5.7}$$

$$\mathbf{f}_W = F_U^{\circ} - \mathbf{f}_U - \mathbf{f}_V \tag{5.8}$$

where $\mathbf{f}_U$, $\mathbf{f}_V$ and $\mathbf{f}_W$ are the concentration profiles of species U, V, W along the time vector **t,** respectively. Vector **t** is the sequence of uniformly spaced times at which the reaction is monitored and it is represented as: $\mathbf{t} = [t_1 \ t_2 \ \cdots \ t_M]^T$. F_U° is the initial concentration of component U. Equations 5.6 to 5.8 indicate that all of the columns of **F** (and therefore **Y**) are exponentials or combinations of exponential equations. The bilinear decomposition of **Y** can be equivalently expressed as shown in equation 5.9 when equation 5.5 is expanded and the similar terms are grouped:

$$\begin{aligned} \mathbf{Y} &= \mathbf{F}\mathbf{G}^T = \mathbf{A}\mathbf{B}^T \\ &= \mathbf{a}_1\mathbf{b}_1^T + \mathbf{a}_2\mathbf{b}_2^T + \mathbf{a}_3\mathbf{b}_3^T \end{aligned} \tag{5.9}$$

where $\mathbf{a}_1 = e^{-k_1 t}$, $\mathbf{a}_2 = e^{-k_2 t}$, $\mathbf{a}_3 = e^{0t}$, $\mathbf{b}_1 = \mathbf{g}_U + k\mathbf{g}_V - (k\text{-}1)\mathbf{g}_W$, $\mathbf{b}_2 = k(\mathbf{g}_W - \mathbf{g}_V)$ and $\mathbf{b}_3 = \mathbf{g}_W$ with $k = k_1/(k_2 - k_1)$. The concentration profiles $\mathbf{f}_U$, $\mathbf{f}_V$ and $\mathbf{f}_W$ are rearranged into exponentially decaying functions $\mathbf{a}_1$, $\mathbf{a}_2$ and $\mathbf{a}_3$, and the columns of **B** are linear combinations of the spectra of U, V and W, designated by $\mathbf{g}_U$, $\mathbf{g}_V$ and $\mathbf{g}_W$, respectively. Therefore, the response matrix **Y** consists of linear combinations of exponentially decaying functions, which justifies the implementation of DECRA. The matrix **Y** (M x N) is used to build two matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$, partitioning **Y** with a constant time shift S as shown:

$$\mathbf{X}_1 = \begin{bmatrix} y_{11} & \cdots & y_{1N} \\ \vdots & & \vdots \\ y_{(M-S)1} & \cdots & y_{(M-S)N} \end{bmatrix} \tag{5.10a}$$

$$\mathbf{X}_2 = \begin{bmatrix} y_{(1+S)1} & \cdots & y_{(1+S)N} \\ \vdots & & \vdots \\ y_{M1} & \cdots & y_{MN} \end{bmatrix} \tag{5.10b}$$

Equations 5.11 and 5.12 show the factorization of both matrices using the bilinear model shown in equation 5.5:

$$\mathbf{X}_1 = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \vdots & \vdots & \vdots \\ a_{(M-S)1} & a_{(M-S)2} & a_{(M-S)3} \end{bmatrix} \mathbf{B}^{\mathrm{T}} = \mathbf{A}_1 \mathbf{B}^{\mathrm{T}} \tag{5.11}$$

$$\mathbf{X}_2 = \begin{bmatrix} a_{(1+S)1} & a_{(1+S)2} & a_{(1+S)3} \\ \vdots & \vdots & \vdots \\ a_{M1} & a_{M2} & a_{M3} \end{bmatrix} \mathbf{B}^{\mathrm{T}} = \mathbf{A}_2 \mathbf{B}^{\mathrm{T}} \tag{5.12}$$

Equation 5.12 can be rewritten as equation 5.13 by using the intrinsic structure present in the exponential profiles shown in equation 5.3:

$$\mathbf{X}_2 = \mathbf{A}_1 \begin{bmatrix} \mathrm{e}^{-k_1 S} & & \\ & \mathrm{e}^{-k_2 S} & \\ & & \mathrm{e}^{0S} \end{bmatrix} \mathbf{B}^{\mathrm{T}} \tag{5.13}$$

This transformation makes $\mathbf{X}_1$ and $\mathbf{X}_2$ appears as two parallel matrices with different contributions to the same concentration and spectral profiles. In other words, this partitioning strategy produces trilinear data (as represented in equation 5.14) that can be decomposed using different algorithms.

$$\mathbf{X}_a = [\mathbf{X}_1 \; \mathbf{X}_2] = \mathbf{A}_1 \mathbf{I}_a (\mathbf{C}^{\mathrm{T}} \otimes \mathbf{B}^{\mathrm{T}}) \tag{5.14}$$

Here $\mathbf{X}_a$ is the (M-S) x $2N$ matrix representing the M-S x N x 2 array $\underline{\mathbf{X}}$ unfolded to retain mode A. $\mathbf{I}_a$ (P x P^2) is the unfolded superdiagonal "identity" matrix of order P used to obtain the column wise Kronecker product of the loadings for modes B and C. Matrix $\mathbf{C}$ holds a new set of loadings pertaining the reaction rate constant as shown in equation 5.15.

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 \\ e^{-k_1 S} & e^{-k_2 S} & e^{0S} \end{bmatrix} \tag{5.15}$$

This methodology has been further extended to cases in which three or more slices are generated [9] using the same shift S. The argument given in the original paper is that this can be advantageous in terms of noise reduction. The procedure is similar to the two-slice case, but in this case (K-1) estimates (*i.e.* K is the number of created slices) can be obtained for each rate constant, and these can be averaged to obtain a better estimate. For example, the composition matrix of the **C** matrix for $K = 3$ and $P = 3$ is shown in equation 5.16.

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 1 \\ e^{-k_1 S} & e^{-k_2 S} & e^{0S} \\ e^{-k_1 S} & e^{-k_2 S} & e^{0S} \end{bmatrix} \tag{5.16}$$

So far, the authors have used the standard notation employed in previous papers to introduce DECRA in which mode A (rows) of the three-way array (I x J x K) represent the concentration profiles (time), mode B (columns) represents the spectra (wavelength) and mode C (slices) represents the rate constants. However, in order to maintain a consistent notation with the previous MLPARAFAC papers [21-23], in this paper the trilinear data will be arranged in a fashion such that the time domain is represented by mode B, mode A will carry spectral information and mode C will carry information about the rate constants. This permutation pre-step is due to the loss of symmetry of the PARAFAC model due to the introduction of the error information as explained in reference 22. The value for I will be equal to the number of spectral channels N, and K will be the number of slices that will be used. The value of J will be a function of the original number of time points M, the data shift value S, and the number of slices K via equation 5.17:

$$J = M - (K-1)S \tag{5.17}$$

5.3.2 Noise considerations

Up to this point, the DECRA procedure has been illustrated by using the part of the data that is related to the underlying deterministic structure. However, the manipulations are carried out over all components of the acquired data, and therefore they are consequently carried out over the noise structure as well, leading to some artificial correlations among the elements when the trilinear data are analyzed. This is a

consequence of the shifting strategy, which copies measurement errors along with the data when creating a new slice of the three-way array, leading to perfectly correlated noise among the slices. Even for the simplest scenario, where noise in the original matrix is independent and identically distributed, the pattern of correlation is going to be considerable. This is illustrated in Figure 5.1, which shows the relationship among the measurements (and therefore measurement errors) in the folded and unfolded arrays for a simple 3 x 5 matrix with $S = 1$ and $K = 3$.

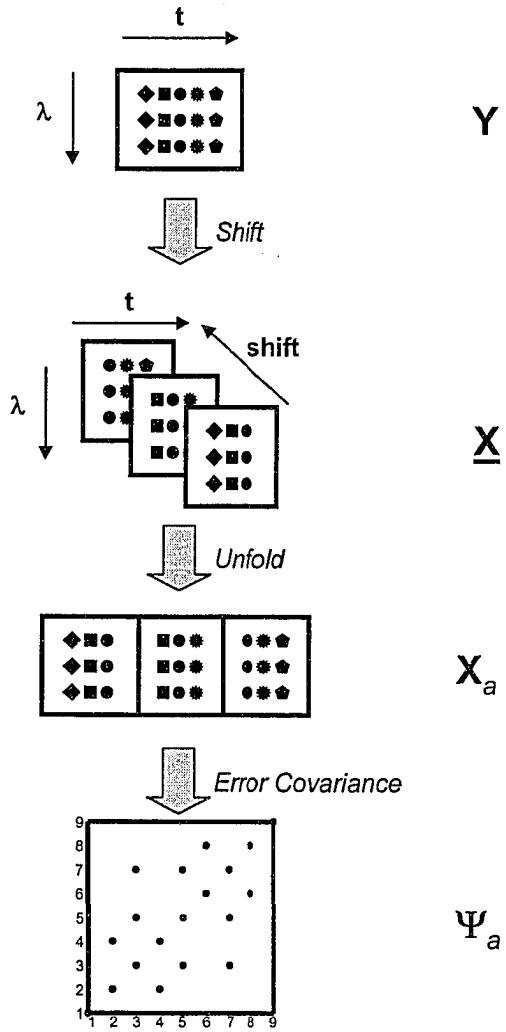

Figure 5.1. Illustration of the DECRA procedure and its effects on the error structure.

The I x J x K array is unfolded to retain mode A, which is the spectral order. This gives an I x JK matrix. Because of the shifting process, a certain number of measurements will be repeated among the slices (unless $S > (M/K) + 1$). This means that the measurement

errors for repeated measurements will be perfectly correlated, leading to a near band-diagonal structure in the JK x JK error covariance matrix for the rows of the unfolded matrix, $\mathbf{X}_a$, as shown in Figure 5.1. The error covariance matrix is not perfectly band-diagonal because there are some unrepeated measurements in each slice, so the bands off the main diagonal have some uncorrelated measurements. For *iid* measurement noise, all of the non-zero elements of this matrix will be equal to σ^2, the variance of the measurement errors. In practice σ^2 would not need to be known, since only relative magnitudes are important in maximum likelihood estimation. For *iid* noise case, the error covariance matrix will be the same for all of the rows of $\mathbf{X}_a$, and the total number of bands (including the main diagonal) will be $2K$-1 (except when $S > (M/K) + 1$, where there will only be the main diagonal).

Figure 5.2 shows the effect of various values of the number of slices ($K = 2, 3, 4$) and shift parameter ($S = 3, 6, 11$) on the structure of the error covariance matrix for $\mathbf{X}_a$ in the *iid* noise case. This figure assumes that the total number of time points is 36, which together with K and S, will define the dimension J of $\underline{\mathbf{X}}$ (as given in equation 5.17) and hence the dimension of the unfolded matrix, $\mathbf{X}_a$, and the error covariance matrix, Ψ_a. The reduction in size for the time mode is the most important effect related to the increment of the shift step S, although it will also affect the length of the bands off the main diagonal and eliminate them when S is larger than M/K. A case where the value of S is larger than this limit is represented at the bottom right of Figure 5.2. As mentioned earlier, such a case where the shift was conveniently made to eliminate correlation did not produce good estimates, as reported in reference 19.

Real error structures are much more complex than the *iid* scenario depicted above. PGSE-NMR data and spectrophotometric-kinetic data, which are the most common types of bilinear data used in conjunction with DECRA, are already permeated by correlated noise due to the continuous way in which the data are recorded. This produces a temporal noise correlation related to experimental variables (*e.g.* start time, gradient variations, temperature, instrument drift, etc.) along the sample mode, which will also combine with correlation that arises from factors such as cell positioning and cross talk among different channels in the spectral domain. The effects on both orders will readily interact due to

experimental design factors. Therefore, the correlation pattern in these cases will be much more complex than the cases represented in Figures 5.1 and 5.2. This extreme departure

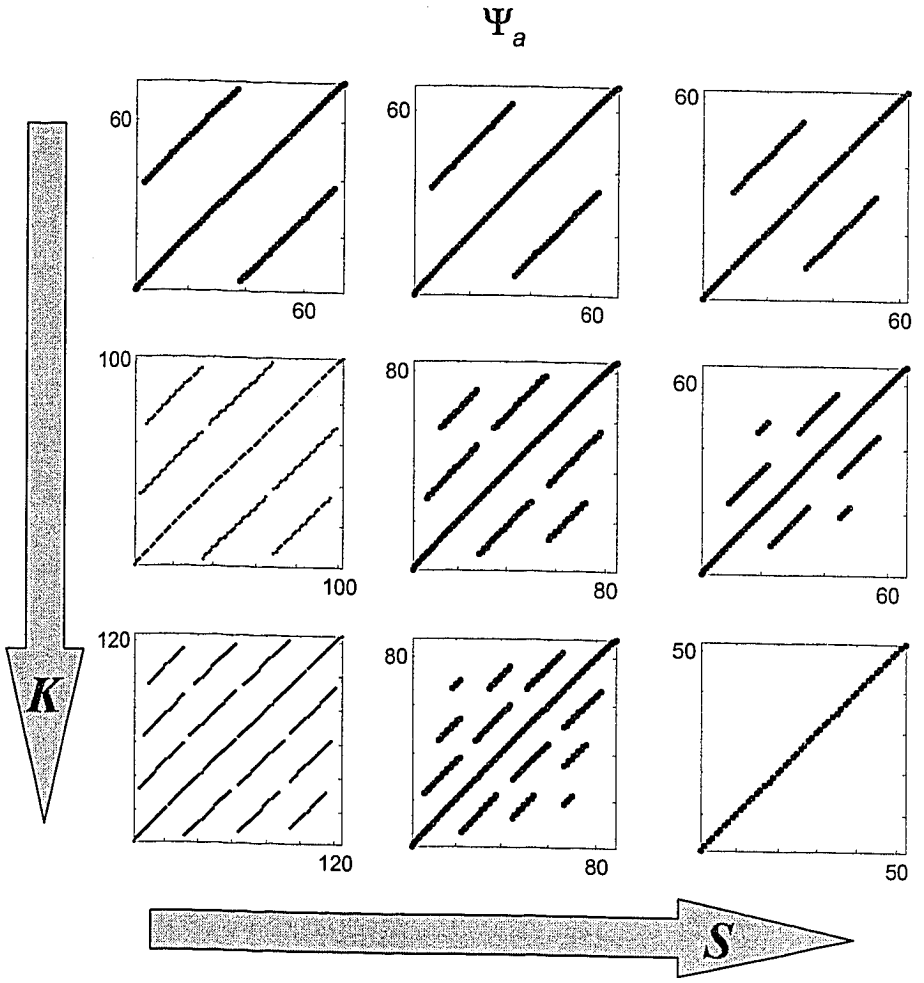

Figure 5.2. Illustration of the relationship between the error structure and the values of the shifting parameter (S) and the number of slices (K). In this example M =36, S = 3, 5 and 11, and K = 2, 3 and 4.

from the expected *iid* noise characteristics should have an impact on the estimation process. The main aim of this paper is to investigate the extent of this impact by comparing the results of different estimation methods with the results of MLPARAFAC for two well-studied data sets and a number of simulated data sets.

It should be noted that there are not widely accepted criteria given in the literature for the selection of values of S and K. Different selection strategies, ranging from empirical expertise to trial-and-error, are usually employed. In general, only anecdotal explanations, such as the relationship between exponential decay rate and the shift

parameter [16] and the relationship between the number of slices and noise reduction during the estimation process [9], have been given as rational reasons for these approaches. For this reason, a variety of S and K values will be examined in this study.

5.3.3 MLPARAFAC

MLPARAFAC is an errors-in-variables modeling method that accounts for measurement errors in the estimation of model parameters for trilinear data. It is an optimal modeling method in a maximum likelihood sense for functional models with no errors in the model equations. It is a natural extension to PARAFAC of the MLPCA method introduced by Wentzell *et al.* [24]. The mathematical aspects of the algorithm have been described in detail elsewhere [21,22] to allow the principles to be readily applied. There are a number of simplified algorithms that can accommodate a variety of error structures, such as heteroscedastic and correlated noise in one and two orders. Also, when correlation is affecting more than two orders, the more general algorithm is normally used after a compression step [22] to make the size of the data manageable. All of the algorithms have excellent convergence characteristics because their core is based on the same alternating least squares procedure as the original PARAFAC. The statistical properties of these algorithms have been tested using simulated data, and the corresponding improvement in the quality of parameter estimates has been demonstrated using simulated and experimental data [22,23].

Two different MLPARAFAC algorithms will be employed in this work due to the expected error characteristics of the trilinear data obtained after the application of DECRA. Initially, it will be assumed that the experimental data are corrupted by *iid* noise and the only source of correlation is related to the data shift. In such a case, elements along the time domain in different slices will be affected only by the repetition, as shown in Figure 5.1. In this case, a theoretical error covariance matrix can be formulated from knowledge of the shift value, S, the number of slices used, K, and a global estimate of the error variance. This error covariance matrix will be used with the MLPARAFAC algorithm designed to handle the same correlation structure affecting two orders (*i.e.* simplified algorithm 1C in reference 22). This algorithm, which will be referred to as S-MLPARAFAC in this work, minimizes the following objective function:

$$f = trace[(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C} \otimes \mathbf{B})^T)\boldsymbol{\Psi}_a^{-1}(\mathbf{X}_a - \tilde{\mathbf{A}}(\mathbf{C} \otimes \mathbf{B})^T)^T] \quad (5.18)$$

Here $\mathbf{X}_a$ is an I x JK matrix representing the trilinear data unfolded by preserving mode A (spectral mode). $\tilde{\mathbf{A}}$ is an I x P^2 matrix, which is a compact representation of mode A (spectral information, I x P) multiplied by the P x P^2 unfolded superdiagonal matrix of order P (in this case $P = 3$), $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{I}_a$. $\mathbf{B}$ is a J x P matrix carrying the exponential decay profiles. $\mathbf{C}$ (K x P) will have information regarding the rate constants. Finally, $\boldsymbol{\Psi}_a$ is the JK x JK error covariance matrix describing the variance/covariance structure among the elements in mode B along different slices. It is assumed for this algorithm that the error covariance matrix is the same for all the rows. Usually, this error covariance matrix is calculated using replicate measurements, as has been discussed elsewhere[21-24], but in this case, a theoretical error covariance matrix considering only the shifting process will be used, since it is assumed that the measurements are corrupted by *iid* noise. It is likely that this will not be the most optimal choice, but it will provide an excellent pivotal benchmark to compare methods that do not include any information about the error structure with a method that includes all the available information about the error structure.

The general MLPARAFAC (G-MLPARAFAC) algorithm will be the algorithm of choice to include all the available information about the error structure. This is motivated by the anticipated existence of sources of correlation affecting all the orders. The general algorithm minimizes the following objective function:

$$f = \left\{vec(\mathbf{X}_a^T) - \left[(\mathbf{A} \otimes \mathbf{C} \otimes \mathbf{B})vec(\mathbf{I}_a^T)\right]\right\}^T \boldsymbol{\Omega}_a^{-1} \left\{vec(\mathbf{X}_a^T) - \left[(\mathbf{A} \otimes \mathbf{C} \otimes \mathbf{B})vec(\mathbf{I}_a^T)\right]\right\} \quad (5.19)$$

The only difference of this equation with respect to the previous equation is that, due to the generalized error structure, the error covariance matrix $\boldsymbol{\Omega}_a$ must be IJK x IJK in order to include all the necessary information. Therefore, the estimation process has to be carried out using the vectorized representation of the measurements and the PARAFAC model. Even for small arrays, an IJK x IJK matrix is generally very large, making necessary the use of a compression step. In this work, the previously formulated operational procedure depicted in Table 5.1 of reference 22 will be used. The MLPARAFAC estimates will be compared with estimates obtained by a number of

methods previously used with the experimental data sets. For the case of simulations, the estimates will be compared with TLD estimates (equivalent to GRAM when $K = 2$) and standard PARAFAC, since they represent the sub-optimal but fast methods, and the optimal methods assuming *iid* noise characteristics, respectively.

5.4 Experimental

5.4.1 Simulated data

Simulations are used in this work to characterize the performance of different representative estimation methods in terms of accuracy and precision, since the true values of the parameters are known in advance. Therefore, the data were constructed to closely resemble the experimental data sets used in this paper, which follows the three-component bilinear system as shown in equation 5.5. For the true signal part of the simulated data, the first-order consecutive reaction, already described in Section 5.2.1 of this paper, was used as a model kinetic system. The concentration profiles are decaying functions governed by the reaction constants $k_1 = 0.30$ and $k_2 = 0.05$ respectively. The time range was 0 to 20 min with increments of 0.2 min ($M = 101$). Pure spectra were simulated with Gaussian peaks for the three individual species over a range of 100 channels with increments of 1 nm. The peak maxima of the species U, V and W were at channels 22, 72 and 52 respectively and widths corresponded to $\sigma = 15$ to simulate a strong spectral overlap. The concentration profiles and pure spectra of the individual species are shown in Figure 5.3.

For the noise contribution of the simulated data, different scenarios were generated. Initially, a matrix **E** of random numbers with elements characterized by *iid*-normal noise was created. Three levels, with standard deviations equal to 0.1%, 1% and 2% of the maximum absorbance of the simulated spectrum at time zero, were used. As mentioned in Section 5.2.2, the kinetic data usually used in conjunction with DECRA are corrupted by correlated noise affecting both time and spectral modes. Therefore, in order to mimic these cases, a two-dimensional moving average filter block was convolved with the error matrix **E** before it was added to the error-free measurements. At the boundaries

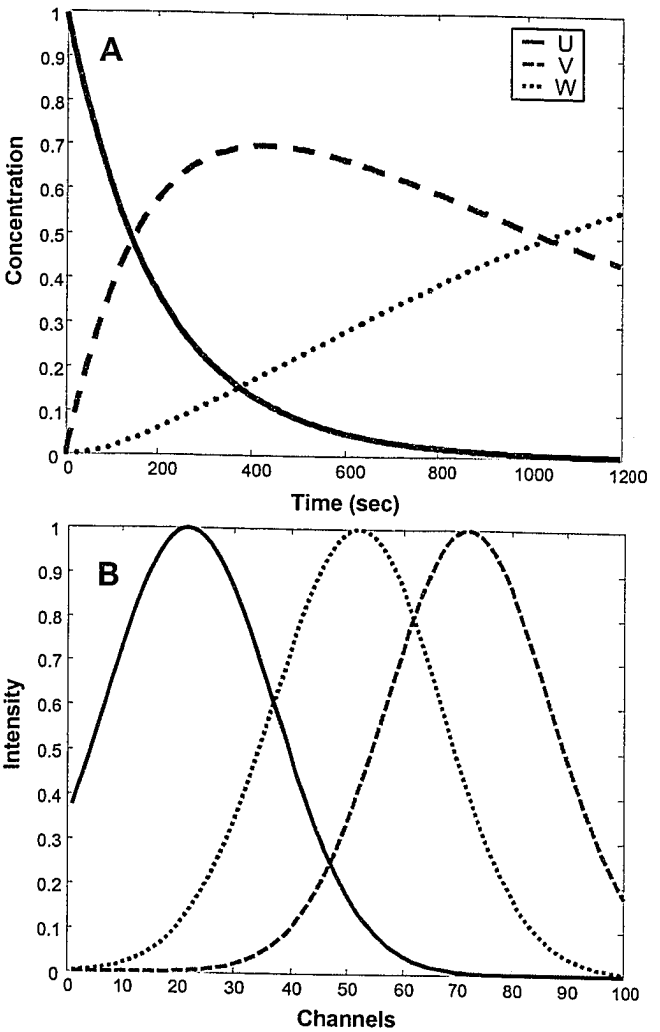

Figure 5.3. Concentrations profiles (A) and spectra (B) of the compounds employed in the simulations.

of the error matrix, the filter was wrapped around to the opposite side in order to eliminate edge effects. Although this approach is not likely to reflect the actual correlation structure in experimental data, it represents a general case for which the degree of correlation and the resultant level of noise can be controlled. The filtering procedure was carried out in an element-wise manner by defining a M x N matrix $\mathbf{F}_{mn}$ containing the filter coefficients applied to the error matrix $\mathbf{E}$ to give the filtered error element $\mathbf{E}_{mn}$. Figure 5.4 shows a pictorial representation of $\mathbf{F}_{11}$, where the filled squares show the position of the filter coefficients for a 3 x 3 filter matrix. For $\mathbf{F}_{12}$ the squares shift right and for $\mathbf{F}_{21}$ they shift down. The correlated noise matrix can be obtained by iterating through equation 5.20:

$$\mathbf{E}_{mn}^{correlated} = vec(\mathbf{F}_{mn})^{T} vec(\mathbf{E}) \tag{5.20}$$

It is worth noting that the convolution with this filter will produce some attenuation of the error variance in addition to the introduction of correlation. The level of attenuation is given by equation 5.21:

$$\frac{\sigma^2_{filtered}}{\sigma^2_{unfiltered}} = \sum c_{ij}^2 \tag{5.21}$$

Here $\sigma^2_{filtered}$ and $\sigma^2_{unfiltered}$ are the noise variances of the filtered and unfiltered noise matrix, respectively, and the c_{ij} represent the coefficients for the smoothing filter used. For a simple moving average filter, all of the coefficients are the same and equal to the reciprocal of the number of coefficients. For example, for a 5 x 5 two-dimensional moving average filter, c_{ij} = 1/25 = 0.04. Therefore, to keep the level of noise variance consistent among the simulations, equation 5.21 was used to normalize the correlated noise levels. In order to introduce different levels of correlation, two-dimensional moving average filter blocks of sizes 9 x 9 and 43 x 43 were used.

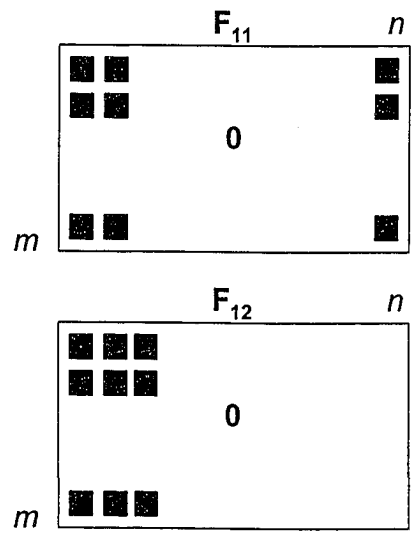

Figure 5.4. Pictorial representation of the elements $\mathbf{F}_{11}$ and $\mathbf{F}_{12}$ of a 3 x 3 wrap-around moving average filter.

Each simulated data set consists of ten replicates calculated using the same noiseless data and different realizations of the error structure. The replicates will have the double aim of allowing calculation of the error covariance matrices and the bias and variance components of the estimates.

5.4.2 Experimental data

The two experimental data sets employed in this work were obtained from previously published studies from the laboratory of Prof. Age Smilde [9]. Data Set 1 consisted of short-wavelength near-infrared (SW-NIR) spectra sequentially recorded in time during the epoxidation of 2,5-di-*tert*-butyl-1,4-benzoquinone using *tert*-butyl hydroperoxide and Triton B as a catalyst. The proposed reaction mechanism consisted of two steps:

$$A + B \xrightarrow{k_1} C + D$$

$$C + B \xrightarrow{k_2} E + D$$

Species A, B, C, D and E are specified in Table 5.1. No distinction is made between the *cis* and *trans* isomers (E) of the second step of the reaction since they are spectroscopically indistinguishable in SW-NIR. Species B is present in large excess, therefore the first and second steps of the reaction both become pseudo-first-order instead of second-order. Hence, equations 5.6 to 5.8 can be used to describe the concentration profiles of the reactant (A), intermediate (C) and main product (E) of the reaction respectively. Species A, C and E were monitored spectroscopically.

Data Set 2 is a collection of UV-VIS spectra recorded in a consecutive fashion during the reaction of 3-chlorophenylhydrazonopropane dinitrile (U) with 2-mercaptoethanol (V). The proposed reaction mechanism consisted of two steps:

$$U + V \xrightarrow{k_1} W$$

$$W \xrightarrow{k_2} Y + Z$$

Species U, V, W, Y and Z are also specified in Table 5.1. Similar to the first case, V is present in large excess, so the first step of the reaction becomes pseudo-first-order instead of second-order. Hence, equations 5.6 to 5.8 can be used in order to describe the concentration profiles of U, W and Y respectively. In this work, species U (reactant), W (intermediate) and Y (product) were monitored spectroscopically.

Table 5.1. Species involved in Data Set 1 and Data Set 2

Data Set 1	Data Set 2
A: 2,5-di-*tert*-butyl-1,4-benzoquinone	**U:** 3-chlorophenylhydrazonopropane dinitrile
B: *tert*-butyl hydroperoxide	**V:** 2-mercaptoethanol
C: 2,5-di-*tert*-butyl mono-epoxide-1,4-benzoquinone	**W:** Intermediate adduct
D: *tert*-butyl alcohol	**Y:** 3-chlorophenylhydrazonocyanoacetamide
E: *cis* and *trans* 2,5-di-*tert*-butyl di-epoxide-1,4-benzoquinone	**Z:** ethylene sulphide

Both data sets used the same experimental set-up. A Hewlett Packard 8453 UV-Vis spectrophotometer with diode array detection was used to measure spectra of the reacting system. For Data Set 1, a quartz cuvette with 10 cm path length was used to obtain spectra of the reaction mixture. For Data Set 2, a quartz cuvette with 1 cm path length was used. The experimental conditions for the two data sets are summarized in Table 5.2.

Table 5.2. Experimental conditions used for Data Set 1 and Data Set 2

Experimental conditions	Data Set 1	Data Set 2
Reaction temperature (°C)	17	25
Integration time (s)	1	1
Sampling time (s)	5	10
Total run time (s)	1200	2700
Wavelength range (nm)	860-880	300–500
Wavelength interval (nm)	1	1
Number of recorded spectra	241	271

To stress the dynamic spectral features for Data Set 1, second-derivative spectra were estimated using a third-order Savitzky–Golay filter [25] with a window size of 15

data points. Second-derivative spectra were calculated after subtracting the fourth recorded spectrum from all the other spectra remaining. The first three recorded spectra were not used for data processing because of the unsatisfactory reproducibility of these spectra as described in an earlier paper [26]. The small wavelength range 860–880 nm was used for data processing, since spectral features caused by the three species which were monitored (species A, C and E) are located in this region. Data Set 2 was obtained by subtracting a spectrum of KH_2PO_4 buffer solution used as blank. The wavelength range 300–500 nm was used for data processing since there are only spectral features caused by species U, W and Y in this spectral range. Figure 5.5 shows spectra for the first replicate (*i.e.* batch) in each of these data sets. Spectra in both cases are shown after blank subtraction and, in the case of Data Set 1, after second-derivative filtering. More details about the reactions, experimental set-up and preprocessing steps can be found in the original papers [9, 16].

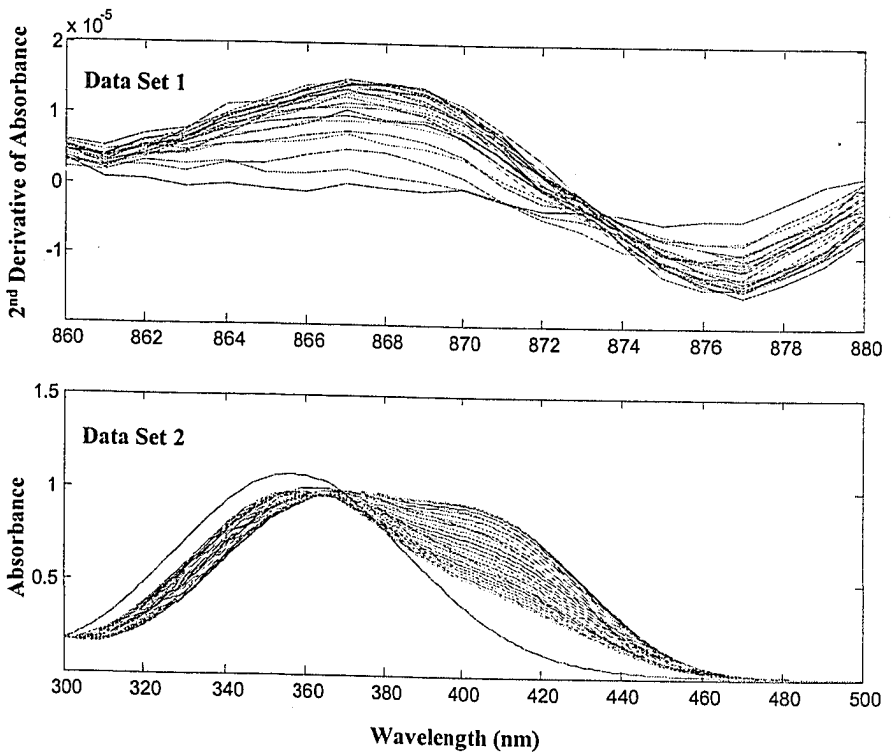

Figure 5.5. Representative spectra of the first replicate for Data Set 1 (top panel) and Data Set 2 (bottom panel). Only every tenth spectrum is shown for clarity.

5.4.3 Computational Aspects

All the calculations performed in this work were carried out on a Sun Ultra 60 workstation with 2 x 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 TLD, PARAFAC and TUCKER3 functions that were run using the N-Way Toolbox [27].

5.5 Results and Discussion

Both simulated and experimental data were used to investigate the performance of different estimation methods for DECRA transformed data under a variety of conditions. The metric chosen for this comparison was the mean squared error (MSE), as shown in equation 5.22, for the case of simulated data:

$$\begin{aligned} MSE &= \frac{1}{R}\sum_{r=1}^{R}(\theta_r - \theta^\circ)^2 \\ &= \frac{1}{R}\sum_{r=1}^{R}(\theta_r - \bar{\theta})^2 + (\bar{\theta} - \theta^\circ)^2 \end{aligned} \tag{5.22}$$

In this equation, R is the number of replicate experiments and θ represents the parameter estimated (k_1 and k_2 in this work). For simulated data, the true value of the parameter (θ°) is known, so the MSE can be further broken down into the two terms representing estimates of the precision (variance) and accuracy (bias), respectively. For experimental data, the true parameter value is unknown, so MSE is estimated from the mean parameter:

$$MSE = \frac{1}{R-1}\sum_{r=1}^{R}(\theta_r - \bar{\theta})^2 \tag{5.23}$$

5.5.1 Simulated data

Different contributors can affect the quality of the rate constants estimated through the DECRA approach. These contributors can be grouped into two different but related classes. The first class consists of attributes related to intrinsic properties of the original data set, such as the dimensions of the data, the level and structure of the noise, the spectral characteristics, and the values of the decay constant imbedded in the data. The second class consists of features related to the application of the mathematical

methods, such as the choice of S and K, the decomposition algorithm employed and (in the case of MLPARAFAC) the quality of the error covariance estimates. These two classes of parameters are related to one another. For instance, aspects such as the level and structure of the noise will have an impact on the quality of results obtained by different decomposition algorithms. The structure of the measurement errors in the three-way array, as well as its size, will depend on the dimensions of the original matrix and on the choice of the shift parameter, S, and the number of slices, K. Other parameters, such as number of replicates used to estimate the error covariance matrix and the use of compression can also play an important role in the quality of the estimates. Therefore, in order to analyze all of these factors, the performance of TLD, PARAFAC and MLPARAFAC (as representative of eigenvalue, least squares and maximum likelihood formulations, respectively) will be evaluated for scenarios with different levels of noise and correlation that have been modified with different combinations of S and K.

Two types of plots are used to analyze the influence of these parameters in the estimation of k_1 and k_2. The first type of plot, exemplified by Figures 5.6 and 5.7, is a relative plot to more readily compare the quality of the estimates for PARAFAC and TLD with respect to MLPARAFAC. The logarithm of the ratio of the root mean square error (RMSE) for the method in question and MLPARAFAC is shown for different scenarios with different values of S and K. The logarithm is used for scaling convenience since it can accommodate wide differences in magnitude and is symmetric in the visualization of relative performance. The second type of plot, exemplified by Figures 5.8 and 5.9, shows the MSE divided into its contributors, variance and bias, for different values of K and S to assess the influence of different parameters in the absolute value of the errors.

The first important conclusion that can be drawn from the results of the simulations shown in Figures 5.6-5.9 is the obviously poor performance of conventional PARAFAC for parameter estimation. This is clearly shown in Figure 5.6, where the *RMSE* for PARAFAC can be as much as three orders of magnitude worse than that for MLPARAFAC. In some cases, PARAFAC failed to converge to real-number solutions, hence some of the lines in Figure 5.6 are discontinuous. The poor performance of

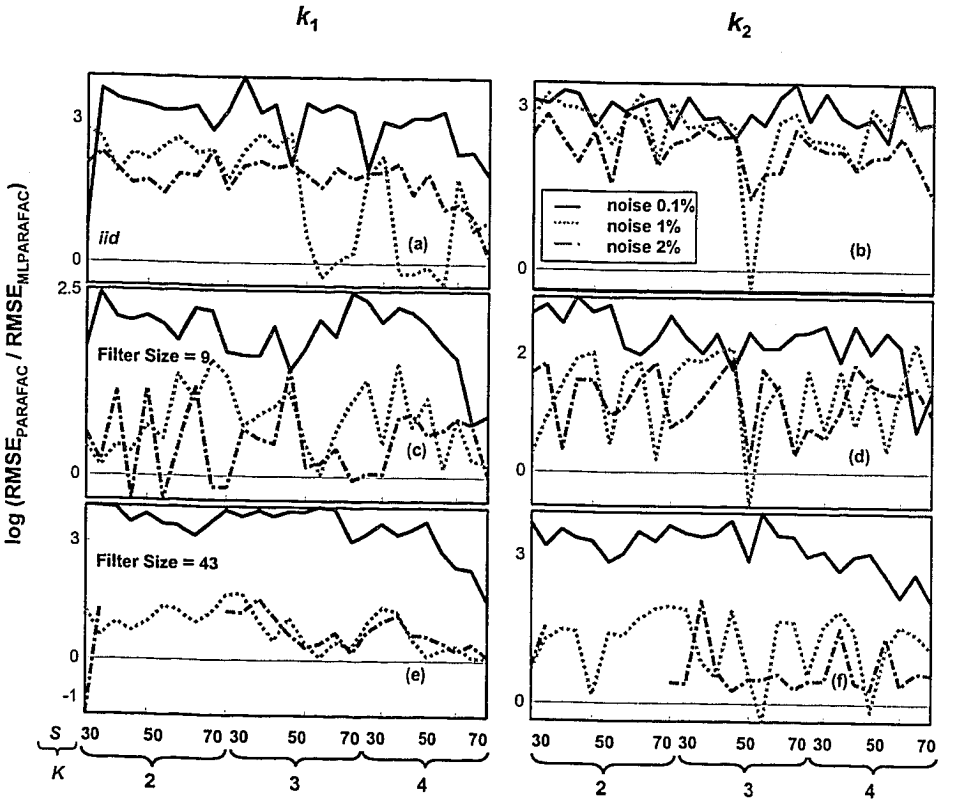

Figure 5.6. Comparative plots illustrating the relative performance of PARAFAC with respect to MLPARAFAC in the form of the logarithmic ratio of the root-mean-squared error of both methods. The different plots show the results for different levels of noise and error structures in the estimation of k_1 (left column) and k_2 (right column).

PARAFAC was found to be a general trend that was encountered in all cases (*e.g.* k_1, k_2, different combinations of K, S, error structures and noise levels). This indicates that measurement error correlation introduced by DECRA (even in the case of *iid* noise in the original data) will have an enormous negative impact on the estimates. This is consistent with the fact that conventional PARAFAC is rarely used for DECRA, and more constrained algorithms (*e.g.* LM-PARAFAC) are typically employed. Also, for the PARAFAC estimates, Figures 5.8 and 5.9 suggest that variance, as opposed to bias, was the major contributor to MSE in most scenarios. Because of the generally poor results obtained by PARAFAC, this decomposition approach will not be treated in detail in the remainder of this discussion.

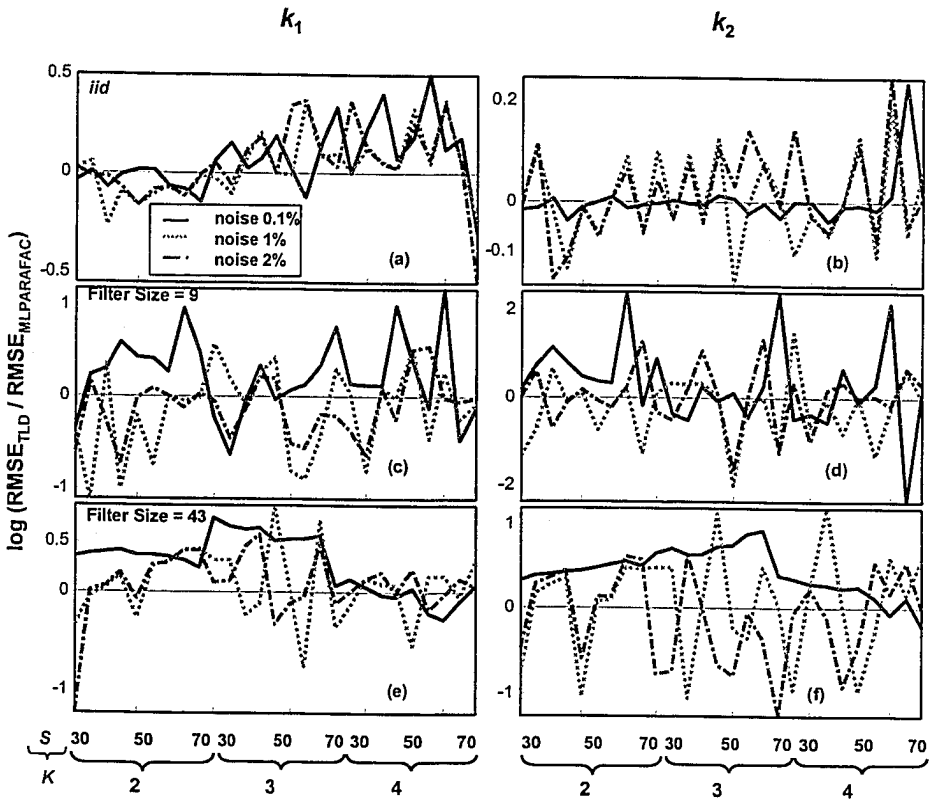

Figure 5.7. Comparative plots illustrating the relative performance of TLD with respect to MLPARAFAC in the form of the logarithmic ratio of the root-mean-squared error of both methods. The different plots show the results for different levels of noise and error structures in the estimation of k_1 (left column) and k_2 (right column).

A second important observation that can be made from the results shown in Figure 5.7 is that MLPARAFAC and TLD provided similar results for different combinations of K and S. Even though this robustness in the results was expected for MLPARAFAC due to its use of error information, it was completely unexpected for TLD. It has been extensively reported in the literature [28] that GRAM and TLD are suitable for cases where the signal-to-noise ratio is very high, since they give an exact solution for the noise free data, but they lack a clear objective function to be minimized when noise is present. Although Figure 5.7 shows some stochastic variations in relative performance, as well as some systematic differences that are elaborated upon below, the differences between MLPARAFAC and TLD are not dramatic.

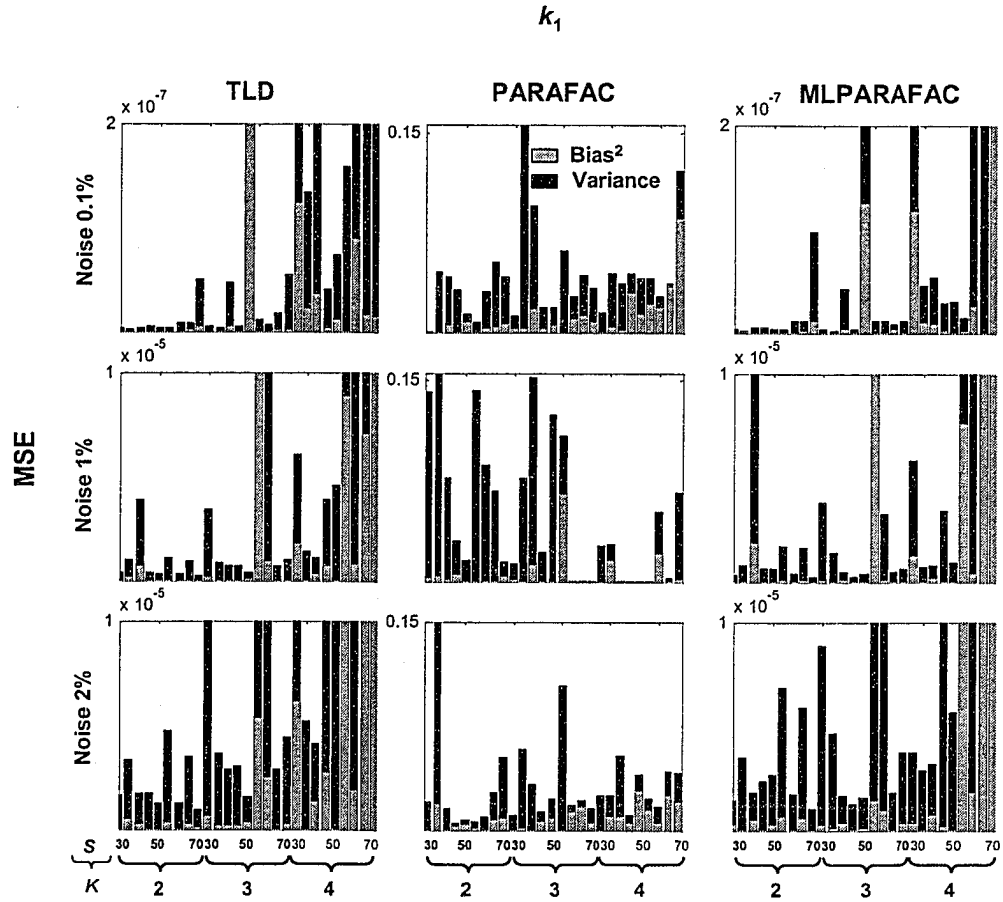

Figure 5.8. Comparative plots illustrating the absolute performance of TLD (left column), PARAFAC (center column) and MLPARAFAC (right column) in the form of the mean-squared error dissected into its two components, bias^2 and variance. The different plots show the results in the estimation of k_1 for different levels of noise and an *iid* error structure affecting the two-way data.

In Figures 5.7a and 5.7b, which represent cases where the noise in the original data is *iid*, no important differences in performance can be observed for different levels of noise. This indicates the introduction of non *iid*-noise (correlation) due to the inherent shifting in DECRA does not seem to adversely affect TLD. A probable explanation is that the amount of noise introduced was not high enough to reach a point in which TLD becomes significantly suboptimal. Since the structure of the error covariance matrix for these simulated cases is exactly known (because of the assumed *iid* structure of the original data), these results indicate that the similarity in performance of TLD and MLPARAFAC is a merit of TLD, and not connected with a poor estimation of the error covariance matrix used with MLPARAFAC due to the limited number of replicates used.

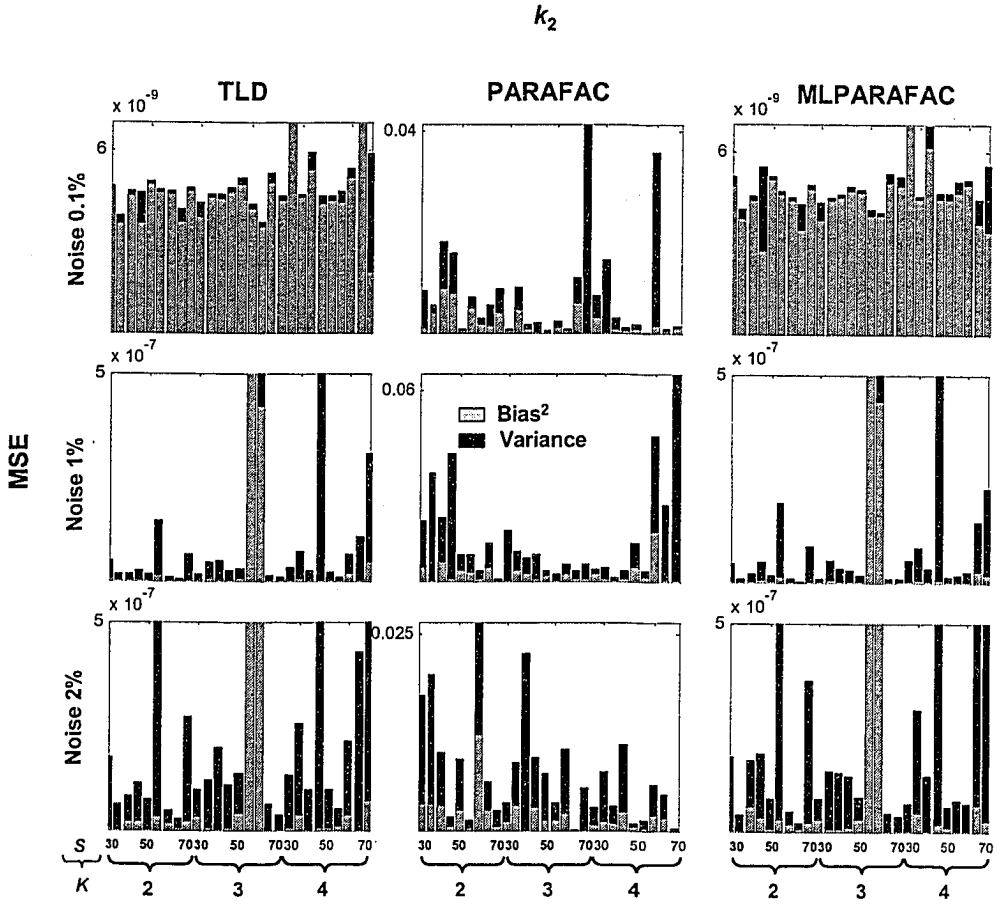

Figure 5.9. Comparative plots illustrating the absolute performance of TLD (left column), PARAFAC (center column) and MLPARAFAC (right column) in the form of the mean-squared error dissected into its two components, bias^2 and variance. The different plots show the results in the estimation of k_2 for different levels of noise and an *iid* error structure affecting the two-way data.

In contrast to Figures 5.7a and 5.7b, Figures 5.7e and 5.7f show the effect of extensively correlated noise (filter size = 43) on the estimation for different settings of K and S. In this case, the existence of highly correlated noise affecting the bilinear data and the introduction of correlation due to the shifting affect the TLD estimates, yielding cases where MLPARAFAC performed better than TLD. These differences are observed for both rate constants at different levels of noise, although the best examples are for the scenarios with high S/N ratios. Although error covariance matrices in these cases were estimated from replicates, it seems that the amount of correlated noise is sufficiently important to influence the results and diminish the effects of a limited number of replicates and the use of a compression preprocessing step.

Figures 5.7c and 5.7d show the results for intermediate cases where correlated noise affecting the bilinear data was also present, but was not very pervasive (filter size equal to 9). MLPARAFAC outperformed TLD for the scenario with the highest S/N (noise equal to 0.1% of the maximum), but differences in the two methods are not distinguishable as the S/N decreases. These results are consistent with those observed in the preceding paragraphs and suggest that MLPARAFAC will provide better performance than TLD for cases where there is significant correlated noise in the original data and the S/N is relatively high.

While Figures 5.6 and 5.7 show the *relative* performance of the three methods considered, information about the absolute performance is lost. Figures 5.8 and 5.9 show the absolute mean-squared error, broken down into bias and variance contributions, for the estimation of k_1 and k_2, respectively, for different settings of K and S in cases where different levels of *iid* noise are corrupting the original data. As noted above, the performance of PARAFAC was comparatively very poor and, although it is included in the figures for completeness, it will not be discussed in detail. In the examination of the results for TLD and MLPARAFAC, there is a significant variation in the results due to the complex relationship between the selection S and K and features such as the size of the three-way data matrix, its error structure, and the information contained in the individual slices. Because of this, some combinations of S and K produce estimates that do not correspond to apparent trends. These cases will be easier to observe in the absolute plots, since the relative plots will diminish this behavior.

One of the first noteable features of Figures 5.8 and 5.9 are the differences in the quality of the estimates of k_1 and k_2 for different values of K and S. As the values of S and K increase, the MSE values for k_1 also increase, while the MSE values for k_2 remain relatively stable. It is likely that this behavior arises from the difference in the magnitudes of k_1 and k_2. Since k_1 is much larger (faster decay) than k_2, increasing the values for K and/or S will employ a proportionately larger amount of data at extended times, where the variation in the signal due to k_1 becomes very small in comparison to the level of noise, introducing a large amount of variance in the estimates. On the other hand, the small value of k_2 will provide a slow decay, which in turn will translate into good S/N ratios and estimates of similar quality for different combinations of K and S.

Figures 5.8 and 5.9 also show that, as expected, the quality of the estimates decreases with the level of noise introduced in the original data. In most cases, however, the quality of the estimates was good, typically below 1% error in the corresponding parameter, although there is a wide variation. The plots show that most of the *MSE* arises from variance contributions in most cases, with the exception of the estimation of k_2 in the lowest noise case, which was dominated by bias for both TLD and MLPARAFAC.

It is important to note that, although Figures 5.8 and 5.9 show only the cases where the original data was affected by *iid* noise, the same trends were found for cases where correlated noise was affecting the original data. However, these results are not shown for conciseness.

5.5.2 Experimental data

5.5.2.1 Analysis of the error covariance matrices

Figures 5.10 and 5.11 show mesh plots of the pooled error correlation matrices for each mode of Data Sets 1 and 2, respectively. In both cases a very strong pattern of correlation for the spectroscopic and the time modes is observed. The sources of this correlation are likely to include offset effects due to variability in cell positioning, long- term instrument drift, temporal correlation during scanning, temporal variation in reaction initiation, and (for Data Set 1) effects related to derivative filtering. No attempt was made to decompose these complex error covariance structures into individual contributions, but some cursory observations can be made. For Data Set 1, the correlations in the spectral mode show a systematic variation, with both positive and negative correlations in the measurement errors. Correlation in the time domain was essentially flat, with values off the main diagonal around 0.75, which suggests a significant offset contribution coupled with a contribution from *iid* noise. For Data Set 2, the error correlation in the spectral domain again shows a complex structure, similar in nature to the spectral error correlation in Data Set 1, but with no negative correlation. Some regions of small correlation are evident. In the time domain, Data Set 2 also shows a large flat region similar to the case of Data Set 1, but in this case the correlation values are smoother and approach unity, likely because the contribution of *iid* noise is smaller, as is evident in Figure 5.11. Also, there is a band

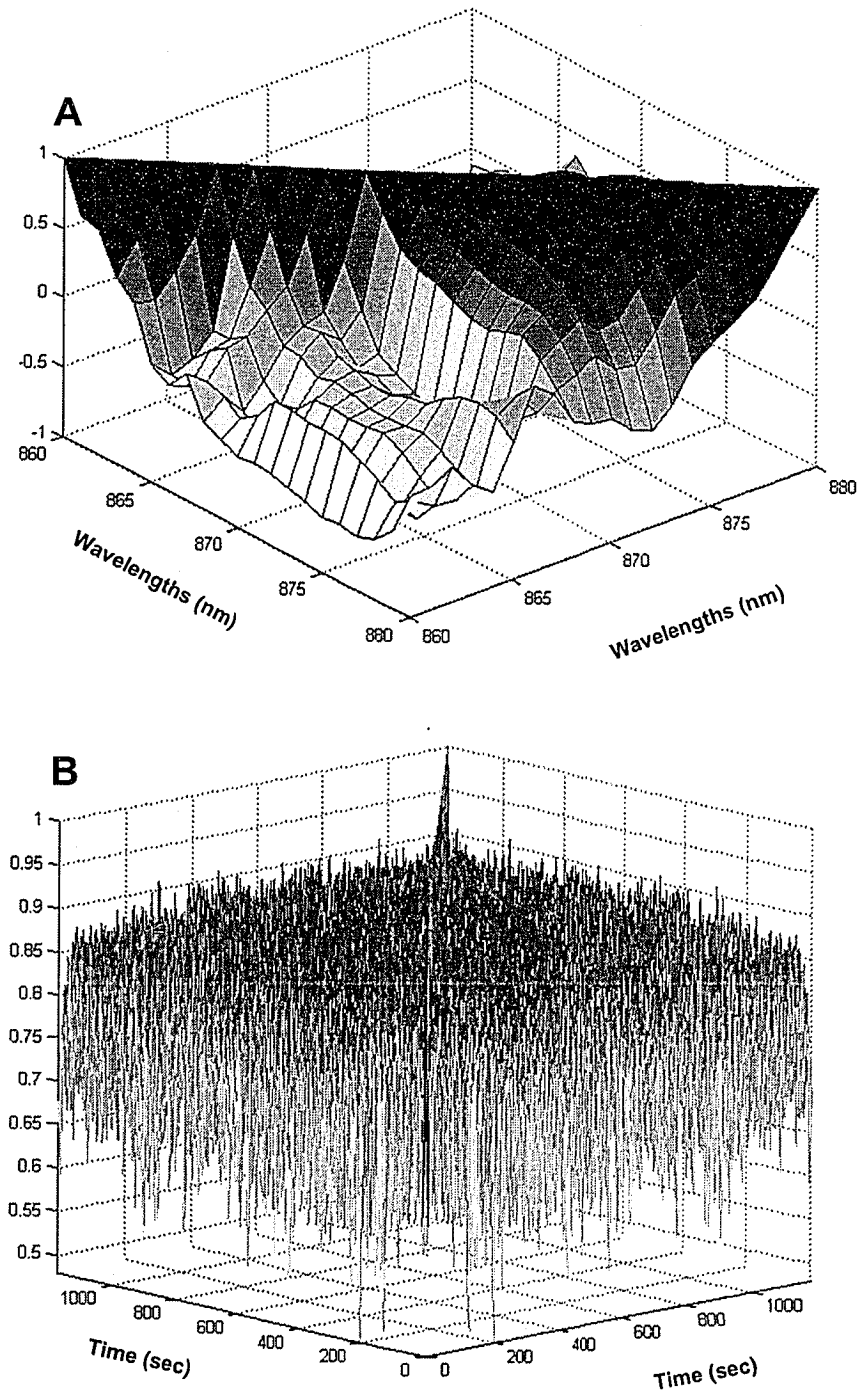

Figure 5.10. Pooled correlation matrices for each mode of Data Set 1 (SW-NIR). Plots A and B show the correlation matrix for the spectral and the time domain of the two-way data, respectively.

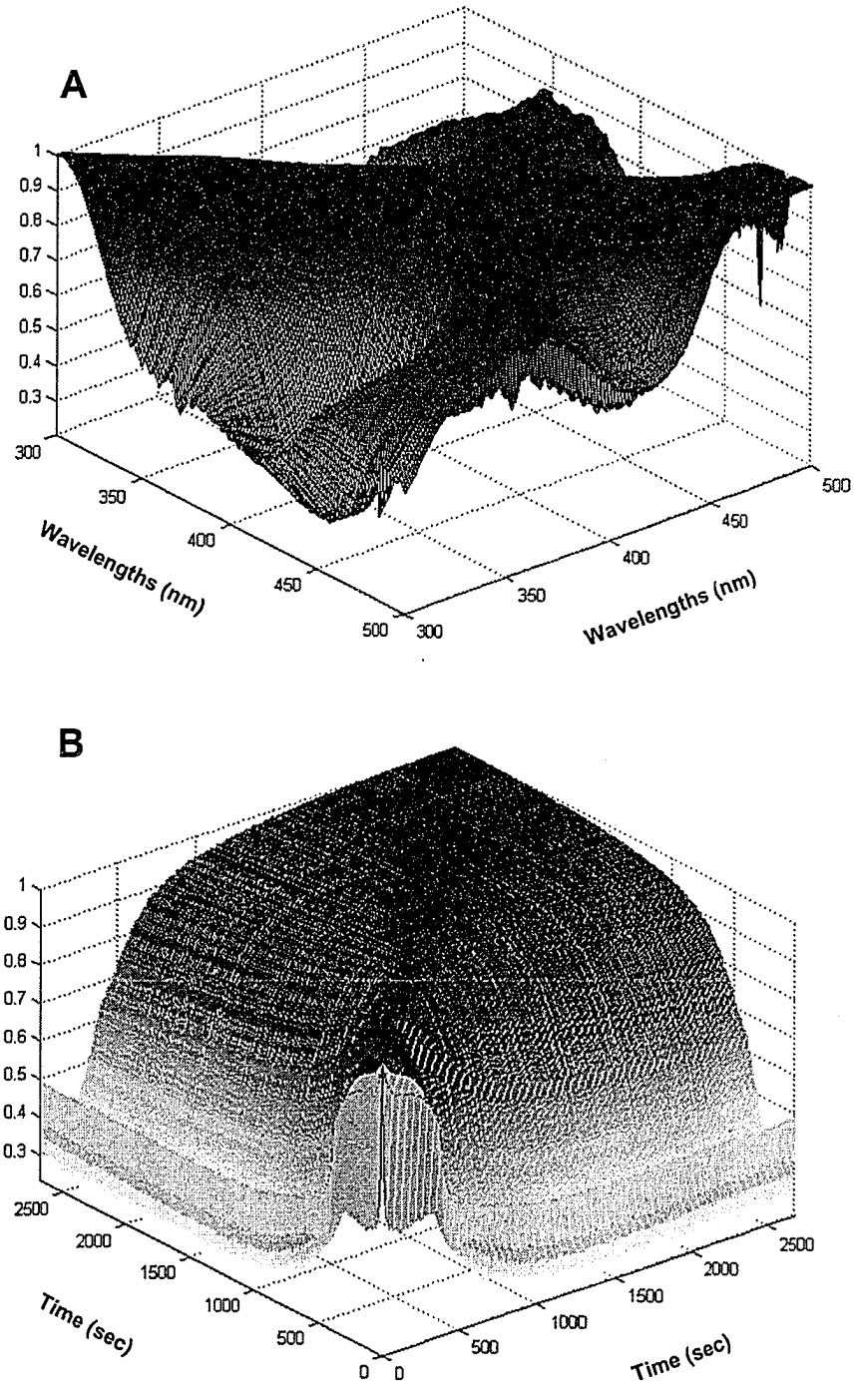

Figure 5.11. Pooled correlation matrices for each mode of Data Set 2 (UV-Vis). Plots A and B show the correlation matrix for the spectral and the time domain of the two- way data, respectively.

of low correlation near the edges of the correlation matrix. Although the source of this behavior is unknown, it may be related to variations in the start time of the reaction.

These correlation patterns indicate that the use of the general MLPARFAC algorithm will be necessary, since both modes are affected by correlated noise and the shifting strategy will introduce correlation in the third mode. The existence of correlation in more than two modes precludes the use of any of the published simplifications of the general algorithm [22] to obtain an optimal solution. However, for comparison purposes, the deterministic model of the error covariance matrix assuming *iid* errors (*i.e.* only considering the correlation related to the shifting procedure) was also employed. This simplified algorithm will be referred to as S-MLPARAFAC, while the generalized algorithm will be referred to as G-MLPARAFAC. It is important to note that, in order to use G-MLPARAFAC, a compression of the trilinear data set, as explained in reference 22, needs to be carried out. For both data sets, this was done using 21 Tucker3 factors for the spectral and the time modes. The mode related to the shifting scheme was not compressed due to its small dimension.

5.5.2.2 Data Analysis

Figures 5.12 and 5.13 depict the estimates for the rate constants and their standard deviations for Data Sets 1 and 2, respectively, when different values of K (2 to 5) and S (20 to 70 for $K = 2$ and 3, and 20 to 30 for $K = 4$ and 5) are used in conjunction with the G-MLPARAFAC, S-MLPARAFAC and TLD algorithms. It is important to note that the variations in the estimates from different replicates (represented by the standard deviations) can arise from a variety of sources which include model errors (*e.g.* kinetic model, Lambert-Beer Law), experimental errors (*e.g.* concentration errors, time errors) and instrumental noise (*e.g.* variation of the spectrometer). The combined effect of all of these sources is known as the upper error limit [15]. There are a number of approaches [15, 20] to obtain estimates that represent the lower error limit, which is mainly related to instrumental noise, but none of these were used in this work, since the upper error limit is more relevant in the context of maximum likelihood estimation, which considers all sources of error.

Estimated values for Data Set 1 for different combinations of K and S show substantial standard deviations for the individual cases, as well as among the different

combinations, suggesting that the quality of the data was poor. The variability was much lower for the estimates of k_1 than for the estimates of k_2 for all the estimation methods. In general, the estimates of the rate constants for G-MLPARAFAC are smaller than the other two methods, although they are typically within one standard deviation of those methods. In addition G-MLPARAFAC appears to provide the most precise estimates for both rate constants, but in only a few cases were these differences appreciably important.

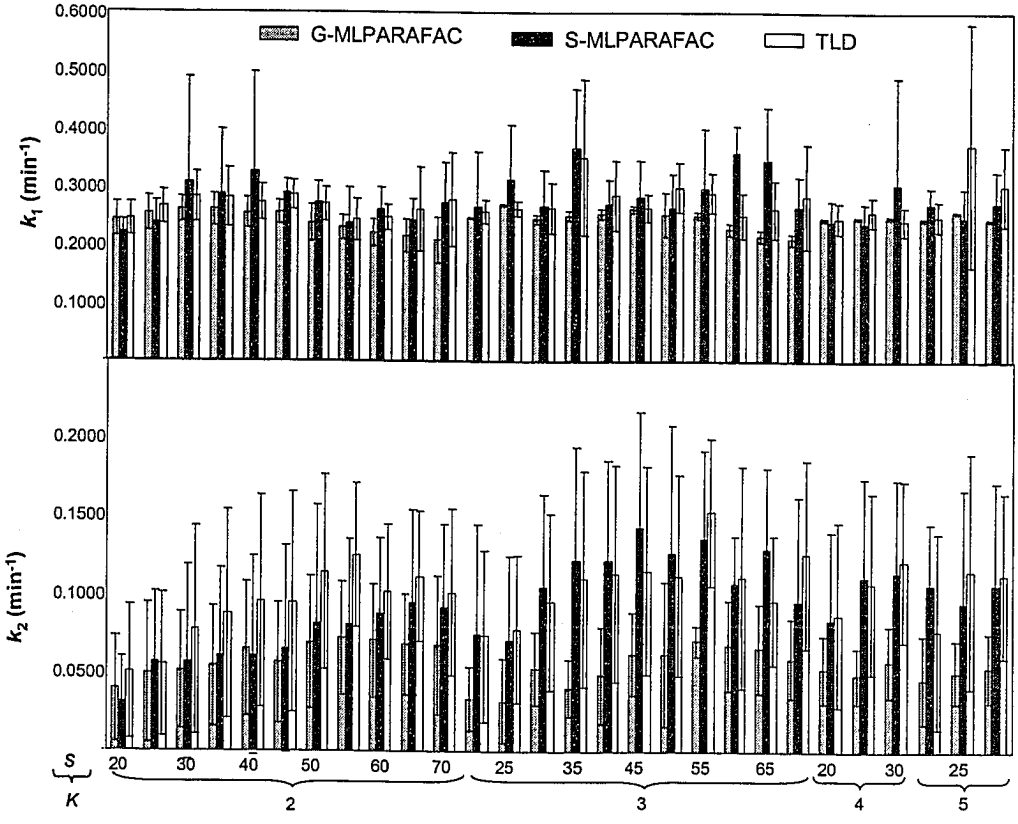

Figure 5.12. Comparison plots showing the estimates of k_1 and k_2 and their precision (as one standard deviation) obtained by G-MLPARAFAC, S-MLPARAFAC and TLD when different combinations of S and K are used for Data Set 1.

S-MLPARAFAC estimates seemed to be of similar magnitude and quality to those provided by TLD in many cases, indicating that, despite the introduction of some weighting information, the results are not completely optimal, since other sources of variation (*i.e.* correlation among samples and channels in the bilinear data) are not included. However, it should be noted that, in spite of the relatively simple error model, the performance of S-MLPARAFAC was still far superior to conventional PARAFAC (results not shown), which typically converged to very poor or imaginary solutions.

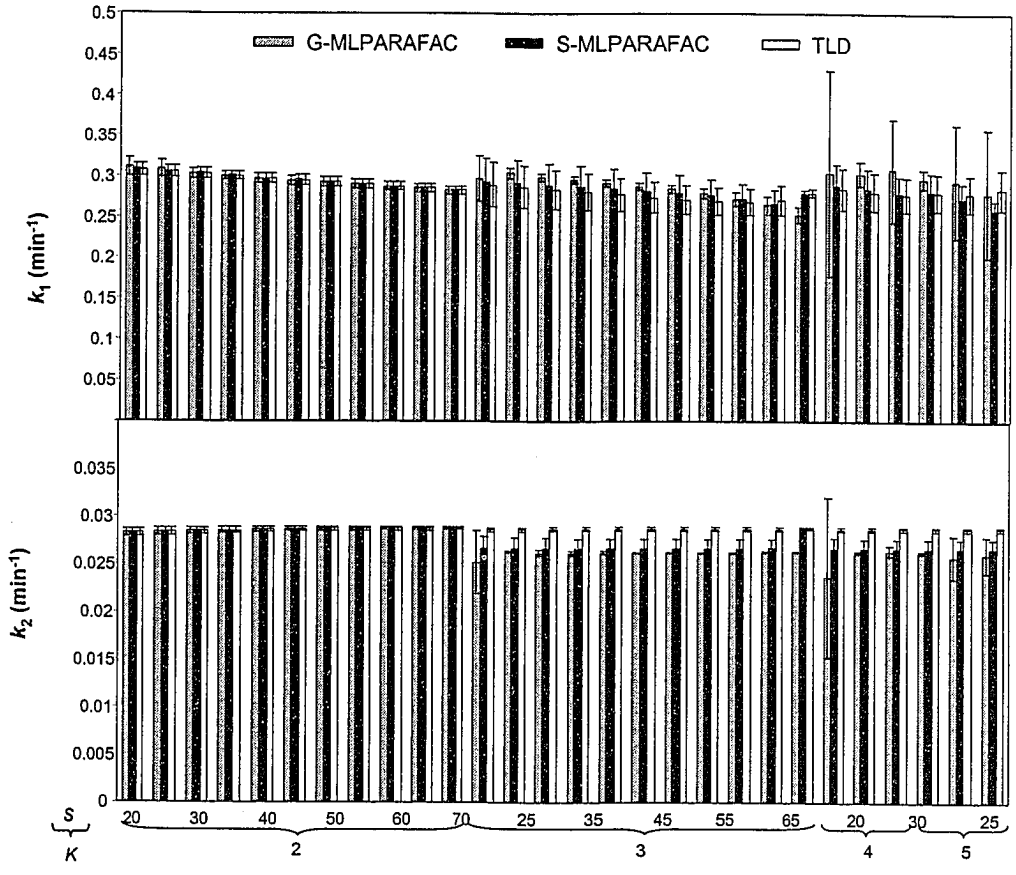

Figure 5.13. Comparison plots showing the estimates of k_1 and k_2 and their precision (as one standard deviation) obtained by G-MLPARAFAC, S-MLPARAFAC and TLD when different combinations of S and K are used for Data Set 2.

For Data Set 2, estimated values for k_1 and k_2 are not affected by a high variability indicating that the quality of the data was excellent, as has been recognized in the literature [20]. For both rate constants, the estimates were very similar for different values of S when two slices were employed. When more than two slices were used, the standard deviation of the estimates increased for all methods when k_1 was estimated and for both MLPARAFAC methods when k_2 was estimated. In general, G-MLPARAFAC again appeared to provide the most precise estimates in most cases. It is important to note, however, that TLD provided estimates of high quality for both k_1 and k_2 at a small fraction of the time needed for the more computationally involved MLPARAFAC algorithms (seconds vs. hours).

Despite the practical importance of testing new algorithms on experimental data, a drawback is that an absolute assessment based on a known "true" value cannot be made.

To put the results obtained here into an appropriate context, Figures 5.14 and 5.15 have been constructed to compare the MLPARAFAC estimates with other known approaches to kinetic rate constant estimation. Each method is shown by its 95% confidence interval ellipse (based on a bivariate normal distribution) that was constructed using data presented in the literature [9, 18]. It is worth noting that, although Faber has reported [20] that the estimation of k_1 and k_2 could be done independently (different pairs of K and S values for each rate constant) instead of finding a compromise value (same pair of K and S values providing a low variability for both rate constants), the second alternative will be used here to maintain the comparison consistent with previous studies. All of these methods are well-described in the literature [9], therefore only a brief summary is given here.

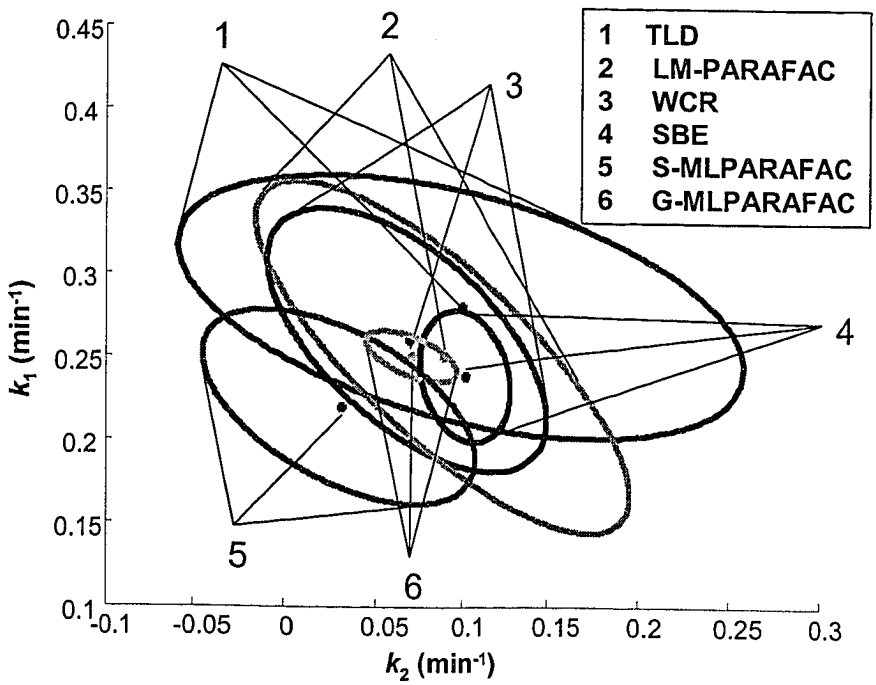

Figure 5.14. Comparison plots showing the estimates of k_1 and k_2 and their confidence interval (as 95 % probability) obtained by different methods when the best combination of S and K values are used for Data Set 1.

Classical curve resolution (CCR) is the successive two-step algorithm that utilizes the intrinsic bilinearity of the data set. Weighted curve resolution (WCR) is also an approach that exploits the intrinsic bilinearity of the data set, but in this case combining a soft model using singular value decomposition of the data set with a hard model

representing the concentration matrix. Successive Bayesian estimation (SBE) [18] solves a sequence of univariate models in which the previous results are used to construct the "prior". The Levenberg-Marquardt-PARAFAC (LM-PARAFAC) algorithm combines one alternating least squares step of the PARAFAC algorithm with a sequence of nonlinear optimization steps that use the Levenberg-Marquardt algorithm. This method, as well as TLD, simplified MLPARAFAC (S-MLPARAFAC) and generalized MLPARAFAC (G-MLPARAFAC), uses the DECRA shifting scheme in order to obtain trilinear data that will provide the rate constants via equation 5.15.

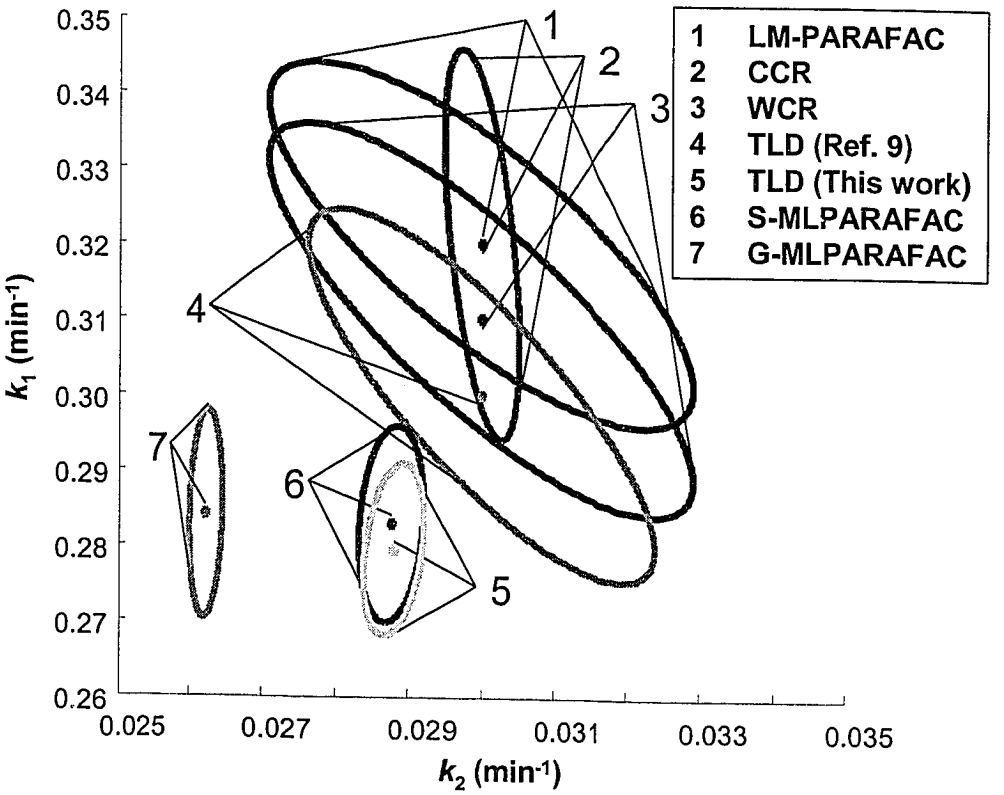

Figure 5.15. Comparison plots showing the estimates of k_1 and k_2 and their confidence interval (as 95 % probability) obtained by different methods when the best combination of S and K values are used for Data Set 2.

Figure 5.14 shows the 95% confidence boundaries for all of these methods applied to Data Set 1. The confidence interval for the G-MLPARAFAC ($K = 3$, $S = 55$, $k_1 = 0.250$, $k_2 = 0.07$, $\sigma(k_1) = 0.006$, $\sigma(k_2) = 0.01$, $r = -0.62$) method is the smallest, indicating that it yielded the most precise estimates for the rate constants. Simple inspection of the figure reveals a strong overlap in the confidence intervals for all

methods, suggesting a consistency in the results obtained. Even though S-MLPARAFAC ($K = 2$, $S = 20$, $k_1 = 0.22$, $k_2 = 0.03$, $\sigma(k_1) = 0.02$, $\sigma(k_2) = 0.03$, $r = 0.55$) gave a smaller confidence region than most of the traditional methods, the size of its region is appreciably bigger than the region for G-MLPARAFAC, probably because some important information (correlation in the spectral and time modes) was not included during the estimation process. Although the precision of the estimates does not guarantee accuracy, these results reinforce confidence in the method.

In contrast to Data Set 1, the different estimates for Data Set 2 were widely spread, as can be seen in Figure 5.15. It was somewhat disturbing to find that the values obtained in this work for the TLD estimates of Data Set 2 and its standard deviations were different from those previously published [9]. The standard deviations obtained were about a factor of two smaller than the previously reported values. This situation was previously reported by Faber [20], and the differences were attributed to the existence of different TLD implementations. The confidence interval region for all traditional methods but the TLD used in this work were much bigger than both MLPARAFAC algorithms, indicating that the introduction of a weighting scheme is necessary to obtain better estimates. The confidence regions of G-MLPARAFAC ($K = 3$, $S = 50$, $k_1 = 0.28403$, $k_2 = 0.026241$, $\sigma(k_1) = 5.1868 \times 10^{-3}$, $\sigma(k_2) = 8.8338 \times 10^{-5}$, $r = 0.14$) and S-MLPARAFAC ($K = 3$, $S = 70$, $k_1 = 0.27951$, $k_2 = 0.028765$, $\sigma(k_1) = 4.3011 \times 10^{-3}$, $\sigma(k_2) = 1.5771 \times 10^{-4}$, $r = 0.32$) were very similar in shape and size, possibly indicating that correlation due to shifting is more important than other sources of correlated errors in this case. On the other hand, the TLD results (from this work) ($K = 2$, $S = 70$, $k_1 = 0.28287$, $k_2 = 0.028767$, $\sigma(k_1) = 4.9091 \times 10^{-3}$, $\sigma(k_2) = 1.6859 \times 10^{-4}$, $r = 0.19$) exhibit a confidence region that was also very similar in shape and size to both MLPARAFAC implementations. Qualitatively, this scenario seems to bear resemblance to the simulated case represented in Figures 5.8 and 5.9, where the level of noise and the amount of correlated noise were not very high. This scenario provided estimates where the main contributor to the mean-squared error for k_2 was the bias and not the variance, with MLPARAFAC and TLD giving very similar results. This hypothesis is reinforced when these values are compared with the bias-corrected TLD estimates published by Faber [20]

(k_1 = 0.29775 min^{-1} and k_2 = 0.028837 min^{-1}), indicating the excellent quality of both MLPARAFAC and TLD estimates, even though they are uncorrected estimates.

5.6 Conclusions

A number of important conclusions can be drawn from this work. First, through the use of simulated and experimental data, it was shown that MLPARAFAC can be applied to DECRA to produce results far superior to those generated by conventional unconstrained PARAFAC. This implies that an important reason for the failure of the traditional PARAFAC algorithm is its inability to accommodate correlated errors that arise from the shifting procedure in DECRA. Unlike TLD, MLPARAFAC has a well-defined and rational objective function which makes use of measurement error information, but unlike some constrained PARAFAC algorithms, it makes no additional assumptions about the underlying model.

A second important result of this study is the similarity of the performance observed for MLPARAFAC and TLD. Although the performance of MLPARAFAC was somewhat better in a number of cases, the quality of results was comparable overall. The much greater computational efficiency of TLD is clearly an important consideration in this comparison, and these results support the widespread application of TLD to DECRA.

Simulation studies suggested that the quality of results can be significantly affected by the selection of S and K, but optimal methods for their selection were not investigated. As expected, the quality of the results diminished as the level of noise (correlated or uncorrelated) was increased.

Two sets of published and widely-studied data were also used to investigate the performance of the MLPARAFAC algorithms, which were compared to a variety of other methods. The experimental data sets exhibited strong correlations in both the time and spectral modes. The application of the generalized MLPARAFAC algorithm produced the most precise estimates for the rate constants for both data sets. A simplified MLPARAFAC algorithm, considering only correlation introduced by the shifting procedure, also produced good results.

Although the results of this work did not indicate a clear superiority of MLPARAFAC over TLD, as had been anticipated, this is nevertheless useful in that it validates the application of TLD, which is a much faster algorithm. Moreover, this study

has demonstrated that the weakness of the conventional PARAFAC algorithm lies in its inability to properly treat correlated errors, and this can be overcome through the use of MLPARAFAC.

Acknowledgements:

The authors gratefully acknowledge the financial support of the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Dow Chemical Company (Midland, MI). The authors also thank the laboratory of Age Smilde for making the experimental data used in this work available.

5.7 References

1. B. Antalek, W. Windig, J. Am. Chem. Soc. 118 (1996) 1031.
2. R.H. Vogel, SPLMOD Users Manual, Data Analysis Group, EMBL-DA09, EMBL, Heidelberg, Germany, 1988.
3. S.W. Provencher, Comput. Phys. Commun. 27 (1982) 213.
4. A. de Juan, R. Tauler, Anal. Chim. Acta 500 (2003) 195.
5. D. Schulze, P. Stilbs, J. Magn. Reson. 105 (1993) 54.
6. M. Kubista, Chemom. Intell. Lab. Syst. 7 (1990) 273.
7. E. Sanchez, B.R. Kowalski, Anal. Chem. 58 (1988) 496.
8. R.A. Harshman, Working Papers in Phonetics. 16 (1970) 1.
9. S. Bijlsma, A.K. Smilde, J. Chemom. 14 (2000) 541.
10. W. Windig, J.P. Hornak, B. Antalek, J. Magn. Reson. 132 (1998) 298.
11. W. Windig, B. Antalek, Chemom. Intell. Lab. Syst. 37 (1997) 241.
12. W. Windig, B. Antalek, L.J. Sorriero, S. Bijlsma, D.J. Louwerse (Ad), A.K. Smilde, J. Chemom. 13 (1999) 95.
13. A. Nordon, C. Meunier, R.H. Carr, P.J. Gemperline, D. Littlejohn, Anal. Chim. Acta 472 (2002) 133.

14. A. Nordon, , P.J. Gemperline, C. MacGill, D. Littlejohn, Anal. Chem. 73 (2001) 4286.

15. S. Bijlsma, D.J. Louwerse (Ad), W. Windig, A.K. Smilde, Anal. Chim. Acta 376 (1998) 339.

16. S. Bijlsma, D.J. Louwerse (Ad), A.K. Smilde J. Chemom. 13 (1999) 311.

17. W. Windig, B. Antalek, M.J. Robbins, N. Zumbulyadis, C.E. Heckler, J. Chemom. 14 (2000) 213.

18. A.L. Pomerantsev, Chemom. Intell. Lab. Syst. 66 (2003) 127.

19. H.T. Pedersen, R. Bro, S.B. Engelsen, J. Magn. Reson. 157 (2002) 141.

20. N.M. Faber, J. Chemom. 15 (2001) 169.

21. L. Vega-Montoto, P.D. Wentzell, J. Chemom. 14 (2000) 213.

22. L. Vega-Montoto, H. Gu, P.D. Wentzell, J. Chemom. **Submitted**.

23. L. Vega-Montoto, P.D. Wentzell, J. Chemom. **Submitted**.

24. P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, J. Chemom 11 (1997) 339.

25. A. Savintzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627.

26. S. Bijlsma, D.J. Louwerse, A.K. Smilde, AIChE Journal 44 (1998) 2713.

27. C.A Andersson R. Bro, Chemom. Intell. Lab. Syst. 52 (2000) 1.

28. N.M. Faber, J. Ferré, R. Boqué, Chemom. Intell. Lab. Syst. 55 (2001) 67.

Chapter 6
Conclusions

6.1 Summary

The evolution of analytical instrumentation has increased the amount and complexity of the measurements produced. This increment in complexity of the analytical data has been accompanied by a departure from the standard assumptions about the errors corrupting these measurements. A few examples of the sources producing these ubiquitous and pernicious effects are cell positioning errors, multiplicative noise effects, spatial or temporal correlation of detectors in spectroscopy, temporal correlation of pump noise in chromatography, and electronic or digital signal processing. These deviations make least squares estimation methods suboptimal from a statistical point of view. The only optimal means to account for the correlation in measurement errors is using a maximum likelihood approach to estimate model parameters that are most likely to give rise to the observed measurements. For bilinear data, this problem has been addressed in the literature through the development of maximum likelihood principal component analysis (MLPCA) and related techniques, which have been shown to provide improved results where the effects of noise correlation are significant. The objective of this work was to extend the maximum likelihood approach to treat the parallel factor analysis (PARAFAC) model by formulating a method called maximum likelihood parallel factor analysis (MLPARAFAC).

In Chapter 2, the basic principles of two algorithms for carrying out MLPARAFAC were introduced. The simplest of these algorithms was designed to work with cases where the measurement errors are non-uniform (heteroscedastic) but uncorrelated. On the other hand, analytical data corrupted by errors characterized by any type of error covariance structure can be treated by the most general form of the algorithm. Unfortunately, the vectorized formulation used by the general algorithm precludes its use to treat typical experimental data sets because of memory

considerations. Therefore, two additional algorithms, based on exact simplifications of the error structure, were presented as exploratory alternatives to extend the maximum likelihood approach in a more manageable manner to data sets of typical size. The four algorithms were shown to produce maximum likelihood estimates through a comparison of the distribution of the objective function with the χ^2 distribution. It was also shown that the quality of the estimated loading vectors for MLPARAFAC was significantly better than for the PARAFAC models in cases where the error covariance matrix is known. In this work, the standard practice of expressing the estimation process by minimizing the different formulations of the same objective function was employed. It was later realized by the author that, even though this approach was optimal from a statistical perspective, it was not as efficient as it could be from a computational/numerical point of view, since it accounted for the loss of symmetry caused by the introduction of error information in a naive manner.

Chapter 3 introduced a new approach which, in contrast to the standard practice, involved no alternation amongst objective functions to estimate the loadings for all the modes. The new approach has the benefits of locating the noise information in one or two modes as a simple representation and using it equivalently to obtain the estimation equations for each mode. Using this approach, four algorithms for carrying out simplified variations of general MLPARAFAC when the data at hand are corrupted by correlated noise affecting one or two orders were described. For completeness, a compression step was included prior to the use of general MLPARAFAC for cases where the noise structure is affecting three modes and the volume of data precludes the use of general MLPARAFAC on the raw data. To test the optimality of these algorithms, a comparison of the distribution of the objective function with the χ^2 distribution was carried out for data sets corrupted by noise realizations that follow the corresponding error structure. A number of simulated data sets were employed to illustrate that the use of simplified algorithms when the data at hand merit the simplification is beneficial from a computational point of view, and the quality of the estimates is improved when the data are treated with the algorithm designed to handle the corresponding error structure. Also, two simulated scenarios where the error structure assumed departs from the actual error structure were used to illustrate the importance of a thorough characterization of the error

covariance. Regarding the compression step used with the general formulation of MLPARAFAC, it was found that no significant differences were detected between Tucker3 and Tucker1 basis sets, at least for the data used in the simulation studies.

In Chapters 2 and 3, a number of algorithms were developed and tested using simulated data. Even though simulated data are very convenient to test the statistical properties of these algorithms, the simulations had a very well defined error structure, making the process of choosing the simplification extremely simple, since information about the number of orders affected by the correlated noise and its structure were accurately known in advance. Practical applications are not characterized by this simplicity, making the decision process a more complex task. Therefore, in Chapter 4 practical aspects related to the application of the different simplifications of MLPARAFAC to experimental data were explored using three sets of fluorescence EEM data from mixtures of three polycyclic aromatic hydrocarbons. Because of the different experimental design protocols used to acquire these data sets, 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 error structures were confirmed by the use of a set of novel tools, previously introduced for the analysis of the error structure affecting two-way data, and extended to three-way data in this work. This characterization process allowed estimation of an optimal representation of the error covariance matrix for each data set. The use of these optimal representations in conjunction with the corresponding algorithm yielded the best models in each case. In addition, different error structures and algorithms were employed, showing that the inclusion of statistically meaningful error information always produced an improvement in the estimates, by as much as a factor of three, over conventional PARAFAC, even in cases where the error covariance information was incomplete.

In Chapter 5, the testing of MLPARAFAC algorithms was taken a step further when one of the simplifications and the general algorithm were used to estimate the rate constants of reactions for two widely-studied experimental data sets and a number of simulated ones. The algorithms were used in conjunction with the Direct Exponential Curve Resolution Algorithm (DECRA), a novel method to exploit the intrinsic structure present in exponential-decay bilinear data to produce three-way data. In all cases

(simulated and experimental) when MLPARAFAC was applied to DECRA, it produced results far superior to those generated by conventional unconstrained PARAFAC. These results were a clear indication that the failure of the traditional PARAFAC algorithm was linked to its inability to accommodate correlated errors that arise from the shifting procedure in DECRA. In contrast, TLD, which is usually regarded as a suboptimal estimation approach, provided estimates of comparable overall quality to MLPARAFAC, supporting the widespread application of TLD to DECRA. The experimental data sets employed in this work were characterized by correlations in both the time and spectral modes. The quality of the estimates obtained by the MLPARAFAC algorithms was compared with a number of algorithms previously used with these data sets. For both data sets, the generalized MLPARAFAC algorithm produced the most precise estimates for the rate constants. Also, the simplified MLPARAFAC algorithm, considering only correlation introduced by the shifting procedure, produced good results.

6.2 Future Avenues of Investigation

The present work explored the introduction of measurement errors in the estimation of the PARAFAC model. As a result, a number of possible scenarios were typified and corresponding algorithms were formulated. However, in all of the cases, simulated as well as experimental, the data employed followed the structural model closely. Unfortunately, this is not always the case for experimental data sets. Depending on the severity of these deviations, mathematical constraints can be added to the estimation process. It is properly argued that the addition of constraints should be superfluous in the case of the PARAFAC model, since the presence of trilinearity assures the uniqueness of the solution. However, a number of practical motivations, such as accounting for deviations of the data from the trilinear model, avoiding degeneracy and numerical problems, achieving uniqueness in cases where ambiguous solutions are possible, and obtaining estimates that do not contradict *a priori* knowledge (*i.e.* require chromatographic profiles to be unimodal), can justify the use of constraints. The work presented here can be improved and/or extended by implementing the use of constraints in the formulated algorithms.

Constraints can be added in two different manners, constraints affecting the structural model and constraints affecting the structural factors. Constraints can be added to (or perhaps, more accurately, partially removed from) the structure of the PARAFAC model. It is well-known that the PARAFAC model is the most constrained model of the Tucker-model family. At one extreme is located the Tucker3 model, and at the other is the PARAFAC model. In between are a number of other models, such as Tucker1 [1], Tucker2 [1], PARATUCK2 [2], and PARAFAC2 [3]. However, only the PARATUCK2 and PARAFAC2 models are relevant for spectral data, since these preserve some uniqueness in the decomposition. Structurally, these models can be expressed as a more constrained Tucker3 model or as a less constrained PARAFAC model. Mathematically, the formulation of any of these models differs from the PARAFAC model in the structure of the core array, $\underline{\mathbf{G}}$, which is expressed for the PARAFAC model as a super-identity array. For instance, the PARATUCK2 model, which is well-suited for some analytical scenarios that involve interaction among factors (*i.e.* rank deficient problems) can be expressed in its more restricted form as in eqquation 6.1.

$$\mathbf{X}_a = \mathbf{A}\mathbf{H}\mathbf{I}_a(\mathbf{C} \otimes \mathbf{B})^{\mathrm{T}}) \tag{6.1}$$

Here, $\mathbf{X}_a$ is an I x JK matrix representing the trilinear data unfolded by preserving mode A, $\mathbf{I}_a$ is a P x P^2 unfolded superdiagonal matrix of order P, and $\mathbf{B}$ (J x P) and $\mathbf{C}$ (K x P) matrices represent the loadings for mode B and C, respectively. $\mathbf{A}$ (I x Q with $Q < P$) is a matrix formed by the loadings of mode A and $\mathbf{H}$ (Q x P) is the interaction matrix, formed by ones and zeros. For example a 2 x 3 $\mathbf{H}$ matrix could appear as follows:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.2}$$

When the product $\mathbf{AH}$ is evaluated, the following matrix is obtained:

$$\mathbf{AH} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix} \tag{6.3}$$

Different arrangements of $\mathbf{H}$ will reproduce the interactions in different ways, and therefore lead to different manifestations of the rank-deficiency for the model. When the model is solved using ALS in conjunction with the standard methodology of expressing the different conditional linear problems by an equivalent but differently expressed

objective function, $\mathbf{H}$ can be absorbed by the product $\mathbf{Z}_a = \mathbf{H}\mathbf{I}_a(\mathbf{C} \otimes \mathbf{B})^T)$ when $\mathbf{A}$ is estimated, and expressed as $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{H}$ when $\mathbf{B}$ and $\mathbf{C}$ are estimated. If the rank deficiency is located in mode B or C, the model is equivalently permuted to have the rank deficiency in mode A. Often, the interaction matrix is known in advance and it is introduced in the estimation process as a deterministic entity, but in many practical cases, it will be estimated to explore the interaction of the given rank-deficiency problem.

Unfortunately, if the *iid* assumption is not fulfilled by the data and a maximum likelihood approach is more suitable, a more complex problem has to be solved. As shown in this work, the introduction of the error structure destroys the symmetry of the structural model. Therefore, since the arrangement of the array is going to be primarily determined by the error structure, two additional estimation equations (one for the rank deficient mode and the other for the interaction matrix) are going to be needed for each error structure scenario. In this way, the full modes can still be estimated using the equations used for each mode in each error structure scenario. In general, this is the more recommended procedure to implement maximum likelihood algorithms for the other modes, since only a small number of equations will need to be derived.

On the other hand, the estimation algorithms can add constraints to the estimation of the loading describing each mode. Constraints such as non-negativity, symmetry, monotonicity, unimodality, smoothness, and orthogonality are a few examples. These types of constraints are easily implemented by a column-wise estimation, since the influence of each component in a mode can be determined separately, conditionally on all remaining components in that mode and the remaining modes [4]. An extreme example of these constraints is the N-PLS model [5], which can be seen as a rotated version of the PARAFAC model to better describe the variation given by an external vector or matrix ($\mathbf{y}$ or $\mathbf{Y}$). To conclude this topic, it is important to say that a considerable effort needs to be invested into understanding and implementing the use of constraints in algorithms involving compressed data, which could then be extended to the compressed version of the generalized MLPARAFAC algorithm.

In this work, a number of tools were introduced to characterize the error structure of three-way arrays, since the correct use of this information translated into better quality estimates throughout the data sets investigated. A rational extension of these techniques

can ultimately lead to a deterministic modeling of the error covariance structure for a particular experimental system, which in turn will greatly reduce the need for extensive replication and make the error structure more reliable by removing "statistical effects" within the structure.

Finally, a number of figures of merit have been introduced in the literature to assess the quality and statistical properties of the estimates [6] for three-way data. However, the established figures of merit depend highly on the assumptions made about the error structure. Therefore, an important goal of future research should be the development of a more fundamental approach to incorporate more realistic measurement error information into these metrics. Only with the refinement of such tools can the variety of multi-way methods now available be reliably evaluated and compared.

6.3 References

49. Kroonenberg P.M. and de Leeuw J. *Psychometrika* 1980; **45**: 69.

50. Harshman R.A. and Lundy M.E. *Comp. Stat. Data Anal.* 1994; **18**: 39.

51. Kiers H.A.L., ten Berge J.M.F. and Bro R. *J. Chemom.* 1999; **13**: 275.

52. Bro R. *Multi-way Analysis in the Food Industry: models, algorithms and applications* (Doctoral Thesis) 1998.

53. Bro R. *J. Chemom.* 1996; **10**: 47.

54. Booksh K.S. and Kowalski B.R. *Anal. Chem.* 1994; **66**: 782

References

Adcock R.J. *The Analyst* 1878; **5**: 53.

Alsberg B.K., Kvalheim O.M. *Chemom. Intell. Lab. Syst.* 1994; **24**: 43.

Anderson G.G., Dable B.K. and Booksh K.S. *Chemom. Intell. Lab. Syst.* 1999; **49**: 195.

Andersson C.A. and Bro R. *Chemom. Intell. Lab. Syst.* 1998; **42**: 93.

Andersson C.A. and Bro R. *Chemom. Intell. Lab. Syst.* 2000; **52:** 1.

Andrews D.T. and Wentzell P.D. *Anal. Chim. Acta* 1997; **350**: 341.

Antalek B. and Windig W. *J. Am. Chem. Soc.* 1996; **118:** 1031.

Apellof C.J. and Davidson E.R. *Anal. Chem.* 1981; **53**: 2053.

Arnold M.A., Burmeister J.J. and Small G.W. *Anal. Chem.* 1998; **70**: 1773.

Baunsgaard D., Andersson C.A., Arnadal A. and Munck L. *Food Chem.* 2000; **70**: 113.

Beltran J.L., Ferrer R. and Guiteras J. *Anal. Chim. Acta* 1998; **373**: 311.

Beltran J.L., Guiteras J. and Ferrer R. *Anal. Chem.* 1998; **70**: 49.

Beltran J.L., Guiteras J. and Ferrer R. *J. Chromatogr. A* 1998; **802**: 3.

Bezemer E. and Rutan S.C. *Anal. Chem.* 2001; **73**: 4403.

Bijlsma S. and Smilde A.K. *J. Chemom.* 2000; **14:** 541.

Bijlsma S., Louwerse D.J. (Ad) and Smilde A.K. *J. Chemom.* 1999; **13:** 311.

Bijlsma S., Louwerse D.J. (Ad), Windig W. and Smilde A.K. *Anal. Chim. Acta* 1998; **376:** 339.

Bijlsma S., Louwerse D.J. and Smilde A.K. *AIChE Journal* 1998; **44:** 2713.

Booksh K.S. and Kowalski B.R. *Anal. Chem.* 1994; **66**: 782A.

Booksh K.S., Lin Z., Wang Z. and Kowalski BR. *Anal. Chem.* 1994; **66**: 2561.

Bro R. and Kiers H.A.L. *J. Chemom.* 2003; **17**: 274.

Bro R. and Smilde A.K. *J. Chemom.* 2003; **17**: 16.

Bro R. *Chemom. Intell. Lab. Syst.* 1997; **38**: 149.

Bro R. *Chemom. Intell. Lab. Syst.* 1999; **46**: 33.

Bro R. *J. Chemom.* 1996; **10**: 47.

Bro R. Multi-way analysis in the food industry: models, algorithms and applications. PhD Thesis, Univerteit van Amsterdam. 1998.

Bro R., Sidiropoulos N.D. and Smilde A.K. *J. Chemom.* 2002; **16**: 387.

Brouwer P., Kroonenberg P.M. *J. Classification* 1991; **8**: 93.

Brown C.D. and Wentzell P.D. *J. Chemom.* 1999; **13**: 133.

Brown C.D., Vega-Montoto L. and Wentzell P.D. *Appl. Spectrosc.* 2000; **54**: 1055.

Carroll J.D. and Chang J. *Psychometrika.* 1970; **35**: 283.

Carroll J.D., Pruzansky S. and Kruskal J.B. *Psychometrika* 1980; **45**: 3.

Chen Z.P., Li Y. and Yu R.Q. *J. Chemom.* 2001; **15**: 149.

Chen Z.P., Wu H.L., Jiang J.H., Li Y. and Yu R.Q. *Chemom. Intell. Lab. Syst.* 2000; **52**: 75.

da Silva J.C. and Novais S.A. *Analyst* 1998; **23**: 2067.

de Juan A. and Tauler R. *Anal. Chim. Acta* 2003; **500:** 195.

de Juan A., Rutan S.C., Tauler R. and Massart D.L. *Chemom. Intell. Lab. Syst.* 1998; **40**: 19.

de Leew J. Block relaxation algorithms in statistics, In: Information systems and data analysis, (Eds. H.H. Bock, W. Lenski, M.M Ritcher) Spinger, New York, 1994, 308.

Durell S.R., Lee C., Ross R.T. and Gross E.L. *Arch. Biochem. Biophys.* 1990; **278:** 148.

Faber N.K.M., Bro R. and Hopke P.K. *Chemom. Intell. Lab. Syst.* 2003; **65**: 119.

Faber N.M. *J. Chemom.* 2001; **15:** 169.

Faber N.M., Ferré J. and Boqué R. *Chemom. Intell. Lab. Syst.* 2001; **55:** 67.

Gui M., Rutan S.C. and Agbodjan A. *Anal. Chem.* 1995; **67**: 3293.

Harshman R.A. and Lundy M.E. *Comp. Stat. Data Anal.* 1994; **18**: 39.

Harshman R.A. and Lundy M.E. The PARAFAC model for three-way factor analysis and multidimensional scaling, In: research methods for multimode data analysis, Praeger, New York, 1984, 122.

Harshman R.A. and Lundy M.E., Data preprocessing and the extended PARAFAC model, In: Research methods for multimode data analysis, (Eds. H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald) Praeger, New York, 1984, 216.

Harshman R.A. *UCLA Working Papers in Phonetics* 1970; **16**: 1.

Harshman R.A. *UCLA Working Papers in Phonetics* 1972; **22**: 111.

Harshman R.A. *UCLA Working Papers in Phonetics*. 1970; **16**: 1.

Henshaw J.M., Burgess L.W., Booksh K.S and Kowalski B.R. *Anal. Chem.* 1994; **66**: 3328.

Hirschfeld T. *Anal. Chem.* 1980; **52**: 297A.

Ho C., Christian G.D. and Davidson E.R. *Anal. Chem.* 1978; **50**: 1108.

Ingle J.D. and Crouch S.R. *Spectrochemical Analysis* Prentice Hall: New Jersey, 1972.

Jiang J.H., Wu H.L., Li Y. and Yu R.Q. *J. Chemom.* 1999; **13**: 557.

Jiang J.H., Wu H.L., Li Y. and Yu R.Q. *J. Chemom.* 2000; **14**: 15.

JiJi R.D. and Booksh K.S. *Anal. Chem.* 2000; **72**: 718.

JiJi R.D., Andersson G.G. and Booksh K.S. *J. Chemom.* 2000; **14**: 170.

JiJi R.D., Cooper G.A. and Booksh K.S. *Anal. Chim. Acta* 1999; **397**: 61.

Kalivas J.H. *J. Chemom.* 1999; **13**: 111.

Karukstis K.K., Krekel D.A., Weinberger D.A., Bittker R.A., Naito N.R. and Bloch S.H. *J. Phys. Chem.*1995; **99**: 449.

Kiers H.A.L. *Psychometrika*. 1997; **62**: 251.

Kiers H.A.L. *Statistica Applicata* 1992; **4**: 659.

Kiers H.A.L., and Harshman R.A. *Chemom. Intell. Lab. Syst.* 1997; **39:** 31.

Kiers H.A.L., and Krijnen W. P. *Psychometrika*. 1991; **56**: 147.

Kiers H.A.L., ten Berge J.M.F. and Bro R. *J. Chemom.* 1999; **13**: 275.

Kiers H.A.L., ten Berge J.M.F. and Rocci R. *Psychometrika* 1997; **62**: 349.

Kroonenberg P.M. and de Leeuw J. *Psychometrika* 1980; **45**: 69.

Kruskal J.B. *Linear algebra and its applications* 1977; **18**: 95.

Kruskal J.B. Rank decomposition and uniqueness for 3-way and N-way arrays, In: Multiway data analysis. Elsevier: North-Holland,1989, 7, (Eds. R. Coppi, S. Bolasco).

Kubista M. *Chemom. Intell. Lab. Syst.* 1990; **7:** 273.

Laeven J.M. and Smith C. *J. Chromatogr.* 1985; **176:** 77.

Lawton W.H. and Sylvestre E.A. *Technometrics* 1971; **13**: 617.

Lee J.K., Ross R.T. and Leurgans T.S. *J. Phys. Chem.* 1992; **96**: 9158.

Leger M., Vega-Montoto L. and Wentzell P.D. *Chemom. Intell. Lab. Syst.* 2004; **In press.**

Leger M.N. and Wentzell P.D. *Appl. Spectrosc.* 2004; **58**: 855.

Leurgans S. and Ross R.T. *Statist. Sci.* 1992; **7**: 289.

Leurgans S., Ross R.T. and Abel R.B. *J. Matrix Anal. Appl.* 1993; **14**: 1064.

Li Y., Jiang J.H., Wu H.L., Chen Z.P. and Yu R.Q. *Chemom. Intell. Lab. Syst.* 2000; **52**: 33.

Lin Z., Booksh K.S., Burgess L.W. and Kowalski B.R. *Anal. Chem.* 1994; **66**: 2552.

Lorber A. *Anal. Chem.* 1985; **57**: 2395.

Lorber A. *Anal. Chim. Acta* 1984; **164:** 293.

Magnus J.R. and Neudecker H. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley: Chichester, 1988.

Mardia K., Kent J.T. and Bibby J.M., Multivariate Analysis Academic Press: 1979.

Martens H. and Næs T. *Multivariate Calibration*; John Wiley & Sons; New York, 1989.

Martens H., Hoy M., Wise B.M., Bro R. and Brockhoff P.B. *J. Chemom.* 2003; **17:** 153.

Martins J.A., Sena M.M., Poppi R.J. and Pessine F.B.T. *Appl. Spectrosc.* 1999; **53**: 510.

Millican DW and McGown LB. *Anal. Chem.* 1990; **62**: 2242.

Moberg L., Robertsson G. and Karlberg B. *Talanta* 2001; **54**: 161.

Nordon A, Meunier C., Carr R.H., Gemperline P.J. and Littlejohn D. *Anal. Chim. Acta* 2002; **472:** 133.

Nordon A., Gemperline P.J., MacGill C. and Littlejohn D. *Anal. Chem.* 2001; **73:** 4286.

Paatero P. and Tapper U. *Chemom. Intell. Lab. Syst.* 1993; **18**: 183.

Paatero P. and Tapper U. *Environmetrics* 1994; **5**: 111.

Paatero P. *Chemom. Intell. Lab. Syst.* 1997; **38**: 223.

Pedersen D.K., Munck L. and Engelsen S.B. *J. Chemom.* 2002; **16**: 451.

Pedersen H.T., Bro R. and Engelsen S.B. *J. Magn. Reson.* 2002; **157:** 141.

Phillips G.R. and Georghiou S. *Biophys. J.* 1993; **65**: 918.

Pomerantsev A.L. *Chemom. Intell. Lab. Syst.* 2003; **66:** 127.

Provencher S.W. *Comput. Phys. Commun.* 1982; **27:** 213.

Rao C.R. and Mitra S. *Generalized Inverse of Matrices and its Applications*. Wiley: New York ,1971.

Ross R.T. and Leurgans S. *Methods Enzymol.* 1995; **246**: 679.

Ross R.T., Lee C.H., Davis C.M., Ezzeddine B.M., Fayyad E.A. and Leurgans S.E. *Biochim. Biophys. Acta* 1991; **1056**: 317.

Sanchez E. and Kowalski B.R. *Anal. Chem.* 1988; **58:** 496.

Sanchez E. and Kowalski B.R. *J. Chemom.* 1990; **4**: 29.

Savitzky A. and Golay M.J.E. *Anal. Chem.* 1964; **36**: 1627.

Schreyer S., Bidinosti M. and Wentzell P.D. *Appl. Spectrosc.* 2002; **56**: 789.

Schulze D. and Stilbs P. *J. Magn. Reson.* 1993; **105:** 54.

Sheperd P. *J. Chemom.* 1987; **1**: 3.

Sidiropoulos N.D. and Bro R. *J. Chemom.* 2000; **14**: 229.

Tauler R., Marques I. and Casassas E. *J. Chemom.* 1998; **12**: 55.

Tauler R., Smilde A.K., and Kowalski B.R. *J. Chemom.* 1995; **9**: 31.

Tomiši V. and Simeon V. *J. Chemom.* 1993; **7**: 381.

Van Huffel S. and Vanderwalle J. *The total Least Squares Problem: Computational Aspect and Analysis*; SIAM; Philadelphia, 1991.

Vega-Montoto L. and Wentzell P.D. *J. Chemom.* 2003; **17**: 237.

Vega-Montoto L. and Wentzell P.D. *J. Chemom.* **Submitted**.

Vega-Montoto L., Gu H. and Wentzell P.D. *J. Chemom.* **Submitted.**

Vogel R.H. SPLMOD Users Manual, Data Analysis Group, EMBL-DA09, EMBL, Heidelberg, Germany, 1988.

Weisberg S. *Applied Linear Regression* Wiley: Chichester, 1985.

Wentzell P.D. and Lohnes M.T. *Chemom. Intell. Lab. Syst.* 1999; **45**: 65.

Wentzell P.D., Andrews D.T. and Kowalski B.R. *Anal. Chem.* 1997; **69**: 2299.

Wentzell P.D., Andrews D.T., Hamilton D.C., Faber K. and Kowalski B.R. *J. Chemom.* 1997; **11**: 339.

Wentzell P.D., Nair S.S. and Guy R.D. *Anal. Chem.* 2001; **73**: 1408.

Windig W. and Antalek B. *Chemom. Intell. Lab. Syst.* 1997; **37:** 241.

Windig W., Antalek B., Robbins M.J., Zumbulyadis N. and Heckler C.E. *J. Chemom.* 2000; **14:** 213.

Windig W., Antalek B., Sorriero L.J., Bijlsma S., Louwerse D.J. (Ad) and Smilde A.K. *J. Chemom.* 1999; **13:** 95.

Windig W., Hornak J.P. and Antalek B. *J. Magn. Reson.* 1998; **132:** 298.

Wold S. *Kem. Tidskr.* 1972; **3**: 34.

Wu H.L. and Shibukawa M.O.K. *J. Chemom.* 1998; **12**: 1.

Wu H.L., Shibukawa M. and Oguma K. *Anal. Sci.* 1997; **13**: 53.

Wu H.L., Shibukawa M. and Oguma K. *Anal. Sci.* 1997; **13**: 53.

Xie Y.L., Baeza-Baeza J.J. and Ramis-Ramos G. *Chemom. Intell. Lab. Syst.* 1996; **32**: 215.

Yates F. *Empire J. Exp. Agric.* 1933; **1**: 129