TENSORIAL CALIBRATION:
THE GENERALIZED RANK ANNIHILATION METHOD

by

EUGENIO SANCHEZ

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

1987

Approved by _____

(Chairperson of Supervisory Committee)

Program Authorized
to Offer Degree _____

Date _____

University of Washington

Abstract

TENSORIAL CALIBRATION:
THE GENERALIZED RANK ANNIHILATION METHOD

by Eugenio Sanchez

Chairperson of the Supervisory Committee: Professor Bruce R. Kowalski
Department of Chemistry

An integrated approach to analytical calibration from a tensorial perspective is introduced. Tensorial theory provides a proper language to address the calibration problems for first, second and higher order analytical instruments. A instrument that yields an n- dimensional matrix of data per sample is defined as a nth-order instrument.

First order tensorial calibration (FOTC) is used to describe multivariate calibration methods and illustrate tensorial concepts. It is shown that the most essential step in FOTC is building the calibration model. Several methods for multivariate calibration are them discussed from this perspective.

Second Order Tensorial Calibration (SOTC) is defined as calibration of instruments that produce two-dimensional (2D) data. Focus is given to bilinear instruments, for which the 2D data can be modeled as an outer product of the intrinsic patterns in each order, such as liquid chromatography with a diode-array UV detector (LC/DA-UV) where the data can be modeled as the outer product of the UV spectrum and the chromatographic profile. The Generalized Rank Annihilation Method (GRAM) is introduced as a method for calibration and resolution of the pure spectra. Only one calibration mixture and its bilinear data are necessary to determine all the analytes of

interest. Then the bilinear data of each unknown sample is processed with GRAM yielding the pure spectra and their concentrations relative to the calibration sample.

Computer simulations and experimental LC/UV data are used to illustrate GRAM. The possibilities and limitations of GRAM are discussed in relation to the different factors that affect the quality of the results, including pattern similarity, similarity of the relative concentrations, noise level and number of components.

Doctoral Dissertation

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Request for copying or reproduction of this dissertation may be referred to University Microfilms, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature _____

Date _____

TABLE OF CONTENTS

Page

List of Figures.....iv

List of Tables.....vi

Introduction.....1

Chapter I: First Order Tensorial Calibration.....7

Multivariate Calibration.....8

Vectors.....9

Bases and Components, Covariant and Contravariant.....12

Direct Calibration.....25

Indirect Calibration.....27

Least Squares.....27

Principal Components Regression.....29

Partial Least Squares.....23

Conclusion.....30

Chapter II: Second Order Tensorial Calibration.....32

Theory.....34

Characteristics of Bilinear Calibration.....40

Factors that affect the quality of the GRAM results: Theory....43

Factors that affect the quality of the GRAM results: Computer Simulations....47

Model Chosen for Projection....48

Similarity of the Intrinsic Vectors....53

Effect of the relative concentrations....55

Effect of the number of analytes present....58

Application of GRAM to LC/UV Data....61

Simulated LC/UV Data....64

Experimental LC/UV Data....73

Second order and beyond....86

Bibliography....89

Appendix A: Tensors....94

Appendix B: Propagation of Error for Rank Annihilation....100

Appendix C: GRAM Algorithm....104

Appendix D: Synchronization of LC/UV Data....106

Appendix E: Experimental Details....110

LIST OF FIGURES

Number *Page*

1.1 Two Sensor Vector.....11

1.2 Concentrations as components.....13

1.3 Finding components on a orthonormal base.....14

1.4 General base vector.....16

1.5 Covariant base.....17

1.6 Contravariant base.....18

1.7 Covariant components.....19

1.8 Contravariant components.....20

2.1 Effect of the Model on the Error.....51

2.2 Effect of the Similarity of the Vectors.....54

2.3 Effect of the Calibration Concentration.....57

2.4 Effect of the Number of Analytes.....59

2.5 Total Wavelength Chromatogram of Simulated Complex Mixture.....65

2.6 GRAM Resolved Chromatogram Profiles for Simulated Complex Mixture 68

2.7 GRAM Resolved Chromatogram, Simulation with 4% Noise.....70

2.8 Normalized Chromatographic and Spectral Uncertainty Bands for Simulation with 4% Noise 71

2.9 GRAM Resolved Chromatograms for 1st Experimental Sample.....75

2.10 GRAM Estimated Spectra for 1st Experimental Sample.....76

2.11 GRAM Resolved Chromatograms for 2nd Experimental Sample.....77

2.12 GRAM Estimated Spectra for 2nd Experimental Sample.....78

2.13 Total Wavelength Chromatogram for Complex Environmental Sample.....81

2.14 GRAM Resolved Chromatograms for Complex Environmental Sample....82

2.15 GRAM Estimated spectra for first group of analytes.....84

2.16 GRAM Estimated spectra for second group of analytes.....85

LIST OF TABLES

Number *Page*

2.1 Effect of Incorrect Dimensionality of the Model.....49

2.2 Effect of Eigenvalue Similarity.....56

2.3 Complex Mixture Simulation Results.....66

2.4 4% Noise Simulation Details.....72

2.5 GRAM Results for Real Samples.....74

2.6 GRAM Results for Complex Environmental Sample.....83

ACKNOWLEDGEMENTS

The author specially desires to acknowledge Bruce Kowalski for his guidance throughout these years, not only in the academic and research environment, but in all aspects of life.

He is also very grateful to the "Fundacion Gran Mariscal de Ayacucho" from Caracas, Venezuela, that made it possible for him to come to the University of Washington by awarding him with a two year scholarship.

Scott Ramos is specially thanked for all his support throughout the development of this work. He also provided all the experimental data.

All the members of the laboratory for chemometrics are also thanked, specially those that shared almost every day of these years, Scott, Pat and Ken. The author also wants to thank Peter and Nan for their support in simply everything; Sandy for providing a more humane environment; Avi for inspiration that resulted in this work; and Brice, Bruce, David, Debbie, Larry, Meg, Paul, Randy, Rena and Walter.

And last but not least, to Clara, whom the author married while doing this work, who was always there to provide her caring support, specially during the hardest moments. In few words, thanks for her support.

To my Parents
Ines and Eugenio

INTRODUCTION

The analytical chemist is confronted frequently with the problem of analyzing a chemical sample to identify some of its constituents (qualitative analysis) and determine their relative amounts (quantitative analysis). To solve this problem, an instrument is usually necessary to analyze the sample. The instrument produces signals (e.g., voltages or currents) which ideally are related to the amount of the constituents of interest (analytes) present in the sample. The process of building a model that predicts amounts (or concentrations) from the instrument responses is called calibration.

An example of calibration is the classical calibration curve. The instrument responses are measured with set of calibration standards of known concentration. If a linear model is appropriate, then a least-squares regression may be used to estimate the model parameters. This is a univariate problem (only one variable). Concentrations of future samples may then be estimated by using the linear model. Univariate calibration requires an instrument response to be dependent only on the analyte of interest. In order to fulfill this condition, the analyst either separates the analyte from the other constituents of the sample that interfere with the instrument response (complete resolution) or uses a highly selective instrument (complete selectivity).

Classical univariate calibration forces the chemist to make sure there are no interfering species. If inadvertently an interferent component is present, there is no way to detect the error, much less to correct it. On the other hand, univariate calibration is well understood, and it is the easiest to adopt. Even when an instrument is available that

produces multivariate information, on many occasions, most of data is discarded in order to use the more comfortable and intuitive univariate calibration.

With the development of instruments that produce many responses per sample, e.g., a UV-Vis absorption spectrometer, it is no longer necessary to achieve complete resolution or selectivity. Using multivariate calibration methods, such us multiple linear regression, the analyst could estimate concentrations of multicomponent samples if the spectra of all the constituents present in the samples is known (Direct Calibration). Another possibility is to record the spectra of multicomponent calibration samples for which the concentrations of the analytes of interest is known (Indirect Calibration) [Martens and Naes, 1984].

Multivariate calibration relates a set of signals from a multi-channel instrument to the concentrations of one or several analytes in a sample. In general, instrument responses are recorded from a set of multicomponent calibration samples from which the concentrations of the analyte(s) of interest are known. A *prediction model* is built with this information. Then the instrument responses from the unknown sample(s) are obtained. If the responses can be fitted to the model, then the concentrations of the analytes can be estimated. If the responses cannot be fitted to the model, the sample is rejected as an outlier; it probably has interferent components. Interferences can be detected using multivariate *linear* calibration, but they usually cannot be corrected [Oste and Kowalski, 1984].

In spite of the limitations of multivariate calibration, Ho and coworkers presente a calibration technique that allowed quantitation even in the presence of interferences not accounted for in the calibration set, namely, the rank annihilation (RA) method [Ho *et al*, 1978, 1980, 1982]. The implications of this discovery were remarkable: most research in analytical chemistry is designed to increase resolution, selectivity and

eliminate interferences, but rank annihilation worked without all those restrictions. They showed that this is possible only if the instrument responses can be bilinearly modeled, i.e., the experimental responses are a matrix, $\mathbf{M} = \mathbf{X}\,\mathbf{Y}^T$, where $\mathbf{X}$ and $\mathbf{Y}^T$ are matrices that represent the intrinsic factors responsible for the observations* .

The RA method requires one calibration matrix per analyte of interest, $\mathbf{N}_k$, and the unknown, or test sample data matrix, $\mathbf{M}$. The method consists of iteratively computing the matrix $\mathbf{D} = (\mathbf{M} - \beta\,\mathbf{N}_k)$ and varying β until the rank of $\mathbf{D}$ decreases in one unit. In practice, they used the value of β for which the smallest significant eigenvalue of the matrix $\mathbf{D}$ was a minimum. This method has been successfully applied to emission-excitation fluorescence [Ho *et al*, 1978, 1980, 1982], LC/UV [McCue and Malinowski, 1983], and TLC-reflectance imaging spectrophotometry [Gianelli *et al*, 1983; Burns *et al*, 1986] with good results. Appellof and Davidson extended RA to tridimensional arrays [Appellof and Davidson, 1983].

Burns and coworkers have shown, that when only information in one order is available for the analyte of interest (e.g., only its UV spectrum is available for LC/UV data), quantitation using RA is still feasible if the data in the other order can be restricted in some way, e.g., a non-negativity constrain can be applied to a chromatographic concentration profile for LC/UV data [Burns *et al*, 1986]. In ordinary RA, the matrix $\mathbf{N}_k$ can be modeled as the outer product of the intrinsic factors belonging to one analyte, $\mathbf{N}_k = \mathbf{x}_k\,\mathbf{y}_k^T$. If only one of the vectors is known, e.g., $\mathbf{x}_k$, they showed that all possible vectors $\mathbf{y}_k$ that reduce the rank of the matrix $(\mathbf{M} - \beta\,\mathbf{N}_k)$ by one unit (for certain β) are limited to a single line in the vectorial space of $\mathbf{y}_k$. Combining this fact with the non-negativity constrain gives a unique solution to $\mathbf{y}_k$, which is later used to compute β.

* A typical example of this is the data generated in liquid chromatography, using a multi-wavelength UV absorbance detector (LC/UV). The columns of the matrix $\mathbf{X}$ represent the pure spectra of each component present in the sample. The rows of $\mathbf{Y}^T$ represent the pure component elution profiles.

Lorber and Kim independently introduced a non-iterative version of the RA method, presenting the problem as a generalized eigenvalue-eigenvector equation for which a direct solution is found by using the singular value decomposition (SVD) [Lorber, 1984, 1985; Kim, 1984]. The SVD of the matrix **M**, i.e., $\mathbf{M} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$ is used to compute its pseudoinverse, $\mathbf{M}^+$, and then the trace of the matrix $\mathbf{M}^+\mathbf{N}_k$ is equal to the ratio of concentrations.

Multivariate Calibration cannot account for the results of the RA method, mainly because multivariate calibration deals with vectorial quantities, and RA deals with matrices. As multivariate calibration is a generalization of univariate, RA could probably be explained within the context of a generalization of multivariate calibration. Tensorial Calibration provides such generalization.

Using a tensorial description of rank annihilation, the generalized rank annihilation method (GRAM) has been developed. GRAM allo or simultaneous determination of several analytes from a multicomponent unknown using a single calibration sample containing a mixture of standards of all the analytes of interest (Quantitative analysis). GRAM also extracts the intrinsic spectra of the two orders for each analyte that is determined (Qualitative analysis). E.g., if the bilinear data comes from an emission-excitation fluorescence instrument, then GRAM will estimate the emission and the excitation spectrum for each individual analyte shared by the samples, and their ratio of concentrations. If the data comes from an LC/diode-array UV instrument, then GRAM estimates the UV spectrum and the chromatographic concentration profile for each analyte shared by the samples and their ratio of concentrations.

Tensors

Multivariate calibration can be described with the *vectorial language*, but rank annihilation (RA) cannot. It could be said that vectorial theory does not "span" the space that explains RA. This work shows that *tensorial* theory does. Simultaneously, by being a generalization of vectors, tensors can also describe multivariate calibration.

A vector is analogous to a *first order* tensor. A matrix of data can be considered to be the components of a *second order* tensor. A single number can be thought of as the component of a *zero* order tensor. Similarly, chemical measurements can be classified into zero, first, second and higher order, according to the data that they produce. A pH meter is a zero order instrument, a UV spectrometer is a first order instrument, and a video fluorometer [Warner *et al*, 1975; Johnson *et al*, 1977] is a second order instrument.

Tensorial theory, or the tensorial language, provides a unified approach to calibration and chemometrics. When different chemometric methods are explained from the same ground, their advantages and disadvantages, similarities and differences become more clear. It also provides a framework for what could be the beginning of a fundamental theory of Analytical Chemistry.

No extensive description of the tensorial theory will be presented here, but an appendix describing the essentials of tensorial theory is included, adapted from Budiansky's introduction to tensors [Budiansky, 1974]. The difficulty of the tensorial notation, with multiple subindexes and superindexes has been avoided, in favor of a more intuitive notation to facilitate the understanding of the equations.

Accomplishments

The most important accomplishments of this work are itemized below:

• The Generalized Rank Annihilation Method (GRAM) has been developed. GRAM takes two multicomponent bilinear matrices **M** and **N**, and extracts the underlying factors $\mathbf{x}_i$ and $\mathbf{y}_i$ that are common to both samples, and their respective concentration ratios.

• An integrated approach to multivariate, multi-order calibration is presented, that is an appropriate framework for a formal theory of measurement in chemometrics.

• The effect of the experimental noise in the results of GRAM have been studied theoretically and with computer simulations.

• The first application of GRAM to experimental data (LC/UV) is presented.

Organization

The first chapter uses multivariate calibration to introduce some of the useful tensorial concepts and at the same time provides interesting insights into the subject. The second chapter, second order tensors, shows how tensorial theory is used to explain rank annihilation (RA) and naturally develop a new method denominated the generalized rank annihilation method (GRAM), a calibration and curve resolution method for bilinear data. Examples of the application of the theory to simple and complex simulated and real experimental samples will be given. The last section of the second chapter presents a short discussion of the potentials of third and higher order instrumentation.

CHAPTER I

FIRST ORDER TENSORIAL CALIBRATION

First order tensorial calibration is known in the literature as multivariate calibration. The intention of this chapter is to present known concepts from linear multivariate calibration from a tensorial perspective. This will provide a familiarity with the subject that will make the next chapter more readable. It also provides certain insights into calibration that are worth mentioning.

There will be no discussion of classical univariate (single sensor) calibration, that would correspond to zero order calibration. The only aspect worth mentioning in this context is that zero order tensors, like vectors, need a base of reference. For example, it is not very useful to say "the concentration is 5.3". Units should be specified, or at least they must be implicit. It is more useful to say "the concentration is 5.3 *g/l*". The number 5.3 by itself does not have significance if not referred to its base, in this case *g/l*. If the base is changed, the component 5.3 must also change. For reviewing univariate calibration see e.g. [Garden *et al*, 1980; Mandel and Linnig, 1957].

Multivariate Calibration

The task of multivariate calibration is to find a predictive model that relates two vectorial spaces. Interest in multivariate calibration in analytical chemistry has grown from very few publications in the last decade to a very important topic during the last few years. The widespread availability of computers and the increasing amounts of data that the experimentalist obtains has prompted this change in interest in multivariate methods. For recent reviews see, e.g., [Carey *et al*, 1986; Ramos *et al*, 1986; Martens, H. and Naes, T., 1984; Brown, P. J., 1982].

An example of a specific problem for multivariate calibration is multicomponent analysis by infrared spectrophotometry [see, e.g., Maris *et al*, 1983]. The infrared spectrum of a group of calibration samples is recorded; then a prediction model is built that relates the spectra with the concentrations of the analytes of interest. Finally, the spectrum of an unknown sample is recorded, and the concentrations of the analytes are estimated.

In this work, calibration will refer to calibration of an analytical instrument. The instrument is calibrated to estimate some property of an object or sample. In first order calibration, an instrument that produces a vector of responses for a given sample is used. This discussion will be based on the assumptions that the objects are homogeneous multicomponent samples of matter and the property to estimate is the concentration of one or several constituents (analytes) present in the samples. Not to reduce generality, but to make the topic less abstract, it will be assumed that the instrument is an array of sensors, such us a diode-array UV (DA/UV) detector.

Vectors

A vector is an *n*-tuple of real numbers or *components*, represented as $\mathbf{r} = (r_1, r_2 \ldots r_n)$. These components are the coordinates of a point in a *n*–dimensional abstract space. They are relative to the *base* vectors, i.e., a set of independent vectors that span the vectorial space. Multivariate calibration and, in general, multivariate analysis, use vectorial quantities, and a vector space approach has already been proved to be very useful in describing these subjects [Eaton, 1983].

The numerical value of the responses of a sensor array can be arranged as the components of a vector **r**. Therefore, this sensor array response "pattern" can be considered a vector in a multidimensional vectorial space, where the base vectors of the space are the unitary responses for each sensor, and the components are the actual numbers that the sensor array has provided. Fig 1.1 provides a simple example for a two-sensor case, where the components of the vector **r** in the base $\{\mathbf{u}_1, \mathbf{u}_2\}$ are (4.0, 3.0) respectively.

The response pattern of a multicomponent sample is a function of many factors, where the concentrations of the analytes present in the sample are potentially the most important ones. A generalized calibration model would relate the instrument responses indirectly to the concentrations of analytes [see, e.g., Friedman, 1986; Friedman and Stuetzle, 1981]; it will be assumed that there is a direct linear relationship between the array responses and the concentrations of the analytes,

$$\mathbf{r} = \sum_{i=1}^{q} c_i \mathbf{x}_i + \mathbf{x}_o + \boldsymbol{\varepsilon} \qquad [1.1]$$

where $\mathbf{r}$ ($p \times 1$) is the response of an array with p sensors, to a sample with q analytes, c_i is the concentration of the analyte i, $\mathbf{x}_i$ ($p \times 1$) is the array's vector of responses of the pure analyte i; $\mathbf{x}_o$($p \times 1$) is the array response of the background, and $\boldsymbol{\varepsilon}$($p \times 1$) is the model error.

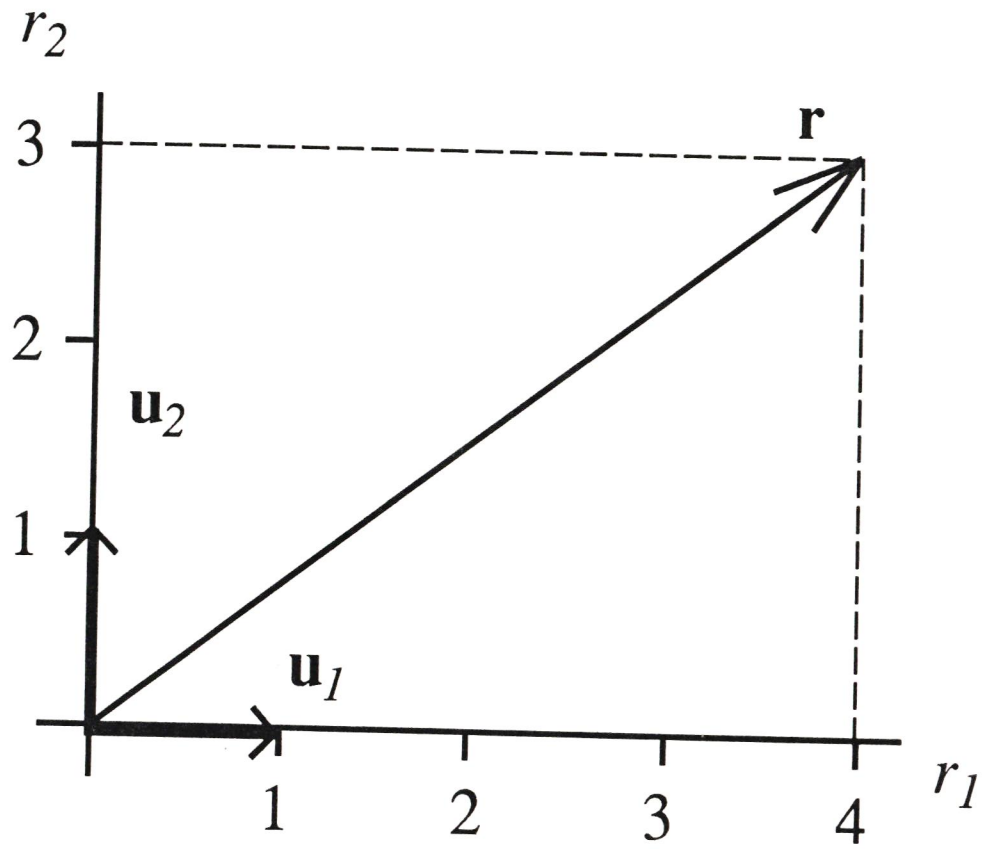

Fig 1.1 Two sensor vector. The response of a two-sensor array to a chemical sample can be represented as a vector, **r**, in a plane, where the base vectors are unitary responses from each sensor and the components are the actual sensor responses to the sample, in this case (4.0,3.0).

*Bases and Components, Covariant and Contravariant**

When there are only two analytes present in a sample, and the background and error terms are not taken into account in Eq 1.1, the array pattern **r** can be modeled as a simple linear combination of the analytes patterns $\mathbf{x}_1$ and $\mathbf{x}_2$ (Fig 1.2),

$$\mathbf{r} = c_1 \mathbf{x}_1 + c_2 \mathbf{x}_2 \qquad [1.2]$$

it is interesting to observe that if the vectors $\mathbf{x}_i$ are defined as the pure analyte patterns at unitary concentration of analyte i, then the vectorial components of **r** in the *base* $\{\mathbf{x}_1, \mathbf{x}_2\}$ are simply the *concentrations* c_1 and c_2.

Therefore, the problem of multivariate calibration involves finding the components of the response vector on an specific set of base vectors, each of them corresponding to the array pattern of responses from a unitary amount of each analyte present in the sample.

It is known that a vector is defined by its base vectors and its components. For an *orthonormal* base of vectors $\mathbf{u}_i$ (orthonormal vectors are defined as unitary length perpendicular vectors), the components c_i of a vector **r** are equally defined by the explicit *projection* formula or the implicit *composition* formula (see Appendix A)

projection $$c_i = \mathbf{r} \cdot \mathbf{u}_i \qquad [1.3]$$

composition $$\mathbf{r} = \sum c_i \mathbf{u}_i \qquad [1.4]$$

where the symbol "·" represents the dot product between two vectors. The components c_i from both Eqs 1.3 and 1.4 are the same, because the base is orthonormal. The projection formula relates concentrations to response patterns directly, but the composition formula requires finding simultaneously all the c_i such that Eq 1.4 is an identity. Naturally, it is more convenient to use the projection formula. Fig 1.3 illustrates this, again for a two-sensor case. If the patterns of the analytes present in the

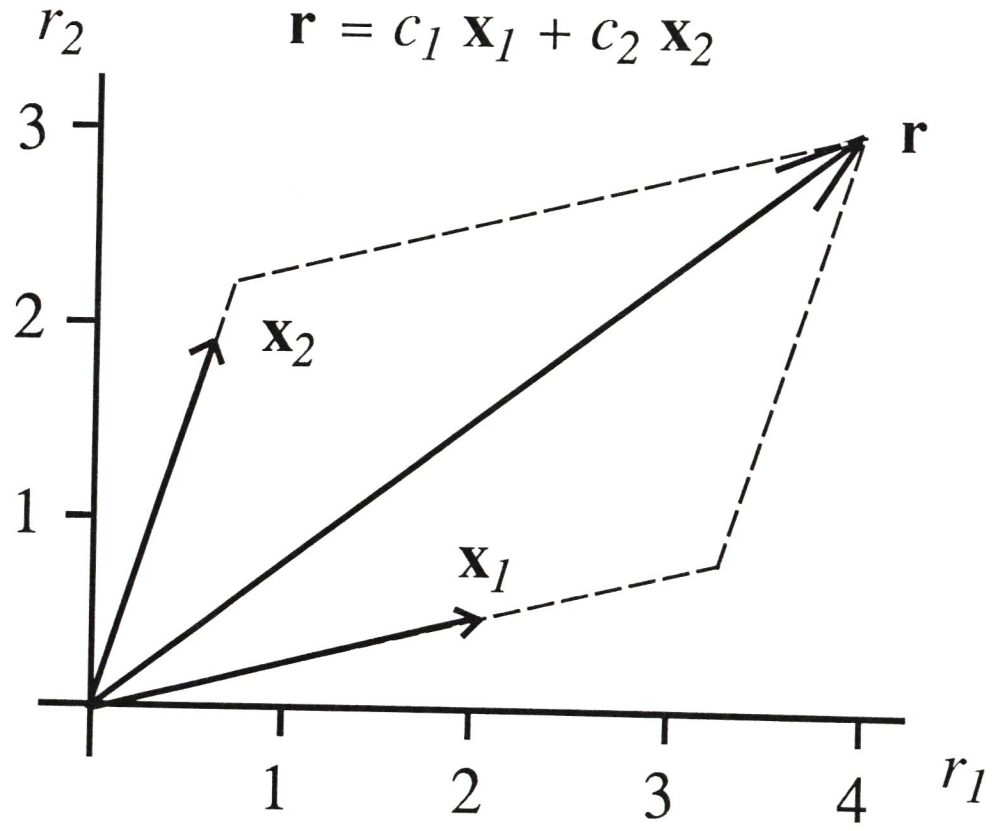

Fig 1.2 Concentrations as components. For a two component sample, if the base selected to express the response vector **r** is the pure patterns of the two analytes, $\mathbf{x}_1$ and $\mathbf{x}_2$, at unitary concentrations, then the components of **r** are simply the concentrations c_1 and c_2.

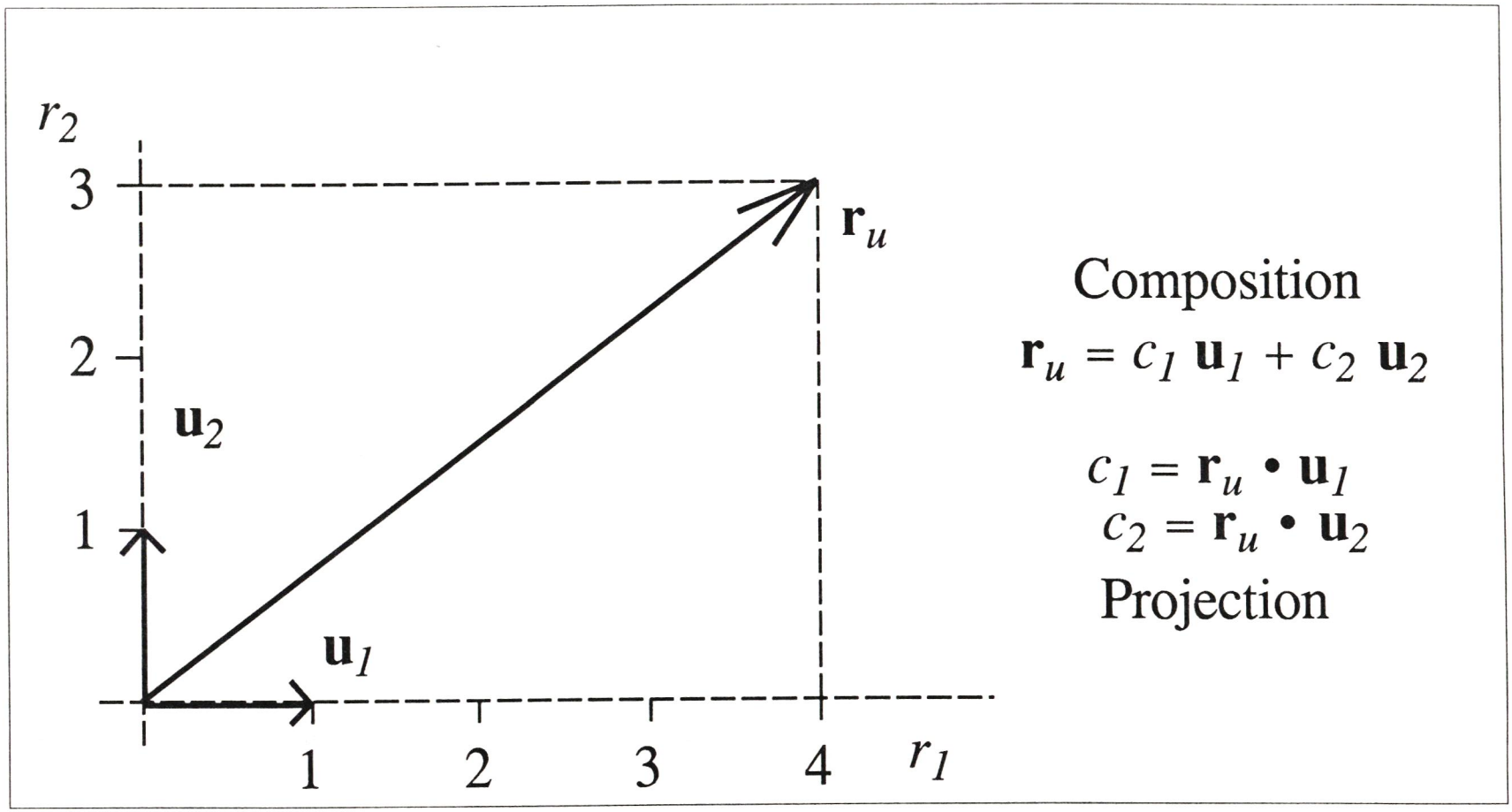

Fig 1.3 Finding components of a vector. For an orthonormal base $\{\mathbf{u}_1,\mathbf{u}_2\}$ the components of a vector $\mathbf{r}$ are equally defined by the projection and the composition formulas.

mixture were orthogonal to each other, the calibration problem could be solved directly, using the projection formula, i.e., Eq 1.3.

Unfortunately, patterns of real pure analytes are seldom orthogonal, and the concentrations have to be indirectly estimated from the composition formula. General bases have two kinds of components: the projection formula gives the *covariant components* and the composition formula gives the *contravariant components*. It is usual to denominate a general base as a *covariant* base. The contravariant components are the most valuable for the analysis, because they are equivalent to the concentrations.

Any general, non-orthogonal base also has an associated *contravariant* base. Fig 1.4 illustrates a non-orthogonal base $\{\mathbf{x}_1, \mathbf{x}_2\}$ and its contravariant form $\{\mathbf{x}_1^+, \mathbf{x}_2^+\}$, in a bivariate space. The contravariant base is useful because the roles of projection and composition are opposite to those for the covariant base. Therefore the contravariant components (concentrations) can be estimated directly using the projection formula on the contravariant base,

$$c_i = \mathbf{r} \cdot \mathbf{x}_i^+ \qquad [1.5]$$

If $\mathbf{x}_1$ is the pattern of analyte 1, then $\mathbf{x}_1^+$ represents its contravariant pattern. The contravariant pattern depends on the patterns of all other analytes $\{\mathbf{x}_{i\neq 1}\}$ present in the sample because it is perpendicular to all of them. Figs 1.5 - 1.8 should illuminate these points with a bivariate case. $\mathbf{x}_i^+$ is also dependent on $\mathbf{x}_1$ because it is constrained by its definition, $\mathbf{x}_i^+ \cdot \mathbf{x}_1 = 1$ (Eq A.12, Appendix A). These dependencies are true if the number of analytes (q), is equal or less than the number of sensors (p). If $p < q$ (fewer sensors than analytes), the patterns $\{\mathbf{x}_i\}$ do not form a base, because they cannot be linearly independent, and the contravariant vector is not defined. If $p > q$ (more sensors than analytes), the contravariant pattern is additionally restricted to lie within the subspace spanned by the base $\{\mathbf{x}_i\}$

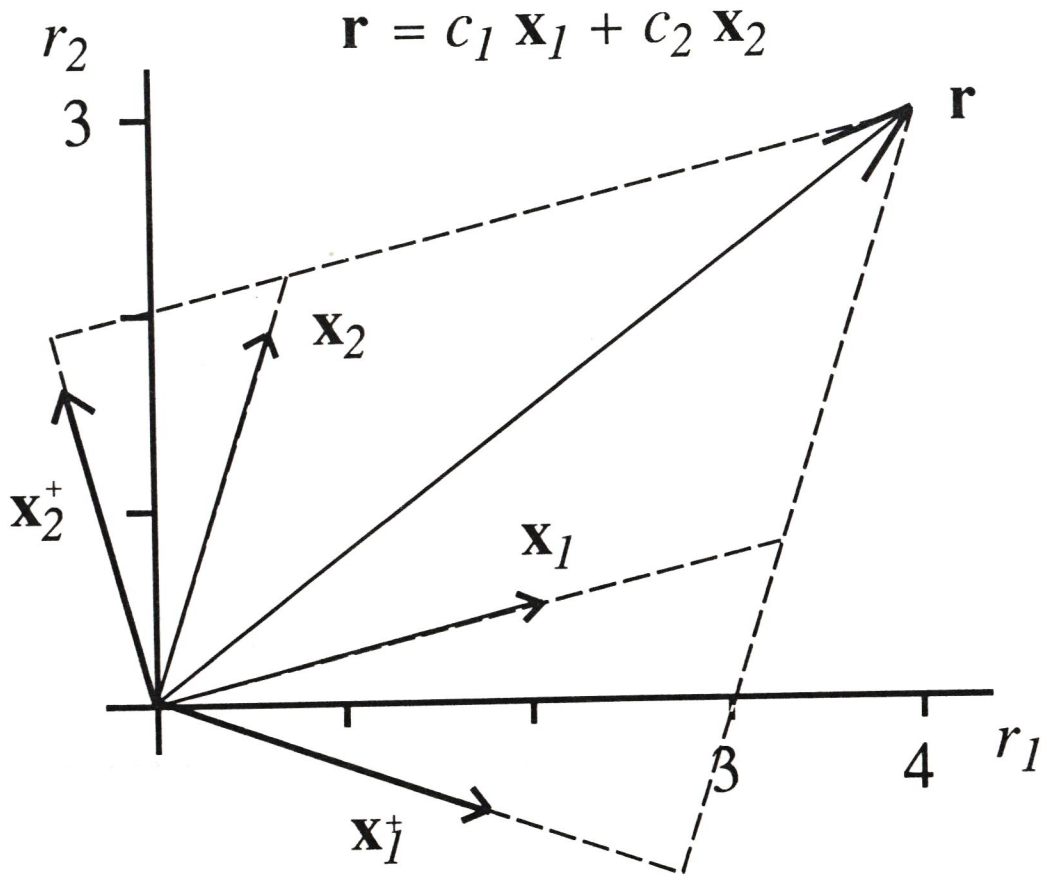

Fig 1.4 General base vector. Any general base (non-orthogonal) covariant vector $\mathbf{x}_i$ has an associate contravariant vector $\mathbf{x}_i^+$. If we know $\mathbf{x}_i^+$, we can estimate the concentration using the direct projection formula, e.g., $c_i = \mathbf{r} \cdot \mathbf{x}_i^+$.

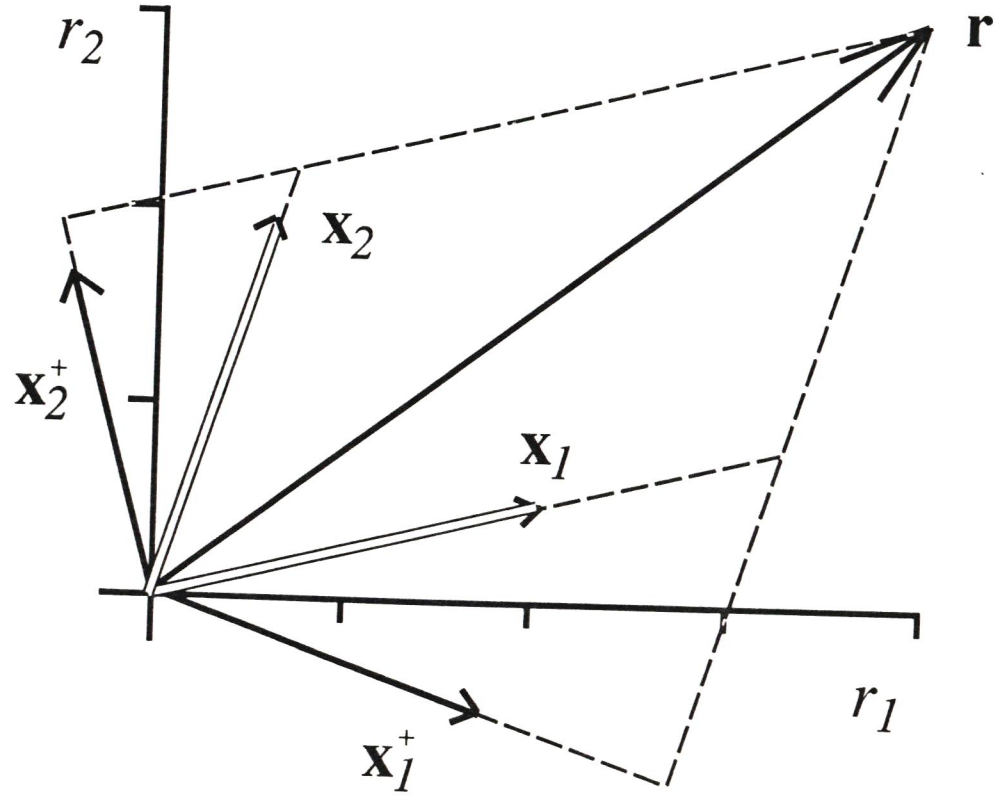

Fig 1.5 Covariant base, $\mathbf{x}_1$ and $\mathbf{x}_2$

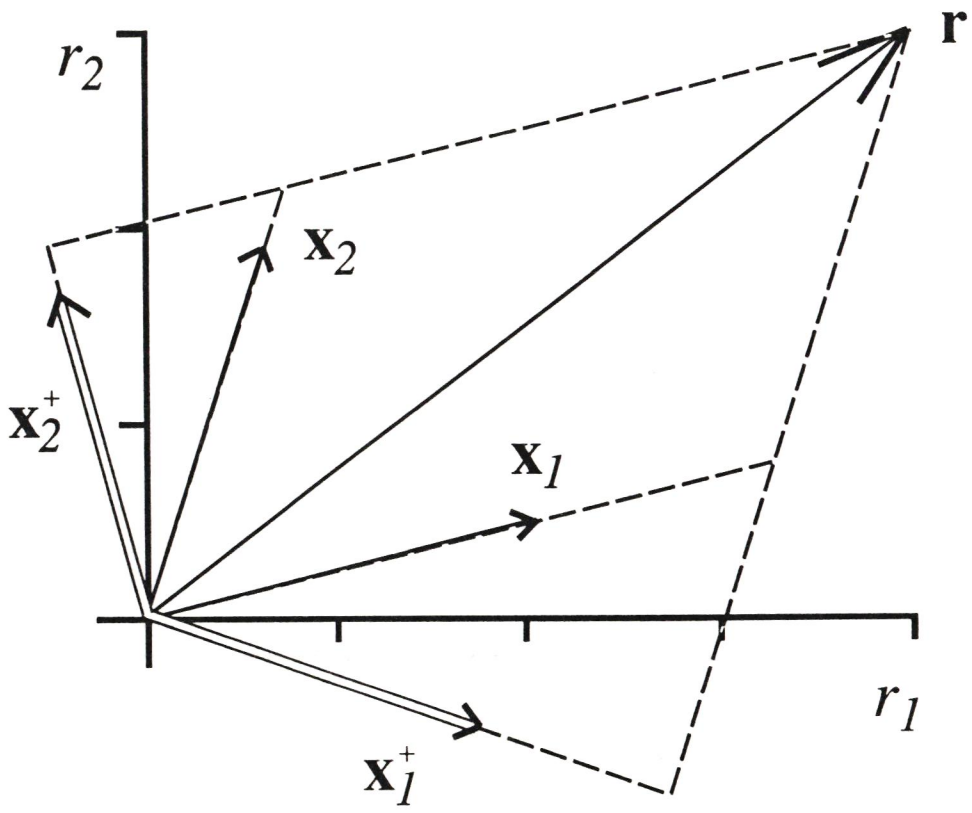

Fig 1.6 Contravariant base, $\mathbf{x}_1^+$ and $\mathbf{x}_2^+$.

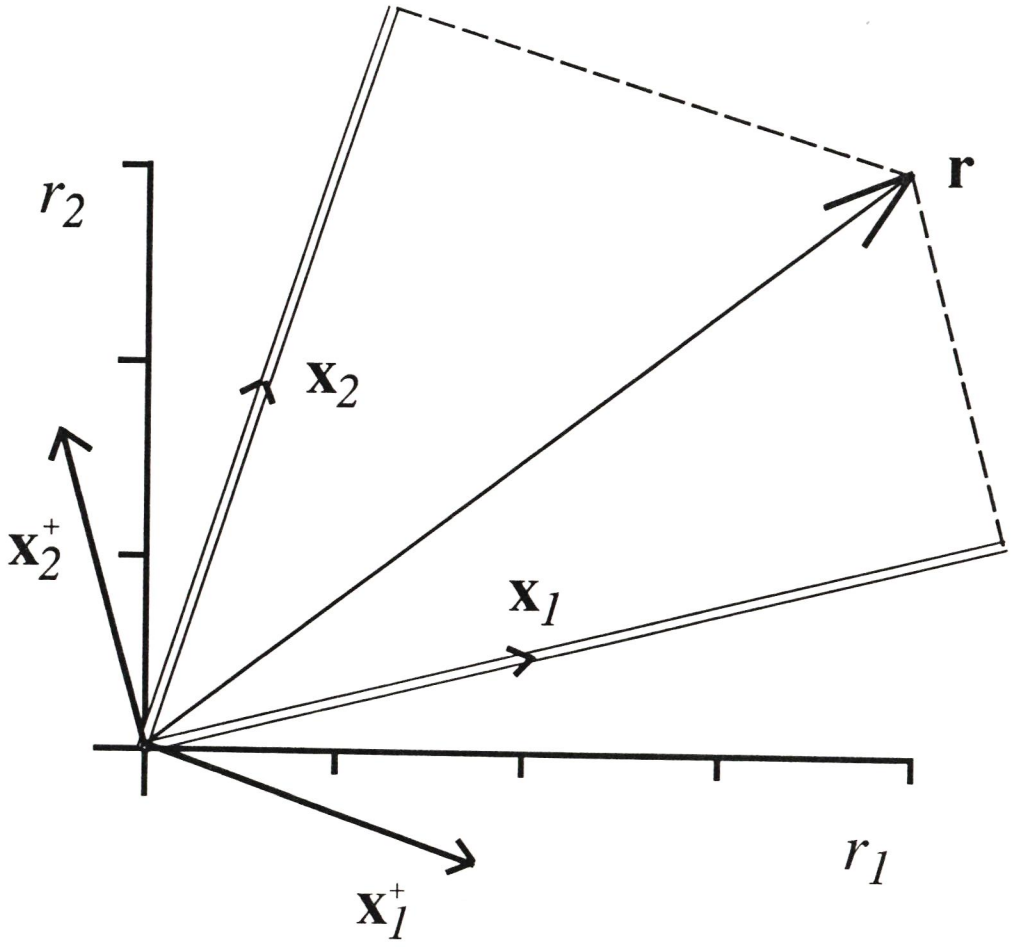

Fig 1.7 Covariant components.

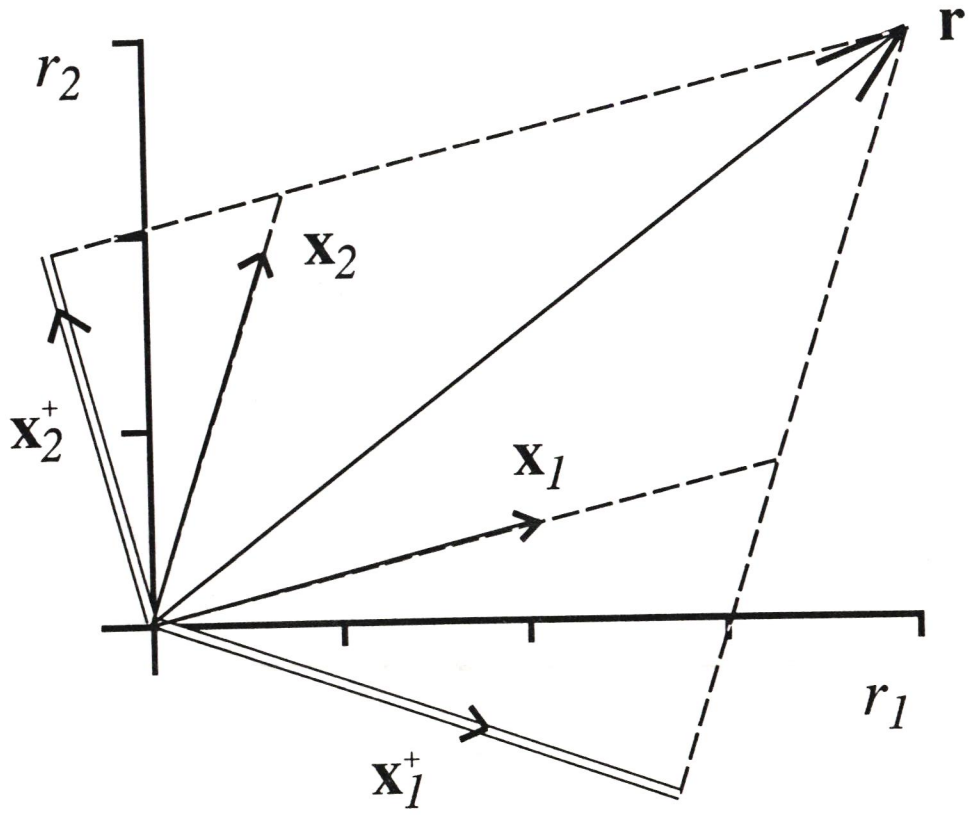

Fig 1.8 Contravariant components.

In summary, each analyte present in the sample has a covariant response pattern, a contravariant pattern and a concentration. The dot product of the sample pattern with an estimated contravariant pattern results in an estimated concentration, therefore, *estimating the contravariant sensor array pattern of an analyte in a sample solves the calibration problem.*

But, how is the contravariant pattern estimated? A set of calibration samples is necessary for which the concentration of the analyte is known in advance. The samples may be analyzed directly (external calibration) or they may be added to the unknown(s) before the analysis (GSAM, generalized standard addition method [Saxberg and Kowalski, 1979]). Only the former case will be discussed, but the concepts presented here may be extended to GSAM. The calibration samples must be selected to contain *all* of the components present in the future unknown samples, with ratios of concentration different enough to fully span all the variations of interest [Zemroch, 1986; Honigs *et al*, 1985]. Calling **R** the matrix whose rows are the response patterns of all the samples in the calibration set, an equation similar to Eq 1.5 can be written,

$$\mathbf{c}_i = \mathbf{R} \cdot \mathbf{x}_i^+ \qquad [1.6]$$

where $\mathbf{c}_i$ is a vector with the concentrations of the i^{th} analyte in every calibration sample. In practice, Eq 1.6 is an approximation, because of the experimental error in the measurement of **R** and $\mathbf{c}_i$, and in some cases also because of possible deviations from the linear model. The concentrations vector $\mathbf{c}_i$ is known and the patterns **R** are measured, therefore the contravariant vector $\mathbf{x}_i^+$ can be estimated by obtaining a *pseudoinverse* [see e.g. Lawson and Hanson, 1974] of the matrix **R**, i.e., $\mathbf{R}^+$,

$$\hat{\mathbf{x}}_i^+ = \mathbf{R}^+ \cdot \mathbf{c}_i \qquad [1.7]$$

where the symbol "^" stands for estimate. It will be shown that the difference between the most common multivariate linear calibration methods used in chemistry lies in the computation of the pseudoinverse of **R**.

The pseudoinverse, $\mathbf{R}^{+}$, of the second order tensor **R** is another tensor such that the dot product (defined in the response space) of the two is an identity tensor within the subspace spanned by the $\{\mathbf{x}_i\}$ base. The product of that identity tensor with any vector in that subspace leaves it unchanged and the product with any vector outside of the subspace simply projects it into the space (Analogous to target factor analysis, [Malinowski and Howery, 1980; Lorber, 1984B]) In general, the pseudoinverse is not equal to the *inverse* $\mathbf{R}^{-1}$, because the components of **R** do not usually form a square matrix. Also, the *generalized inverse* (defined as $\mathbf{R}^{-} = (\mathbf{R}^{T}\mathbf{R})^{-1}\mathbf{R}^{T}$) often can not be used as a pseudoinverse [see, e.g., Mandel, 1982; 1985], because the covariance matrix $\mathbf{R}^{T}\mathbf{R}$ may not be invertible.

The sensor array patterns of the calibration samples can be represented as a group of points in a multidimensional pattern space. For example, in spectroscopy, the number of wavelengths, or dimensions of the space is often as high as one thousand. Clearly, the number of dimensions of the space is probably much higher than the dimensionality of the subspace spanned by the calibration samples spectra. In addition, also the number of calibration samples is usually higher than the intrinsic dimensionality of the calibration. Therefore, in many cases the generalized inverse of the response matrix **R** is ill conditioned or undefined.

Imagine a simple case in which only two components are present in n (>> 2) calibration samples, and the array has only three sensors. If the system has a linear response, the dimensionality of the calibration set will be *two*. Therefore, a plane (two dimensions) that best fits the calibration samples has to be found, and then the generalized inverse within the subspace defined by the plane can be computed.

Naturally, some information is lost when the data is projected onto the plane; due to the experimental noise, it is also clear that there exists no vector $\mathbf{x}_i^+$ that will make Eqs 1.6 - 1.7 an identity for every concentration, but only an approximation. The goal of multivariate calibration is to find $\mathbf{x}_i^+$ such that it has the minimum error of prediction. Prediction error is defined as the error in predicting concentrations of samples that were not included in the calibration set.

In conclusion, first order tensorial calibration is performed in two steps: model building and contravariant pattern calculation. Once the model of the subspace has been defined, the contravariant pattern can be evaluated algebraically. The difference between the different calibration methods is therefore based only on the basis chosen for projecting the data. It is possible to show this mathematically by changing tensor $\mathbf{R}$ in Eqs 1.6 - 1.7 to a given orthonormal base, $\{\mathbf{v}_i\}$ with q $(< n)$ vectors, which spans the same subspace spanned by the q independent vectors $\{\mathbf{x}_i\}$,

$$\mathbf{R}_v \approx \mathbf{R}\,\mathbf{V} \qquad [1.8]$$

$\mathbf{R}_v$ $(q \times n)$ can be further reduced by observing that its columns have n components. But there are only q columns, $q < n$. Therefore, q orthonormal vectors $\{\mathbf{u}_i\}$ can be found that span the row space (e.g., using its singular value decomposition), obtaining

$$\mathbf{R}_{uv} = \mathbf{U}^T \mathbf{R}_v = \mathbf{U}^T \mathbf{R}\,\mathbf{V} \qquad [1.9]$$

with $\mathbf{R}_{uv}$ $(q \times q)$ being an invertible matrix, that approximates the tensor $\mathbf{R}$, and represents information restricted to the subspace defined by $\{\mathbf{v}_i\}$. The pseudoinverse of $\mathbf{R}$ can now be defined as

$$\mathbf{R}^+ = \mathbf{V}\,(\mathbf{R}_{uv})^{-1}\,\mathbf{U}^T \qquad [1.10]$$

The concentration is then estimated combining Eqs 1.5, 1.7, 1.8, 1.9 and 1.10. The only approximation occurs in the initial projection (Eq 1.8). This confirms the fact that the different methods of linear calibration must differ only in the base used for

projection of the calibration response matrix. The next sections will illustrate this for a few well known methods of calibration.

Direct Calibration

Direct calibration methods assume that the array pattern for all pure analytes present in the sample are available, at some unitary concentration, and the patterns are significantly different (linearly independent) [Martens and Naes, 1984]. Rewriting Eq 1.1 in matrix notation,

$$\mathbf{r} = \mathbf{X} \cdot \mathbf{c} + \mathbf{e} \qquad [1.11]$$

where $\mathbf{X}$ ($p \times q$) is a matrix whose columns are the patterns, $\mathbf{x}_i$, of all the analytes present in the sample, including any possible background, and $\mathbf{c}$ is a vector with the concentrations of each analyte in the sample.

The pure patterns form a complete base of vectors to span the calibration subspace. If the pure pattern matrix, $\mathbf{X}$, is known, its *truncated* singular value decomposition (SVD) [Lawson and Hanson, 1974] can be computed. Truncated SVD is here defined as the SVD of the matrix with all its zero singular values discarded, together with their corresponding vectors. This SVD provides us with a base for which the components of the tensor $\mathbf{X}$ are the diagonal matrix $\mathbf{S}$,

$$\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T \qquad [1.12]$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal basis sets and $\mathbf{S}$ is the diagonal matrix of the principal singular values. Substituting in Eq 1.11 and dropping the error term,

$$\mathbf{r} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T\,\mathbf{c} \qquad [1.13]$$

from which $\mathbf{c}$ can easily be estimated

$$\mathbf{c} = \mathbf{V}\,\mathbf{S}^+\,\mathbf{U}^T\,\mathbf{r} \qquad [1.14]$$

The result of Eq 1.14 is equivalent to the least squares (LS) estimate: If $\mathbf{X}$ is known, the LS estimate of $\mathbf{c}$ can be computed by

$$\mathbf{c} = (\mathbf{X}^T\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{r} \qquad [1.15]$$

which is defined if the covariance matrix $\mathbf{X}^T\mathbf{X}$ has an inverse. If one or several of the patterns are a linear combination of other patterns, then this covariance matrix cannot be inverted, and Eq 1.15 is not useful for the particular problem. This points out an advantage of Eq 1.14 over Eq 1.15: the former can still be used for quantitation of analytes whose patterns are *not* linearly dependent, even though the other patterns are.

If $\mathbf{X}$ is expressed in terms of its truncated SVD vectors, $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$, its covariance matrix is

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\,\mathbf{S}\,\mathbf{U}^T\mathbf{U}\,\mathbf{S}\,\mathbf{V}^T = \mathbf{V}\,\mathbf{S}^2\,\mathbf{V}^T \qquad [1.16]$$

and the inverse of the covariance matrix is

$$(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{V}\,\mathbf{S}^{-2}\,\mathbf{V}^T \qquad [1.17]$$

therefore, Eq 1.15 can now be written as

$$\mathbf{c} = \mathbf{V}\,\mathbf{S}^{-2}\,\mathbf{V}^T\,\mathbf{V}\,\mathbf{S}\,\mathbf{U}^T\mathbf{r} = \mathbf{V}\,\mathbf{S}^{-1}\,\mathbf{U}^T\mathbf{r} \qquad [1.18]$$

which is identical to Eq 1.14 when $\mathbf{S}^{-1}$ is equal to $\mathbf{S}^+$, i.e., when all the spectra are linearly independent.

Indirect Calibration

Often the analyst is given a group of mixture samples for building a calibration model, for which no control over the amount of analytes present in each sample is possible. This is the so-called indirect calibration problem [Martens and Naes, 1984]. For example, an accurate, expensive method is used to measure the amount of the analyte(s) of interest in the calibration set, and a model is built that relates those concentrations to the responses from some other, perhaps unexpensive, measuring technique. Three methods will be discussed from the tensorial point of view, namely the least squares approach (LS), the principal component regression method (PCR) and partial least squares (PLS).

Least Squares. The LS method is similar to the example presented in the previous section for direct calibration. The only difference is that now the calibration patterns belong to mixtures rather than pure components. Either the number of samples used for the calibration or the number of sensors should be equal to the number of variant components present in the samples, and they should be representative of the population. It is often difficult to test for these requirements, therefore this method has limited use for complex samples even though it is by far the most commonly employed.

A response matrix $\mathbf{R}$ is measured that contains q columns (one column per sample), and it will be assumed that the number of variant components is also q. To span the subspace of variation, a set of q perpendicular vectors, $\{\mathbf{u}_i\}$, can be generated from the truncated SVD of the $\mathbf{R}$ matrix,

$$\mathbf{R} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{T} \qquad [1.19]$$

whose pseudoinverse is given by

$$\mathbf{R}^{+} = \mathbf{V}\,\mathbf{S}^{+}\,\mathbf{U}^{T} \qquad [1.20]$$

If $\mathbf{c}_i$ is the vector of concentrations of the i^{th} analyte for the samples in the calibration set and $\mathbf{r}$ is the measured pattern of an unknown sample, then Eq 1.5 and 1.7 can be combined to yield

$$c_r = \mathbf{r} \cdot \mathbf{R}^{+} \cdot \mathbf{c}_i \qquad [1.21]$$

which is the desired estimate of the concentration for the unknown pattern $\mathbf{r}$. The least squares approach to this problem is analogous to Eq 1.15

$$c_r = \mathbf{r} \cdot (\mathbf{R}^{T}\,\mathbf{R})^{-1}\,\mathbf{R}^{T}\,\mathbf{c}_i \qquad [1.22]$$

and again, Eq 1.21 is obtained by substituting $\mathbf{R}$ by its truncated SVD.

Principal Components Regression. The PCR method is directly related to the SVD. The model chosen for the spanning of the subspace is given by the first q singular vectors, which are a least squares expansion of the calibration response data matrix. As with many other methods, the most difficult problem is to select how many or which components will be used for the regression. The proper truncation of the SVD is a very important factor in determining the quality of the results [see, e.g., Mandel, 1982],

$$\mathbf{R} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{T} \qquad [1.23]$$

$$c_r = \mathbf{r} \cdot (\mathbf{V}\,\mathbf{S}^{+}\,\mathbf{U}^{T}) \cdot \mathbf{c}_i \qquad [1.24]$$

the only difference between these equations and Eqs 1.19 - 1.21 is that these require that the number of components of singular values be estimated for the calculation, whereas the LS equations take all the possible singular values into account, sometimes amounting to a considerable overfit of the data. This is a problem with no satisfactory solution. The PCR solution spans the space in a generalized least squares sense, with no emphasis on prediction. Nevertheless, PCR prediction results are consistently better than LS when collinearities are present in the calibration samples [Mandel, 1982; Dempster *et al*, 1977].

Partial Least Squares. The PLS method [H. Wold, 1982; S. Wold *et al*, 1984; Naes and Martens, 1984] can also be seen from a tensorial point of view. The PLS model is usually written as

$$\mathbf{R} = \mathbf{TP} + \mathbf{e} \qquad [1.25]$$

where the orthonormal matrix $\mathbf{T}$ of vectors $\{\mathbf{t}_i\}$ is computed according to an algorithm that repeatedly projects the matrix $\mathbf{R}$ onto the concentration vector, $\mathbf{c}_i$. [see, e.g., Geladi and Kowalski, 1986]. These vectors usually span the response space biased toward the analyte of interest. Lorber and coworkers have shown that the comparison of PLS and PCR is very useful in understanding the PLS method [Lorber *et al* , 1987]. The first PLS component is a weighted sum of the principal component vectors, where the weights are the amount of the projection of the concentration vector on principal components of the profile space. In other words, rather than spanning the subspace starting with the first principal component, another vector is chosen that is biased toward the concentration vector, in an effort to explain more relevant variance in the lower vectors. Similarly to other techniques, if the $\mathbf{T}$'s are orthonormal, the concentration predictor is given by

$$c_r = \mathbf{r} \cdot (\mathbf{P}^T \mathbf{P})^{-1} \mathbf{P}^T \mathbf{T}^T \cdot \mathbf{c}_i \qquad [1.26]$$

The PLS method has consistently been found to be slightly better or at least as good as PCR for calibration in the literature. In fact, there is controversy over why should PCR discard smaller components simply based on their smaller variance, when they may as well be better for prediction [Jolliffe, 1982]. From the discussion presented herein, it can be suggested that in order to obtain good prediction, the contravariant pattern, or the prediction vector, must be at least included in the subspace generated by the spanning vectors. Obviously, if it is not in the space, poor results are expected. It is natural to understand then why the PLS method usually needs fewer factors than PCR to produce accurate concentrations: By projecting repeatedly on the concentration vector,

the model is forced to first span the region around the concentration vector, that in turn will span the region that is in the neighborhood of the contravariant pattern for that analyte. In the extreme case, if the first vector chosen to span the subspace was the contravariant pattern itself, it would obviously be the only important factor, and the rest could be discarded.

Conclusion

As stated earlier in the chapter, the intention of introducing multivariate calibration from a tensorial perspective was to provide a unifying theory of calibration. It has been seen that a considerable insight is acquired by the geometrical representation of some of the basic tensorial concepts in multivariate calibration. As a final note for this chapter, it is suggested that tensorial description of multivariate calibration should be pursued for the understanding of this topic. Two possibilities for future research are the following:

- First, given that the particular choice of vectors that span the space has crucial importance to the results, the tensorial approach could be used in the future to find the optimal set of spanning vectors for prediction. Optimal should be understood in relation to prediction results, rather than optimal model for description.
- Second, given that non-linearities and deviations from the model are one of the most important sources of errors in collinear calibration, this may be another reason why biased methods such as PLS give better results by spanning the neighborhood of the regions of the subspace that are more important for prediction. Future research in this area should start to give more importance to the samples from the calibration whose sensor array patterns are in the neighborhood of the unknown sample's

pattern, e.g., by using proximity weighting. This would be biased not only toward the concentration vector, but also toward the response vector of the unknown sample, that is not usually considered for the calibration model. It would also be similar to a generalization of univariate local estimation methods to multivariate data [Tibshirani, 1984].

Finally, it is important to remark that whenever interferences are present in an unknown sample that were not included in the calibration set, the results of first order calibration will be unreliable. Therefore it is important to validate the prediction model for the unknown sample before estimating the analyte concentration. Unfortunately, if the model is not valid, the concentration can not be estimated. However, second order calibration can handle this problem, and it will be discussed in the next chapter.

CHAPTER II

SECOND ORDER TENSORIAL CALIBRATION

Second order calibration is defined as calibration of an instrument that produces a two dimensional (2D) array of data per object or sample that is analyzed. The discussions in this chapter are specific to a special kind of 2D data: bilinear or dyadic data. Bilinear data from a single pure chemical component can be expressed as the outer product of two vectors. Bilinear data for n constituents can be expresed as the sum of n outer products. Examples of bilinear techniques are emission-excitation fluorescence or chromatography-spectroscopy combinations. Examples of non-bilinear techniques are two dimensional nuclear magnetic resonance (2D-NMR) and mass spectrometry - mass spectrometry (MS/MS) [see, e.g., Johnson and Yost, 1985]. Even though it will not be presented here, tensorial concepts could also be applied to non-bilinear data.

The idea behind second order calibration is the following: bilinear data, collected from a sample with a second order instrument, can be interpreted as the components of a second order tensor. If two samples have the same intrinsic constituents, there exists at least one change of the two second order bases that will transform the components of both tensors to diagonal matrices. If there is only one such pair of bases, they can be estimated, and will correspond to the intrinsic vectors or the true spectra in both orders, and the tensor diagonal components will be directly proportional to the amounts of each component present, allowing for quantitation.

The last chapter was an introduction to the tensorial concepts using well-known first order (multivariate) calibration. This chapter will use some of those ideas and

concepts to develop second order calibration for bilinear data, and will outline some possibilities for third and higher order calibration.

The structure of this chapter is as follows. First, GRAM theory from a second order calibration perspective is presented. Then, computer simulations are used to illustrate the effect on GRAM results from different factors like noise level, similarity of the analytes spectra, etc. Finally, some examples of the application of GRAM to real chromatographic data are presented, including an interesting, complex sample with more than 300 constituents.

Theory

There are many instruments in chemistry that generate second order data, e.g, all the "hyphenated" methods [Hirschfeld, 1980, 1985] including chromatography-spectroscopy conbinations, MS/MS, 2D-NMR, etc. All these techniques generate, from a single sample, data that can be represented in a matrix. From tensorial theory, it is known that the components of a second order tensor are also a matrix. Therefore, data acquired from a second order instrument can be defined as the components of a tensor. Designating M_{ij} the element at the i row and j column of a the data matrix, then there is an associated tensor **M** whose components are M_{ij}:

$$\mathbf{M} = \sum_{i,j} M_{ij} \ \tau_i \ \lambda_j \qquad [2.1]$$

where τ_i represents the basis vectors for one order and λ_j represents the set of base vectors in the other orther. They are the standard bases; e.g., for a LC/UV instrument, τ_i represents measurement at time i, and λ_j represents measurement at wavelength j. Using the summation convention, Eq 2.1 can be expressed as* ,

$$\mathbf{M} = M_{ij} \ \tau_i \ \lambda_j \qquad [2.2]$$

for vectors and tensors. The summation convention invariably applies to summation over a repeated index in one side of the equation.

The data matrix M_{ij} from a second order instrument has an associated tensor **M**. From tensorial theory, the tensorial object **M** is invariant to changes in base, therefore any base that spans the spaces defined by the standard bases τ_i and λ_j could be used to represent the tensor **M**, and it would be describing the same mathematical object. The standard bases used in Eq 2.2 are a convenient choice because their vectors are

* See appendix A for a description of the summation convention.

orthonormal, therefore, the data matrix can be directly related to the instrument variables.

The standard base is not the only useful base. A generally appropriate base for a tensor can be obtained from the singular value decomposition (SVD) of the components matrix of the tensor. The SVD decomposes any matrix into the product of three matrices,

$$\mathbf{M} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{T}, \qquad [2.3]$$

where **U** and **V** are orthonormal matrices, i.e., their respective columns $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_j\}$ are orthogonal, unitary vectors; and **S** is a diagonal matrix. These two sets of vectors can be used as the base for the tensor **M**, and the covariant and contravariant components are equivalent, because they are orthonormal, and the projection formula is enough for the change of base. In matrix notation, by left multiplying **M** by $\mathbf{U}^T$ and right multiplying by **V**, the components of **M** in the new base are obtained.

$$\mathbf{U}^T\,\mathbf{M}\,\mathbf{V} = \mathbf{U}^T\,\mathbf{U}\,\mathbf{S}\,\mathbf{V}^T\,\mathbf{V} = \mathbf{S} \qquad [2.4]$$

But **S** is a diagonal matrix, therefore the singular value decomposition vectors provide a base where the components of the tensor are a diagonal matrix, and Eq 2.2 is reduced to

$$\mathbf{M} = S_{ii}\ \mathbf{u}_i\ \mathbf{v}_i \qquad [2.5]$$

Thus, the matrices **M** and **S** represent the same tensor under different bases. There are an infinite number of matrices that represent the tensor **M**, but in most cases there is only one that is, at the same time, diagonal and its bases are orthonormal, and this is given by the singular value decomposition.

Another important base with diagonal components arises when the second order data comes from a bilinear instrument. As stated earlier, such data has the interesting property that if there is only one analyte present in the sample, then the matrix can be approximately expressed as the outer product of two vectors, which also means that the rank of the matrix should be one, within the noise level. If the technique is, e.g.,

emission-excitation fluorescence, then one of the vectors would correspond to the emission spectrum of the analyte and the other to its excitation spectrum,

$$\mathbf{N}_i = \mathbf{x}_i \mathbf{y}_i^T + \mathbf{e}_i \qquad [2.6]$$

where $\mathbf{x}_i$ is the spectrum in one order, e.g., excitation; $\mathbf{y}_i$ is the spectrum in the other order, e.g., emission; and $\mathbf{e}_i$ is the error of the approximation. Of course, this is the base that is of greatest interest to the chemist. If a sample has q analytes, then the matrix can be modeled by a sum of q unitary rank matrices, and will have rank q,

$$\mathbf{M} = \sum_i^q \mathbf{N}_i = \sum_i^q \mathbf{x}_i \mathbf{y}_i^T \qquad [2.7]$$

where the error in the model has been dropped for simplicity. Rewritting Eq 2.7 in matrix notation, by considering the vectors $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ as the columns of the matrices $\mathbf{X}$ and $\mathbf{Y}$, yields

$$\mathbf{M} = \mathbf{X}\,\mathbf{Y}^T \ , \qquad [2.8]$$

forcing the $\{\mathbf{x}_i\}$ vectors and the $\{\mathbf{y}_i\}$ vectors to be of unitary length, then a third matrix is necessary, with the normalization constants in its diagonal, and zeros in the rest,

$$\mathbf{M} = \mathbf{X}\,\mathbf{ß}\,\mathbf{Y}^T . \qquad [2.9]$$

This equation is similar to Eq 2.3, with a diagonal matrix ß, but the matrices $\mathbf{X}$ and $\mathbf{Y}$ are not orthonormal. They are the pure analyte spectra, or sensor array patterns, in each order. By analogous reasoning, expressing the tensor $\mathbf{M}$ in the bases $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$, its components will be again a diagonal matrix, in this case ß. It is important to realize that Eq 2.9 is only an approximation. For normal experimental conditions, factors such as noise, or systematic deviations from the model are always present. Again, because ß is a diagonal matrix, the tensor $\mathbf{M}$ can be expressed as

$$\mathbf{M} = \beta^{ii}\ \mathbf{x}_i\ \mathbf{y}_i \qquad [2.10]$$

Note the position of the indexes ii as a superscript, indicating that the β^{ii} are contravariant components for both bases. If the instrument is linear, the vectors $\mathbf{x}_i$ and

$\mathbf{y}_i$ do not change with a change in concentration of the analyte i, i.e., c_i. The only thing that changes is β^{ii}, therefore, for a linear response instrument, the β^{ii}'s are directly proportional to the concentrations, and if they are determined, they can be used for quantitation.

It has been found that there are at least two important advantages derived from calibration with bilinear data [see also Sánchez and Kowalski, 1986]:

1.- The concentrations of the analytes of interest can be estimated even if linearly independent background and other factors are present in the unknown sample but absent in the calibration sample(s).

2.- The intrinsic bilinear spectrum, or array pattern, of all the components in common between the unknown sample and the calibration sample(s) can be estimated.

Assume that there is only one multicomponent sample in the calibration set. Calling **N** the tensor of its responses, it can modeled with an equation similar to Eq 2.9,

$$\mathbf{N} = \mathbf{X}\,\boldsymbol{\xi}\,\mathbf{Y}^{T} \qquad [2.11]$$

N is the tensor in the standard base and $\boldsymbol{\xi}$ is the diagonal tensor in the ***xy*** base. The components of **N** are known only in the standard base. Similarly, an unknown or test sample has a response tensor **M**, that can also be modeled with Eq 2.9, repeated here for convenience.

$$\mathbf{M} = \mathbf{X}\,\boldsymbol{\beta}\,\mathbf{Y}^{T} \qquad [2.12]$$

To describe a general case, assume that the **X** and the **Y** are matrices that contain a superset of all the components present in both **M** and **N**. There will be corresponding β_i or ξ_i elements that will be zero if they are not present in one or the other sample. It is evident that the bases **X** and **Y** can express both tensors **M** and **N** with a diagonal matrix of components, respectively $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$. It will be shown that under certain conditions,

To find the base, Eqs 2.10 -2.11 can be used as a system of two matrix equations with four unknowns, namely **X**, **Y**, **ß** and **ξ**. To solve for these parameters, **N** is expressed as a function of **M**,

$$\mathbf{N} = \mathbf{X}\,\boldsymbol{\xi}\,\mathbf{Y}^{T} = \mathbf{X}\,\boldsymbol{\beta}^{-1}\,\boldsymbol{\xi}\,\boldsymbol{\beta}\,\mathbf{Y}^{T}$$

$$\mathbf{N} = \mathbf{X}\,(\boldsymbol{\beta}^{-1}\,\boldsymbol{\xi})\,\mathbf{X}^{+}\,(\mathbf{X}\,\boldsymbol{\beta}\,\mathbf{Y}^{T})$$

$$\mathbf{N} = \mathbf{X}\,(\boldsymbol{\beta}^{-1}\,\boldsymbol{\xi})\,\mathbf{X}^{+}\,\mathbf{M} \qquad [2.13]$$

these equations are valid only if all the elements of the diagonal matrix **ß** are non-zero. This implies that **M** contains all the components of the superset **X** and **Y**. For a case that this is not true, the new matrix **W** = **M** + **N** can be used instead of **M**, that by definition would include all the components. $\mathbf{X}^{+}$ represents the contravariant form of the base **X**, and is simply the generalized inverse of the **X** matrix. Defining $\boldsymbol{\lambda} \equiv \boldsymbol{\beta}^{-1}\,\boldsymbol{\xi}$, Eq 2.13 can now be factorized to result into a non-symmetric eigenvalue-eigenvector problem, after right multiplying by $\mathbf{M}^{+}$ and then **X** as,

$$\mathbf{N}\,\mathbf{M}^{+} = \mathbf{X}\,\boldsymbol{\lambda}\,\mathbf{X}^{+}\,\mathbf{M}\,\mathbf{M}^{+} \qquad [2.14]$$

$$(\mathbf{N}\,\mathbf{M}^{+})\,\mathbf{X} = \mathbf{X}\,\boldsymbol{\lambda}\,\mathbf{X}^{+}\,\mathbf{X}$$

$$(\mathbf{N}\,\mathbf{M}^{+})\,\mathbf{X} = \mathbf{X}\,\boldsymbol{\lambda} \qquad [2.15]$$

The spectra **X** are the *right* eigenvectors of the square non-symmetric matrix $(\mathbf{N}\,\mathbf{M}^{+})$, and the eigenvalues **λ** are the ratios of concentrations, because **ß** and **ξ** are proportional to concentrations. Once **X** is known, $\mathbf{Y}^{T}$ can be estimated using

$$\mathbf{Y}^{T} = \boldsymbol{\beta}^{-1}\,\mathbf{X}^{+}\,\mathbf{M} \qquad [2.16]$$

Eqs 2.15-2.16 summarize the Generalized Rank Annihilation Method (GRAM) [Sánchez and Kowalski, 1986]. A particular case of Eq 2.15 arises when **N** has only one component, $\mathbf{N} = \mathbf{x}_1\,\xi_1\,\mathbf{y}_1^{T}$. All the eigenvalues will be nearly zero with the exception of one, $\lambda_1 = \beta_1/\xi_1$. Then Eq 2.15 can be rewritten as

$$(\mathbf{x}_1\,\xi_1\,\mathbf{y}_1{}^{T}\mathbf{M}^{+})\,\mathbf{x}_1 = \mathbf{x}_1\,\lambda_1 \qquad [2.17]$$

dropping $\mathbf{x}_1$ and changing sides,

$$\lambda_1 = \xi_1 \, \mathbf{y}_1^{\mathrm{T}} \mathbf{M}^{+} \mathbf{x}_1 \qquad [2.18]$$

which is equivalent to the non-iterative Rank Annihilation equation introduced by Lorber [Lorber, 1984, 1985]. Actually it is not necessary to estimate $\mathbf{y}_1^{\mathrm{T}}$ and $\mathbf{x}_1$ from **N**. Recognizing that $N_{ij} = \xi_1 \, x_i \, y_j$, Eq 2.18 can be simplyfied to

$$\lambda_1 = N_{ij} \, (M^{+})_{ij} \qquad [2.19]$$

where a double summation over i, j applies, and $(M^{+})_{ij}$ represents the i^{th} row and jth column component of the pseudoinverse of **M**.

Characteristics of Bilinear Calibration

Eqs 2.18 - 2.19 provide information useful to understand the possibilities and limitations of Bilinear Calibration using GRAM. First of all, they not only represent a calibration method, but also a curve resolution method, because the intrinsic factors are extracted. It is not the same curve resolution as described in the literature [see e.g. Lawton and Sylvestre, 1971; Osten and Kowalski, 1984], in which an uncertainty region is defined where the intrinsic factors are present, and further constraints must be used to choose a solution within the region. GRAM estimates a unique solution without empirical assumptions.

It is a fact that when two (or more) eigenvalues are identical (or very close to each other for experimental data), their corresponding eigenvectors are not unique. Therefore, one of the restrictions of GRAM is that if two components have the same ratio of concentrations between the samples, their eigenvalues will be very similar, and the estimated spectra will be unreliable. Nevertheless, the eigenvalues should still provide quantitative information, but it will be more difficult to match the eigenvalue with its corresponding analyte, because no estimated spectrum is available. If a library spectrum is available, simple target factor analysis on the space generated by all the eigenvectors with the same eigenvalue should confirm the identity of at least one of the eigenvalues [Malinowski and Howery, 1980; Lorber, 1984].

A limitation arises when the intrinsic spectra in one of the orders are not linearly independent. Eq 2.17 is no longer valid, because $\mathbf{X}\,\boldsymbol{\lambda}\,\mathbf{X}^{+}\,\mathbf{M}\mathbf{M}^{+}$ is not equal to $\mathbf{X}\,\boldsymbol{\lambda}\,\mathbf{X}^{+}$. Assuming that the $\mathbf{Y}$ are linearly dependent, the matrix $\mathbf{M}$ will have lower rank than the matrix $\mathbf{X}$, which is made up of linearly independent vectors. The matrix ($\mathbf{M}\mathbf{M}^{+}$) is a projection matrix that behaves like the identity matrix for vectors in the subspace spanned

by **M**. But if **X** has higher rank than **M**, it is necessarily true that the vectors in **X** will have some component outside of the space spanned by **M**, therefore $\mathbf{MM}^+$ does not leave $\mathbf{X}\,\boldsymbol{\lambda}\,\mathbf{X}^+$ unchanged. The results then should be unpredictable when such a dependency exists. Fortunately there is a simple test that will detect, but not correct this problem: the projection of **N** on $\mathbf{M}\,\mathbf{M}^+$ should leave **N** unchanged within the noise level.

The estimation of the pseudoinverse of **M** is the most important step in GRAM. In a similar way to first order tensorial calibration, the selection of the proper subspace for the pseudoinverse is a determining step in the quality of the GRAM results. Eq 2.18 shows that **N** is literally projected onto **M**, therefore the best way to span the space is to find a set of vectors that express both **M** and **N** in an unbiased way. The approach used in this work was to join the matrices to form a larger matrix $\mathbf{W} = (\mathbf{M}|\mathbf{N})$, and then obtain the truncated singular value decomposition of the joined matrix,

$$\mathbf{W} = (\mathbf{M}|\mathbf{N}) = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T \qquad [2.20]$$

By joining the matrices **M** and **N** into the matrix **W**, there will be twice as many columns in **W** as there are columns in the original matrices. In this way, the columns of the matrix **U** will be vectors that are an unbiased estimation of the column subspace of both **M** and **N**. This and other models for projection of the data will be discussed in the following sections. The number of relevant singular values is first estimated by cross validation in order to truncate the SVD [Eastment and Krzanowski, 1982; Wold, 1978]. Then, new projected matrices are used for the calculations,

$$\mathbf{N} = \mathbf{U}\mathbf{U}^T\,\mathbf{N} \qquad [2.21]$$

$$\mathbf{M}^+ = (\mathbf{U}^T\,\mathbf{M})^+\,\mathbf{U}^T \qquad [2.22]$$

$$\mathbf{N}\,\mathbf{M}^+ = \mathbf{U}\mathbf{U}^T\,\mathbf{N}\;(\mathbf{U}^T\,\mathbf{M})^+\,\mathbf{U}^T$$

$$\mathbf{N}\,\mathbf{M}^+ = \mathbf{U}\,(\mathbf{U}^T\,\mathbf{N})\,(\mathbf{U}^T\,\mathbf{M})^+\,\mathbf{U}^T$$

$$\mathbf{N}\,\mathbf{M}^+ = \mathbf{U}\,\mathbf{N}_u\;\mathbf{M}_u{}^+\,\mathbf{U}^T \qquad [2.23]$$

where $\mathbf{N}_u$ and $\mathbf{M}_u$ stand for the **N** and **M** tensor components in the **U** subspace. For this case $\mathbf{M}_u^+$ is equivalent to the generalized inverse, $\mathbf{M}_u^+ = \mathbf{M}_u^T (\mathbf{M}_u \mathbf{M}_u^T)^{-1}$. Finally, The new matrix $\mathbf{N}\,\mathbf{M}^+$ is then substituted in Eq 2.15 for a GRAM calculation.

Next sections will present theoretical and computer simulation results that illustrate and expand on the characteristics of GRAM and its limitations.

Factors that affect the quality of the GRAM results: Theory

It is a difficult problem to estimate the error of the GRAM results, considering that there is error in **M**, **N** and in the concentrations of the calibration sample. Ho and coworkers [Ho *et al*, 1980] and Appellof [Appellof, 1981] have shown that for small, homoscadastic errors in the matrix **M**, the error in estimating the concentration ratio for one analyte is given by

$$\sigma^2(c_M/c_N) = (\xi^2 q_x q_y)^{-1} \sigma_M^2 \qquad [2.24]$$

where σ_M^2 is the average square residuals from **M**, c_M/c_N is the concentration ratio, ξ is the only nonzero element of the diagonal matrix $\boldsymbol{\xi}$ as defined in Eq 2.11, and $q_x q_y$ are the so called *uniqueness* factors, for **x** and **y**, and define the degree of overlap that the analyte's spectrum has with all the other spectra in each order. It can be shown that $(\xi^2 q_x q_y)^{-1}$ is an extension to second order data of the net analyte signal concept developed by Lorber [Lorber, 1986, 1987]. q_x (q_y) represents the component of the $\mathbf{x}_i$ ($\mathbf{y}_i$) vector that is perpendicular to the rest of the vectors present in **M**. Mathematically,

$$q_x = sin^2(\alpha_x) , \qquad q_y = sin^2(\alpha_y) \qquad [2.25]$$

where α_x is the angle between $\mathbf{x}_i$ and the subspace spanned by $\{\mathbf{x}_{j\neq i}\}$ and α_y is the angle between $\mathbf{y}_i$ and the subspace spanned by $\{\mathbf{y}_{j\neq i}\}$.

It is difficult to estimate the uniqueness factors unless all vectors $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ are known. Fortunately, GRAM estimates these vectors, providing a direct computation of q_x and q_y. Eq 2.18 provides the $\{\mathbf{x}_i\}$ in the form of the matrix **X**, and Eq 2.19 provides the $\{\mathbf{y}_i\}$. Then the formula $q_x = (1 - \mathbf{x}_i^T \mathbf{S}\mathbf{S}^+ \mathbf{x}_i)$ can be used (where **S** is **X** matrix excluding the column corresponding to $\mathbf{x}_i$.) [Ho *et al*, 1980].

Eq 2.24 is useful for understanding the nature of error propagation in Rank Annihilation with an error free **N** matrix. Unfortunately, the experimental errors in **M**

and **N** are usually of the same order, therefore to assume error in **M** and not in **N** is rather arbitrary. In spite of the difficulty of considering both errors simultaneously, it is possible to estimate the relative concentration variance for a simple **M**(*2x2*) with two components and a **N**(*2x2*) with one component, yielding*

$$\sigma^2 (c_{M1})/c_{M1}^2 = \sigma_c^2 /c_{N1}^2 + \beta_2^2/D^2 \sigma_M^2 +$$

$$+ (c_{M1} /c_{N1})^2/D^2 [\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2 \cos(\alpha_x) \cos(\alpha_y)] \sigma_N^2 \qquad [2.26]$$

$$D = \beta_1\beta_2 \sin(\alpha_x) \sin(\alpha_y) \qquad [2.27]$$

where σ_N^2 is the error in **N**; in many cases it can be considered to be equal to σ_M^2.

The first term is the relative variance of the concentration estimate in the calibration matrix, setting a lower bound for the error. It implies the logical result that the estimate of c_{M1} cannot be better than c_{N1}.

The second term in Eq 2.26 is the contribution from the σ_M^2 error, and it can be shown that it is equivalent to a *2x2* case of Eq 2.24. The coefficient of σ_M^2 is actually $(\beta_1 \sin \alpha_x \sin \alpha_y)^{-2}$, therefore, the smaller the angles, the greater the error propagation. Also, higher concentrations of the analyte in the unknown sample, i.e., higher c_{M1}, decrease the error of the estimation.

The effect of the **N** error is more difficult to interpret (third term), but it can be compared with the effect of the **M** error by dividing the third by the second term, using $\beta_1 = c_{N1}/c_{M1}$, and assuming that $\sigma^2 \equiv \sigma_N^2 = \sigma_M^2$, and that $\cos(\alpha_x) \cos(\alpha_y) \approx 1$ results in

$$(\beta_1 / \xi_1)^2 (\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2) / \beta_2^2$$

$$= (\beta_1 / \xi_1)^2 (\beta_1 + \beta_2)^2 / \beta_2^2 \qquad [2.28]$$

which is usually bigger than one , e.g., if all variables have similar magnitude, this ratio is *4*. This means that in general, the error due to the calibration matrix is bigger than the error in **M**, or at least of the similar magnitude. In a real experiment, only the

* See appendix B for the derivation of this formula.

ξ_1 term may be controlled, which is proportional to the concentration in the calibration, c_N. Therefore, increasing c_N, decreases the coefficient of σ_N^2, as long as the bilinear model holds.

The previous discussion applies for the simple *2x2* case with two components in **M** and one component in **N**. The use of Eq 2.26 for cases with more components or higher dimensionality is qualitatively useful, but presents several problems when applied to GRAM:

• Eq 2.26 is valid for the $c_{N1}/c_{M1} = \xi_1 \mathbf{y}^T_1 \mathbf{M}^+ \mathbf{x}_1$ RA estimator (Eq 2.18), but GRAM estimates c_{N1}/c_{M1} as an eigenvalue of the $\mathbf{NM}^+$ matrix. They provide the same result for noise free data, but that is not necessarily the case when noise is present. Another problem with Eq 2.26 is that any bias introduced by the use of Eq 2.18 is not taken into account, and it will be seen later how choosing the wrong model affects the GRAM result much more significantly than the standard deviation of the estimator.

• If q $(q > 2)$ components are present in a full rank $\mathbf{M}(qxq)$ matrix, the equation does not apply. Higher order terms may be necessary that are not present in the *2x2* case. Nevertheless, as more parameters are added to the computation, more accurate results will be obtained due to signal averaging.

• Even for an **M** matrix with *2* components, if the dimensions are higher than *2*, a model has to be built to project both **M** and **N** onto it. The error of the model introduces a bias in the result that is not predicted by Eq 2.26. An example of this error occurs when the model is generated from the singular value decomposition of **M**. The space spanned by the true intrinsic factors $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ will form an angle with the model spanned by **M**, and there will always be an angle, whose value will be correlated with the value of σ_N. Part of **N** will be lost when projecting into the **M** space, producing average estimated concentrations that are higher than the expected value. Using unbiased models, such as (**M**|**N**), should solve that problem.

• GRAM may use more than one component in the calibration matrix. In general, the estimation of the intensity ratio of a component present in both **M** and **N** is a symmetrical problem, that ideally should have a symmetrical solution. Therefore, if **N** has more than one component, considerations similar to those for **M** should apply.

• The intrinsic factors $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ common to **M** and **N** are also estimated with GRAM, but Eq 2.26 gives no estimation of their error. Nevertheless, considering that the error of the eigenvalues is correlated to the error in the eigenvectors, Eq 2.26 should provide a qualitative idea of the quality of the estimated spectra.

These problems do not necessary imply that error estimations using those equations are useless. This equations are useful to obtain a qualitative and semi-quantitative idea of error propagation in GRAM for many cases. The results are probably expected: GRAM will perform better with less noise and less overlap between the intrinsic vectors of the samples. Nevertheless, next section will show how for several simulations, errors vary linearly with the inverse of the uniqueness.

Factors that affect the quality of the GRAM results: Computer Simulations

Several computer simulations were performed to study the effect of noise upon the GRAM results under different conditions. For noise free data, the results were consistently perfect up to the level of the computer precision, therefore all simulations included some noise. UV spectra from a library of polyaromatic hydrocarbons were used coupled with gaussian chromatographic profiles, to simulate the data. The spectra had ninety seven (*97*) wavelengths each (matrix **X**, Eq 2.8) and the chromatograms had *ca.* thirty (*30*) scans (matrix **Y**, Eq 2.8). All the spectra were normalized to a maximum absorption of 100 *units*, and the gaussian peaks to unitary peak height. A constant 2% noise was added to all the data. 2% noise means that the standard deviation of each measurement is 2 *units*. Therefore, for unitary concentration, the relative noise is 2% only at the maximum of the chromatogram and at the maximum of the spectrum. On the average, the noise is much higher than 2%.

In light of the discussion in the previous section, simulations were performed testing the effect of several factors on the results of GRAM . The most important factors to consider include:

- Model chosen for projection: (1) Number of components (2) Spanning subspace.
- Similarity of the intrinsic vectors: (1) Similarity of the $\{\mathbf{x}_i\}$ (2) Similarity of the $\{\mathbf{y}_i\}$ (3) Similarity of the ratio eigenvalues $\{\lambda_i\}$.
- Effect of the relative concentrations: (1) Concentrations of the calibration sample (2) Concentrations of the unknown sample.
- Effect of the number of analytes present: (1) In the calibration sample (2) In the unknown sample.

The following discussion will go into the details of each simulation and will present its results.

Model chosen for projection. Model selection is the most difficult and important factor to be considered. The model can be chosen, whereas other factors are usually already established. Two sets of simulations were used to illustrate the problems associated with improper model selection.

- The first set of simulations shows the effect of choosing an incorrect dimensionality for the subspace that models the matrices **M** and **N**. Both samples were simulated with four chemical constituents each. Table 2.1 shows the results when selecting three, four, five and six mathematical components (singular values) for the singular value decomposition of (**M**|**N**) (see Eq 2.20). Only one calculation per dimensionality was sufficient to illustrate this effect. The noise level was fixed at two *units* as described in the previous section.

For the three component selection very poor results are obtained, with a best estimate for BbFl, which is the major constituent. By definition, only three spectra are estimated when three components are used in the model. The three-dimensional model, in attempting to fit a four-dimensional data set, loses non-random information. It is expected that the estimated spectra will be linear combinations of the true spectra. Naturally, the correlation coefficients between the estimated spectra and the true spectra are very low (maximum = 0.9533) when contrasted with the results for models with higher dimensionalities.

It is interesting to note that the estimates using five and six components are reasonably good in spite of the fact that only four chemical constituents are present. This is very useful when there is uncertainty in the number of components, because it is safe to select more components than necessary for the model. Therefore, the dimensionality

TABLE 2.1

EFFECT OF MODEL DIMENSIONALITY ON THE ESTIMATED CONCENTRATIONS AND SPECTRA

The first number in the groups of two is the estimated concentration and the second is the correlation coefficient between estimated spectrum and library spectrum. The proper dimensionality should be four (4). The missing values occur because only three estimates are obtained with three components.

Analyte	*Concentration*	*Dimensionality of the Model*			
		3	**4**	5	6
BbFl	3.0000	1.3167	2.9842	2.9890	2.9915
		0.9533	0.9999	0.9999	1.0000
BaP	2.0000	-	2.0008	2.0004	2.0018
		-	0.9944	0.9942	0.9942
BkFl	1.5000	2.8887	1.4871	1.4883	1.4904
		0.9063	0.9995	0.9996	0.9995
BeP	1.0000	1.7198	0.9996	0.9998	1.0023
		0.2810	0.9991	0.9990	0.9991

of the model must be equal to or higher than the true dimensionality in order to get the best results. However, this is not true for noise free data, for which estimating a pseudoinverse $\mathbf{M}^+$ with more components than the true number produces computer overflow or unpredictable results. But this is not a problem, because for this kind of data (very low or no noise data) there is no problem in selecting the right number of components for the calculation. There is no perfect method for determining the number of principal components for noisy data, but it is apparent from the table that observing the behavior of the eigenvalues as more components are selected for the model provides a useful method for GRAM, because the important eigenvalues do not change significantly beyond the correct number of components.

• To compute the pseudoinverse of the **M** matrix, the data must be projected on a representative set of vectors, here refered to as the model. The second set of simulations compares four different models for a two component sample **M** and a one component sample **N**. For the first model the SVD of **M**, $\mathbf{M}=\mathbf{USV}^T$, and the columns of the **U** matrix where used to span the space and then project the matrices into it. The second model used the SVD of (**M**+**N**) and the third used the SVD of (**M**|**N**). Finally, the fourth model used the SVD of the true intrinsic vectors used to simulate the data in the first place, i.e., the matrix **X**. Five concentration levels c_M were used, and the GRAM estimation was performed five times at each level, with different noises but again constant $\sigma_M = 2\ units$, to estimate the standard deviation of the estimated concentrations. A value of 0.01 in the plot is equivalent to a 1% error in the concentration estimation of c_M.

There are two distinct errors to consider: the standard deviation (SD), or precision of the estimated concentration, and the accuracy of the mean estimation (bias),

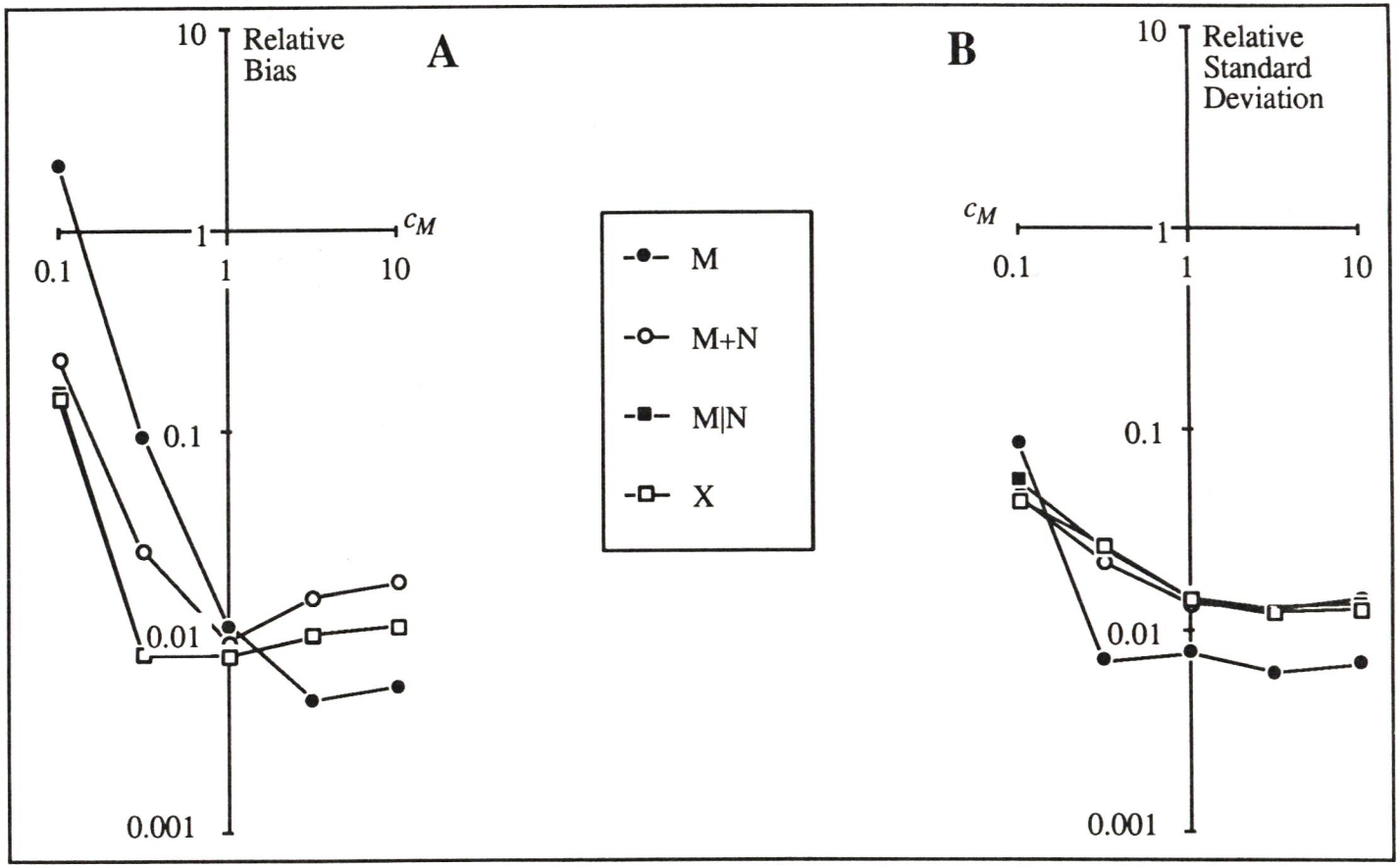

Fig 2.1 Model effect on GRAM. Relative standard deviation and bias of GRAM results at different c_M concentrations are compared for four different models computed from the singular value decomposition of four different matrices: **M**, **M+N**, **M|N** and **X**. (A) Relative Bias; (B) Relative Standard Deviation. (A value of 0.01 is equivalent to a 1% error).

or difference between the true value and the mean estimate. Fig 2.1A shows the relative bias, and Fig 2.1B the relative SD, for the four models. The most important point to illustrate here is that the bias is smaller and it is in the same order of magnitude as the precision for high concentrations (as it should be: for a normal distribution, the distribution of many calculations of the bias is simply the standard deviation of the mean), but as the concentration decreases, the bias turns out to be the most important factor, being even two orders of magnitude larger than the precision in the worst case (with the **M** model). Therefore, the proper choice of a model is an important step in GRAM. If the model was chosen improperly, it does not matter how many times the experiment is repeated, it will not give a better average estimate of the concentration. In this case, the best model was the pure spectra used for the simulation; in practice, those factors are probably not available, and the models using (**M**+**N**) and (**M**|**N**) are the next best choice. The use of **M** alone for the model should be avoided, especially for low c_M concentrations. The only exception is for high c_M concentrations, where the **M** model does marginally better than the others. In this particular case, with a concentration of ten *units* in **M** and a concentration of one *unit* in **N**, it is possible that **M** is a better model, because the signal in **N** is ten times weaker, and adding **N** to the model is almost like duplicating the noise without significant increment in the signal. In practice it is not advisable to have a ten-fold difference in concentrations due to possible instrument deviations from linearity, unless several calibration samples are used that cover the dynamic range.

Eq 2.26 predicts that a higher concentration c_M results in a lower relative SD, and that holds true for all the models shown in Fig 2.1B. That equation is not valid for the prediction of the bias, but the bias also has the same general trend of the standard deviation.

For a concentration c_M of 0.1 *units*, the maximum point of the data has a 20% noise level, and the calculated average for this data set at that concentration is 60% noise. Most of the simulated response is noise, and it is not surprising that the **M** model fails with and error above 100%. By adding (**M**+**N**) or adjoining (**M**|**N**) the matrices, the effective noise is reduced by a factor of the square root of two (ca. 1.42), and as a consequence, projecting in these models reduces the error of the estimation.

Finally, the only difference between the (**M**+**N**) and the (**M**|**N**) models for this particular simulation is in the bias estimate, with the (**M**|**N**) model consistently less bias than the (**M**+**N**) model. Even though the noise averaging effect of both models is similar, the (**M**|**N**) model is an unbiased (least squares) model for both matrices simultaneously, and the result shown in the figure should be expected. In fact, the bias for the (**M**|**N**) model and the true factors model (**X**) are completely overlapped within the resolution of the plot.

Similarity of the intrinsic vectors. The effect of the similarity between the intrinsic vectors is the same for the $\{\mathbf{x}_i\}$ and the $\{\mathbf{y}_i\}$ vectors. Two component simulations using fixed $\{\mathbf{x}_i\}$ vectors and varying the $\{\mathbf{y}_i\}$ vectors were performed. Fig 2.2 shows how the SD error in the calibration increases linearly with the inverse uniqueness, $1/sin(\alpha_y)$, as defined in Eq 2.25, and showing for this case that Eqs 2.24 and 2.26 are a good approximation. The values near to the curve correspond to the correlation coefficient, calculated as $cos(\alpha_y)$. Even though the correlation coefficient has no linear relation with the error, it is a useful measure of high the degree of similarity between the vectors used for the simulation

As the similarity increases, the importance of the bias relative to the standard deviation also increases, but it does not reach the level of the SD within the limits of the

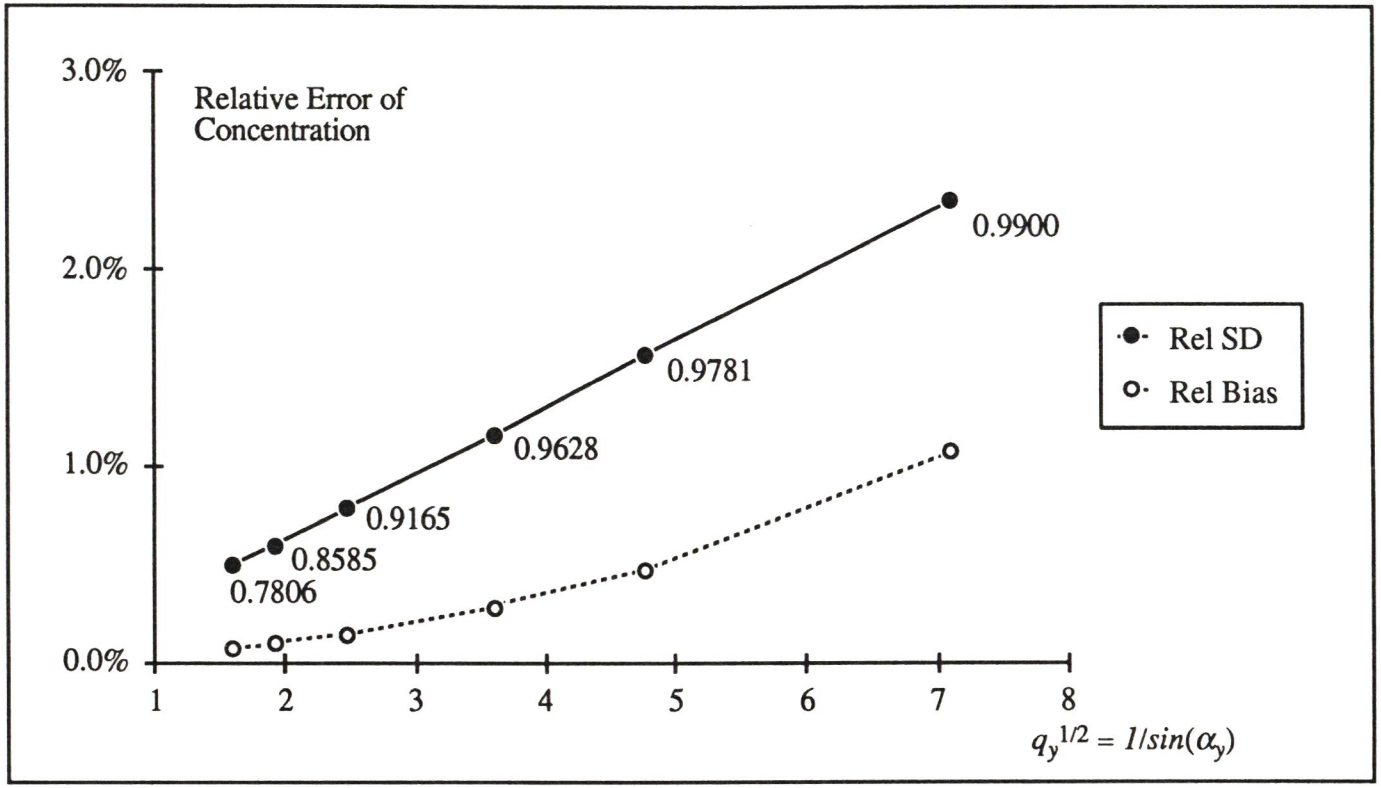

Fig 2.2 Effect of the similarity of the intrinsic vectors. The relative standard deviation (σ) and bias of the GRAM concentration results are compared for different degrees of correlation (numbers adjacent to the points). The σ increases linearly with the inverse of the uniqueness, $1/\sin(\alpha_y)$, as predicted by Eq 2.23, and the bias remains as minor contribution to the total error (**M|N** model used for calculations, based on 40 simulations per point).

simulation. If the correlation between the vectors continues increasing, the difference between the vectors will eventually reach the noise level, and the model will be totally random for the second component. It is reasonable to expect the bias contribution to the error to increase, because the quality of the model decreases. In the other hand, It is important to point out the remarkable linearity of the SD with the inverse uniqueness. It will later be seen that other effects like the number of analytes present, can also be integrated under the uniqueness coefficient.

Another important effect occurs when two or more eigenvalues (Eq 2.18) become very similar. Even though the accuracy of the eigenvalues is not affected by their similarity, the eigenvectors suffer degeneracy, and become perpendicular to each other for identical eigenvalues. Table 2.2 shows the correlations between estimated and true spectra for a two component case, as the eigenvalues get more similar. It is apparent that when the eigenvalues get as close as their SD bounds, the estimated spectra start to degrade. Therefore, comparing the SD of the eigenvalues with their difference is a useful qualitative criterion to infer the quality of the obtained spectra.

Effect of the relative concentrations. GRAM has been presented as a one point calibration technique, and when used in this way, the concentration of calibration must be as similar as possible to that of the test sample. Nevertheless, it is possible to have several calibration matrices $\mathbf{N}_1, \mathbf{N}_2 \ldots \mathbf{N}_n$, and obtain a calibration curve. But if the linear model is correct, one point calibration is safer for second order calibration than for zero order, because the background is taken into account. Fig 2.1 showed the effect of the concentration c_M on the error and the bias. For high concentrations, the relative error is almost constant, but as the concentration decreases, the errors increase dramatically. A

TABLE 2.2

EFFECT OF EIGENVALUE SIMILARITY

Δ is the difference between the two eigenvalues. When Δ is smaller than the S.D. of the eigenvalues, spectra degrade. Spectral correlation is between estimated and library spectra.

Δ	*Standard deviation of eigenvalues*		*Spectral correlation*	
	1st	2nd	1st	2nd
0.0010	0.0023	0.0044	0.9664	0.7938
0.0030	0.0028	0.0048	0.9545	0.6901
0.0100	0.0018	0.0051	0.9964	0.9919
0.0300	0.0018	0.0052	0.9993	0.9989
0.1000	0.0016	0.0052	0.9998	0.9998

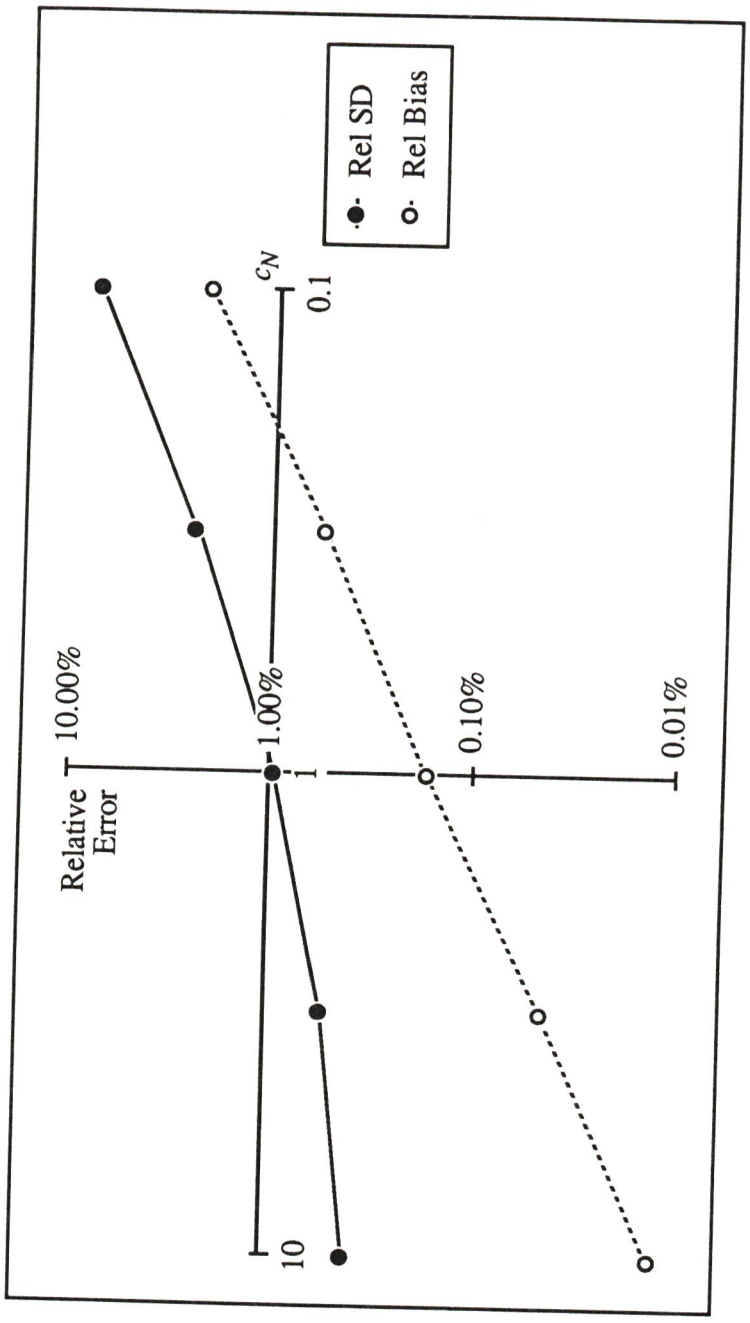

Fig 2.3 Effect of the calibration concentration. The relative standard deviation ($\hat{\sigma}$) and bias of the GRAM concentration results are compared for different concentrations in the calibration sample. There exists an aproximate linear relationship for both kinds of errors. ($c_M = 1.0$).

concentration of 0.1 represents a S/N value of 5 for that component at the maximum of its spectrum. A similar plot is shown in Fig 2.3, this time with varying concentration c_N. Two different effects work in parallel to produce the errors, again, the bias and the standard deviation. The bias is caused mostly by the inadequacy of the model, and for lower concentrations c_M, the model is worst. It is also important to observe that the error in **M** has greater impact in the bias than the error in **N**.

Effect of the number of analytes present. This problem is closely related to the effect of the similarity and the choice of the model. The more components present, the less unique they will be when compared with the subspace spanned by the rest of the components, resulting in a larger net similarity. Fig 2.4 shows the error versus the inverse uniqueness factors (based on five replicates per case), for different numbers of components, illustrating that the error varies almost linearly with the similarity, accounting for the number of components effect.

Another problem arises in model selection to decide how many principal components are used, because the greater the number of analytes, the more difficult it is to find the proper number of components. Therefore, calibration matrices with too many components should be avoided. Of course, in chromatography, where well separated clusters of components can be quantitated separately, many more components can be successfully analyzed.

In summary, there are many factors that affect the results of rank annihilation. How can the analyst use those factors to obtain the best possible quantitation? The calibration sample **N** should have concentrations close the those present in **M** but different enough to avoid similar concentration ratios (eigenvalues). The model of the

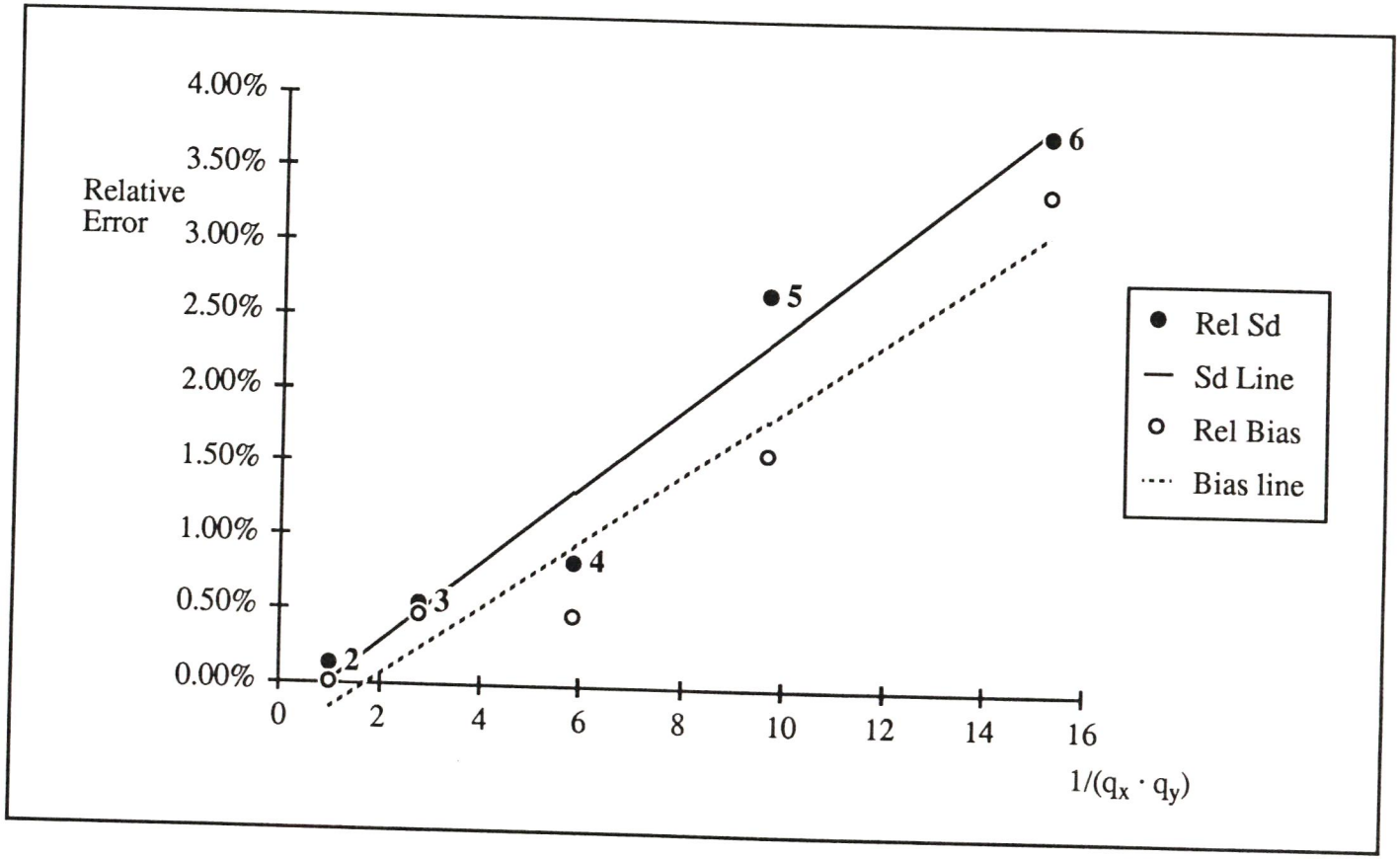

Fig 2.4 Effect of the number of analytes. The relative standard deviation (σ) and bias of the GRAM concentration results are compared for different number of analytes (2-6) present in the test sample. The error is almost linearly proportional to the inverse of the uniqueness factors, which increases as more components are present in the sample.

space should be the best unbiased estimate of the subspace spanned by both matrices, and it is suggested to use the SVD of (**M**|**N**). If both samples have components that are not present in the other, then **M**+**N** should be used instead of **M** for the estimation [Sanchez and Kowalski, 1986]. **X**, **Y** and $\mathbf{c}_M$ will be obtained from the calculation, and the uniqueness factors can be estimated for each component quantitated, to have an idea of the error in the estimation. Even better, if some of the spectra $\{\mathbf{x}_i\}$ are available, e.g., from a library of spectra, then the analyst can compare the estimated and the true spectrum, and have an idea of how good is the estimated concentration.

Application of GRAM to LC/UV Data

Data from a liquid chromatography (LC) system coupled with a diode array ultraviolet - visible (UV-Vis) detector was used to test GRAM with real measurements* . This data is two dimensional, and it will be shown that it can be factorized in the three matrices of Eq 2.9, e.g. it is bilinear.

The experiment consists in injecting the multicomponent sample in the inlet of the LC column, and collecting the UV-Vis spectra of the effluent from column at regular intervals of time, e.g., every second. Eventually, the components reach the detector cell, but because they move at different speeds through the column, their residency times (Retention Times, t_R) in the column are different. The position at a given time of an analyte in the column is not a single point, because as it moves along the column, diffusion and other processes occur that broaden its concentration band. Therefore, even if two analytes have different t_R, they may overlap at the detector, and their pure spectra may not be collected. The situation is worse when there are many coelluting components, because multiple overlaps occur, and the possibilities of identification or quantitation are greatly reduced. GRAM can resolve this problems under certain conditions.

Beer's law describes the absorption of light by homogeneous solutions. It assumes a linear relation between the concentration of a single analyte in a homogeneous solution and the absorption of light by that solution at a given wavelength, λ_o:

$$A = a\,b\,c \qquad [2.29]$$

* All experimental data courtesy of L. Scott Ramos.

Where A is the absorption of light at the wavelength λ_o, a is the absorptivity constant at that wavelength, b is the path length, and c is the concentration of the analyte in the solution. In general, Eq 2.29 can be expressed for every wavelength λ as

$$A(\lambda) = a(\lambda)\, b\, c \qquad [2.30]$$

where $a(\lambda)$ is a continuous function of the wavelength, and represents the spectrum of the analyte.

The concentration c of an analyte at the output of a chromatographic column is not a constant, but a function of time, therefore, the absorption of light is a function of both the wavelength and time:

$$A(\lambda,t) = a(\lambda)\, b\, c(t) \qquad [2.31]$$

$A(\lambda,t)$ is function of λ and t independently, i.e., it is a *bilinear function.* In practical terms, this means that the spectrum $a(\lambda)$ is the same (within the noise level) no matter what time, t, it is measured in a chromatographic peak. Similarly, the shape of the chromatographic peak is the same at every wavelength, λ.

The data collected from a liquid chromatograph with a DA-UV spectrometer as a detector is not continuous. The absorption is measured at certain wavelengths (λ_1,λ_2 ...λ_m) obtaining a absorption vector, $\mathbf{a} = (a_1,a_2 \ldots a_m)$, every scan. Similarly, a full spectrum is measured at certain times ($t_1,t_2 \ldots t_n$) to get a concentration profile vector, $\mathbf{c} = (c_1,c_2 \ldots c_n)$, for every wavelength. Then Eq 2.31 takes the form

$$A(\lambda_i,t_j) = a(\lambda_i)\, b\, c(t_j) \qquad [2.32]$$

The data can be assembled into a matrix, $\mathbf{A}$ ($m \times n$). The rows are assigned to the different wavelengths and the columns to the different scans when a spectrum is acquired. In matrix notation Eq 2.32 can be expressed as an outer product,

$$\mathbf{A} = \mathbf{a}\, b\, \mathbf{c}^T \qquad [2.33]$$

which is equivalent to Eq 2.6. Any analytical technique for which the data of a single component can be factorized like Eq 2.33, was defined as a bilinear technique,

$$\mathbf{M} = \mathbf{x}\ \beta\ \mathbf{y}^{T} \qquad [2.34]$$

where **M** is the data matrix, **x** and **y** are data vectors in the two different orders and β is the proportionality constant. For chromatography/spectroscopy combinations, the vector **x** corresponds to the normalized spectrum of the analyte (normalized **a**) and **y** corresponds to the normalized concentration profile or peak shape (normalized **c**). By defining **y** as normalized, its absolute concentration information is lost, but now the constant β is directly proportional to the concentration.

For multicomponent samples, with p components, the resultant data matrix can usually be approximated by the sum of the p individual bilinear contributions,

$$\mathbf{M} = \sum_{k=1}^{p} \mathbf{x}_k \beta_k \mathbf{y}_k^{T}. \qquad [2.35]$$

or similarly to Eq 2.9,

$$\mathbf{M} = \mathbf{X}\ \mathbf{ß}\ \mathbf{Y}^{T} \qquad [2.36]$$

where the k^{th} column of the matrix **X** (m x p) corresponds to the spectrum $\mathbf{x}_k$, and the k^{th} row of the matrix $\mathbf{Y}^{T}$ (p x n) corresponds to the chromatogram $\mathbf{y}_k^{T}$ and **ß** is a diagonal matrix with $\mathbf{ß}_{kk}=\beta_k$, that are proportional to the concentrations; m is the number of wavelengths and n is the number of scans in the chromatogram.

A problem arises when GRAM is used with LC/UV data. The problem is that the retention times may change from the calibration to the test sample, because of small fluctuations in the column characteristics. Therefore, the **Y** matrix from **M** is different from the **Y** matrix from **N**. Kim has called this problem the *synchronization error* [Kim, 1984]. If the change in the column that causes the error is not large, it can be assumed that the error exist only in the time index (j) of the **N** matrix,

$$M_{i,j} <=> N_{i,j+\Delta j} \qquad [2.37]$$

Kim assumed that the error Δj was an integer, and a simple integer iteration could find a Δj for which the projection of **N** on the subspace spanned by **M** changed **N** the least. Unfortunately, when sampling occurs at the order of *1 Hz*, in general, Δj is not an integer. Furthermore, if the change in the column is because of a small change in the flow rate, Δj is not constant, but changes linearly for every j. Appendix D describes three algorithms that have been developed in this work to correct the synchronization error, each useful under different circumstances. For each algorithm, a corrected **N** matrix is generated, which is used for the GRAM calculations.

Simulated Data

Two simulations were performed to illustrate the use of GRAM for multicomponent analysis in chromatography [see also Sanchez *et al*, 1987]. Real spectra and gaussian chromatographic peak profiles were used for the simulations. In every case, the peak width of the chromatograms was set equal to 20 seconds.

As mention earlier, the quality of the results obtained with GRAM is a function of several factors, the most important for LC/UV being noise level, number of overlapped components, similarity of the concentration ratios, similarity of the spectra and degree of chromatographic overlap (resolution).

Fig 2.5 shows the total wavelength chromatogram (TWC) of the first simulation samples. TWC is defined as the chromatogram resulting from summing the absorptions at all wavelengths for each scan. Table 2.3 presents a summary of the details of this simulation. Gaussian distributed noise was added to the matrices to simulate experimental noise as 1% of the average signal value.

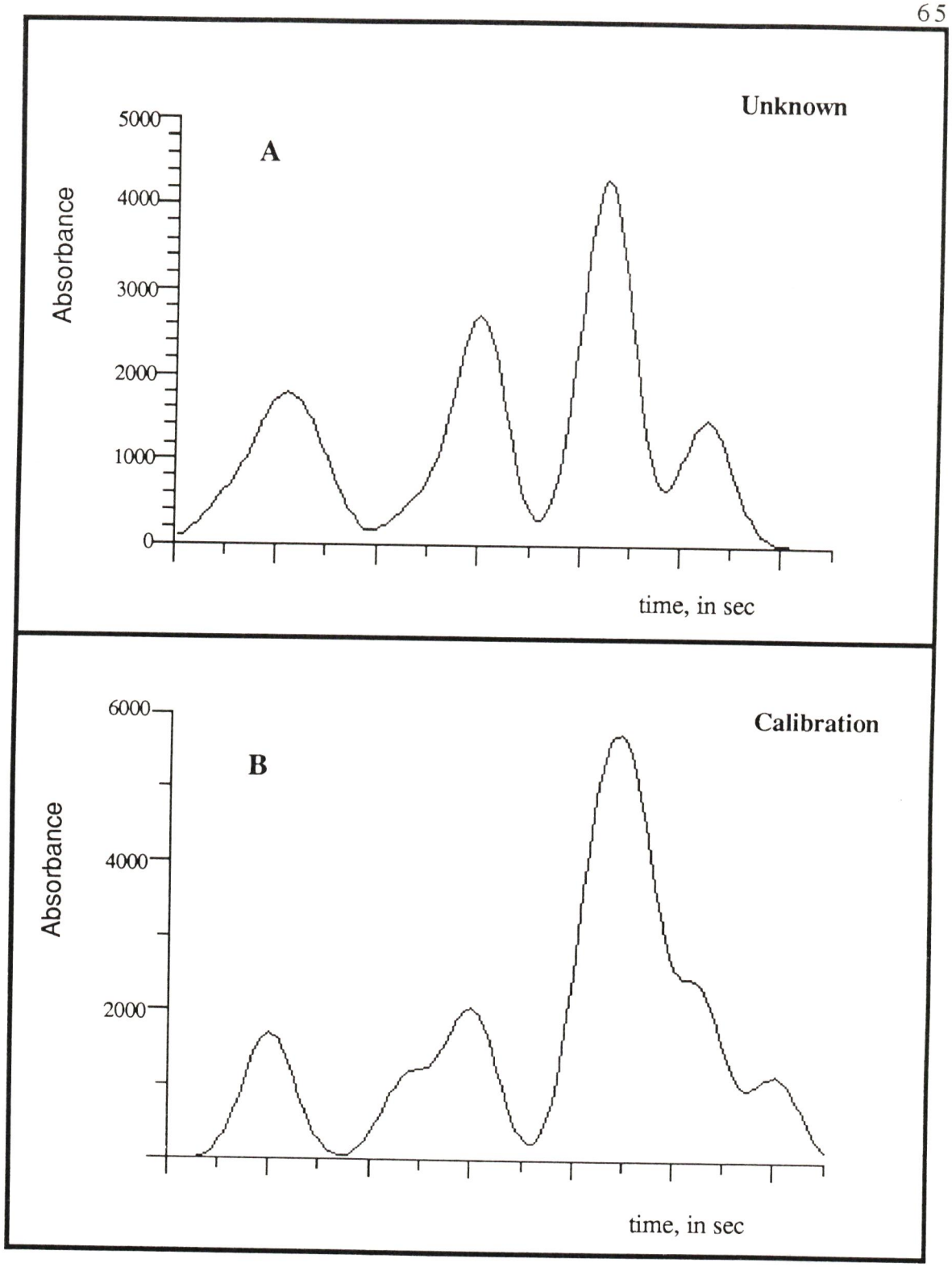

Fig 2.5 Total wavelength chromatograms (TWC) for the first simulation: (A) "unknown", or test sample; (B) calibration sample.

TABLE 2.3

MULTIPLE COMPONENT SIMULATION WITH 1% NOISE

Noise is 1% of the average absorbance. Concentrations in arbitrary units. Retention times (RT) in sec. The groups correspond to peak clusters.

*Component**	*RT*	*Input Concentrations*		*Estimated Concentration*
		Calibration	*Test*	
Acen	10.0	0.000	0.500	-
Phen	20.0	1.000	0.800	0.801
Anth	27.5	0.000	1.000	-
BaA	47.5	0.500	0.200	0.200
Chry	60.0	1.500	2.000	2.000
BbFl	85.0	0.900	1.000	1.000
BkFl	92.5	1.000	0.000	0.000
BeP	105.0	0.600	0.400	0.400
BaP	120.0	0.300	0.000	0.000

* For abbreviations see appendix E

The "unknown", or test, sample and the calibration sample have several components in common (Phen, BAA, Chry, B*b*F and BeP); two components are present only in the test sample, namely Acen and Anth; and two components are present only in the calibration sample, B*b*F and BaP. Therefore, these samples represent the most general case that a chromatographer might face in real conditions (The identities of these components are not important here, see appendix E for their names).

The data matrices were cut into three windows, corresponding to points near the valleys of the TWC. GRAM was applied to each of these windows, and the resultant resolved chromatograms of the test sample are presented in Fig 2.6.

Fig 2.6A represents the the first peak cluster, in the 0-40 sec. range. Table 2.3 shows that the calibration sample has only one component in this range, Phen, whereas, the test sample has three: Acen, Phen and Anth. Therefore, the only component that is correctly resolved is the one present in both samples, i.e., Phen. The two other curves represent linear combinations of the other two components, and could be resolved using two-component self modeling curve resolution [Lawton and Sylvestre, 1971]. This cluster illustrates the most important feature of GRAM: the ability to quantitate an unresolved component, in this case Phen, which is overlapped with other, unknown components in the test sample. The other two components are not resolved because they violate one of GRAM's requirements, i.e., the concentration ratios must be different; for both components the ratio of concentrations calibration/unknown is zero.

Fig 2.6B represents the second peak cluster, in the 40-70 sec. range. Both samples have the same two components with different concentration ratios (ca. 3/1), and both are successfully resolved in the test sample.

Finally, Fig 2.6C represents the third cluster of the test sample. This is the opposite case of the first cluster, i.e., the calibration sample has more components than

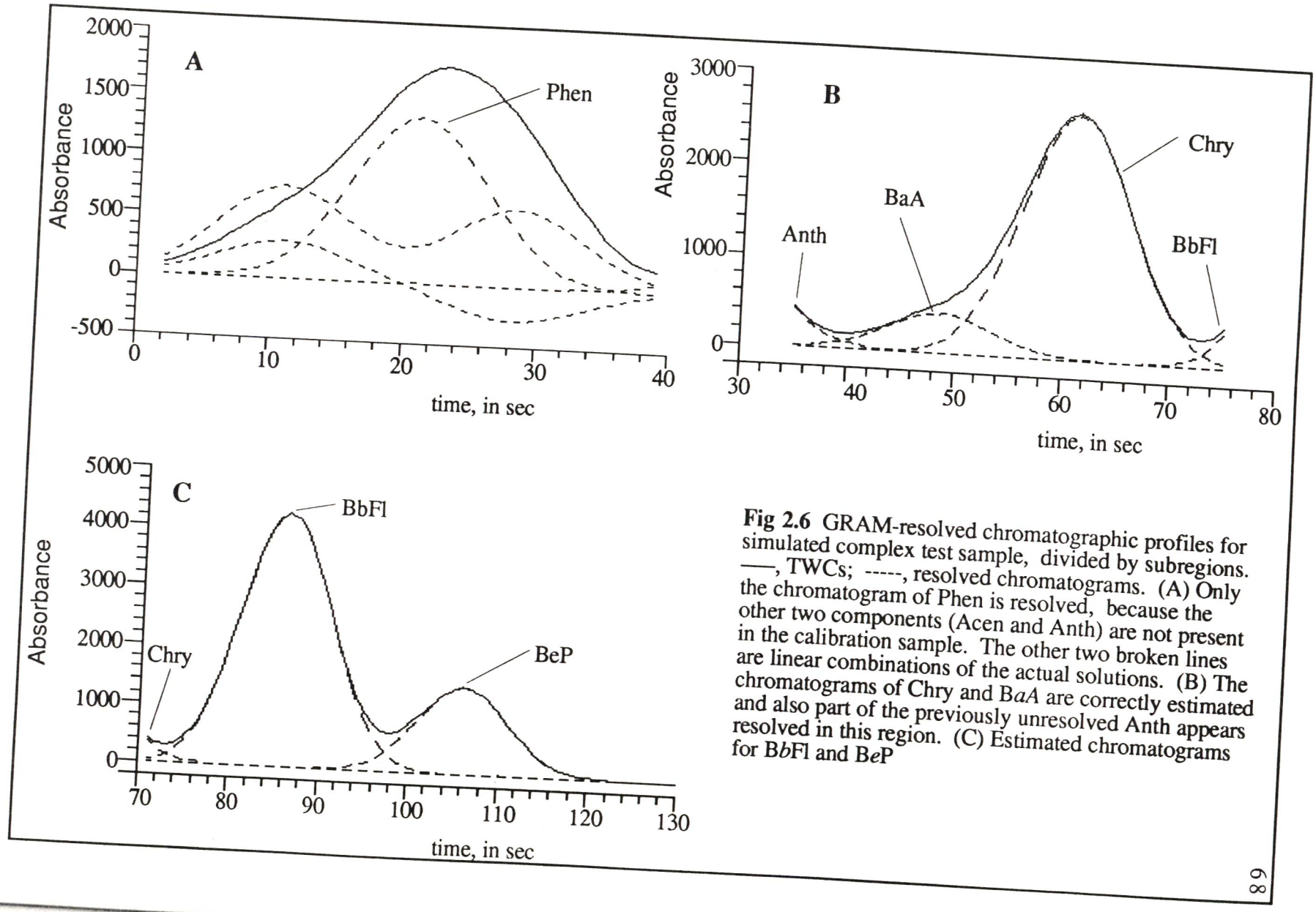

Fig 2.6 GRAM-resolved chromatographic profiles for simulated complex test sample, divided by subregions. ——, TWCs; -----, resolved chromatograms. (A) Only the chromatogram of Phen is resolved, because the other two components (Acen and Anth) are not present in the calibration sample. The other two broken lines are linear combinations of the actual solutions. (B) The chromatograms of Chry and B*aA* are correctly estimated and also part of the previously unresolved Anth appears resolved in this region. (C) Estimated chromatograms for B*b*Fl and B*e*P

the test sample. Because both components in the test sample are present in the calibration, they are correctly resolved.

Careful observation of Fig 2.6 reveals that the arbitrary division into windows occurs at points where two components overlap, and in each case, GRAM uncovered these borderline components. For example, Fig 2.6B shows four resolved components, two of which are the main Gaussian peaks, and the other two, at the beginning and end of the cluster, represent the tailing edges of components in adjacent windows.

Note also that Anth, the trailing component at the beginning of the second cluster, could not be resolved in the first cluster, but its tail portion is correctly solved in the second cluster. This can be explained with the same argument that was used to explain why it could not be resolved in the first cluster: Anth is the only component in the second cluster that has a concentration ratio (unknown/calibration) equal to zero. Therefore, it can be correctly resolved at its trailing edge, where it no longer overlaps with Acen from the first cluster, which was the other component with a zero ratio.

The second simulation was intended to test the effect of a lower signal to noise level on chromatographic data. An average 4% noise (S/N = 25) was added to a three-component simulation. The spectra used for the simulation were those of BaP, B*b*Fl and BeP. Fig 2.7 shows the GRAM-resolved chromatograms of both samples. This GRAM calculation was repeated 10 times to estimate the error in the results. Fig 2.8 shows the normalized chromatographic and spectral solutions and their respective confidence bands. Table 2.4 compares the estimated and expected concentrations and the predicted error in the results.

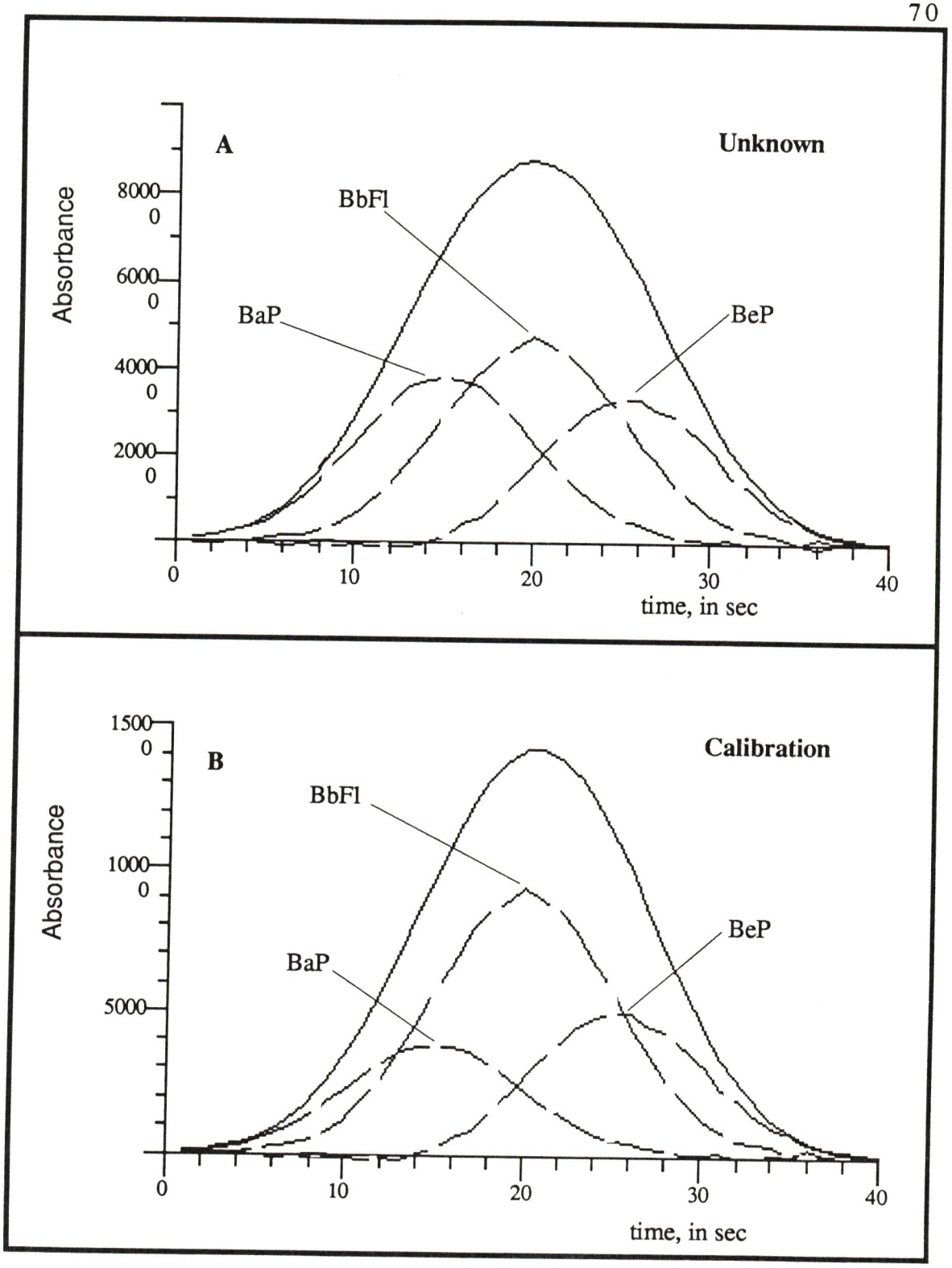

Fig 2.7 Results of the second simulation, with 4% noise: GRAM resolved concentration profiles ofr both the "unknown" and the calibration samples. ——, TWCs; -----, resolved chromatograms.

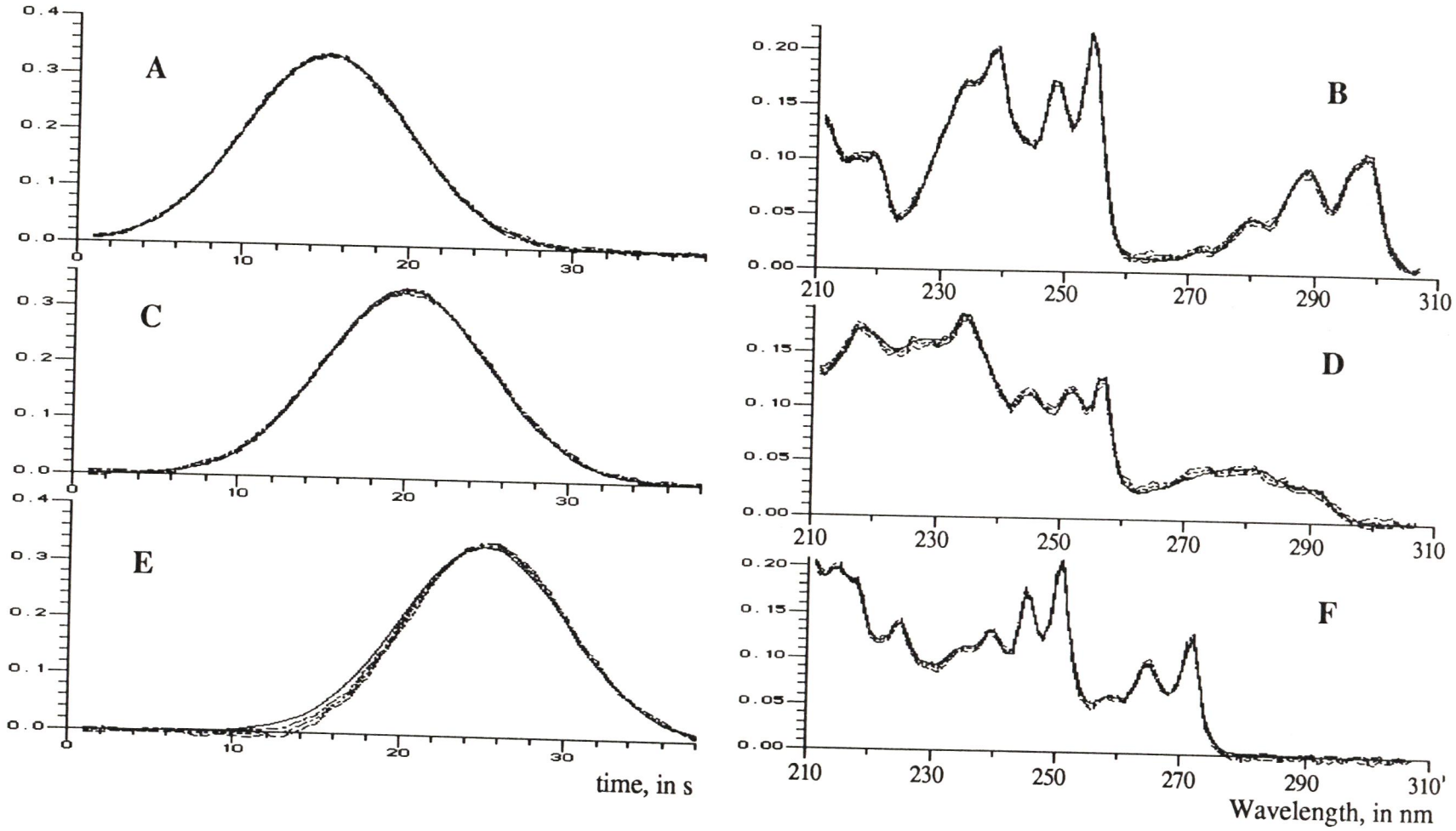

Fig 2.8 Second simulation. The individual normalized chromatograms and spectra compared with their uncertainty regions. Solid line is the expected result, broken lines represent the average estimated spectra and the confidence regions; based on ten simulations. (A) and (B) = B*a*P; (C) and (D) = B*b*Fl; (E) and (F) = B*e*P

TABLE 2.4

SIMULATION WITH 4% NOISE

Noise is 4% of the maximum absorbance in the spectrum measured at the apex of the Gaussian peak. Concentrations are relative to the calibration sample. Retention times (RT) in sec. Standard deviation based on ten calculations with the same 4% noise level.

Component	*RT*	*Expected Concentration*	*Estimated Concentration*	*Standard Deviation*
BaP	15	2.000	1.953	0.008
BbFl	20	1.000	0.999	0.001
BeP	25	1.500	1.485	0.005

Experimental Data

Three sets of samples were used to test the GRAM method with real data. All the experimental data, including the GC analysis results, were courtesy of L. Scott Ramos. The experimental and computational details of the GRAM calculations with this data are presented in appendix E [Sanchez *et al*, 1987].

The first two sets contained three components in both the calibration and the "unknown" (test) samples (both sets of samples were prepared from pure standard solutions, therefore, the test samples were actually known). Table 2.5 presents the details of these two sets of samples and compares the expected with the estimated concentrations. The samples of the same set were analyzed sequentially, under the same chromatographic conditions, to minimize changes in the relative retention times. The minimal residuals projection method (MRP) was used for the correction of the synchronization error for this case (See appendix D for details).

The resolved chromatogram of the first unknown sample is presented in Fig 2.9. This sample and its corresponding calibration sample were analyzed with a mobile phase containing 20% water in CH_3CN. The expected solutions were estimated with a multiple linear regression program (MLR) for comparison with the GRAM results; the MLR solutions do not necessarily represent the actual true solutions, but they are a good approximation of the underlying chromatographic profiles. Note that the MLR solutions and the GRAM-resolved solutions are very similar. This sample was originally thought to contain only two components, but GRAM uncovered the presence of a third component, an impurity. Target factor analysis [Malinowski and Howery, 1980; Lorber, 1984B] was used to test for the presence of several possible impurities, and only BbFl gave a positive test, and its spectrum was used for the MLR estimation. Fig 2.10 shows the GRAM-recovered spectra for this sample, together with the expected spectra. Due to the

TABLE 2.5

REAL SAMPLES

Spectral similarities are the dot product of the estimated spectrum and the real spectrum. A value of 1 indicates perfect match and a value of 0 total dissimilarity. A dash (-) indicates data not available.

Sample No.	*Component*	*Expected Concentration*	*Estimated Concentration*	*Spectral Similarity*
1	BkFl	4.1	4.5	0.9998
	PER	8.6	9.0	0.9996
	BbFl	-	-	0.8145
2	BbFl	6.0	6.4	0.9997
	BkFl	5.4	5.5	0.9997
	PER	4.1	4.1	0.9980

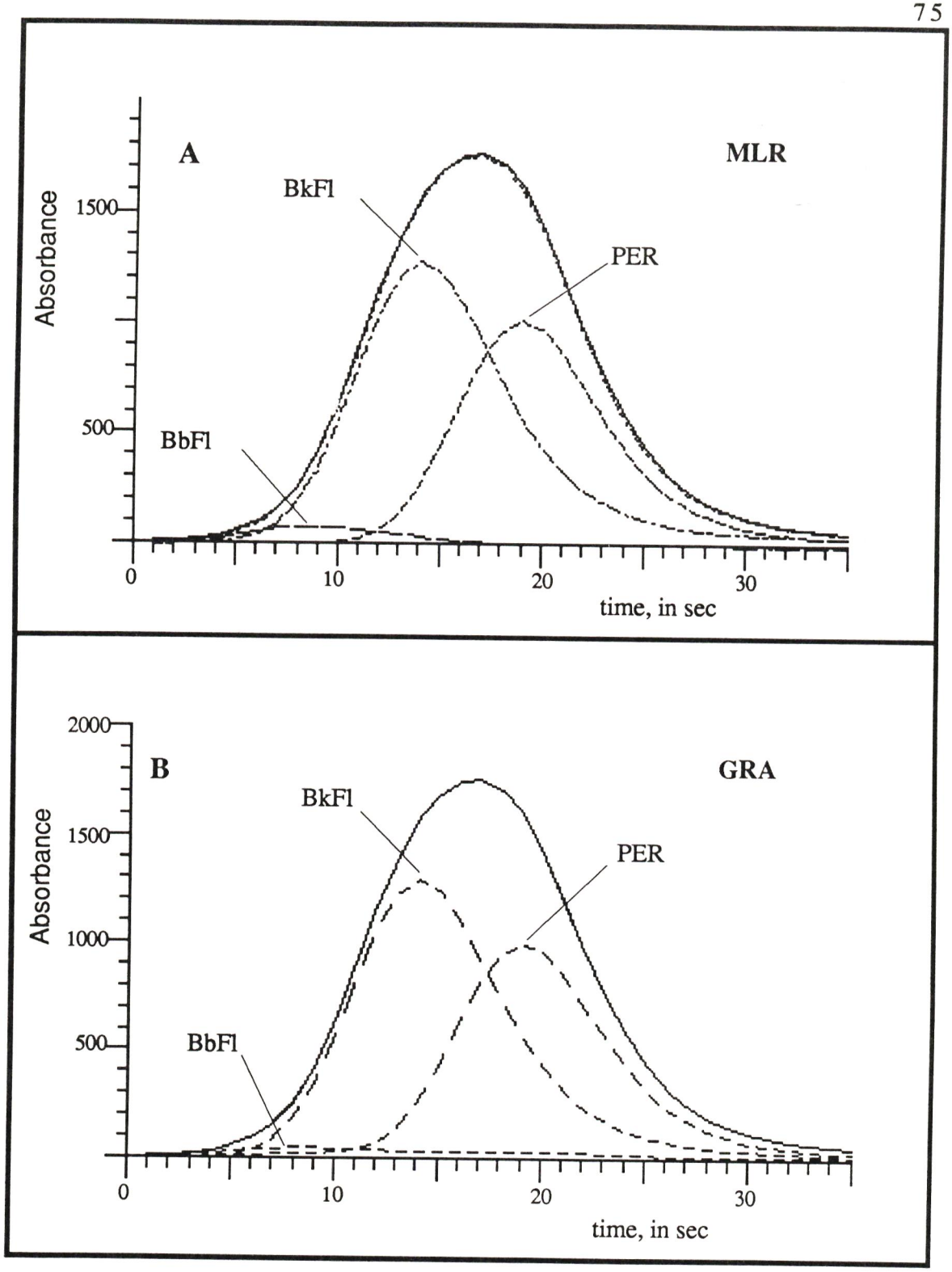

Fig 2.9 Comparison of (A) MLR and (B) GRAM estimations of concentrations profiles for the real sample 1. GRAM has uncovered the presence of an impurity. The components are B*b*Fl (impurity, t_R = 8.8 s), B*k*Fl (t_R = 14.1 s) and PER (t_R = 19.1 s).

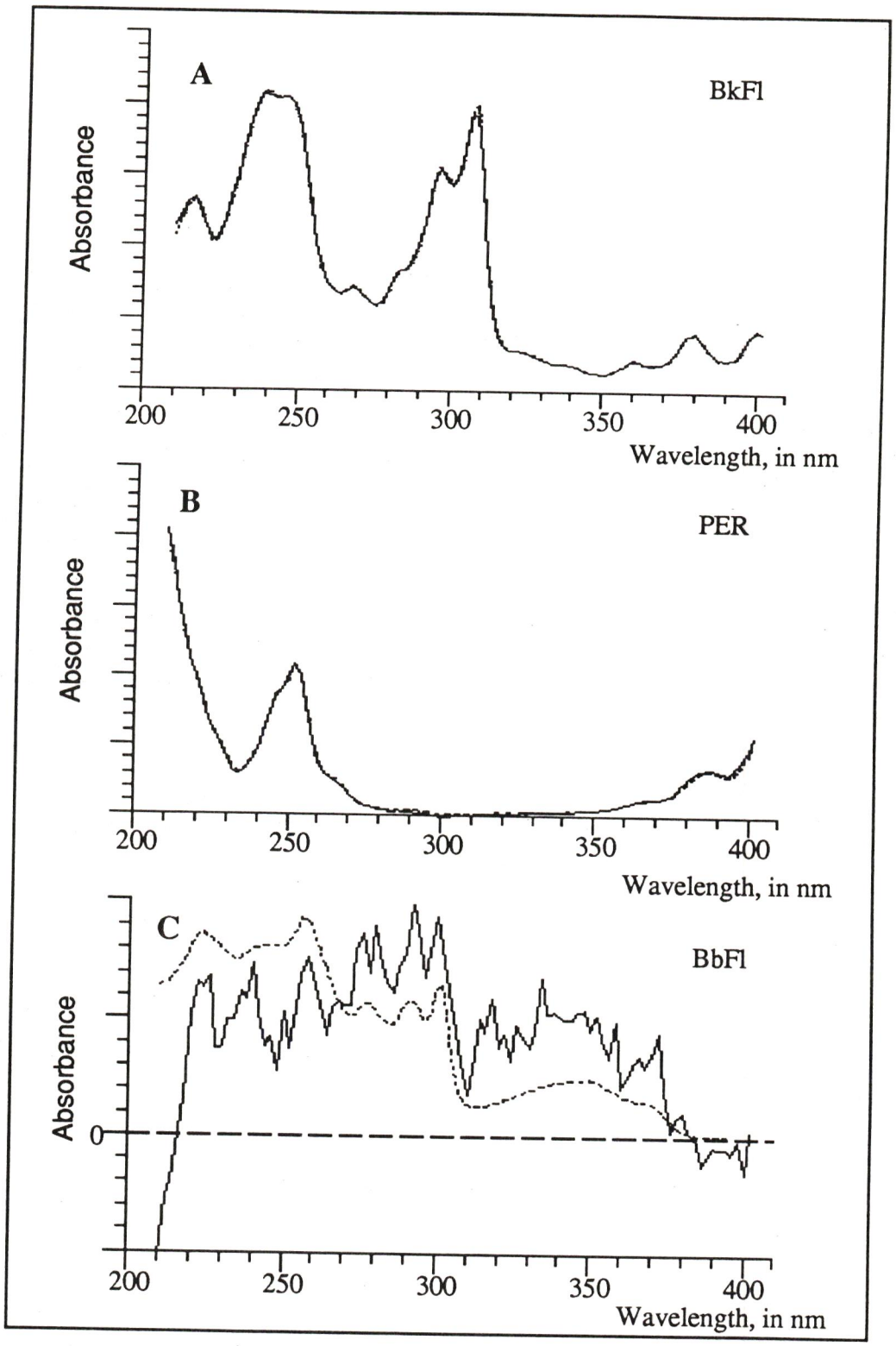

Fig 2.10 Reconstructed spectra from GRAM analysis fo sample 1. ——, predicted spectra; -----, spectra from pure standards. For (A) and (B) the predicted and the pure spectra are so similar that they can not be distinguished. (C) Impurity spectrum, plotted together with the BbFl stnadrad, the most similar in the library.

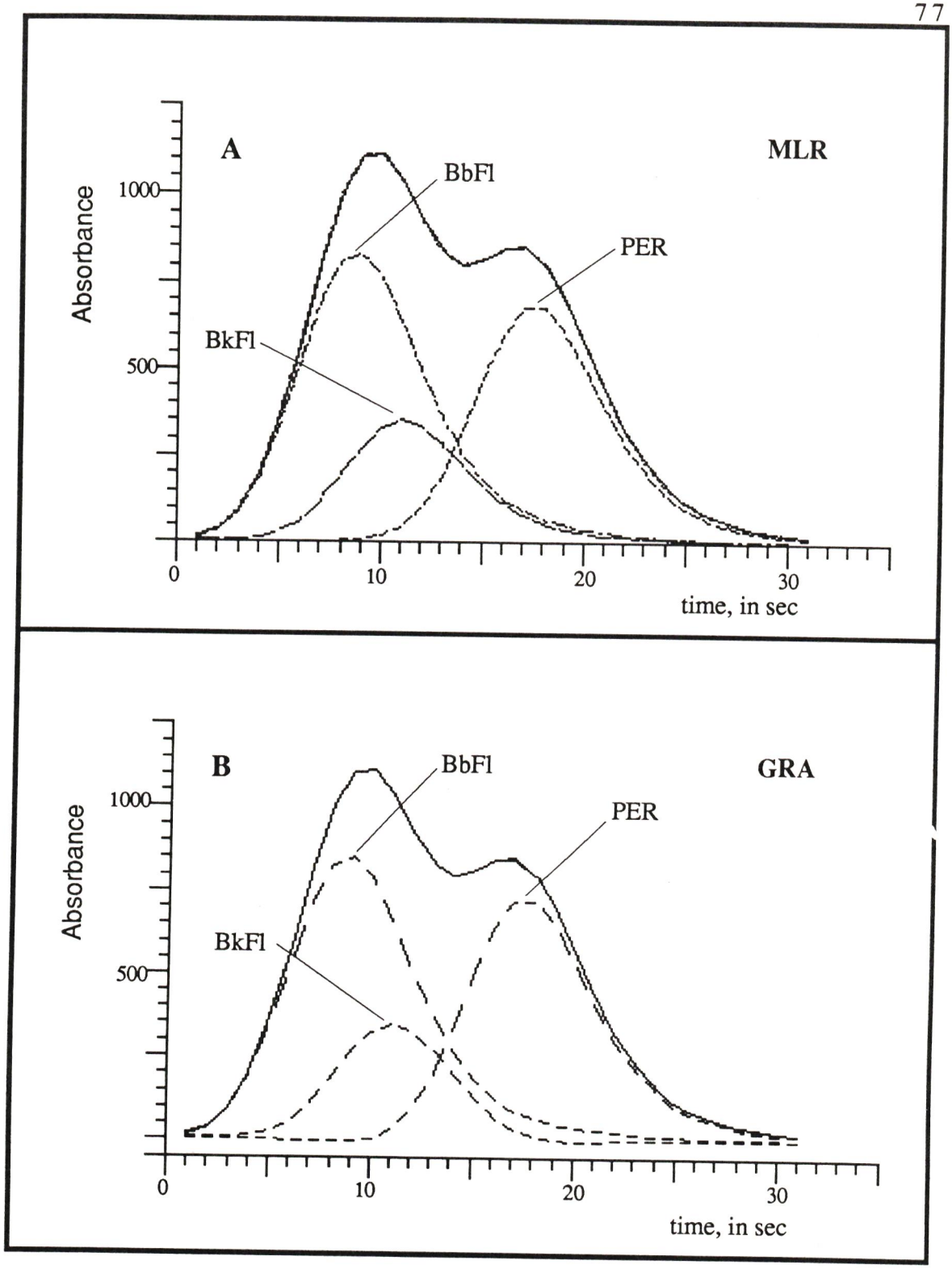

Fig 2.11 Comparison of (A) MLR and (B) GRAM estimated concentration profiles for the real sample 2. The components are B*b*Fl (t_R = 8.8 s), PER (t_R = 11.1 s) and B*k*Fl (t_R = 17.5 s).

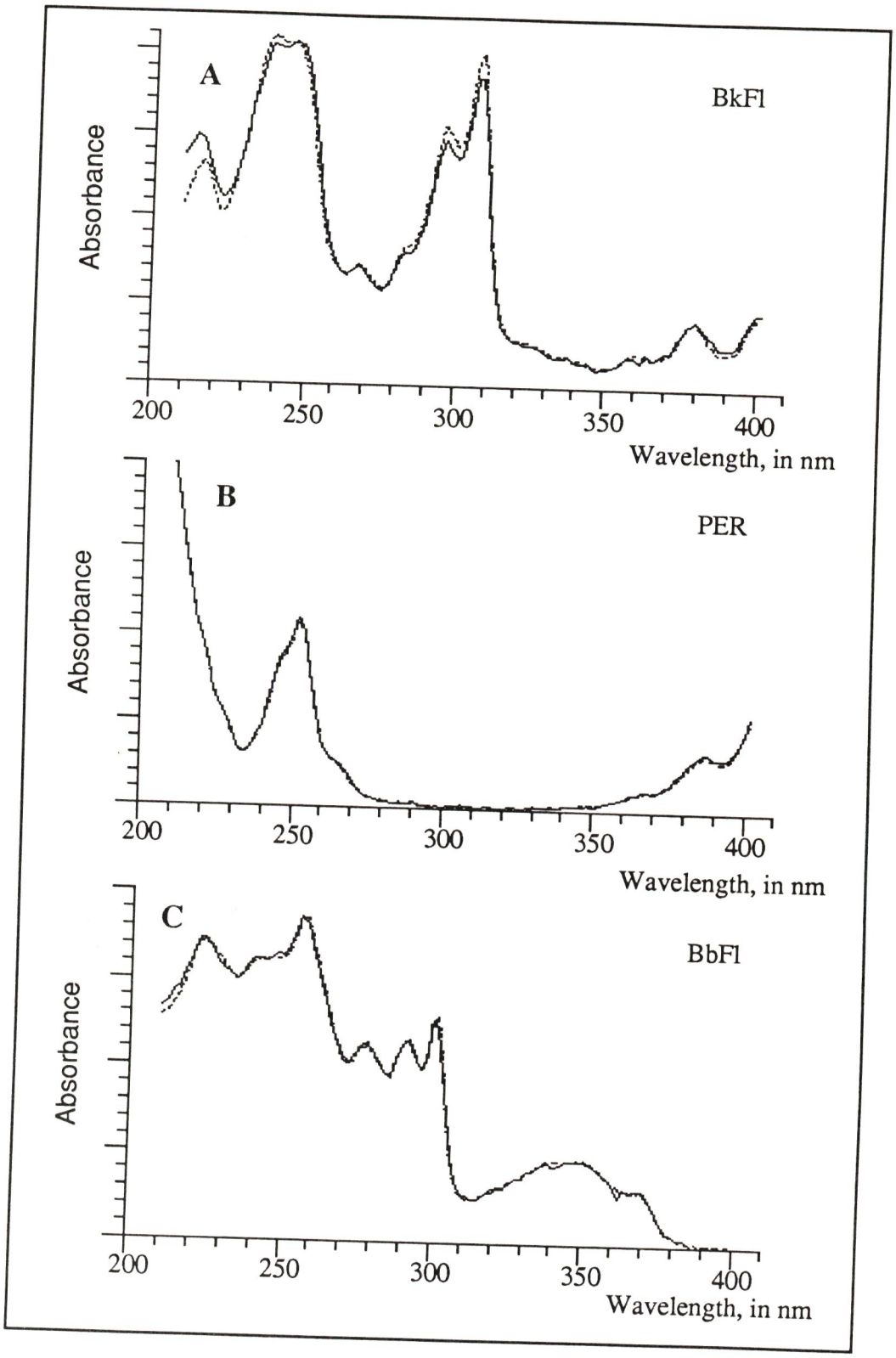

Fig 2.12 Reconstructed spectra from GRAM analysis for sample 2. ——, predicted spectra; -----, spectra from pure standards.

low intensity of the impurity signal, the noise in the recovered spectrum is too high to show clearly its identity. But the estimated spectra of the other two components match very well with their actual spectra.

The solutions for the second set of three components are presented in Fig 2.11. These samples were analyzed using 30% water in CH_3CN as the mobile phase. The expected and the estimated spectra can be compared in Fig 2.12. The resolution between BbFl and PER is very low ($R_s \approx 0.15$), but the GRAM-resolved chromatograms are a very good approximation of the MLR estimates. The best estimated spectra also correspond to BbFl and PER.

The third set was a complex unknown environmental sample, and a mixture of standards as the calibration sample. Fig 2.13 shows a portion of the TWC for the two samples. The unknown sample came from the sediment of a local river contaminated with polyaromatic hydrocarbons. Gas chromatography analysis resolved more than three hundred peaks. Due to the complexity of the sample and the availability of the pure spectra, the synchronization error was corrected using the iterative GRAM method. The calibration sample included eight polyaromatic hydrocarbons of interest, and Table 2.6 compares the GRAM LC/UV with the GC results. Fig 2.14 illustrates the resolved concentration profiles for the eight analytes and Figs. 2.15-2.16 show their resolved spectra, compared with the library spectra.

The quantitation results show that the difference between GRAM and the GC analysis is in average 15%, with the notable exception of Benzo[*a*]pyrene. The estimated spectrum of the latter also has remarkable differences with the library spectrum. Because the error in concentration ratios and the error in the eigenvectors is related, it can be used as a diagnostic tool: when the estimated spectrum is significantly different from

that of the pure analyte, the error of quantitation will also be significant. The other spectra are more accurate, and their concentrations are estimated more accurately.

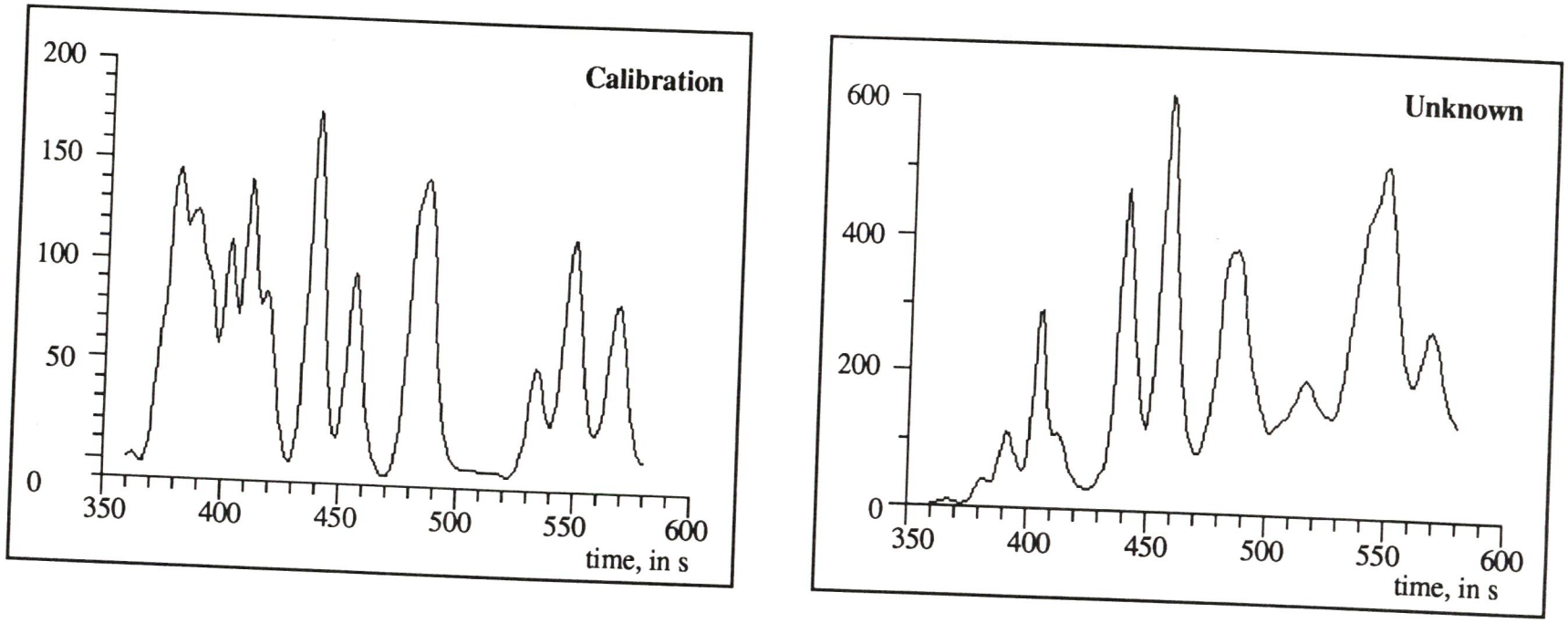

Fig 2.13 TWC for the complex enviromental sample and its corresponding calibration sample. GC analysis of this sample showed over 300 peaks. Experimental data courtesy of L. Scott Ramos.

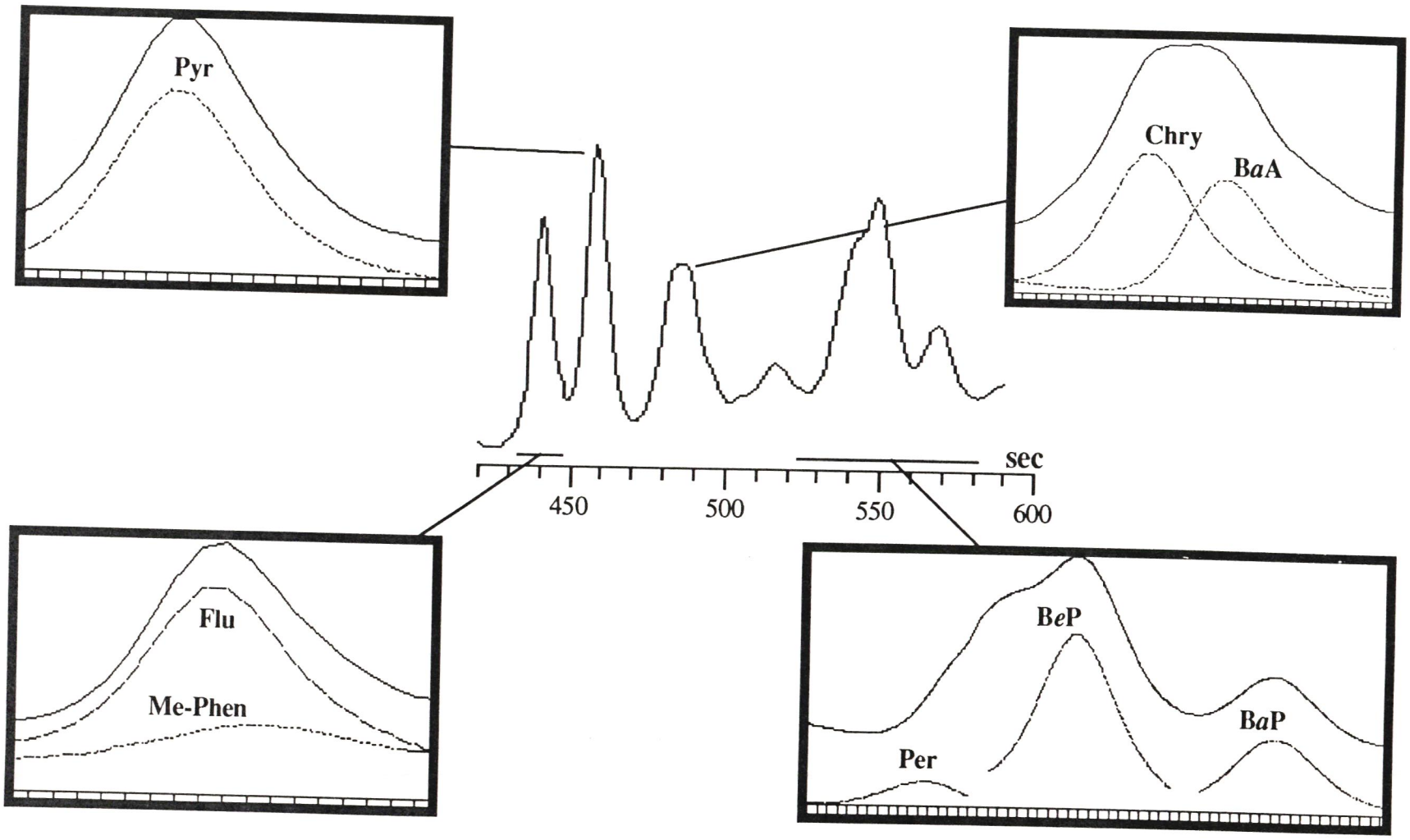

Fig 2.14 Estimated concentration profiles for some analytes in the complex enviromental sample.

TABLE 2.6

COMPLEX ENVIROMENTAL SAMPLE

Comparison of the GRAM estimation results with the Gas chromatography analysis.

Compound Name	*GC/ESTD (ng/μl)*	*LC/GRAM (ng/μl)*	*Difference (%)*
1-Methylphenanthrene	3.83	3.07	20
Fluoranthene	18.39	19.99	9
Pyrene	21.47	24.16	13
Benz[a]anthracene	8.77	7.82	11
Chrysene	10.50	9.21	12
Benzo[e]pyrene	7.63	13.35	75
Benzo[a]pyrene	7.06	9.10	29
Perylene	3.82	4.52	18

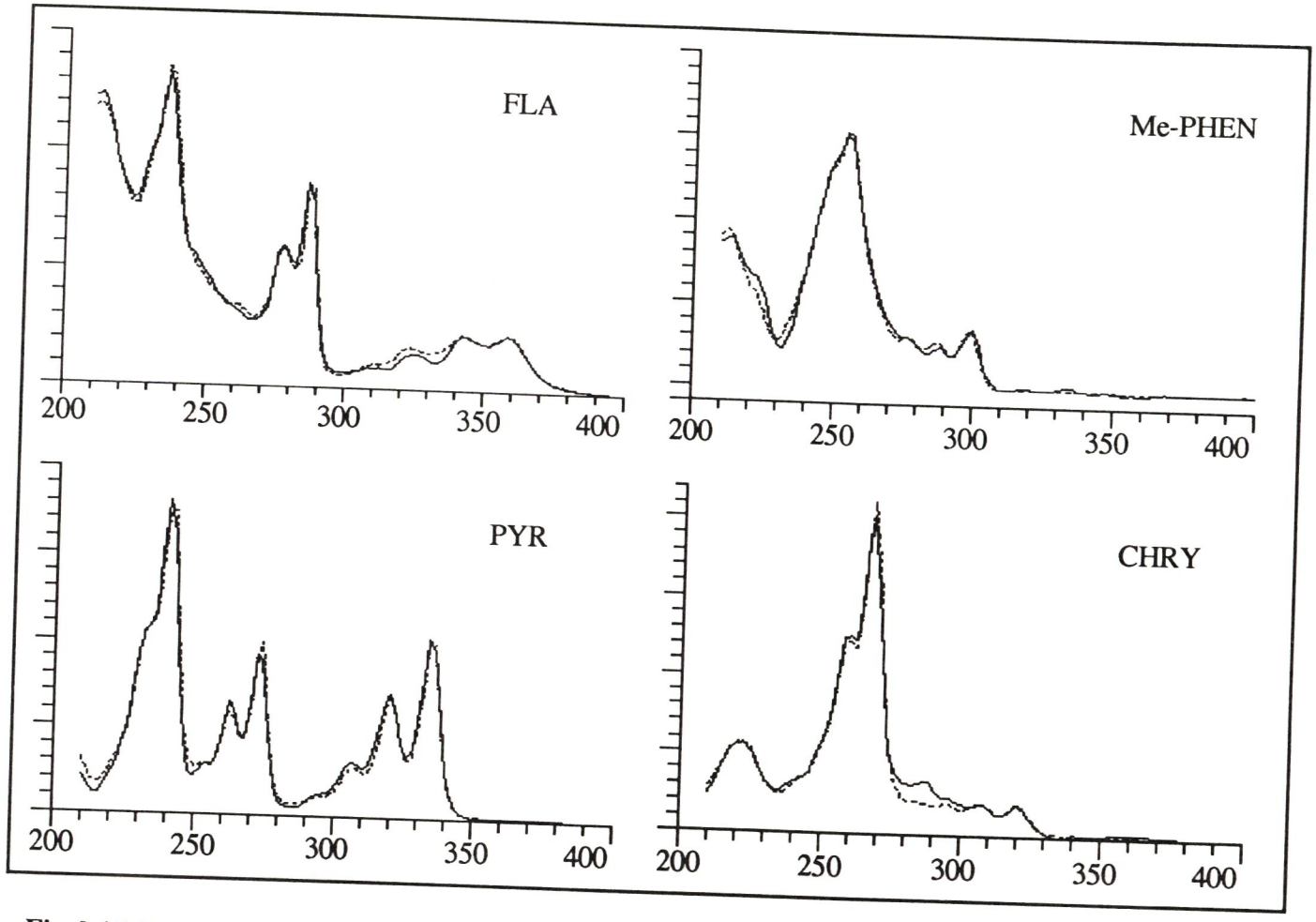

Fig 2.15 Reconstructed spectra from GRAM analysis of complex sample, first set. Solid lines are the estimated spectra and broken lines are the library spectra.

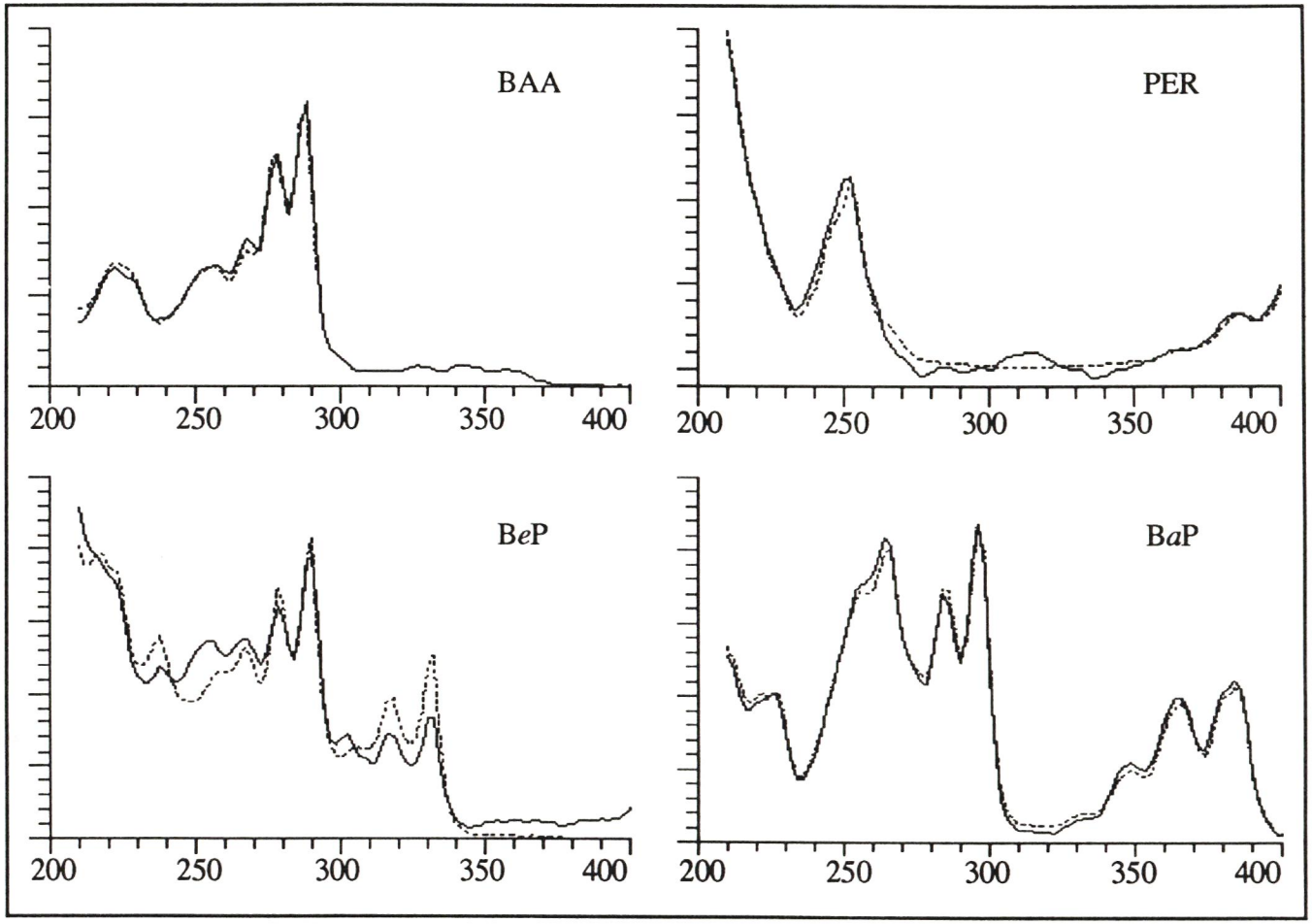

Fig 2.16 Reconstructed spectra from GRAM analysis of complex sample, second set. Solid lines are the estimated spectra and broken lines are the library spectra. Note that for B*e*P the differences are very significant, suggesting that the concentration estimates probably are not very accurate.

Second Order and beyond

The possibilities for quantitation using GRAM in LC/UV have been presented. Any other chromatography/spectroscopy combination will of course be suitable for GRAM. Combinations such us GC/MS should prove to be extremely efficient due to the capabilities of GRAM combined with the more unique MS spectra. Another area that will benefit from GRAM is fluorescence spectroscopy, by using spectral, lifetime, polarization and other measurements combined in diverse ways to generate second and higher order data [Warner *et al*, 1985].

The application of GRAM is not restricted to calibration. Whenever two samples that have some constituents in common are available, GRAM can be applied. Solvent extraction could be used to generate two samples out of a single unknown, and with a proper selection of the solvent the analytes will have different ratios of concentration. Then a bilinear instrument and GRAM can be used to extract the spectra of those analytes for identification.

The greatest potential for GRAM and in general second order methods is perhaps in future second order instruments which are yet to be built. When using a bilinear instrument with GRAM, the analyst need not worry about interferences and/or contaminants in the analysis, and simultaneous determination of several components is possible with only one calibration sample. Until recently, the main goal in analytical chemistry has been to increase the resolution more and more, without considering data in other orders, simply because no advantage was seen in doing so. That trend has started to change, and everyday more second order instruments become available, as well as the mathematical tools to handle their data.

Given the fundamental differences between first order and second order calibration, it is reasonable to wonder what happens in third order and beyond. With the knowledge of the possibilities of first order, it was impossible to predict the properties of bilinear calibration. Therefore, a similar difficulty is faced when trying to predict the possibilities of third order instruments. Appellof and Davidson have extended the rank annihilation method to third order data [Appellof and Davidson, 1981A; 1983]. They also showed that for trilinear data (analogous to bilinear for third order) it was possible to obtain the intrinsic vectors uniquely, by using a minimization algorithm and at least as many slices (matrices of data) in the third order as components were present in the mixture [Appellof and Davidson, 1981B] (GRAM can be seen as a particular case of decomposition of a trilinear matrix, with only two slices in the third order, the calibration and the test bilinear data respectively). Aparently, they did not attempt to use fewer slices in the third order in their calculations, because a unique solution exists for as few as two slices.

The problem of trilinear decomposition was studied in the early seventies by researchers in the area of psycometrics [Harshman, 1970]. It was discovered that the number of factors, n, that could be extracted uniquely from third order data of trilinear nature is related to the rank in each order by the formula:

$$n \leq (q_x + q_y + q_z - 2)/2 \qquad [2.38]$$

where q_x, q_y, and q_z are the ranks in each order as defined by Kruskal [Kruskal, 1976] and the formula sets an upper limit for the number of factors that can be extracted. For example, for $q_z = 2$ (two slices in the third order), and $q_x = q_y$, it is concluded that $n \leq q_x$. This means that the upper limit for the number of unique factors that can be extracted is equal to the rank of the first two orders, which is the result found for GRAM, where the number of extracted components has to be equal or less than the rank in each order (by definition, for bilinear data, the rank in the two orders is the same).

Similarly, Appellof and Davidson assumed that for $q = q_x = q_y = q_z$ was necessary for resolving q factors, but in fact $(3/2)q - 1$ could be solved, if $q = 10$ (e.g., a $10 \times 10 \times 10$ matrix full rank in every order), then $n \leq 14$!! A maximum of fourteen sets of intrinsic vectors can be uniquely extracted from a $10 \times 10 \times 10$ matrix. That necessarily implies that within each order, the vectors are linearly dependent. This means that even completely collinear spectra can be solved in the third order under certain conditions.

Unfortunately, the available algorithms for trilinear decomposition are based in iterative minimizations of residuals, e.g., Alternating Least Squares procedures. Convergence is not always achieved, and the more components that are present, the more difficult it is to find the right solutions [Appellof and Davidson, 1981]. This is a fundamental difference with GRAM, for which the intrinsic vectors are eigenvectors, and no iterations are necessary.

BIBLIOGRAPHY

Appellof, C. J. (1981A) "Multicomponent Mixture Data Analysis: Applications of Three-Dimensional Techniques". *Ph.D. Dissertation.* (University of Washington, Seattle, WA).

Appellof, C. J. and Davidson, E. R. (1981B) "Strategies for Analyzing Data from Video Fluorometric Monitoring of Liquid Chromatographic Effluents". *Anal. Chem.* **53**, 2053-2056.

Appellof, C. J. and Davidson, E. R. (1983) "Three-dimensional rank annihilation for multi-component determinations". *Anal. Chim.* Acta **146**, 9-14.

Brown, P. J. (1982) "Multivariate Calibration". *J. R. Statist. Soc. B* **44**, 287-321.

Budiansky, B. (1974) "Tensors". In: Pearson, C. E., Ed. *Handbook of Applied Mathematics* (Van Nostrand Reinhold Company, New York) ch. 4, 179-225.

Burns, D. H.; Callis, J. B. and Christian, G. D. (1986) "Robust Method for Quantitative Analysis of Two-Dimensional (Chromatographic / Spectral) Data Sets". *Anal. Chem.* **58**, 1415-1420.

Carey, W. P.; Beebe, K. R.; Sanchez, E.; Geladi, P. and Kowalski, B. R. (1986) "Chemometric Analysis of Multisensor Arrays". *Sensors and Actuators* **9**, 223-234.

Dempster, A. P.; Schatzoff, M. and Wermuth, N. (1977) "A Simulation Study of Alternatives to Ordinary Least Squares". *J. Am. Stat. Ass.* **72**, 77-91.

Eaton, M. L. (1983) *Multivariate Statistics. A Vector Space Approach* (John Wiley & Sons, Inc., New York, N.Y.)

Eastment, H. T. and Krzanowski, W. J. (1982) "Cross-Validatory Choice of the Number of Components From a Principal Component Analysis". *Technometrics* **24**, 73-77.

Friedman, J. H. (1986) "Beyond the Linear Model". *III C.A.C Meeting of the Chemometrics Society*, (Lerici, Italy).

Friedman, J. H. and Stuetzle, W. (1981) "Projection Pursuit Regression". *J. Amer. Statist. Ass.* **76**, pgs 817-823.

Garden, J. S.; Mitchell, D. G. and Mills, W. N. (1980) "Nonconstant Variance Regression Techniques for Calibration-Curve-Based Analysis". *Anal. Chem.* **52**, 2310-2315.

Geladi, P. and Kowalski, B. R. (1986) "Partial Least-Squares Regression: A Tutorial". *Anal. Chim. Acta* **185**, 1-17.

Gianelli, M. L.; Burns, D. H.; Callis, J. B.; Christian, G. D. and Andersen, N. H. (1983) "Multichannel imaging spectrophotometer for direct analysis of mixtures on thin-layer chromatography plates". *Anal. Chem.* **55**, 1858-1862.

Harshman, R. A. (1970) "Foundations of the PARAFAC Procedure: Models and Conditions for an 'Explanatory' Multi-Modal Factor Analysis". *UCLA Working Papers in Phonetics* **16**, 1-84.

Hirschfeld, T. (1980) "The Hyphenated Methods". *Anal. Chem.* **52**, 297A-312A.

Hirschfeld, T. (1985) "Instrumentation in the Next Decade". *Science* **230**, 230-291.

Ho, C-N.; Christian, G.D. and Davidson, E.R. (1978) "Application of the method of rank annihilation to quantitative analyses of multicomponent fluorescence data from the video fluorometer". *Anal. Chem.* **50**, 1108-1113.

Ho, C-N.; Christian, G. D. and Davidson, E. R. (1980) "Application of the method of rank annihilation to fluorescent multicomponent mixtures of polynuclear aromatic hydrocarbons". *Anal. Chem.* **52**, 1071-1079.

Ho, C-N.; Christian, G. D. and Davidson, E. R. (1981) "Simultaneous multicomponent rank annihilation and applications to multicomponent fluorescent data acquired by the video fluorometer". *Anal. Chem.* **53**, 92-98.

Honigs, D. E.; Hieftje, G. M.; Mark, H. L.; Hirschfeld, T. B. (1985) "Unique-sample selection via near-infrared spectral subtraction". *Anal. Chem.* **57**, 2299-2303.

Johnson, D. W.; Callis, J. B. and Christian, G. D. (1977) "Rapid Scanning Fluorescence Spectroscopy". *Anal. Chem.* **49**, 747A-757A.

Johnson, J. V. and Yost, R. A. (1985) "Tandem Mass Spectrometry for Trace Analysis". *Anal. Chem.* **57**, 758A-768A.

Jolliffe, I. T. (1982) "A Note on the Use of Principal Components in Regression". *Appl. Statist.* **31**, 300-303.

Kim, R. R. (1985) "Matrix Algorithms for Bilinear Estimation Problems in Chemometrics". *Ph.D. Dissertation.* (Massachusetts Institute of Technology, Cambridge, MA).

Kruskal, J. B. (1976) "More Factors than Subjects, Tests and Treatments: an Indeterminacy Theorem for Canonical Decomposition and Individual Differences Scaling". *Psychometrika* **41**, 281-293.

Lawson, C.L. and Hanson, R.J. (1974) *Solving Least Squares Problems* (Prentice-Hall, Inc., Englewood Cliffs, N.J.).

Lawton, W. H. and Sylvestre, E. A. (1971) "Self Modelling Curve Resolution". *Technometrics* **13**, 617-633.

Lorber, A. (1984A) "Quantifying chemical composition from two-dimensional data arrays". *Anal. Chim. Acta* **164**, 293-297.

Lorber, A. (1984B) "Validation of hypothesis on a data matrix by target factor analysis". *Anal. Chem.* **56**, 1004-1010, 1984.

Lorber, A. (1985) "Features of quantifying chemical composition from two-dimensional data array by the rank annihilation factor analysis method". *Anal. Chem.* **57**, 2395-2397.

Lorber, A. (1986) "Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations". *Anal. Chem.* **58**, 1167-1172.

Lorber, A. (1987) "Error propagation in quantifying chemical composition from two dimensional data array by the rank annihilation factor analysis method". *In Preparation.*

Lorber, A.; Wangen, L. E. and Kowalski, B. R. (1987) "A Theoretical Foundation for the PLS Algorithm". *J. Chemometrics* **1**, in press.

Malinowski, E. R. and Howery, D. G. (1980) *Factor Analysis in Chemistry* (Wiley, New York).

Mandel, J. (1982) "Use of the Singular Value Decomposition in Regression Analysis". *Am. Stat.* **36**, 15-24.

Mandel, J. (1985) "The Regression Analysis of Collinear Data". *J. Research National Bureau Standards.* **90**, 465-478.

Mandel, J. and Linnig (1957) "Study of Accuracy in Chemical Analysis Using Linear Calibration Curves". *Anal. Chem.* **29**, 743-749.

Maris, M. A.; Brown, C. W. and Lavery, D. S. (1983) "Nonlinear Multicomponent Analysis by Infrared Spectrophotometry". *Anal. Chem.* **55**, 1694-1703.

Martens, H. and Naes, T. (1984) "Multivariate Calibration. I. Concepts and Distinctions". *TrAC, Trends Anal. Chem.* **3**, 204-210.

McCue, M. and Malinowski, E.R. (1983) "Rank annihilation factor analysis of unresolved LC peaks". *J. Chromatogr. Sci.* **21**, 229-234.

Naes, T. and Martens, H. (1985) "Comparison of Prediction Methods for Multicollinear Data". *Commun. Statistics.- Simula. Computa.* **14**, 545-576.

Osten, D. W. and Kowalski, B. R. (1985) "Background Detection and Correction in Multicomponent Analysis". *Anal. Chem.* **57**, 908-917.

Osten, D. W. and Kowalski, B. R. (1984) "Multivariate Curve Resolution in Liquid Chromatography". *Anal. Chem.* **56**, 991-995.

Ramos, L. S.; Beebe, K. R.; Carey, W. P.; Sanchez M. E.; Erickson, B. C.; Wilson, B. E.; Wangen, L. E. and Kowalski, B. R. (1986) "Chemometrics". *Anal. Chem.* **58**, 294R-315R.

Sanchez, E. and Kowalski, B. R. (1986) "Generalized Rank Annihilation Factor Analysis". *Anal. Chem.* **58**, 496-499.

Sanchez, E.; Ramos, L. S. and Kowalski, B. R. (1987) "Generalized Rank Annihilation Method. I. Application to Liquid Chromatography / Diode Array UV Data". *J. Chromatogr.* **385**, 151-164.

Saxberg, B. E. H. and Kowalski, B. R. (1979) "The Generalized Standard Addition Method". *Anal. Chem.* **51**, 1031-1038

Tibshirani, R. (1984) "Local Likelihood Estimation". *Technical Report No. 4*, Dept. of Statistics, Standford University, September 1984.

Warner, I. M.; Callis, J. B.; Davidson, E. R.; Gouterman, M. and Christian, G. D. (1975) "Fluorescence Analysis: a new approach". *Anal. Lett.* **8**, 665-681.

Warner, I. M.; Patonay, G. and Thomas, M. P. (1985) "Multidimensional Luminescence Measurements". *Anal. Chem.* **57**, 463A-483A.

Wold, S. (1978) "Cross-Validatory estimation of the Number of Components in Factor and Principal Component Models". *Technometrics* **20**, 397-405.

Wold, S.; Geladi, P.; Esbensen, K. and Öhman, J. (1986) "Principal Components- and PLS-Analyses generalized to multi-way (multi-order) data arrays". *J. Chemometrics* **1**, in press.

Wold, S.; Ruhe, A.; Wold, H. and Dunn III, W. J. (1984). *SIAM J. Stat. Comput.* **5**, 735.

Zemroch, P.J. (1986) "Cluster Analysis as an Experimental Design Generator, with Application to Gasoline Blending Experiments". *Technometrics* **28**, 39-49.

APPENDIX A

Tensors

This appendix is an introductory description of tensors that has been adapted from Bernard Budiansky's excellent presentation [Budiansky, 1974]. Assuming familiarity with vectorial concepts, they are presented in the context of tensor theory for multidimensional spaces. No attempt has been made to expose a rigorous presentation of the subject, which can be found elsewhere in the literature.

Vectors in Multidimensional Space

Orthonormal Base Vectors Let $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ be a set of n mutually perpendicular vectors of unit length. These *orthonormal* vectors will be used as *base vectors* in the definition of the Cartesian components of arbitrary vectors. Therefore,

$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij} \qquad [A.1]$$

where δ_{ij} is the Kronecker delta.

Cartesian Components of Vectors The Cartesian components F_i of a vector $\mathbf{F}$, referred to the orthonormal base vectors $\mathbf{e}_i$, are defined, equivalently, by either the projection formula

$$F_i = \mathbf{F} \cdot \mathbf{e}_i \qquad i = 1, 2, \ldots, n \qquad [A.2]$$

or the composition formula

$$\mathbf{F} = F_i \, \mathbf{e}_i \qquad [A.3]$$

where a summation convention has been used that dictates that when two indexes are repeated at the same side of an equation, it implies that the indexed variables are summed over the range of the indexes. Therefore, Eq A.3 is equivalent to

$$\mathbf{F} = \sum_{i=1}^{n} F_i \, \mathbf{e}_i$$

General Base Vectors The components of a vector do not have to be defined with respect to orthonormal vectors. Let $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, ..., \boldsymbol{\varepsilon}_n$ be any n linearly independent vectors that will be our general base vectors. Eq A.1 no longer holds, but generalizes to the form

$$\boldsymbol{\varepsilon}_i \cdot \boldsymbol{\varepsilon}_j = g_{ij} \qquad [A.4]$$

where g_{ij} is called the metric tensor.

General Components of Vectors One natural way to define the components of an arbitrary vector **F** with respect to the general base vectors is to follow Eq A.2 and write

$$F_i \equiv \mathbf{F} \cdot \boldsymbol{\varepsilon}_i \qquad i = 1, 2, ..., n . \qquad [A.5]$$

A different kind of component arises from a generalization of the composition formula (Eq A.3), requiring a different notation

$$\mathbf{F} = F^i \, \boldsymbol{\varepsilon}_i \qquad [A.6]$$

the F_i are called the *covariant* components of **F**, and the F^i are the *contravariant* components. Cartesian components are both covariant and contravariant.

For any given set of general base vectors, the two kinds of components can be related with the help of the metric tensor g_{ij}. Substituting A.6 and A.4 into A.5 yields

$$F_i = g_{ij} F^j \qquad [A.7]$$

These relations may be inverted. Use g^{ij} to denote the $(i,j)^{\text{th}}$ element of the inverse of the matrix [g]; then

$$g^{ip} \, g_{pj} = \delta^i_j \qquad [A.8]$$

where the indices in the Kronecker delta have been placed in superscript and subscript position to conform to their placement on the left-hand side of the equation. Accordingly

$$F^i = g^{ij} F_j \qquad [A.9]$$

The metric tensor g_{ij} can be used to introduce an auxiliary set of base vectors $\boldsymbol{\varepsilon}^i$ ($i = 1, 2, ..., n$) by means of the definition

$$\boldsymbol{\varepsilon}^i = g^{ij} \boldsymbol{\varepsilon}_j \qquad [A.10]$$

The $\boldsymbol{\varepsilon}^i$ and $\boldsymbol{\varepsilon}_j$ are called, respectively, the covariant and contravariant base vectors, and the following relations are readily established:

$$\boldsymbol{\varepsilon}_j = g_{ij} \boldsymbol{\varepsilon}^j \qquad [A.11]$$

$$\boldsymbol{\varepsilon}^i \cdot \boldsymbol{\varepsilon}_j = \delta^i_j \qquad [A.12]$$

$$\boldsymbol{\varepsilon}^i \cdot \boldsymbol{\varepsilon}^j = g^{ij} \qquad [A.13]$$

$$\mathbf{F} = F_i \boldsymbol{\varepsilon}^i \qquad [A.14]$$

$$F^j = \mathbf{F} \cdot \boldsymbol{\varepsilon}^j \qquad [A.15]$$

Equations A.14 and A.15 show the use of the earlier roles of projection and composition in the definition of covariant and contravariant components of a vector.

Consider, finally, the question of calculating the new components of a vector with respect to a new base vectors $\underline{\boldsymbol{\varepsilon}}_i$. A direct calculation gives

$$\underline{F}_j = (F^i \boldsymbol{\varepsilon}_i) \cdot \underline{\boldsymbol{\varepsilon}}_i = F^i (\boldsymbol{\varepsilon}_i \cdot \underline{\boldsymbol{\varepsilon}}_j) \qquad [A.16]$$

We also find easily that

$$\underline{F}_j = F_i (\boldsymbol{\varepsilon}^i \cdot \underline{\boldsymbol{\varepsilon}}_j) \qquad [A.17]$$

$$\underline{F}^j = F^i (\boldsymbol{\varepsilon}_i \cdot \underline{\boldsymbol{\varepsilon}}^j) = F_i (\boldsymbol{\varepsilon}^i \cdot \underline{\boldsymbol{\varepsilon}}^j) \qquad [A.18]$$

Tensors in Multidimensional Space

Dyads, Dyadics and Second-Order Tensors The mathematical object denoted by **ab**, where **a** and **b** are given vectors, is called *dyad*. The meaning of **ab** is simply that the operation

$$(\mathbf{ab}) \cdot \mathbf{v}$$

where $\mathbf{v}$ is any vector, is understood to produce the vector

$$\mathbf{a}\,(\mathbf{b} \cdot \mathbf{v})$$

A sum of dyads, of the form

$$\mathbf{T} = \mathbf{ab} + \mathbf{cd} + \mathbf{ef} + \cdots$$

is called a *dyadic*, and this just means that

$$\mathbf{T} \cdot \mathbf{v} = \mathbf{a}\,(\mathbf{b} \cdot \mathbf{v}) + \mathbf{c}\,(\mathbf{d} \cdot \mathbf{v}) + \mathbf{e}\,(\mathbf{f} \cdot \mathbf{v}) + \cdots$$

Any dyadic can be expressed in terms of an arbitrary set of general base vectors $\boldsymbol{\varepsilon}_i$; since

$$\mathbf{a} = a^i\,\boldsymbol{\varepsilon}_i\,, \quad \mathbf{b} = b^i\,\boldsymbol{\varepsilon}_i\,, \quad \mathbf{c} = c^i\,\boldsymbol{\varepsilon}_i\,, \quad \ldots$$

it follows that

$$\mathbf{T} = a^i b^j\,\boldsymbol{\varepsilon}_i\,\boldsymbol{\varepsilon}_j + c^i d^j\,\boldsymbol{\varepsilon}_i\,\boldsymbol{\varepsilon}_j + e^i f^j\,\boldsymbol{\varepsilon}_i\,\boldsymbol{\varepsilon}_j + \cdots$$

Hence, $\mathbf{T}$ can always be written in the form

$$\mathbf{T} = T^{ij}\,\boldsymbol{\varepsilon}_i\,\boldsymbol{\varepsilon}_j \qquad [A.19]$$

in terms of the nxn numbers T^{ij}.

A dyadic is the same as a *second-order tensor*, and the T^{ij} are called the *contravariant components* of the tensor. These components depend, of course, on the particular choice of base vectors. The basic meaning of $\mathbf{T}$ should be re-emphasized by noting that, for all vectors $\mathbf{v}$

$$\mathbf{T} \cdot \mathbf{v} \equiv T^{ij}\,\boldsymbol{\varepsilon}_i\,(\boldsymbol{\varepsilon}_j \cdot \mathbf{v})$$

$$= (T^{ij}\,v_j)\,\boldsymbol{\varepsilon}_i$$

Thus, if v_j is the j^{th} covariant component of a vector, then $(T^{ij}\,v_j)$ is the i^{th} contravariant component of another vector. Similarly, we can define the operation

$$\mathbf{v} \cdot \mathbf{T} \equiv T^{ij}\,(\mathbf{v} \cdot \boldsymbol{\varepsilon}_i\,)\,\boldsymbol{\varepsilon}_j$$

$$= (T^{ij}\,v_i)\,\boldsymbol{\varepsilon}_j$$

which produces yet another vector having $T^{ij}\,v_i$ as its contravariant components.

By introducing the contravariant base vectors $\boldsymbol{\varepsilon}^i$ we can define other kinds of components of the tensor **T**. Thus, substituting

$$\boldsymbol{\varepsilon}_i = g_{ip}\,\boldsymbol{\varepsilon}^p \;,\quad \boldsymbol{\varepsilon}_j = g_{jq}\,\boldsymbol{\varepsilon}^q$$

into Eq A.16, we get

$$\mathbf{T} = T_{pq}\,\boldsymbol{\varepsilon}^p\,\boldsymbol{\varepsilon}^q \qquad [A.20]$$

where the nine quantities

$$T_{pq} \equiv g_{ip}\,g_{jq}\,T^{ij} \qquad [A.21]$$

are called the *covariant* components of the tensor.

Transformation rules Suppose new base vectors $\underline{\boldsymbol{\varepsilon}}_i$ are introduced; what are the new contravariant components $\underline{T}^{ij}$ of **T**? Substitution of the representations

$$\underline{\boldsymbol{\varepsilon}}_i = (\underline{\boldsymbol{\varepsilon}}_i \cdot \underline{\boldsymbol{\varepsilon}}^p)\,\underline{\boldsymbol{\varepsilon}}_p \qquad [A.22]$$

$$\underline{\boldsymbol{\varepsilon}}_j = (\underline{\boldsymbol{\varepsilon}}_j \cdot \underline{\boldsymbol{\varepsilon}}^q)\,\underline{\boldsymbol{\varepsilon}}_q \qquad [A.23]$$

into [A.19] gives

$$\mathbf{T} = T^{ij}\,(\underline{\boldsymbol{\varepsilon}}_i \cdot \underline{\boldsymbol{\varepsilon}}^p)\,(\underline{\boldsymbol{\varepsilon}}_j \cdot \underline{\boldsymbol{\varepsilon}}^q) \qquad [A.24]$$

whence

$$\underline{T}^{pq} = T^{ij}\,(\underline{\boldsymbol{\varepsilon}}_i \cdot \underline{\boldsymbol{\varepsilon}}^p)\,(\underline{\boldsymbol{\varepsilon}}_j \cdot \underline{\boldsymbol{\varepsilon}}^q) \qquad [A.25]$$

which is the desired transformation rule. Many different, but equivalent, relations are easily derived; for example

$$\underline{T}_{pq} = T^{ij}\,(\underline{\boldsymbol{\varepsilon}}_i \cdot \underline{\boldsymbol{\varepsilon}}_p)\,(\underline{\boldsymbol{\varepsilon}}_j \cdot \underline{\boldsymbol{\varepsilon}}_q) \qquad [A.26]$$

$\underline{N^{\text{th}}}$ -Order Tensors A third order tensor, or *triadic*, is the sum of *triads*, as follows:

$$\mathbf{abc} + \mathbf{def} + \mathbf{ghi} + \cdots$$

The meaning of this is that, for any vector **V**, de dot products

$$\mathbf{ab}(\mathbf{c}\cdot\mathbf{v}) + \mathbf{de}(\mathbf{f}\cdot\mathbf{v}) + \mathbf{gh}(\mathbf{i}\cdot\mathbf{v}) + \cdots$$

or

$$\mathbf{a}(\mathbf{b}\cdot\mathbf{v})\mathbf{c} + \mathbf{d}(\mathbf{e}\cdot\mathbf{v})\mathbf{f} + \mathbf{g}(\mathbf{h}\cdot\mathbf{v})\mathbf{i} + \cdots$$

provide second-order tensors. It is easily established that any third-order tensor can be written

$$\mathbf{T} = T^{ijk}\,\boldsymbol{\varepsilon}_i\,\boldsymbol{\varepsilon}_j\,\boldsymbol{\varepsilon}_k \qquad [A.27]$$

The extension to N^{th} -order tensors is now immediate. A general tensor of order N -may be written in the *polyadic* form

$$\mathbf{T} = T^{ijk}{}_{lmn\cdots st}\,\boldsymbol{\varepsilon}_i\,\boldsymbol{\varepsilon}_j\,\boldsymbol{\varepsilon}_k\,\boldsymbol{\varepsilon}^l\,\boldsymbol{\varepsilon}^m\,\cdots\,\boldsymbol{\varepsilon}_s\,\boldsymbol{\varepsilon}_t \qquad [A.28]$$

and this means that the dot product of **T** with any vector **v** produces a tensor of order (N - 1). (The dot product with **v** may be with respect to any one of the base vectors; unfortunately, the notations $\mathbf{v}\cdot\mathbf{T}$ and $\mathbf{T}\cdot\mathbf{v}$ are unambiguous only for second-order tensors, and should therefore be avoided for tensors of higher order.)

APPENDIX B

Error Propagation in Rank Annihilation

To estimate error in the Rank Annihilation results is a difficult problem. Both the unknown matrix **M** and the calibration matrix **N** are usually obtained from the same instrument, therefore there is the same instrumental error for both matrices. We will show the error for a simple case with **M**(2x2) and also **N**(2x2), with two components in **M** and one component in **N**. This is an interesting case, because the RAFA calculations can be express easily, and direct error propagation can be applied to estimate the error.

A matrix **M** bilinear in its chemical constituents can be modeled by three matrices

$$\mathbf{M} = \mathbf{X}\,\mathbf{ß}\,\mathbf{Y}^{T} = \mathbf{x}_1 \beta_1\ \mathbf{y}^{T}_1 + \mathbf{x}_2 \beta_2\ \mathbf{y}^{T}_2 \qquad [B.1]$$

that for a 2x2 **M** matrix, all the matrices **X**, **ß**, and $\mathbf{Y}^{T}$ are also 2x2. **X** contains the two unitary vectors ($\mathbf{x}_1$, $\mathbf{x}_2$) as its columns and $\mathbf{Y}^{T}$ contains the two unitary vectors ($\mathbf{y}_1$, $\mathbf{y}_2$) as its rows. Because **M** has two components, **ß** is a diagonal matrix with elements β_1 and β_2 in the diagonal, both different than zero. Similarly, the one component matrix **N** can be modeled by

$$\mathbf{N} = \mathbf{x}_1\ \xi_1\ \mathbf{y}^{T}_1 \qquad [B.2]$$

The rank annihilation estimation may be expressed as the product of three matrices (Eq 2.18),

$$f = c_{N1}/c_{M1} = \xi_1\ \mathbf{y}^{T}_1\ \mathbf{M}^{+}\ \mathbf{x}_1 \qquad [B.3]$$

where f is the ration of concentrations, c_{N1} and c_{M1} are the respective concentrations of analyte *1* in **N** and **M**, $\mathbf{M}^{+}$ represents a pseudoinverse of **M**, and $\mathbf{x}_1$ and $\mathbf{y}_1$ are unit length vectors that together with ξ are a least squares approximation of **N** such that

$\mathbf{N} = \mathbf{x}_1 \xi_1 \mathbf{y}^T_1 + \mathbf{e}$, with $\mathbf{e}$ minimal in the least squares sense, and the subscript *1* stands for the fact that $\mathbf{N}$ only contains component *1*, and $\mathbf{x}_1$ and $\mathbf{y}_1$ are its spectra in each order.

The problem is to find the variance of c_{M1}, $Var(c_{M1})$, given the variances of $\mathbf{M}$ and $\mathbf{N}$, $Var(\mathbf{M})$ and $Var(\mathbf{N})$. If we assume that the variance is independent of i,j for M_{ij} and N_{ij}, then the problem can be greatly simplified. This is a reasonable assumption for many second order techniques, such as Emission-Excitation spectrometry or LC/DAUV.

For a 2x2 square matrix such as $\mathbf{M}$, assuming that there are two linearly independent components, the pseudoinverse is simply the inverse,

$$\mathbf{M}^+ = \mathbf{M}^{-1}. \qquad [B.4]$$

If we call M_{ij} the element of the i row and j column of $\mathbf{M}$, and D the determinant of $\mathbf{M}$, $D = \det(\mathbf{M})$, then the pseudoinverse $\mathbf{M}^+$ is given by,

$$\mathbf{M}^+ = (1/D)\begin{vmatrix} M_{22} & -M_{12} \\ -M_{21} & M_{11} \end{vmatrix} \qquad [B.5]$$

Doing some algebra, and expressing the vectors $\mathbf{x}_1$ and $\mathbf{y}_1$ as $\mathbf{x}_1 = (x_{11}, x_{12})$ and $\mathbf{y}_1 = (y_{11}, y_{12})$, then we can rewrite Eq B.3 as

$$c_{N1}/c_{M1} = (\xi_1 /D)\,(x_1y_1M_{22} - x_1y_2M_{21} - x_2y_1M_{12} + x_2y_2M_{11}) \qquad [B.6]$$

An advantage of Eq B.6 over Eq B.3 is that the approximation of N_{ij} by $\xi_1 x_i y_j$ is not necessary, we can simply substitute them, obtaining

$$c_{N1}/c_{M1} = (1/D)\,(N_{11}M_{22} - N_{12}M_{21} - N_{21}M_{12} + N_{22}M_{11}) \qquad [B.7]$$

Coming back to our problem, we are interested in $Var(c_{M1})$, using f as c_{N1}/c_{M1}, $Var(f)$ is related to $Var(c_{M1})$ as follows:

$$Var(c_{M1}) = (1/f)Var(c_{N1}) + (c_{N1}/f^2)^2\ Var(f) \qquad [B.8]$$

that partially substituting $f = c_{N1}/c_{M1}$ yields

$$Var(c_{M1}) = c_{M1}^2/c_{N1}^2 \; (Var(c_{N1}) + c_{M1}^2 Var(f)) \qquad [B.9]$$

Then, the relative variance, $Var(c_{M1})/c_{M1}^2$, is

$$Var(c_{M1})/c_{M1}^2 = 1/c_{N1}^2 \; (Var(c_{N1}) + c_{M1}^2 Var(f)) \qquad [B.10]$$

Now we need to estimate $Var(f)$. There are two sets of variables to consider, N_{ij} and M_{ij}. The first order variance of f as a function of the variances σ_N and σ_M is defined as

$$Var(f) = \sum (\partial f/\partial M_{ij})^2 \, \sigma_M^2 + \sum (\partial f/\partial N_{ij})^2 \, \sigma_N^2 \qquad [B.11]$$

where the summations run for i and j from *1* to *2*. After computing the derivatives using Eq B.7, and simplifying we get

$$Var(f) = \sum (c_{N1} M_{ij}/c_{M1} - N_{ij})^2/D^2 \, \sigma_M^2 + \sum (M_{ij})^2/D^2 \, \sigma_N^2 [B.12]$$

The two terms of the right hand side of this equation can be estimated, and correspond respectively to the effect of the instrumental error for the unknown sample and the calibration sample. But further simplification is possible to interpret the $Var(c_{M1})$ as a function of the intrinsic factors **x** and **y**. The first term is a subtraction of all the estimated amount of **N** present in **M**, that in turn can be approximated with the vectors $\mathbf{x}_2$ and $\mathbf{y}_2$,

$$c_{N1} M_{ij}/c_{M1} - N_{ij} \approx (c_{N1}/c_{M1}) \, x_{2i} \, \beta_2 \, y_{2j} \qquad [B.13]$$

Because the vectors $\mathbf{x}_2$ and $\mathbf{y}_2$ are both unitary, the sum of squares of their components x_{2i} and y_2 are equal to *1*, and the first term of Eq B.12 becomes

$$\sum (c_{N1} M_{ij}/c_{M1} - N_{ij})^2/D^2 \, \sigma_M^2 = (c_{N1}/c_{M1})^2 \beta_2^2/D^2 \, \sigma_M^2 \qquad [B.14]$$

The second term of Eq B.12, related to the error in the **N** matrix, can also be approximated using the vectors **x** and **y**. Using Eq B.1, we obtain

$$M_{ij} = x_{1i} \, \beta_1 \, y_{1j} + x_{2i} \, \beta_2 \, y_{2j} \qquad [B.15]$$

then

$$\sum (M_{ij})^2/D^2 \, \sigma_N^2 = \sum (x_{1i} \, \beta_1 \, y_{1j} + x_{2i} \, \beta_2 \, y_{2j})^2/D^2 \, \sigma_N^2$$

$$= 1/D^2 \, (\beta_1^2 \Sigma x_{1i}^{\,2} y_{1j}^{\,2} + \beta_2^2 \Sigma x_{2i}^{\,2} y_{2j}^{\,2} + 2\beta_1\beta_2 \, \Sigma x_{1i} \, y_{1j} \, x_{2i} \, y_{2j}) \qquad [B.16]$$

and finally

$$\sum (M_{ij})^2/D^2\,\sigma_N^2 = 1/D^2\,[\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\,\cos(\alpha_x)\cos(\alpha_y)] \quad [B.17]$$

where α_x is the angle between the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, and similarly, α_y is the angle between the vectors $\mathbf{y}_1$ and $\mathbf{y}_2$. The determinant of **M** itself can be express as a function of the intrinsic factors. Using Eq B.1 and the fact that the determinant of the product of three matrices is the product of the three determinants of the individual matrices, D can be approximated with

$$D = \det(\mathbf{ß})\,\det(\mathbf{X})\,\det(\mathbf{Y}) = \beta_1\beta_2\,\sin(\alpha_x)\sin(\alpha_y) \quad [B.18]$$

Now, substituting Eqs B.14, B.17 into equation B.12, we get for $Var(f)$

$$Var(f) = (c_{N1}/c_{M1})^2\beta_2^2/D^2\,\sigma_M^2 + 1/D^2\,[\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\,\cos(\alpha_x)\cos(\alpha_y)]\,\sigma_N^2 \quad [B.19]$$

and the relative concentration variance is obtained substituting Eq B.19 into Eq B.10,

$$Var(c_{M1})/c_{M1}^2 = \sigma_c^2/c_{N1}^2 + \beta_2^2/D^2\,\sigma_M^2 +$$
$$+ (c_{M1}/c_{N1})^2/D^2\,[\beta_1^2 + \beta_2^2 + 2\beta_1\beta_2\,\cos(\alpha_x)\cos(\alpha_y)]\,\sigma_N^2 \quad [B.20]$$

APPENDIX C

Algorithm for the Generalized Rank Annihilation Method

This appendix presents the algorithm used to do a complete GRAM calculation. The algorithms for synchronization of LC/UV data are presented in appendix D. No major details are given but the information should be enough to write a computer program.

GRAM

Problem: Given the matrices **M** and **N** find the matrices **X**, **Y**, **ß** and **ξ** as defined by Eqs 2.11-2.12. A general algorithm is presented here.

Algorithm:

- Choose a model, **U**, for the column space of **M** and **N** (1-3). Three of the possible models are (the first one is not recommended):

1. $\mathbf{M} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$	Singular Value Decomposition of **M**
2. $(\mathbf{M}+\mathbf{N}) = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$	S.V.D. of the sum of **M** and **N**
3. $(\mathbf{M}\|\mathbf{N}) = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$	S.V.D. of the join matrix **M**|**N**

- Truncate **U** matrix according to the number of significant components. Call q the number of components selected.

- Compute matrix **W** <- **M** + **N**, (warrants all the components present).

- Project the matrices **W**, **M** and **N** onto $\mathbf{U}\mathbf{U}^T$:

  $\mathbf{W} \leftarrow \mathbf{U}\mathbf{U}^T\mathbf{W}$

  $\mathbf{M} \leftarrow \mathbf{U}\mathbf{U}^T\mathbf{M}$

  $\mathbf{N} \leftarrow \mathbf{U}\mathbf{U}^T\mathbf{N}$

- Compute S.V.D. of **W**, $\mathbf{W} = \mathbf{U}_W\mathbf{S}_W\mathbf{V}_W^T$. There should be q non-zero singular values, and the rest should be zero within the machine precision.

- $\mathbf{W}^+ \equiv \mathbf{V}_W\,\mathbf{S}_W^{-1}\,\mathbf{U}_W^T$, where $\mathbf{S}_W$ is a qxq diagonal matrix.

- Compute eigenvalues and *right* eigenvectors of non-symmetric matrix $\mathbf{N}\mathbf{W}^+$

  $(\mathbf{N}\,\mathbf{W}^+)\,\mathbf{E} = \mathbf{E}\,\boldsymbol{\lambda}$

- Compute pseudoinverse of **E**, i.e., $\mathbf{E}^+$.

- Compute **X**, **Y**, ß and $\boldsymbol{\xi}$,

  $\mathbf{X} = \mathbf{E}$ (i.e., $\mathbf{x}_i = \mathbf{e}_i$)

  $\text{ß}\,\mathbf{Y}^T = \mathbf{E}^+\mathbf{M}$ -> $\mathbf{Y}^T$ = Normalized($\mathbf{E}^+\mathbf{M}$)

  ß: β_i = Norm(Column i of $\mathbf{E}^+\mathbf{M}$)

  $\boldsymbol{\xi}$: ξ_i = Norm(Column i of $\mathbf{E}^+\mathbf{N}$)

- End.

APPENDIX D

Synchronization of LC/UV Data

The GRAM model as applied to LC/DA-UV assumes that the chromatographic profiles of each analyte in the calibration sample are identical to those in the test sample. However, small variations in the relative positions of these profiles occur for chromatographic data, i.e., the retention times may change from the calibration to the test sample, due to small fluctuations in the column characteristics. Therefore, the **Y** matrix from **M** is different from the **Y** matrix from **N**. Kim [Kim, 1985] has denominated this problem as the synchronization error. If the change in the chromatography that causes the error is not large, we can assume that the error exist only in the time index (j) of the **N** matrix,

$$M_{i,j} <=> N_{i,j+\Delta j} \qquad [2.34]$$

Kim assumed that the error Δj was an integer, and a simple integer iteration could find a Δj for which the projection of **N** on the subspace spanned by **M** changed **N** the least. Unfortunately, when sampling occurs at a slow rate, Δj is not generally an integer. Furthermore, if the change in the chromatography is due to a small change in the flow rate, Δj is not constant, and changes linearly for every j. This appendix describes three algorithms that have been developed in this work to correct the synchronization error, each useful under different circumstances, that account for a non-integer, variable $\Delta j\,(j)$. For each algorithm, the parameters of the function $\Delta j\,(j) = \Delta j_0 + m\,j$, that relates Δj linearly to j are varied with a simplex until some optimality criterion is reached.

When $\Delta j\,(j)$ has been found, a two-dimensional interpolation algorithm is used to estimate the corrected **N** matrix, i.e., **N′**, that is described at the end of this appendix.

The algorithms are not limited to a linear model- more complex models could be used such us $\Delta j\,(j) = \Delta j_0 + m j + n j^2$. But if the conditions of the chromatography change dramatically between runs, to a point that *relative* retention times are not reproduced, then these algorithms cannot synchronize the matrices and GRAM cannot be used. For long chromatographic runs it is sensible to break the data into small windows to separate synchronizations and calibrations, that will reduce the effect of fluctuations of the conditions within the same run, and also reduce the number of components analyzed at a time.

General Synchronization Algorithm:

- Define $\{\Delta j_0, m\}$ as the parameters of the equation $\Delta j\,(j) = \Delta j_0 + m j$
- With a Simplex algorithm, vary parameters $\{\Delta j_0, m\}$ until the optimal minimum of the optimization criterion is reached. Allow very small variations of m and an initial search value of $m = 1$.
- Compute the new matrix **N′** using the estimated synchronization parameters.

where the difference between the algorithms is in the optimization criterion.

Three possible criteria have been developed: Minimal Residuals Projection, Minimal Subspace Distance and Iterative GRAM.

Minimal Residuals Projection (MRP) If the components present in the calibration, **N**, are also present in the test sample, **M**, then this method can be applied. It is based in the fact that if **N** is properly synchronized, the rows (chromatograms) of the **N** matrix can be well approximated by linear combinations of the rows of **M**.

Algorithm:

- Compute the truncated singular value decomposition of **M**: $\mathbf{U\,S\,V^T}$.
- Optimize Residuals after projection: Norm $\| \mathbf{N'(I - VV^T)} \|$.

To test if all the **N** components are present in **M** we can use the necessary condition $\mathbf{N} \approx \mathbf{UU^T\,N}$.

Minimal Subspace Distance (MSD) In general is not possible to use MRP, because we may find that there are components in **N** that are not present in **M**. But if there is at least one significant component in common between the two samples, the intersection of the row subspaces of **M** and **N'** is not null when **N'** is properly synchronized. In practice, the absolute intersection is null due to the random noise, but a **N'** matrix can be found for which the distance between the subspaces is minimal, and it should correspond to approximately the right synchronization if there is one component in common.

Algorithm:

- Compute the truncated SVD of (**M**|**N**): $\mathbf{U\,S\,V^T}$.
- Change **M** and **N'** to **U** base: $\mathbf{M_u = U^T\,M}$; $\mathbf{N_u = U^T\,N'}$.
- Compute truncated SVD of $\mathbf{M_u}$ and $\mathbf{N_u}$:

  $\mathbf{M_u} = \mathbf{U_M\,S_M\,V_M^T}$.

  $\mathbf{N_u} = \mathbf{U_N\,S_N\,V_N^T}$.

- $\mathbf{t}_0 = \mathbf{v}_{M,1}$; $k = 0$; Iterate until convergence:

  $\mathbf{p}_{k+1} = \mathbf{V_N^T\,t}_k$ | This loop finds the two nearest unitary

  $\mathbf{t}_{k+1} = \mathbf{V_M^T\,p}_{k+1}$ | vectors **p** and **t** that belong respectively

  Normalize $\mathbf{t}_{k+1}$ | to the **N** and **M** subspaces.

- Distance between the subspaces = | Normalized $\mathbf{p}_{k+1}$ - $\mathbf{t}_{k+1}$ |

Iterative GRAM The two previous methods only use the information from the **M** and **N** matrices. A limitation of MSD is that when many components are present in **M** or **N**, it is difficult to find the optimal **N'** because there are multiple minima. In addition, the

proper truncation of the SVD for all the matrices becomes very critical for the quality of the results. If the spectrum of at least one of the analytes in common is available, e.g., from a library of spectra, then it is possible to do iterative GRAM calculations until one of the estimated spectra matches the library spectrum. Usually GRAM results are very far from the real factors when the synchronization is not right, therefore this method could be applied to find the optimal synchronization. It is not recommended when there are very few overlapping chromatograms (e.g., 2-3) because the possibility of correct results with a wrong **N** increases for those analytes at the tails of the chromatogram.

Algorithm:

- Find **X** using GRAM(**M**, **N**)
- Compare library spectrum x_k with all the columns of **X**. Choose the most similar (*ms*) column $\mathbf{x}_{ms}$.
- Distance = $|x_k - \mathbf{x}_{ms}|$

Algorithm for transformation of **N** When the Simplex has selected a parameter pair for testing the optimization criterion, a new matrix **N′** is computed for which

$$N'_{i,j} = N_{i,j+\Delta j}$$

Which is a two dimensional interpolation problem in one variable (standard subroutines for interpolation are widely available, e.g., IMSL). For our data the number of wavelengths was equal to *97*, therefore (*97 x Number of Scans*) interpolations are necessary, unless the magnitude of the problem is reduced, by shrinking the spectral vectors by projecting onto the truncated SVD vectors of (**M**|**N**).

Algorithm:

- Compute truncated SVD of (**M**|**N**): $\mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$.
- Approximate **N** in base **U**: $\mathbf{N}_{(u)} = \mathbf{U}^T\,\mathbf{N}$.
- For every j interpolate $\mathbf{N}_{(u)}$ to obtain $\mathbf{N}'_{(u)}$: $N'_{(u):\,i,j} = N_{(u):\,i,j+\Delta j}$
- Estimate $\mathbf{N}' = \mathbf{U}\,\mathbf{N}'_{(u)}$

APPENDIX E

Experimental Details

These experimental details were a courtesy of L. Scott Ramos and they are included for completeness. For more details see [Sanchez *et al*, 1987]

Equipment

The liquid chromatography hardware consisted of two Beckman (Berkeley, CA) 114M pumps, a Beckman 340 μflow mixer, a Valco (Valco Instruments Co., Houston, TX) 10-port injection valve fitted with a 10 μl injection loop, and a Hewlett-Packard (Palo Alto, CA) 1040A DA/UV detector. Program control was provided by a Beckman 421A LC controller, while data acquisition and storage were accomplished by a Hewlett-Packard 85B computer and 9121 dual floppy disk drive.

The gas chromatography hardware consisted of Hewlett-Packard 5890A gas chromatopraph cut-fitted with a capillary injector, a flame ionization detector, and a Hewlett-Packard 3392A integrator.

Reagents

The mobile phase solvents, UV-grade acetonitrile and water, were obtained from Burdick & Jackson (Muskegon, MI). Polynuclear aromatic hydrocarbon standards were purchased from Chem Services, Inc. (West Chester, PA), and included: acenaphthylene (Acen) [206-96-8], Phenanthrene (Phen) [85-01-8], benz(*a*)anthracene (B*a*A) [56-55-3], anthracene (Anth) [120-12-7], chrysene (Chry) [218-01-9], benz(*a*)pyrene (B*a*P) [50-32

8], benz(*e*)pyrene (B*e*P) [192-97-2], benz(*b*)fluoranthene (B*b*Fl) [205-99-2], benz(*k*)fluoranthene (B*k*Fl) [207-08-9] and perylene (PER) [6364-19-8].

Procedures

A 5 μm C_8, 4.6 x 150 mm column (Brownlee, Santa Clara, CA) was used in the LC analyses. The LC analyses followed a basic procedural outline: an events table was first created in the 421A controller, including % of solvent B, mobile phase flow rate, time of injection and time for initiating data acquisition. Calibration and test samples were analyzed in exactly the same manner, one immediately after the other, to minimize error in retention time reproducibility.

Data acquisition was initiated upon a command from the 421A controller and was terminated at the time entered for stop-time in the HP 85B system. The DA/UV detector was operated in the "periodic spectra" mode, in which full spectral scans from 210-400 nm were acquired at a rate of ca. 1 scan/sec, with a bandwidth of 2 nm.

VITA

Eugenio Sánchez Marqués was born in Caracas, Venezuela on August 20, 1961 to Eugenio and Inés Sánchez. He entered the Universidad Simón Bolívar in 1977 where he began to study chemistry. In 1979 he worked with Prof. Tatsuhiko Nakano as a research assistant in synthesis of natural products in the Instituto Venezolano de Investigaciones Científicas (IVIC). In 1981 he worked with Prof. Carlos Manzanares in theoretical calculation of collision induced vibrational energy transference probabilities. He received a *Licenciatura* in chemistry from the Universidad Simón Bolívar, Caracas, Venezuela, in July, 1982.

In 1983 he received a scholarship from the Venezuelan "Fundación Gran Mariscal de Ayacucho" to obtain a M.S. in chemistry, in the University of Washington, Seattle, Washington, that he completed under Prof. Bruce R. Kowalski in 1985. In March 1987 he completed the requirements for a Doctor of Philosophy in the field of analytical chemistry, in the University of Washington, specialized in chemometrics and process analytical chemistry.