MATRIX ALGORITHMS FOR BILINEAR ESTIMATION PROBLEMS IN CHEMOMETRICS

by

RYAN ROYCE KIM

B.S., Columbia University
(May 1977)

S.M., Massachusetts Institute of Technology
(June 1979)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 1985

© Massachusetts Institute of Technology 1985

Signature redacted

Signature of Author:.....
Department of Mechanical Engineering

Signature redacted

Certified by:.....
Thesis Supervisor

Signature redacted

Accepted by:.....
Chairman, Departmental Committee on Gradute Students

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUL 22 1985

LIBRARIES

Archives

MATRIX ALGORITHMS FOR BILINEAR ESTIMATION PROBLEMS IN CHEMOMETRICS

by

RYAN ROYCE KIM

Submitted to the Department of Mechanical Engineering on May 3rd, 1985 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

Chemometrics is the discipline that applies mathematical and statistical methods to the extraction of qualitative and quantitative information from experimental measurements in analytical chemistry.

An important class of problems in chemometrics is that of **bilinear estimation** for the system of equations $Z=XY$, where Z is a known observation matrix that contains the results of chemical experiments, and X and Y are unknown matrices that model the underlying mechanisms responsible for the observations. A key example, and the one that is focused on in this thesis, arises in liquid chromatography. Here, Z contains multi-wavelength absorbance profiles of a mixture, and X and Y respectively embody absorbance spectra and concentration profiles respectively of the constituent components in the mixture.

In developing their techniques, analytical chemists have traditionally relied upon the statistical literature on factor analysis. The purpose of this thesis has been to undertake a systematic research into these problems from a different point of view, drawing on current advances in signal processing, numerical linear algebra, and related fields to develop new formulations and algorithms for these problems.

Two specific categories of problems are dealt with in this thesis: the **target components problem** is encountered in cases where it is desired to estimate, from measurements on a mixture, the concentrations of a few selected, or target components for which calibration data are available; the **unknown components problem** addresses the task of narrowing down the range of possible constituent components of an unknown mixture using measured data, with no prior information on the possible contents of the mixture.

For the target components problem, we have obtained two non-iterative and numerically stable solution procedures that, in comparison to existing methods, dramatically reduce computational requirements. In addition, our algorithms compensate for data synchronization errors that adversely affect existing methods used in liquid chromatography. The two procedures have been applied to both synthetic and experimental data from liquid chromatography, and are found to be robust in the presence

of sensor noise and other experimental errors such as synchronization error.

In addressing the unknown components problem, we have extended earlier work on the case of two and three unknown components, giving a concise formulation of the problem and a physically and geometrically motivated characterization of all possible solution sets for an arbitrary number of unknown components. We have also devised a reduction and representation scheme for noisy data that delineates fine structural details of such chemical measurements that typically elude existing methods of analysis. The representation scheme reduces to a two-dimensional display for the practically important case of three components, thus facilitating interactive data analysis. Our method has been shown to be particularly effective in analyzing absorbance profiles of chemical mixtures from liquid chromatography.

Thesis Supervisor: George C. Verghese
Title: Associate Professor of Electrical Engineering

TO MY PARENTS

ACKNOWLEDGEMENTS

I am greatly indebted to Professor George Verghese for his dedication and patience in guiding me through this thesis. His enthusiasm and thirst for knowledge was an ever present source of inspiration during the dry-periods of the dissertation. His timely suggetions and references were instrumental in resolving many of the difficult issues addressed here. Furthermore, his detailed revisions and comments on this thesis have greatly improved the quality of the final draft, and is doubly appreciated. Thanks, George!!

I also wish to thank Professor Fred Schweppe for many stimulating discussions and suggestions, which helped to broaden the scope of the thesis. Many thanks to Professor Leonard Gould for his efforts at editing the thesis (twice!) at such short notices. Professors David Wormley and Karl Hedricks's contributions during the committee meetings are also appreciated.

The "gang": John Greschak, Clem Karl, Juan Lavalle, Scott Fenton, Seth Sanders! What would I have done without their sense of humor and zaniness? Special thanks to John for many helpful discussions and endless debates.

Thanks to the administrative and secretarial staffs at the LEES (Gabriele Ivey, Kathleen Lipfin, and Barbara Smith) and the Dept. of Mech. E. (Leslie Regan and Sandy Williams Tepper) for helping me cut through the red tapes.

Contributions from the Waters Chromatography Division of Millipore Corporation, which partly supported this thesis, and its technical staff have been invaluable in completing this thesis. In particular, Dr. Richard King (who first brought the problems addressed in this thesis to our attention, has continued to supply us with new interesting problems, and provided insights into the workings of the liquid chromatography) has been a godsend. Mr. William Carson and his "Coat-hanger" model helped to develop and clarify geometric intuitions. Mr. Steve Deslauriers provided the real data used in Chapter 5 (thanks for bringing it over at difficult hours!)

At a more personal level ... Thanks to Annette, Francesca, Haris, Jon, Shariar, and others who have drifted in-and-out, for their friendship... to my brothers and sisters, wish I could have spent more time with you all...

Last but not least, Mom & Dad ... friends and lovers may come and go, but you two are always there in times of need. THANKS MOM, THANKS DAD.

TABLE OF CONTENTS

	PAGE
TITLE PAGE	1
ABSTRACT	2
DEDICATION	4
ACKNOWLEDGEMENTS	5
TABLE OF CONTENTS	6
NOMENCLATURE	10
CHAPTER I. INTRODUCTION	12
1.0 Outline	13
1.1 Background	14
1.2 Scope and Contributions	16
1.3 Outline of the Thesis	23
CHAPTER II. PROBLEM FORMULATION	25
2.0 Outline	26
2.1 Introduction	27
2.2 Liquid Chromatography-UV Spectroscopy Experiment	29
2.3 Structure of Data Matrices	37
(a) Single (Pure) Component	37
(b) Multiple Components (Mixture)	41
2.4 Mathematical Statements of the Problems	45
(a) Target Components Problem	45
(b) Unknown Components Problem	45
2.5 Effect of Experimental Error/Noise	48
2.6 Current Practice in Analysis of Chromatographic Data	51

PAGE

CHAPTER III. THE TARGET COMPONENTS SOLUTION..... 58

3.0 Outline..... 59

3.1 Introduction..... 60

3.2 Solution Characterization..... 62

3.3 Projection Method..... 66

3.3.1 Noise-Free Data Case..... 66

3.3.2 Noisy Data Case..... 76

(a) Number of Components: Rank Test..... 76

(b) Modes of Projection..... 79

(c) Data Synchronization..... 87

3.4 Generalized Eigenanalysis Method..... 95

3.4.1 Solution Method..... 96

(a) Background: Canonical Form (Kronecker Structure) of a Singular Pencil..... 96

(b) Van Dooren's Algorithm: Noise-Free Case..... 97

3.4.2 Application of Van Dooren's Algorithm to Noisy Data Case..... 103

CHAPTER IV. THE UNKNOWN COMPONENTS SOLUTION..... 111

4.0 Outline..... 112

4.1 Introduction..... 114

4.2 Data Representation and Solution Characterization: Noise-Free Case..... 116

4.2.1 Measurement Space..... 117

(a) Data Representation and Solution Characterization for PROBLEM 2a..... 117

(b) Data Representation and Solution Characterization for PROBLEM 2b..... 124

PAGE

4.2.2 Parametrized signal space..... 129

4.3 Extensions to Noisy Data Case..... 142

4.3.1 Reconstructing the signal space..... 142

4.3.2 Representation in Parameterized Reconstructed Signal (PRS) Space..... 147

4.3.3 Representation in Normalized Parametrized Reconstructed Signal (NPRS) Space..... 150

4.4 Solution Procedures: Two- and Three-Components Cases..... 159

4.4.1 Two-Components Case..... 159

4.4.2 Three-Components Case..... 163

4.5 Examples of Data Display with Synthetic Liquid Chromatography Data..... 173

CHAPTER V. NUMERICAL RESULTS: REAL DATA..... 184

5.0 Outline..... 185

5.1 Description of Experiment..... 186

5.2 Target Components Problem..... 191

5.3 Unknown Components Problem..... 192

CHAPTER VI. CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK..... 196

REFERENCES..... 202

APPENDIX A. MODES OF DECOMPOSITION..... 207

APPENDIX B. GENERALIZED BILINEAR PERTURBATION PROBLEM..... 213

APPENDIX C. EFFICIENT ALGORITHMS FOR THE PROJECTION METHOD..... 223

APPENDIX D. A QUALITATIVE ERROR ANALYSIS OF THE PROJECTION METHOD.. 232

PAGE

APPENDIX E. A FIRST ORDER EIGENVALUE SENSITIVITY FORMULA FOR REGULAR PENCILS..... 235

APPENDIX F. PROOFS OF LEMMAS IN CHAPTER 3..... 237

APPENDIX G. PROOFS OF RESULTS IN CHAPTER 4..... 240

APPENDIX H. A MAXIMUM LIKELIHOOD APPROACH TO THE TARGET COMPONENT PROBLEM: A PARTIAL SOLUTION..... 245

APPENDIX I. AN ERROR ANALYSIS OF VAN DOOREN'S ALGORITHM..... 261

NOMENCLATURE

$x \in A$ means x is an element of the set A.

$A \subset B$ means set A is contained in set B.

$\{x: x \text{ has property a}\}$ denotes the set of all x such that x has property A.

R^n denotes the set of all n-dimensional real vectors.

R^{mxn} denotes the set of all mxn real matrices.

A a matrix is written with a capital letter.

A^T denotes the transpose A.

A^{-1} denotes the inverse of A (assuming it exists).

$|A|$ denotes the determinant of A (assuming A is a square matrix).

$||A||_2$ denotes the spectral norm of A.

tr A denotes the trace of A.

a scalar and vector are written with lower case letters.

$|a|$ denotes the absolute value of a (when a is a scalar).

a^T denotes the transpose of a (when a is a vector).

1 "boldfaced" 1 is a vector with all its elements equal to one: $[1,\ldots,1]^T$.

$\sum_{i=1}^{r} A_i$ denotes the summation of A_i from i=1 to i=r.

$\prod_{i=1}^{r} A_i$ denotes the product of A_i from i=1 to i=r.

$\bigcap_{i=1}^{r} \{A\}_i$ denotes the intersection of the sets $\{A\}_i$ from i=1 to i=r.

[A1],[A1,A2] alphanumeric number (or numbers) enclosed in a bracket generally refers to a reference number; matrices and vectors are sometimes also enclosed in a bracket.

Acronyms:

CLS	complementary least squares.
LC-UVS	liquid chromatography-UV spectroscopy.
LLS	linear least squares.
NPRS	normalized parametrized reconstructed signal.
PRS	parametrized reconstructed signal.
SVD	singular value decomposition.
TLS	total least squares.

CHAPTER I: INTRODUCTION

CHAPTER OUTLINE

	Page
1.1 INTRODUCTION.....	14
1.2 SCOPE AND CONTRIBUTIONS.....	16
- Problems Addressed.....	16
- Existing Approaches.....	18
- Thesis Contributions.....	20
1.3 OUTLINE OF THE THESIS.....	23

1.1 INTRODUCTION

Chemometrics, as defined by the International Chemometrics Society, is "the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments; and (b) to provide maximum chemical information by analyzing chemical data." [F2] As such, it encompasses a wide spectrum of topics: from traditional areas in multivariate statistics to the more recent developments in artificial intelligence. An excellent review of analytical chemometrics (chemometrics for analytical chemistry), with numerous references, is given by Frank and Kowalski [F2].

An important class of problems in analytical chemometrics is that of bilinear estimation. A bilinear estimation problem is an estimation problem in a system of the form

$$Z = XY^T \tag{1.1-1}$$

where Z is an observation matrix, and X,Y are the unknown matrices to be estimated. Equation (1.1-1) is in a bilinear form in the sense that, if one set of unknown variables (X or Y) is held constant, the resulting system is linear in the other (Y or X). Thus, bilinear estimation problems exist naturally whenever the observed variable set is a product of two sets of variables of interest. For example, many of the estimation problems in signal processing and applied linear algebra [W1,G4], as well as problems of factor analysis in chemometrics [M1], are of the form given in Equation (1.1-1), differing only in the *a priori* information on the unknowns, and/or the nature of information to be extracted.

In developing their techniques, analytical chemists have

traditionally drawn upon the statistical literature on factor analysis. It turns out, however, that a host of problems whose mathematical statements are close to (and often identical with) those of factor analysis have been extensively studied of late in the areas of digital signal processing, estimation, numerical linear algebra, and other related fields. In these areas, the emphasis is on methods that are recursive, on-line, computationally efficient, numerically robust, adaptive, and so on, with particular attention given to exploiting any special structure of the problem being studied.

The purpose of this thesis has been to undertake a systematic research, from the point of view of these latter areas, into a class of bilinear estimation problems that are of great importance in analytical chemistry. The problems to be addressed and the thesis contributions are discussed in Section 1.2, and an outline of the thesis is given in Section 1.3.

1.2 SCOPE AND CONTRIBUTIONS

Problems Addressed

An often encountered problem in chemistry is that of identifying the constituent components of an unknown chemical mixture. The goal here is to determine, through chemical analysis, what components are present in a set of samples being studied, and in what amounts or proportions they are present. Typical applications include determining alcohol or drug contents in blood samples, identifying and quantifying toxic chemical levels in food additives, or analyzing the relative composition of a known chemical mixture for quality control.

As a first step towards such quantitative and qualitative identification, the unknown mixture can be subjected to some appropriate chemical experiments, such as liquid chromatography and spectrophotometry [M3] or fluorescence spectroscopy [J1,H2]. The constituent components may then be identified by comparing the results to known calibration results for pure components (assembled in some data base). For example, in fluorescence spectroscopy, the experimental results are in the form of excitation and emission spectra; if the spectra of the constituent components do not overlap, the components are readily identified by comparing the results of the experiment to standard spectra of pure components.

Often, however, a complete spectral or chromatographic separation is unattainable, and one has to resort to analytical methods to delineate the underlying structure of the overlapping experimental results. We address, in this thesis, chemical identification problems where there is an overlap in the contributions of the individual components to the experimental results obtained from the mixture, so

that the individual components are not directly recognizable. The two classes of problems to be addressed are:

- The Target Components Problem, and
- The Unknown Components Problem

The target components problem is encountered in cases where it is desired to estimate the concentrations of a few selected, or target, components in the mixture, given calibration data for them. The unknown components problem, on the other hand, addresses the task of narrowing the range of possible constituent components of an unknown mixture by use of experimental measurements, with no a priori information on the possible contents of the mixture. This narrowing of possibilities may, for example, involve computing upper and lower bounds on the spectra of candidate constituents.

Although the methods developed in this thesis are applicable to many chemical and non-chemical data of bilinear form, the emphasis here is on applications to data from liquid chromatography-UV spectroscopy experiments. We have applied our algorithms to both synthetic and real data from such experiments , and the results are presented in this thesis. The assumptions made on the structure of the experimental data to be analyzed are as follows:

(i) Experimental results for unit concentration of a single component are expressible in a matrix dyad of the form

$$N = xy^T \qquad (1.2\text{-}1)$$

where x and y are vectors that contain "signatures " of that component

(e.g. concentration profile and absorbance spectrum in a liquid chromatography-UV spectroscopy experiment),

(ii) The experimental results on a mixture are expressible as a linear combination of the individual component results, weighted according to the respective concentrations of the components

$$Z = c_1 N_1 + c_2 N_2 + \ldots \qquad (1.2\text{-}2)$$

(e.g. the absorbance profile of a mixture is, according to the Beer-Lambert law [P2], well represented as a linear combination of the weighted individual absorbance profiles.)

Existing Approaches

The target components problem has been addressed by Ho et al. in a series of papers published from 1978 to 1981 [H2,H3,H4]. Their so-called rank annihilation method appears to be the state-of-the-art procedure for solving the target components problem, and has been applied by McCue and Malinowski [M3] to data from a liquid chromatography experiment as recently as 1983. The method, however, requires a numerically expensive iterative search procedure, with eigenanalysis at each step of the search, and with no guidelines on the starting values. In addition, successful application of the rank annihilation method to chemical data from liquid chromatography experiments in particular requires careful registration of data from mixture and calibration runs [M3].

The unknown components problem for the two-component case has been addressed by Lawton and Sylvestre [L1], who gave a method for estimating (via upper and lower bounds) the shapes of two overlapping functions from an observed set of additive combination of these two functions.

Applications of their method can be found in [S2], where it is used to resolve overlapping mass spectra from a gas chromatography-mass spectroscopy experiment, and in [O2] where the spectra to be resolved are absorption spectra from a liquid chromatography experiment. Spjotvoll et al. [S5] propose an alternate (restricted least squares method) approach to the two-component problem and claim that the method is applicable to an arbitrary number of components, but extensions to three or more components have not been made explicit or pursued in the literature. Extensions to the three-component case have been partially addressed by several investigators [C2,M2,O1]. Chen and Hwang's method [C2] appears to work only in special cases where the underlying spectra (functions) are as distinct as possble, and even then some trial-and-error is involved in determining the underlying spectra. Meister [M2] also assumes that the spectra are as distinct as possible, and uses an iterative linear programing approach to fit a "maximum-volume" simplex, within some allowable region, that defines a solution. Ohta's extension [O1] closely follows that of Lawton and Sylvestre in defining the allowable solution region, but the characterization of solution sets is unclear (it appears to be a brute force Monte Carlo method) and numerically expensive.

All the work that we have quoted in this subsection remains in the research literature, and does not appear to have filtered into the analysis packages supplied by, for example, manufacturers of equipment for liquid chromatography-UV spectroscopy. Given the deficiencies in the methods and/or in their expositions, as noted above, this fact is perhaps not surprising. Current practice in liquid chromatography-UV spectroscopy does not involve matrix formulations or matrix algorithms,

proceeding instead via more graphically oriented considerations of the shapes of overlapping concentration profiles, see [S3,P2]. (A brief summary of this approach is given in Section 2.6.) The extension of these graphical methods to multi-wavelength data is not at all obvious. It is mainly this gap, between current practice and what seems possible via matrix formulations and algorithms, that this thesis is aimed at.

Thesis Contributions

For the target components problem, two alternate solution methodologies are proposed in this thesis, namely

- a projection method, and
- a generalized eigenanalysis method.

These alternatives are non-iterative and numerically stable methods that dramatically reduce the computational load from that of the rank annihilation method [H2]. The inherent simplicity and clarity of our methods have suggested supplemental methodologies that minimize the effect of experimental errors and compensate for incorrect registration of mixture and calibration runs.

An important step in developing the projection method was recognizing that the target components problem for the noise free measurements case (as stated in [H2]) can be transformed via matrix decompositions to a single reduced (r-dimensional for r target components) eigenvalue problem. Extension to the case of noisy data essentially involves only two additional data projection steps in the transformed spaces.

The generalized eigenanalysis method is based on an algorithm developed by Van Dooren [V2] for recovering the regular "pencil"

structure of a singular matrix "pencil". We have, in this thesis, developed some aspects of his work by analyzing error propagation along each step of the application of his algorithm to our special problem; this has resulted in improved guidance for our application of his algorithm, especially in selection of thresholds for the numerical rank determination.

For the unknown components problem, we have given a concise formulation of the problem and extended the characterization of solution sets to the case of three or more unknown components. In addition, comprehensive solution methods for the case of two and three unknown components are given. Unlike Meister [M2] and Chen [C2], our methods do not assume orthogonality of the underlying spectra. Instead, based on geometric interpretations of the constraints imposed by the problem, the ideas and algorithms developed in this thesis are used to rule out large groups of possible solution components (thereby reducing the number of components to be searched in a database of possible solution components, for example).

The general solution approach is developed in a physically and geometrically motivated solution space (as opposed to the abstract spaces used in the existing methods [L1,C2,M2,O1]) that is easier to visualize, and then transformed into an abstract space in a natural way, so that the solution characterization in this space becomes more transparent. Furthermore, transformation to the abstract space is done with the singular value decompositions [K2] of the data matrices, which is often much more efficient than the principal component analysis of the data covariance matrix generally used in the literature. In the process, we also detail a mapping of the uncertainty regions in the

physical space down to the abstract space.

For the case of three unknown components, we have also devised a reduction and display scheme for noisy data that delineates the fine structural details of liquid chromatography measurements that typically elude the current methods of analysis. The representation scheme reduces to a two-dimensional display for the practically important case of three components, thus facilitating interactive data analysis. Furthermore, as an application of our data display, we have shown that the sampling offsets among the detector channels have certain characteristic patterns that are easily detected in our form of display.

The methodologies developed in this thesis have been successfully applied to both synthetic data (synthesized to simulate, as closely as possible, real liquid chromatography measurements) and experimental data (provided by the Waters Chromatography Division of Millipore Corp.). The results indicate that the contributions in this thesis represent a significant advance in the analysis of data from liquid chromatography, and that our algorithms hold strong potential as tools that can be routinely included in analysis packages for such data.

1.3 OUTLINE OF THE THESIS

The problems to be addressed in this thesis are motivated and stated, with all the underlying assumptions, in Chapter 2. In Section 2.2, a brief overview of a standard liquid chromatography-UV spectroscopy (LC-UVS) experiment is given. Using examples from the LC-UVS experiment for motivation, Section 2.3 presents an idealized model for the structure of the data matrices that describe the experimental results. This model is then then used in Section 2.4 to mathematically state the two chemical problems (target components and unknown components) to be addressed in this thesis. In Section 2.5, effects of experimental errors/noise are factored into our idealized model of the data matrices. Section 2.6 briefly reviews current practice in deconvolution of LC-UVS data, to serve as background for, and contrast to the matrix algorithms developed in the remainder of the thesis.

Solutions to the target component problem are given in Chapter 3. A characterization of the solution is stated in Section 3.2, and two methodologies (projection and generalized eigenanalysis) that solve this problem are given in Sections 3.3 and 3.4, respectively. A brief tutorial on matrix decompositions (used heavily in both methods, and throughout the thesis) is given in Appendix A. Different modes of projection that can be used in the projection method with noisy data are given in Appendix B, as special cases of the generalized bilinear perturbation problem. Several efficient algorithms for implementing the projection methodology are given in Appendix C. An error analysis of the projection method is developed in Appendix D, and a first order eigenvalue sensitivity formula for regular pencils is derived in Appendix E. Proofs lemmas for Chapter 3 that are not in the body of the

text are given in Appendix F. In Appendix H, a maximum-likelihood approach to the target components problem is formulated for the general case, and a closed form solution for a special case is given. Appendix I details an error analysis of Van Dooren's algorithm as applied to our case.

In Chapter 4, solutions to the unknown components problem are given. Section 4.2 provides geometric insight into the structure of the data matrices, and gives a characterization of the solution set for an arbitrary number of components in the noise-free data case. Extensions to the case of noisy data are then presented in Section 4.3. In Section 4.4, solution methodologies for the two and three-component cases are detailed. Finally, in Section 4.5, an efficient data display scheme for a special case (three channel data) is given. Proofs of results in Chapter 4 that are not in the body of the text are given in Appendix G.

The material in Chapters 2 to 4 is illustrated throughout with data synthesized according to the model in Chapter 2. We turn, in Chapter 5, to numerical results with real data from liquid-chromatography-UV spectroscopy experiments. Results on the target components problem are presented in Section 5.2, and those on the unknown components problem in Section 5.3.

In Chapter 6, we outline some related problems and other open questions, and present concluding remarks and recommendations for future work.

CHAPTER II: PROBLEM FORMULATION

CHAPTER OUTLINE

	Page
2.1 INTRODUCTION	27
2.2 LIQUID CHROMATOGRAPHY-UV SPECTROSCOPY	29
2.3 STRUCTURE OF DATA MATRICES	37
(a) Single (Pure) Component	37
(b) Multiple Components (Mixture)	41
2.4 MATHEMATICAL STATEMENTS OF THE PROBLEM	45
(a) Target Components Problem	45
(b) Unknown Components Problem	45
2.5 EFFECTS OF EXPERIMENTAL ERROR/NOISE	48
2.6 CURRENT PRACTICES IN ANALYSIS OF CHROMATOGRAPHIC DATA	51
- Qualitative Analysis	51
- Quantitative Analysis	54

2.1 INTRODUCTION

In addressing the problem of quantitative and qualitative identification of unknown chemical mixtures, analytical chemists have relied heavily upon the methods of chromatography and spectroscopy to identify and quantify unknown chemical mixtures. Typically, an unknown mixture is subjected to some appropriate chemical experiment, such as liquid chromatography and spectrophotometry [M3] or fluorescence spectroscopy [J1,H2], and the results of the experiment are compared to known standard results for pure components (assembled in some data base). For example, in fluorescence spectroscopy, the experimental results are in the form of excitation and emission spectra; if the spectra of the constituent components do not overlap, the results of the experiment are compared to standard spectra of pure components for identification.

Often, however, a complete spectral or chromatographic separation is unattainable, and one has to resort to analytical methods to delineate the underlying structure of the overlapping experimental results. The problem formulations here are designed to address chemical identification problems where the experimental results of the individual components in a mixture ovelap, so that the individual components are not directly recognizable. The two classes of problems to be addressed are:

- The Target Components Problem, and
- The Unknown Components Problem

The target components problem is encountered in cases where it is desired to estimate the concentrations of a few selected, or target,

components in an unknown chemical mixture. The unknown components problem, on the other hand, addresses the task of identifying the constituent components of an unknown mixture, with no a priori information on the possible contents of the mixture.

In Section 2.2 we describe a chemical procedure, liquid chromatography-UV spectroscopy, that is widely used to examine the composition of chemical mixtures. In Section 2.3, we give a mathematical model of the experimental results, showing how these may be assembled in structured data matrices. The data matrices are then used in Section 2.4 to give mathematical statements of the chemical problems to be addressed in this thesis. The effects of experimental error/noise are treated in Section 2.5. Finally, Section 2.6 reviews current practice in deconvolution of chromatographic data, to serve as a background for, and contrast to the matrix algorithms developed in the rest of the thesis.

2.2 LIQUID CHROMATOGRAPHY-UV SPECTROSCOPY

In this section we give a brief overview of a standard liquid chromatography-UV spectroscopy experiment. Most of the material presented in this section is drawn from sources such as [G2,S4,V6,H1]. For more detailed descriptions, the reader is referred to these references.

A schematic of a typical liquid chromatography-UV spectroscopy experiment is shown in Figure 2.2-1. The four major subsystems of a liquid chromatograph-UV detector apparatus are:

- High pressure pumping system
- Sample injection system
- Separation column
- Detector unit

We discuss these below.

Pumping System

The pumping system consists of either a constant flow or a constant pressure pump that pressurizes the solvent drawn from a reservoir. The two main types of constant flow pump are: single stroke (syringe type) pumps, and reciprocating pumps having either a diaphragm or a piston. Constant flow pumps deliver constant flow rate to the columns, with pressure variations compensating for changes in the operating conditions. When reciprocating pumps are used, a "pulse dampener" (typically some form of fluidic capacitor and resistor) is added after the pump to lessen the pulsing effects. Constant pressure pumps generally consist of a pneumatic device that directly pressurizes

the mobile phase. They provide pulse-free flow but are generally less accurate than constant flow pumps, so their use is limited to cases where flow rates and reproducibility are less critical.

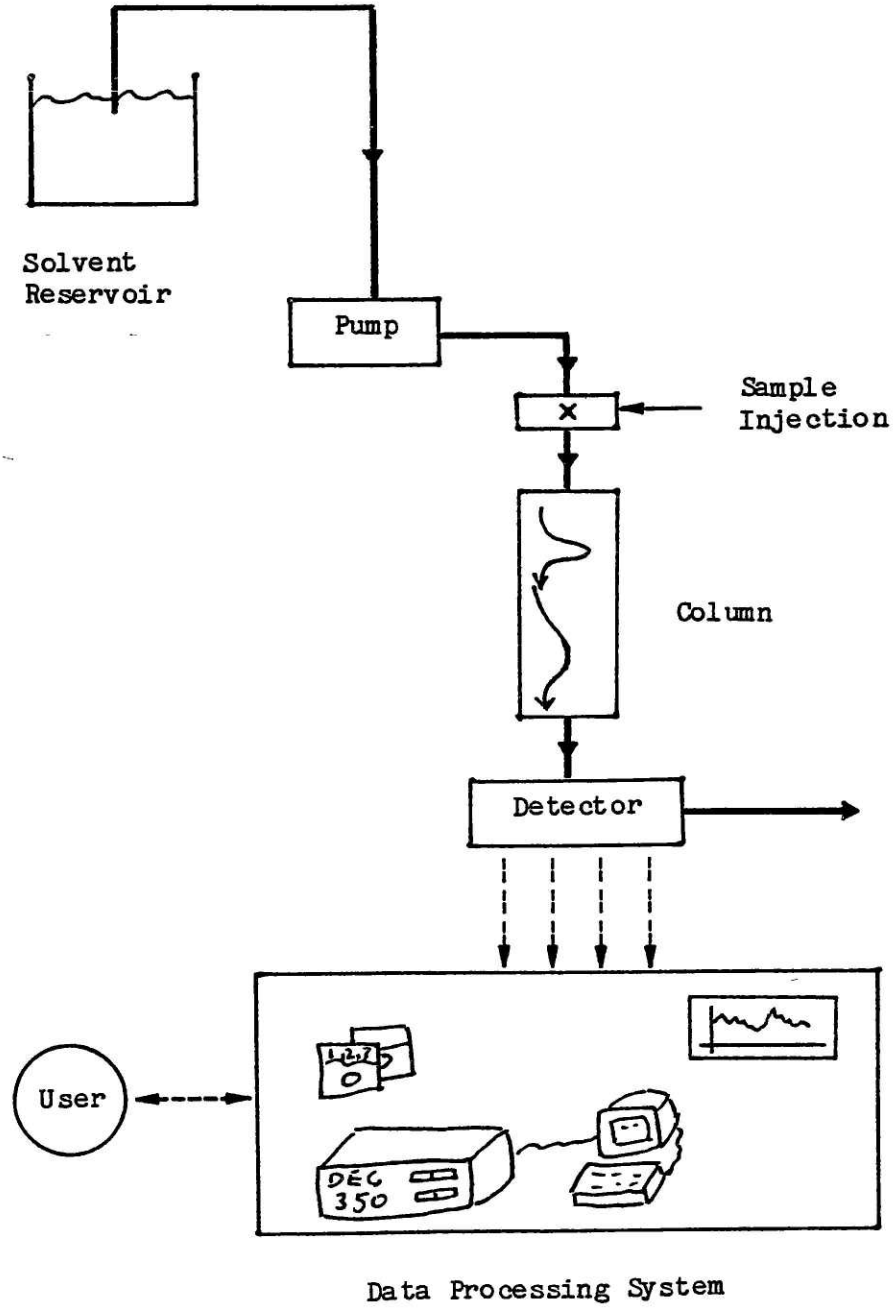


Figure 2.2-1. A schematic of a liquid chromatography-UV spectroscopy experiment.

Sample Injection System

Just before the solvent enters the column, the sample to be analyzed is mixed with the solvent in the injection port. Two frequently used injection modes are syringe and valve-loop injection methods. In the syringe injection method, the sample to be tested is directly injected into the mobile phase of the solvent with a syringe, via a septum injector (a device that allows the needle of a syringe to be inserted into the mobile phase without producing leakage, up to an operating pressure of approximately 3000 psi). With a valve-loop injection method, on the other hand, the solvent flow is bypassed into the column while an external loop is filled with the sample, which is then introduced into the column by switching a valve so that the solvent is redirected through the external loop. The main advantages of the valve-loop injection method are its ability to inject a wide range of sample volumes with a high degree of reproducibility, and the capability of injecting at high pressure (~7000 psi).

Separation Column

At the heart of any chromatographic system is the column that separates the sample mixture into its constituent components. The dominant mechanism that effects the separation in a chromatographic column is the diffusion process of a sample through the column. This diffusion is governed by the Fick's law, which (for the one-dimensional flow case) is [G2]

$$\frac{\partial C}{\partial t} = -V \frac{\partial C}{\partial x} + D \frac{\partial^2 C}{\partial x^2} \qquad (2.2.1)$$

where $\partial C/\partial t$ is the rate of change in concentration at any point along

the column, V the solute zone velocity, D the diffusion coefficient of the solute in the mobile phase, and x the direction of the flow. Thus, any sample introduced into the column in the form of a "plug" (an idealized point source) will undergo diffusive broadening as shown in Figure 2.2-2. The differential rates at which the different components of the sample undergo diffusive broadening effects the separation in the column. Note that, since the area under the profile remains constant due to conservation of mass, the height of the peak decreases as the peak passes through the column.

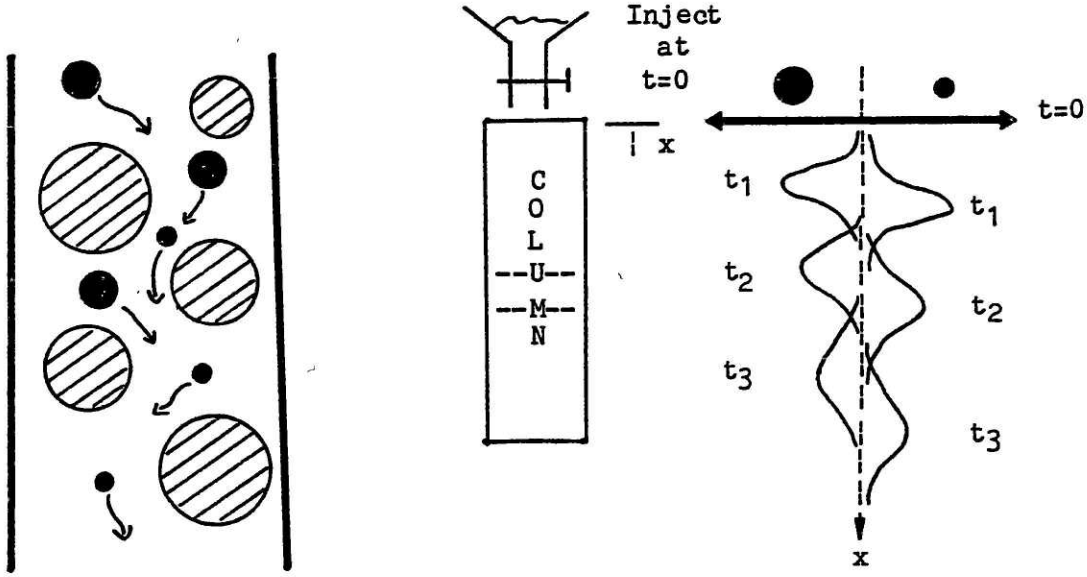


Figure 2.2-2. Diffusion broadening in a column.

When the concentration of the solute is low, so that the linear-ideal case is approximated, the spatial concentration profile down the column (neglecting end effects), at a fixed t, can be described by a Gaussian distribution function, for a pure component. The concentration profile of a pure sample eluting at the end of the column as a function of time, can then also be modeled as a Gaussian [V6]:

$$C(t) = \frac{M}{F} \frac{1}{w(2\pi)^{1/2}} \text{EXP}(-0.5((t_r-t)/w)^2) \qquad (2.2\text{-}2)$$

where M is the total mass of sample injected, F is the flow rate, w is the standard deviation of the distribution, and t_r is the retention time (see Figure 2.2-3).

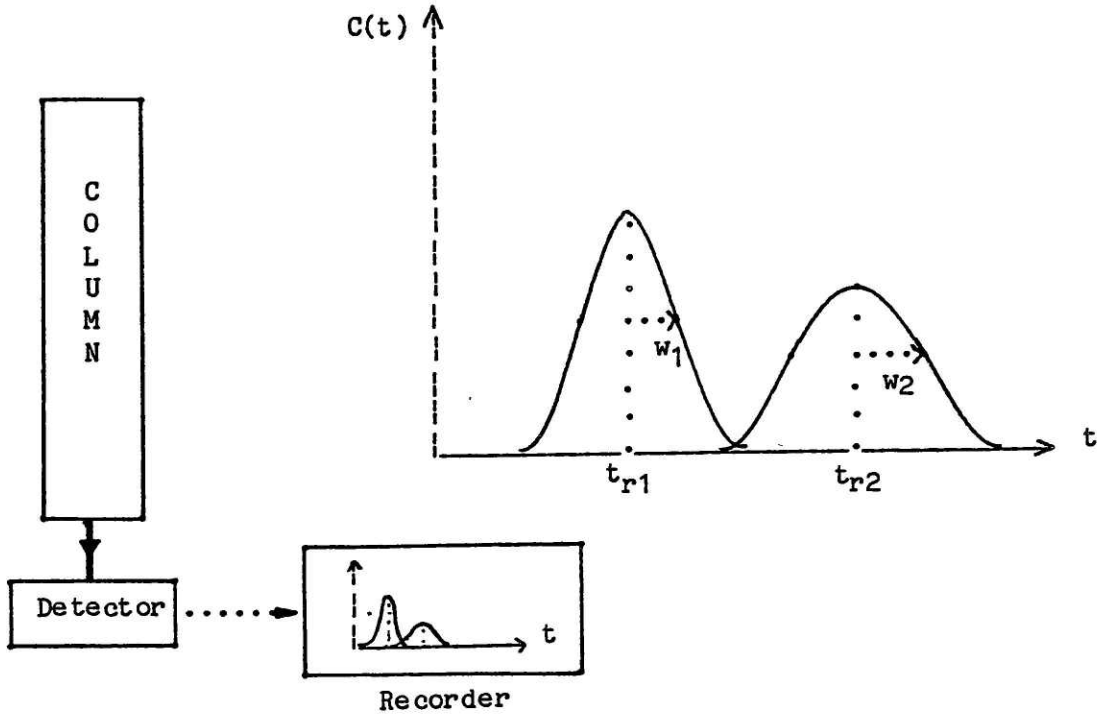


Figure 2.2-3. Concentration profiles of pure samples.

Other more refined models (e.g., the Gaussian-plus-exponential-tail model that takes into account the effect of "tailing") of the concentration profile exist (see [B1]), but we shall not address these in this thesis.

With a fixed amount of a sample, a column that gives a smaller w for a given t_r (hence a better separation among the constituent components in a mixture) is said to be a more efficient column. In fact, the efficiency of a column is measured by its plate number

$$\text{Plate Number} = \frac{t_r^2}{w^2} \tag{2.2-3}$$

The higher the plate number, the higher the efficiency. The reason this measure is useful is that the plate number of a column remains approximately constant for a wide range of operating conditions and samples. It follows that the longer the retention time t_r of a sample, the greater the standard deviation w of its concentration profile.

Detector Unit

One of the major factors that has contributed to the development of the modern high performance liquid chromatograph (HPLC) has been the availability of sensitive, on-line detectors. Two basic type of detectors used are:

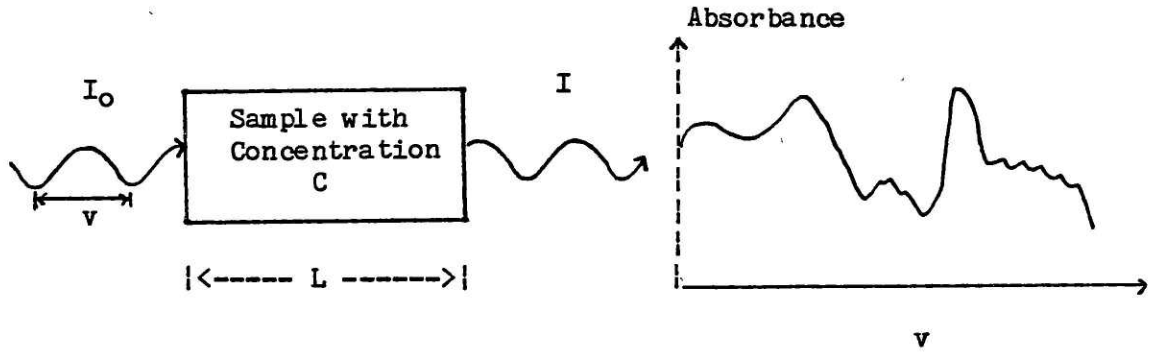
(i) bulk property detectors

(ii) solute property detectors

The first type of detector uses differential measurements of a property that is common to both the sample and the mobile phase. Examples of

bulk property detectors are differential refractometers, conductivity and dielectric constant detectors. The second type of detector measures a property that is specific to the sample. Commonly used solute property detectors that do not require removal of the mobile phase before detection are ultraviolet (UV) absorption, polarographic, and radioactivity detectors.

For the UV absorption detectors, the detector responses are governed by the the Beer-Lambert law of absorbance, which states that the light absorbance of a chemical species in solution is proportional to the concentration (C) of the sample and the optical path length (L) traversed (see Figure 2.2-4).



$$\text{Absorbance at } v = \log_{10} \frac{I_o}{I} = e(v)LC$$

Figure 2.2-4. Beer-Lambert law of absorbance.

When the concentration is expressed in moles per liter, this linear relationship between concentration and absorbance is expressed as

$$\text{Absorbance at } v = e(v)LC \qquad (2.2\text{-}4)$$

where e(v) is the molar extinction coefficient of the species at wavelength v, L is the optical path length, and C is the concentration. We shall define ABSORPTIVITY as the absorbance of a pure sample with unit concentration (C=1).

If there are two or more species in the sample cell, the total absorbance is taken to be the linear sum of the individual absorbances, i.e.

$$\text{Absorbance} = L(e_1C_1+e_2C_2+ \ldots) \qquad (2.2\text{-}5)$$

where e_i and C_i are the molar extinction coefficients and concentrations of the species.

Ancillary Subsystem

The output of the detector is logged in an analog recorder and/or may be stored in a digitized form for use in quantitative analyses. A modern HPLC typically also includes a computer system, for both control of operation and for qualitative/quantitative data analysis that further enhances the chemical identification process. Section 2.6 outlines some of the signal processing that is typically carried out by the computer in current systems. One of the background criteria used in developing the algorithms in this thesis was that they should hold the potential for efficient implementation on such a computer.

2.3 STRUCTURE OF DATA MATRICES

Analytical approaches to chemical identification problems generally involve statistical and linear algebraic methods operating on experimental results that are summarized in the form of data matrices. Typically, these matrices contain digitized measurements of certain chemical or physical properties (such as absorption spectra) of mixtures. Some examples of chemical procedures that yield data amenable to the sort of analytical development carried out in this thesis are:

- Liquid Chromatography-UV Spectrophotometry [M3, Section 2.2]
- Fluorescent Exitation-Emission Spectroscopy [J1, H2]
- Gas Chromatography-Mass Spectrometry [R2, S2]
- Nuclear Magnetic Resonance Spectrometry [K3]

The key feature of all these procedures is that they involve two bilinearly coupled sets of measurements, in a sense that will be made clear as we describe the structure of the data matrices generated by these procedures.

(a) Single (Pure) Component

Data Matrix **N**

We shall deal with the case where the experimental results for unit concentration of a (calibration) sample of a component can be summarized in a data matrix N in the bilinear or dyadic form

$$N = xy^T \qquad (2.3\text{-}1)$$

where x and y are vectors that contain the "signatures" that uniquely

define that component. N is thus of unit rank. Equation 2.3-1 is bilinear in the sense that, when one of the sets of variables (x or y) is fixed, the resulting equation is linear in the other set (y or x).

For a component with non-unit concentration, an assumption of proportionality between the absorbances and concentration is typically justified (e.g., the Beer-Lambert law in the case of LC-UVS). Thus, for a component with concentration c, the data matrix in our work will be taken to be

$$cN = cxy^T \qquad (2.3\text{-}2)$$

The following two examples show how the structure of Equations 2.3-1 and 2.3-2 arises.

Example 1. Liquid chromatography-UV spectroscopy.

In a liquid chromatography-UV spectrophotometry (LC-UVS) experiment (see Figure 2.3-1), the vector x contains the absorbance values of a unit concentration of the sample at chosen wavelengths, and the vector y contains the concentration profile of a unit concentration of the sample eluting at the end of a chromatographic column. The raw data for each run in an LC-UVS experiment consists of a collection of absorbance measurements (at a set of chosen wavelengths) made on the eluting component (at a set of chosen times). These measurements are assembled in a data matrix N, where $N_{i,j}$, the i,j-th element of N, is the j-th measurement of absorbance at the i-th wavelength (see Figure 2.3-2), with injection of unit concentration of the sample; that is, the i-th row of the absorbance matrix N is just the absorptivity of the component at the i-th wavelength, scaled by the concentration profile of the eluting sample (for injection of unit concentration). The number of

columns of N therefore equals the number of time instances at which measurements are taken, and the number of rows of N equals the number of wavelengths at which each measurement is made.

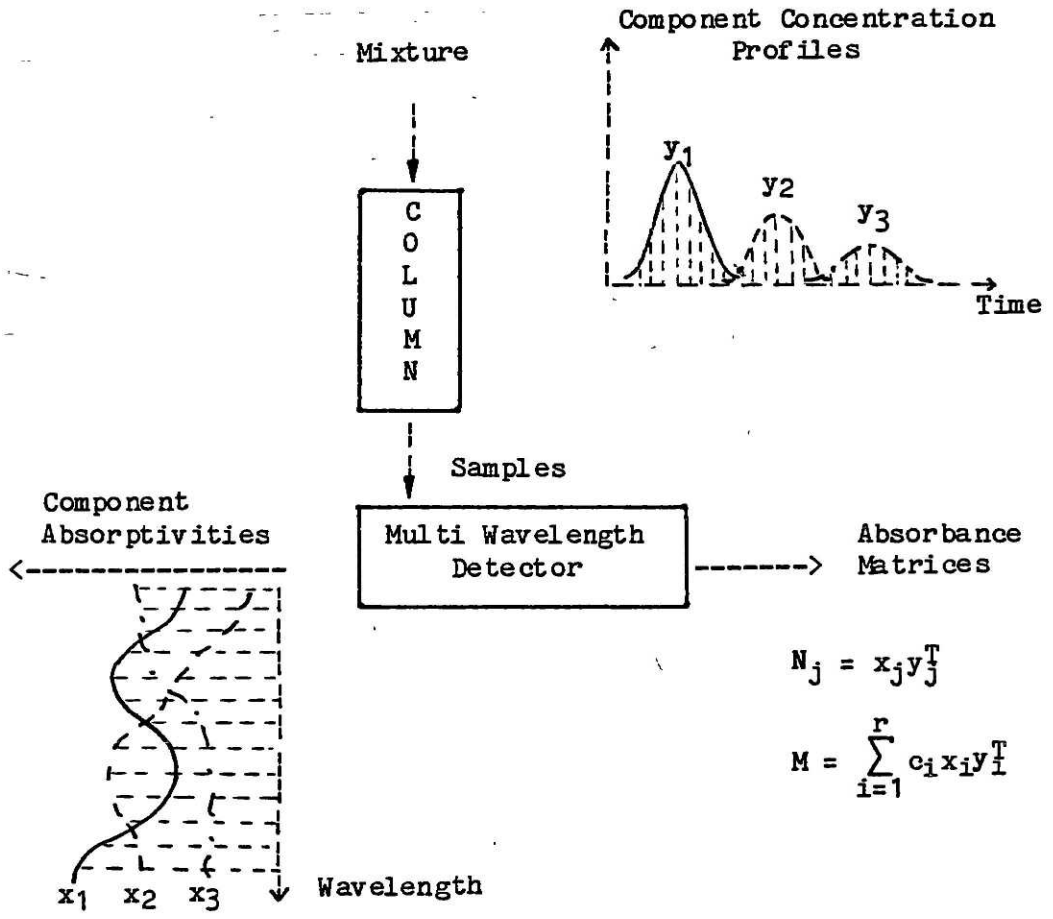
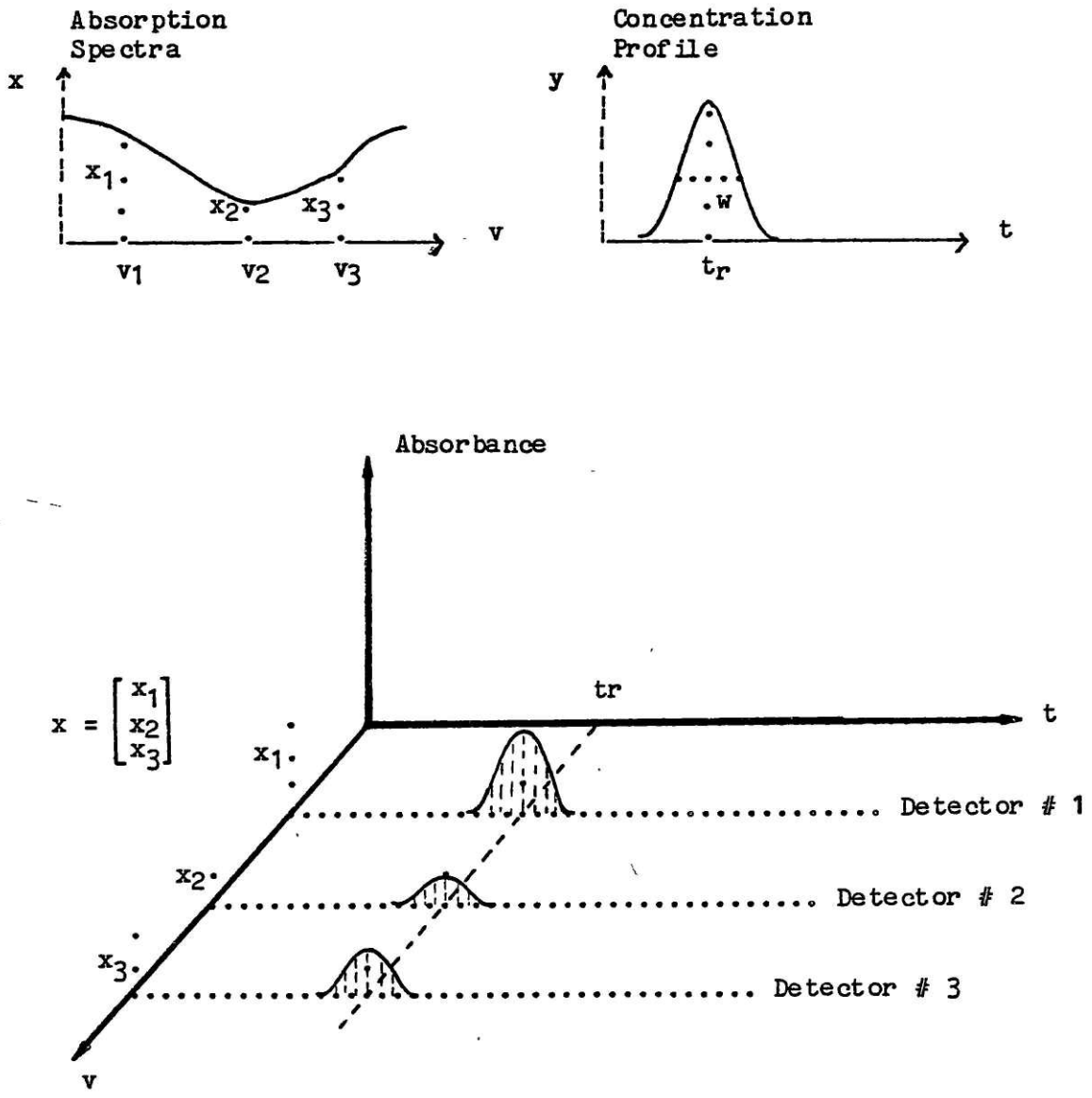


Figure 2.3-1. Data acquisition in an LC-UVS experiment.



$$N = [\, xy(t_1), \ldots, xy(t_r), \ldots, xy(t_q)] = x\, [\, y^T \,]$$

Figure 2.3-2. Formation of a pure component data matrix N in a LC-UVS experiment.

Example 2. Fluorescent excitation-emission.

For fluorescence experiments [J1], the matrix N is the Excitation-Emission Matrix (EEM), and the vectors x and y are the excitation and emission (fluorescence) spectra, respectively, of a pure component. The dimensions of the data matrix are then the numbers of wavelengths at which the sample under test are excited and detected.

(b) Multiple Components (Mixture)

Data Matrix M

For a sample containing r components, the data matrix M is taken to be of the form

$$\begin{aligned} M &= c_1 N_1 + \ldots + c_r N_r \\ &= c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T \end{aligned} \qquad (2.3\text{-}3)$$

where c_i, $i=1,\ldots,r$, are the concentrations of the r constituent components in the sample, and both $\{x_i\}$ and $\{y_i\}$ are independent sets. That is, the experimental results for a mixture are assumed to be a linear combination of the results for the individual components in the mixture, and the individual components are assumed to be linearly independent.

The independence assumption implies that

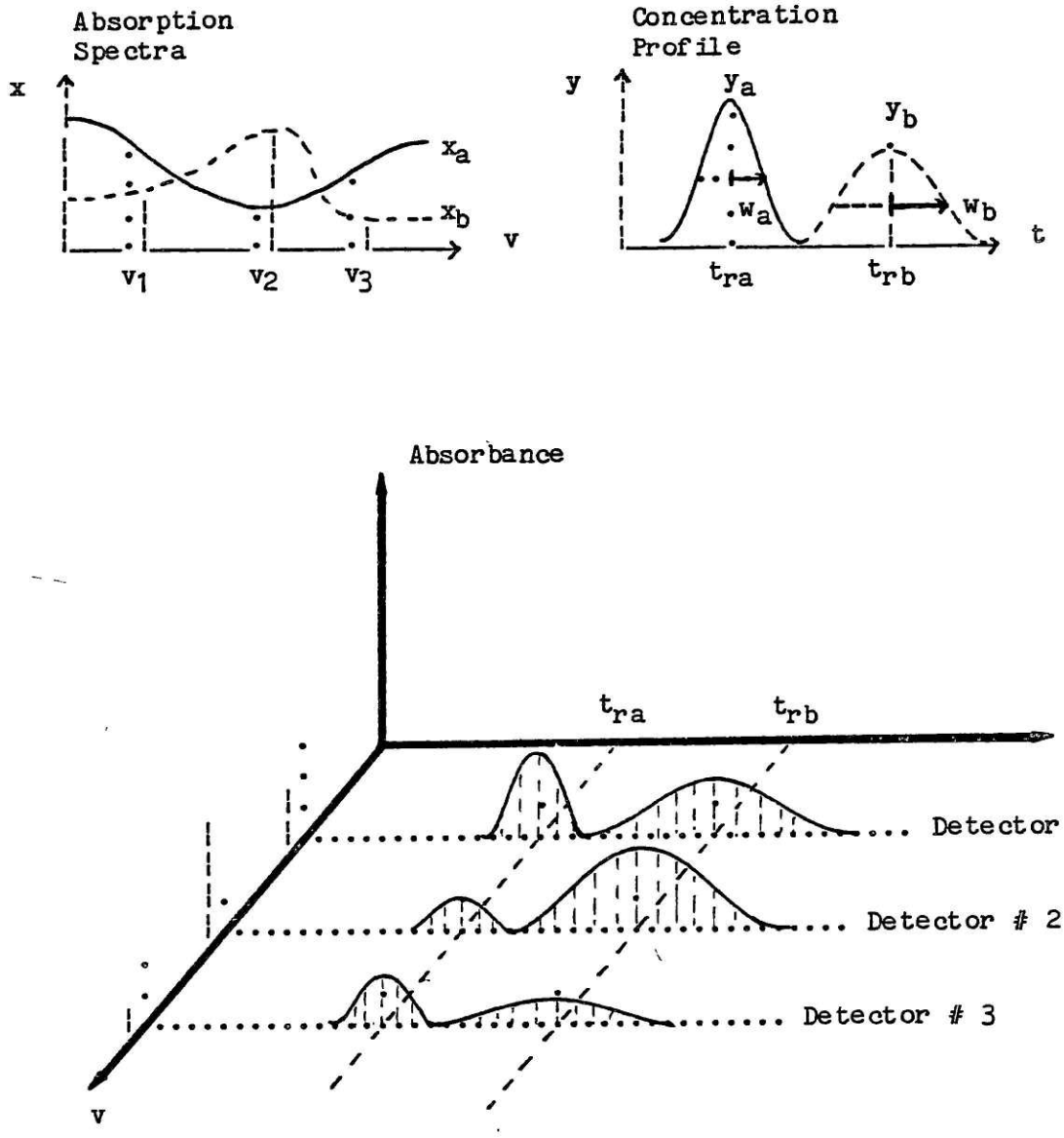
$$\text{rank of } M = r \qquad (2.3\text{-}4)$$

If the dimensions of M and N_i are pxq, then evidently a necessary condition for the independence assumption and Equation 2.3-4 to hold is that $p, q \geq r$.

Example 3. LC-UVS for a two-component mixture.

For a two-component mixture, Figures 2.3-3a,b illustrate two cases with varying degrees of peak overlap. Assuming that the two components satisfy Beer-Lambert law (proportionality between absorbance and concentration), and the mixture absorbance is a linear combination of the two individual absorbances, the mixture data matrix M is simply a linear sum of the two pure-component data matrices N_a and N_b weighted by the respective component concentrations c_a and c_b.

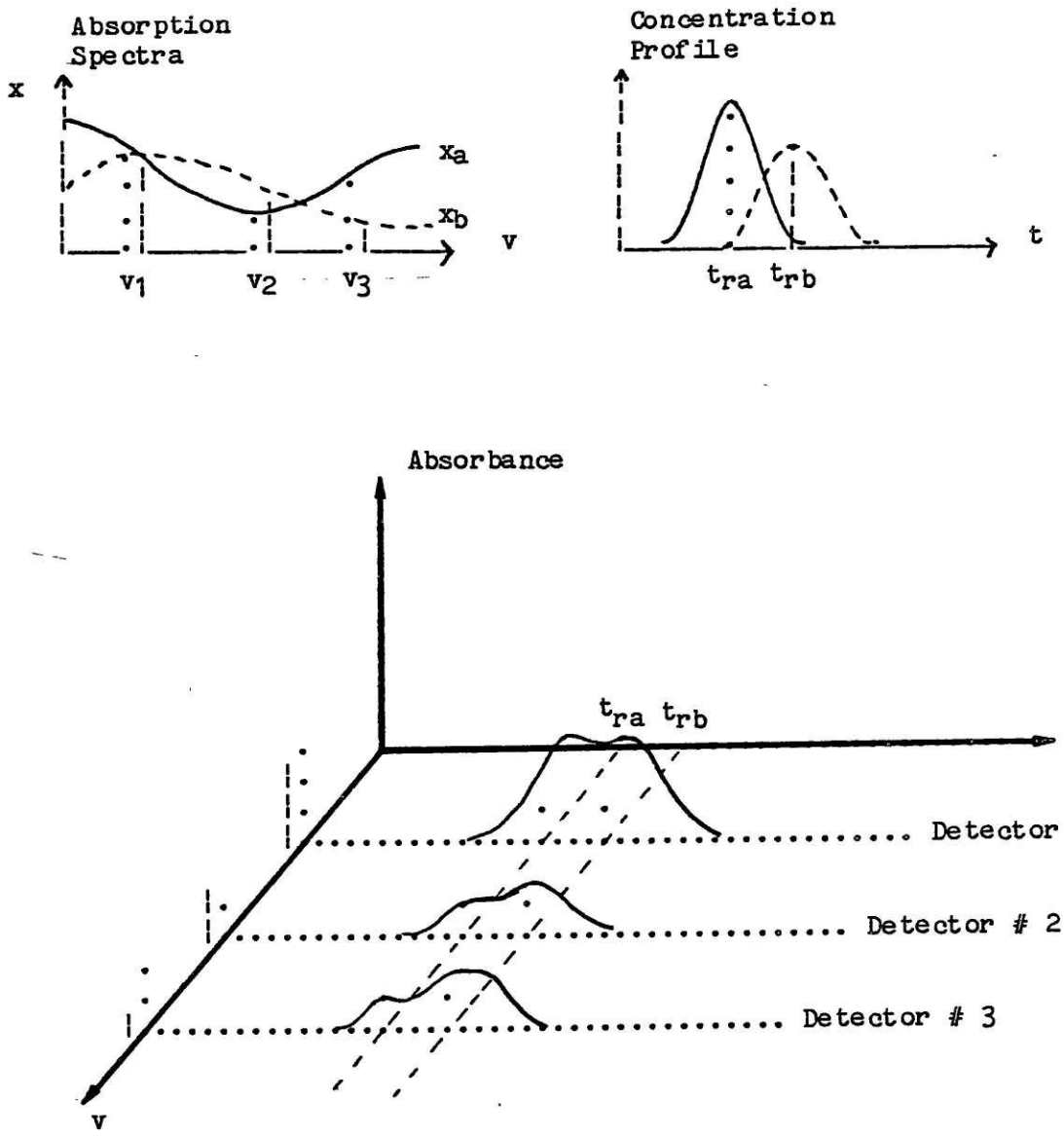
The first case is an example where the two peaks are well resolved, and the data matrix M is essentially composed of two pure-component data matrices stacked next to each other. In the second case, there is some overlap in the concentration profiles of the two components, and the columns of M contain varying degree of contributions from the two components. In both cases, the data matrix M is composed of a linear sum of the data matrices of the constituent components.



$$M = [c_a x_a y_a(t_1)+c_b x_b y_b(t_1),\ldots,c_a x_a y_a(t_q)+c_b x_b y_b(t_q)]$$

$$= c_a x_a y_a^T + c_b x_b y_b^T = c_a N_a + c_b N_b$$

Figure 2.3-3a. Formation of a two-component data matrix M in LC-UVS experiment: resolved peaks case.



$$M = [c_a x_a y_a(t_1)+c_b x_b y_b(t_1),\ldots,c_a x_a y_a(t_q)+c_b x_b y_b(t_q)]$$

$$= c_a x_a y_a^T + c_b x_b y_b^T = c_a N_a + c_b N_b$$

Figure 2.3-3b. Formation of a two-component data matrix M in LC-UVS experiment: unresolved peaks case.

2.4 MATHEMATICAL STATEMENTS OF THE PROBLEMS

(a) Target Components Problem

Here, one is given a pxq data matrix M of a mixture of unknown components, and a pxq data matrix N of a target component that is known or suspected to be in the mixture. The problem is then to estimate the concentration of the target component in the unknown mixture. Mathematically, this is stated as:

PROBLEM 1a.

Given: $M\ (= c_1 N_1 + \ldots + c_r N_r)$ and N_1

Find: c_1 □

For multiple target components, Problem 1a generalizes to:

PROBLEM 1b.

Given: $M\ (= c_1 N_1 + \ldots + c_r N_r)$ and $N_1, \ldots, N_d$, $d \leq r$

Find: $c_1, \ldots, c_d$ □

(b) Unknown Components Problem

In this problem, one tries to estimate the composition of an unknown chemical mixture from its data matrix M, using knowledge of constraints on the underlying structure. Mathematically, this is stated as:

PROBLEM 2a.

Given: $M\ (= c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T)$

Estimate: $x_i, y_i,\ i=1,\ldots,r$

Subject to: $c_i \geq 0$, $x_i \geq 0$, and $y_i \geq 0$, $i=1,\ldots,r$ □

where the inequalities in x and y are componentwise inequalities. The "estimate" may take the form of bounds on the unknowns, or may simply involve reducing the number of candidate solutions from some known database that contains all likely solutions. In liquid chromatography-UV spectroscopy experiments, the positivity constraints on c_i, x_i, and y_i reflect positivity constraints on the concentrations, absorption spectra, and concentration profiles, respectively.

It is shown in Chapter 4 that, as a direct consequence of the scaling problem inherent in our bilinear estimation problem, we can only expect to recover the x_i and y_i in Problem 2a to within some (positive) scale factors. Given these arbitrary scale factors, we can hope to simplify our solution procedure by picking some convenient normalization for the x_i and y_i. One such normalization is the so-called unit-sum normalization, where the elements of the unknown vectors are constrained to sum to one. For example, in an LC-UVS case, this is equivalent to constraining the area under the absorption spectra to be one.

A modified form of Problem 2a that places unit-sum constraints on the unknown vectors is:

PROBLEM 2b.

Given: $M (= c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T)$

Estimate: x_i, y_i, $i=1,\ldots,r$

Subject to: (i) $c_i \geq 0$, $x_i \geq 0$, and $y_i \geq 0$, $i=1,\ldots,r$

(ii) $1^T x_i = 1$ and $1^T y_i = 1$, $i=1,..,r$

□

where $1 (=[1,\ldots,1]^T)$ is a vector with all its elements equal to 1.

2.5 EFFECTS OF EXPERIMENTAL ERROR/NOISE

The effects of experimental errors, including sensor noise, will be modeled as additive terms to the underlying error free data matrices. The general form of the data matrices with the error terms will be as follows:

$$\tilde{M} = M + E_m \qquad (2.5\text{-}1a)$$

$$\tilde{N} = N + E_n \qquad (2.5\text{-}1b)$$

where M, N are the nominal error-free data matrices, and the matrices E_m, E_n are the error terms. Several sources of experimental errors are listed below, along with mathematical models used to describe them.

Short-Term Noise

The short-term or high-frequency fluctuations in the experimental results can be attributed to such sources as the natural noise levels of the transducer (sensor) elements in the detection units of an experiment. They are, generally, random in nature, and for optical transducers, exhibit strong wavelength dependence. The elements of the error matrices E_m and E_n in (2.5-1) can then be modeled as being composed of independent random variables with the following possible structures:

(i) Probabilistic structure: Gaussian profile.

The elements $E_{i,j}$ are independent and normally distributed, having zero means and variances w_{ij}^2. A special case of this model for optical transducers is

- Constant variance along the row: $w_{i,j}^2 = w_i^2$ for all j

to reflect wavelength dependence of the noise (recall that each row contains measurements at a fixed wavelength, e.g, absorbance measurements at a particular wavelength in a LC-UVS test.)

(ii) Unknown but bounded.

In this model, we assume no probabilistic structure on the elements of the noise matrix E, but the magnitudes of the elements are taken to be limited by the componentwise bounds

$$|e_{ij}| \leq d_{ij} \qquad (2.5\text{-}2a)$$

where $|.|$ indicates absolute value operator. The choices for the numerical values of d_{ij} would depend upon the available prior knowledge of the noise. As in (i), for optical transducers, we may reflect the wavelength dependence of the noise with constant bounds along the rows of the noise matrix

$$|e_{ij}| \leq d_i, \text{ all } j \qquad (2.5\text{-}2b)$$

Synchronization Error

Synchronization error refers to the error in registration results from separate but related runs of a chemical experiments. For example, consider a repeated run of a pure component in a liquid chromatography experiment using "identical" experimental conditions for each run. If one assigns the time at which the sample is injected as the starting time, the time at which the concentration profile peaks (retention time) should be the same for each run. However, due to timing errors and other nonidealities, the measured retention time will vary from experiment to experiment, and one may incorrectly infer that each run corresponds to a

different component. Synchronization error refers to the component of error that is caused by timing errors, and is reflected in improper registration of the indices of the data matrices from several experiments. Note that the synchronization error may also occur within an experiment; e.g., if more than one detector is used, such as in multiwavelength detectors, incorrect initialization of detectors would result in synchronization error within a single run.

In terms of our notation in (2.5-1), synchronization error can be written as an error in the indices of the data matrices:

$$\tilde{M}_{i,j} = M_{i+ei,\,j+ej} \qquad (2.5\text{-}3)$$

where the subscripts ei and ej indicate deviations from indices i and j of the underlying error free data matrix. The error indices ei and ej may be modeled as being random, or unknown but bounded, depending on the a priori information that is available.

2.6 CURRENT PRACTICE IN ANALYSIS OF CHROMATOGRAPHIC DATA

A brief overview of standard current practice in the analysis of chromatographic data is given here. The material presented in this section is drawn from [S3,S4,P2,B1], and represents commonly accepted techniques for identifying and quantifying unknown components. These methods are based on graphically oriented considerations of the shapes of the overlapping absorbance profiles, and do not involve matrix formulations or algorithms. As such, they are generally restricted to analyzing single channel absorbance profiles with two overlapping peaks.

Qualitative Analysis

The simplest way of identifying an unknown sample band in a chromatogram is to compare its retention time (see Figure 2.6-1) with the known retention times of substance suspected to be in the sample. If its retention time matches that of some particular substance, then there is a high probability that the unknown component is this substance. When two adjacent sample bands overlap, however, the observed "apparent" peaks do not correspond to the underlying retention times, and thus the accuracy with which the components can be identified is affected. The degree to which the true and apparent peaks deviate depends on the band size ratio and the resolution between the bands (see Figure 2.6-2). (Typically, the band size of a pure component is given by its peak height, and the resolution between two components by the formula:

$$RS_{12} = \frac{t_{r2}-t_{r1}}{2(w_1+w_2)}$$

where t_{r1}, t_{r2} $(t_{r2}>t_{r1})$ are the retention (peak) times, and w_1, w_2 are the standard deviations of the Gaussian profiles that approximate the

pure component chromatograms.) In some extreme cases, one may only observe one peak despite the presence of two components.

The standard procedure in the case of poorly separated peaks is to repeat the experiment with modified chromatographic conditions to improve the resolution. Several guidelines for updating the test conditions based on the preceding results are available in the literature (see [S3, S4]). An alternative that does not require re-experimentation is the use of correction factors to estimate the true retention times (band centers) from the apparent ones [S2]. These correction factors are discussed subsequently.

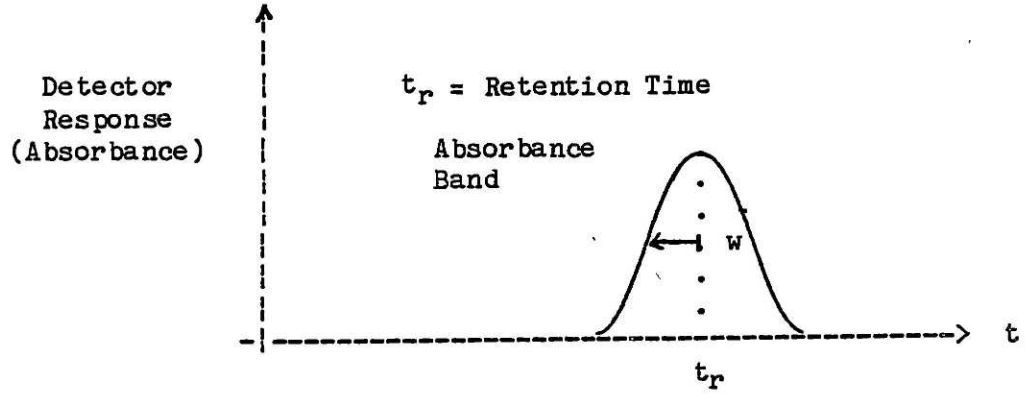
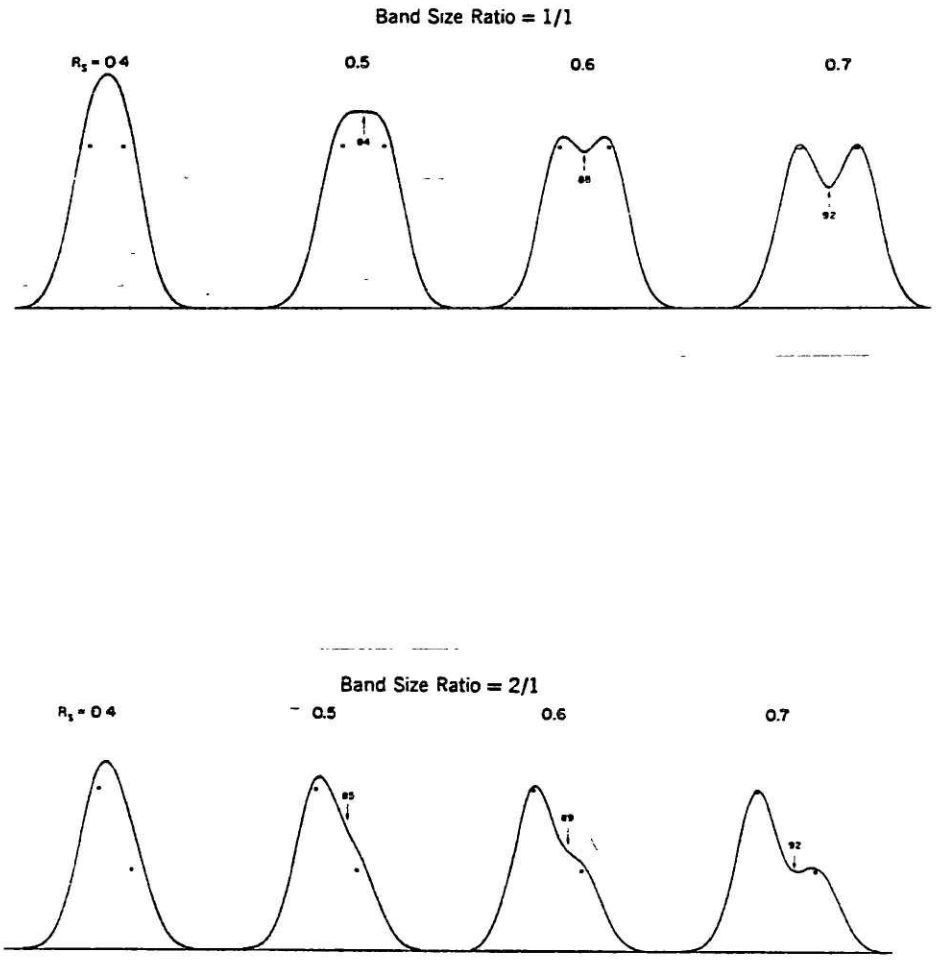


Figure 2.6-1. Chromatogram of a pure component.



The dot "•" indicates true peak.

Figure 2.6-2. Standard Resolution curves for various band-size ratios and resolutions [S3, S4]

Quantitative Analysis

Two commonly used techniques for determining the concentrations of one or more components in a mixture from its chromatogram are

- Band (Peak) Height Method, and
- Band (Peak) Area Method

These two approaches assume that there is a proportionality between the band height/area (see Figure 2.6-3) and the concentration of a given component. With this assumption, the concentrations of the components in a test sample are obtained by comparing the sample band heights/areas with that of calibration runs with known concentrations of the constituent components. Once again, difficulties arise when the adjacent bands overlap so that the band heights/areas of the constituent components cannot be determined with certainty. We discuss below several geometric constructions that are currently used to estimate the "true" band height/area from overlapping bands.

Perpendicular Drop Criterion

The most widely used geometric construction for apportioning the area between two fused peaks is to drop a perpendicular from the minimum point of the valley between the peaks (see Figure 2.6-3). The band areas for the two components are then taken to be the areas under the curves to the left and right of this division. Clearly, this technique is limited to cases where there is a valley between two peaks. Even then, as indicated in [S4,B1], the band areas determined in this fashion may contain significant errors for severely overlapping peaks.

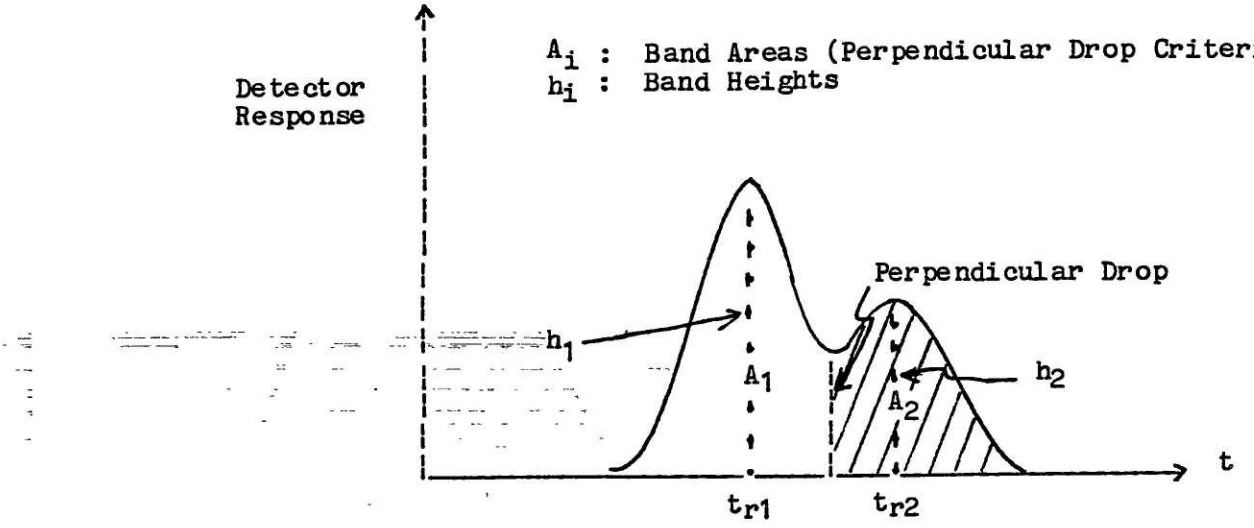


Figure 2.6-3. Quantitation by band height/area.

Tangent Separation Criterion

An alternate criterion for apportioning the areas when there is a smaller peak at the trailing edge of a larger peak is the use of tangent line separation. As illustrated in Figure 2.6-4, a line that is tangent to the valley and the trailing edge of the smaller peak is constructed; the shaded area is then taken to be the band (peak) area of the smaller peak.

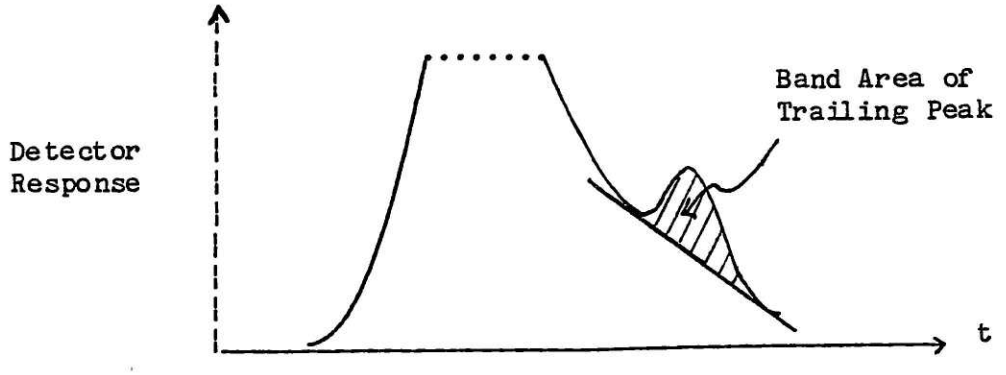


Figure 2.6-4

Peak Height: Baseline Construction

One way to improve the accuracy of an estimated band height is to reconstruct as reasonably as possible the underlying baseline of the peak in question (see Figure (2.6-5). One such construction is the tangent line discussed above; other more sophisticated alternatives are to use curve fitting techniques to find the shape of the preceding peak, which can then be subtracted from the trailing peak [B1].

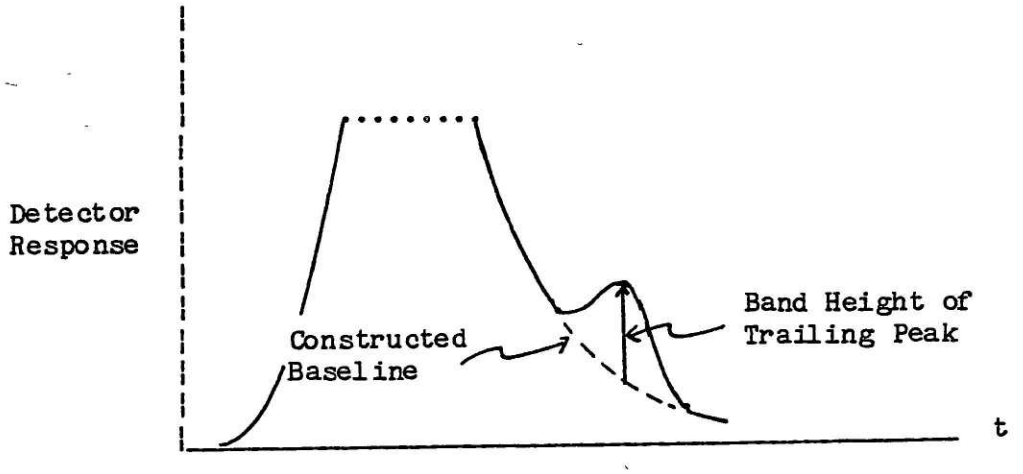


Figure 2.6-5.

Modern chromatographic systems with digital data acquisition and computational capabilities do provide some degree of automation and improved accuracy in generating peak areas, heights, and so on. However, the underlying principles and techniques used to generate these information are based on the rather simplistic geometric construction criteria discussed above. Hence current systems on the market are intrinsically limited in their ability to identify and quantify unknown

components. Furthermore, the extension of these graphical methods to multiple-wavelength data is not at all obvious. It is mainly this gap, between current practice and what seems possible via matrix formulations and algorithms, that this thesis is aimed at.

CHAPTER III: THE TARGET COMPONENTS SOLUTION

CHAPTER OUTLINE

	Page
3.1 INTRODUCTION	60
3.2 SOLUTION CHARACTERIZATION	62
3.3 PROJECTION METHOD	66
3.3.1 Noise Free Case	66
3.3.2 Noisy Data Case	76
(a) Number of Components: Rank Test	76
(b) Modes of Projections	79
- Reconstructed Data Matrix	79
- Alternate Projections	80
(c) Data Synchronization	87
3.4 GENERALIZED EIGENANALYSIS METHOD	95
3.4.1 Solution Method	96
(a) Background: Canonical Form (Kronecker Structure) of a Singular Pencil	96
- Definitions	96
- Basic Result	97
(b) Van Dooren's Algorithm: Noise-Free Case	97
- Motivation	97
- Application to Target Component Problem	99
3.4.2 Application of Van Dooren's Algorithm to Noisy Data Case	103

3.1 INTRODUCTION

The target components problem for bilinear estimation has been formulated and solved in a matrix algebraic setting by Ho et al. in a series of papers published from 1978 to 1981 [H2,H3,H4]. Their so-called "rank annihilation" method appears to be the state-of-the-art in solving the target component problem, and has been applied by McCue and Malinowski to data from liquid chromatography experiments as recently as 1983 [M3]. The method, however, requires a numerically crude and expensive search procedure with eigenvalue analysis at each step of the search, and with no guideline on the starting values. This may be a major reason why the method has not yet found its way into standard packages for analysis of chromatographic data, so methods such as those described in Section 2.6 are what are used at present.

The two methods proposed here are non-iterative and numerically stable alternatives that significantly reduce the computational load in comparison to the rank annihilation method. In addition, the inherent simplicity and clarity of these methods have suggested supplemental methodologies that reduce the effect of experimental errors, such as the data synchronization error (see Section 2.5). The two alternative methods are:

- the projection method, and
- the generalized eigenanlysis method

and are discussed in subsequent sections of this chapter.

A general outline of the chapter is as follows:

A characterization of the target components solution is stated in

Section 3.2. The two methods, projection and generalized eigenanalysis, for finding the solution characterized in Section 3.2 are then presented in Sections 3.3 and 3.4, respectively. In Section 3.3, the projection method is developed for the noise-free case first (Subsection 3.3.1), and then modified in the following subsection (Subsection 3.3.2) to address the noisy data case . In Subsection 3.4.1, we give Van Dooren's algorithm for solving the generalized eigenanalysis problem for the case of noise-free data matrices, and extend it to the case of noisy data matrices in Section 3.4.2.

Efficient algorithms for implementing the projection method are given in Appendix C. Proofs of lemmas that are not in the body of this chapter are given in Appendix F. An alternative maximum-likelihood approach to the target components problem is briefly outlined in Appendix H, and a closed form solution for a special case is given there. Appendix I details an error analysis of Van Dooren's algorithm as applied to our case, and develops guidelines to be used in making rank decisions required by the algorithm.

3.2 SOLUTION CHRACTERIZATION

Consider the pictorial representation of a two component mixture data matrix in Figure 2.3-3 (Chapter 2). If we subtract from the mixture absorbance profile the correct multiple of the absorbance profile from a calibration run of one of the components in the mixture, then the resulting absorbance profile will contain only that of the other component.

Clearly, the "correct" amount is simply the ratio of the concentration of the target component in the mixture to the concentration of the component in the calibration run. If the calibration run is done with unit concentration, the correct amount is the concentration of the target component in the mixture. Lemma 1a summarizes this result, and is a characterization of a solution to PROBLEM 1a.

As stated in Section 2.3, the pure component data matrices N_j are assumed to be of the form

$$N_j = x_j y_j^T \qquad (3.2\text{-}1)$$

where both $\{x_i\}$ and $\{y_i\}$ are independent sets. Note that, in order for the signatures $\{x_i\}$ and $\{y_i\}$ to be independent sets, the minimum dimension of the data matrices N_j (and hence M) must be greater than or equal to r, the number of components in the mixture.

Lemma 1a.

Given

$M\ (=c_1N_1+\ldots+c_rN_r)$

N_1,

c_1 is the value of t for which the composite matrix

$$E(t) = M - tN_1$$

loses rank.

Proof:

We have,

$$E(t) = M - tN_1$$

$$= (c_1 - t)N_1 + c_2N_2 \ldots + c_rN_r$$

Clearly, when t is equal to c_1, the matrix E loses rank from r to r-1.

Q.E.D.

A generalization of solution characterization to p target components case, PROBLEM 1b (Chapter 2.) is given by Lemma 1b.

Lemma 1b.

Given

$$M\ (= c_1N_1 + \ldots + c_rN_r),$$

$$N\ (= N_1 + \ldots + N_p),\ p < r,$$

$c_1, \ldots, c_p$ are the p values of t for which the composite matrix

$$E(t) = M - tN$$

loses rank.

Proof:

We have,

$$E(t) = M - t(N_1 + \ldots + N_p)$$

$$= (c_1 - t)N_1 + \ldots + (c_p - t)N_p + c_{p+1}N_{p+1} + \ldots + c_rN_r$$

Clearly, when the value of t is equal to any one of the c_j, $j = 1, \ldots, p$, the matrix E loses rank.

Q.E.D.

Remark 1: For the multi-target case, if $c_i \neq c_j$ for $i \neq j$, then the matrix E drops rank by one for each of the p values of t; otherwise, the rank drops by the multiplicity of each c_j.

Remark 2: This characterization (Lemma 1b) does not show how to associate the p values of t with the particular target components. However, if calibartion runs of the individual target components are available, one possibility is to systematically match these p values with the pure component data matrices and check for rank deficiency in each of the composite matrices E. For p target components, this requires at most p(p-1)/2 rank tests. This can be seen by noting that the first eigenvalue requires at most p-1 rank tests, the second eigenvalue p-2 rank tests, and so on, so that the total number of rank tests required for the p target components problem is

$$(p-1)+(p-2)+\ldots+(p-(p-1)) = 1+2+\ldots+(p-1) = p(p-1)/2$$

Alternatively, one can process a single p-target components problem as p single-target component problems. Because the methods developed in this thesis (see Section 3.3) for the single-target component case are numerically very efficient, this is the preferred method for the cases with large p.

Remark 3: It may be useful to deal with a generalization of Lemma 1b, in which we pick

$$N = w_1 N_1 + \ldots + w_p N_p,$$

where the w_i are weights that can be chosen to modify the numerical conditioning of the resulting problem. In this case, the values of the c_i are given by $t_i w_i$, where t_i is the appropriate choice from the set of p values of t for which E(t) drops rank.

Remark 4: The solutions characterized in Lemmas 1a,b differ from the traditional generalized eigenvalue problem in that the matrix E may be non-square, or if square, may be singular. However, the current usage of the term "generalized eigenvalue problem" following the work of Van Dooren [V3] and others [K5], includes these more general cases.

3.3 PROJECTION METHOD

The projection method will be developed in Section 3.3.1 for the error/noise free case, and subsquently modified in Section 3.3.2 to take into account the effects of experimental errors/noise.

3.3.1 Noise Free Case

Solution Methodology

The fundamental idea of the method is as follows. Given a composite matrix

$$E(t) = M - tN \tag{3.3.1-1}$$

where N is one of the components that make up M in the sense of Equation 2.3-3, we can express the composite matrix E as

$$E(t) = M_L(I-tab^T)M_R \tag{3.3.1-2}$$

Here M_L and M_R are the factors in a decomposition (see Appendix A) of the matrix M, yielding the row and column spaces of M, and a and b map M_L and M_R down to the row and the column spaces, respectively, of N. Once the composite matrix is decomposed into the form given by Equation 3.3.1-2, the value of t that drops the rank of E is given by the value of t that drops the rank of

$$I - tab^T \tag{3.3.1-3a}$$

or of

$$(1/t)I - ab^T \tag{3.3.1-3b}$$

That is, the value of t for which E(t) drops rank is given by the

reciprocal of the (only) nonzero eigenvalue of ab^T. One can find this value of t by solving

$$|I-tab^T| = 0 \qquad (3.3.1-4)$$

where $|.|$ denotes the determinant operator. Using the fact [K1] that

$$|I-tab^T| = |1-tb^Ta| \qquad (3.3.1-5)$$

we see that t is simply given by

$$t = 1/(b^Ta) \qquad (3.3.1-6)$$

A detailed derivation of this result is given in the proof of Lemma 2a (see Appendix F).

The rank annihilation method of Ho et al. [H4] reduces the target component problem to essentially the form given by Equation 3.3.1-3a (although not in the bilinear form, ab^T), but fails to recognize that it can be further simplified to give the solution in Equation (3.3.1-6). Instead, their method literally searches for the value of t (with no guidance on the starting and incremental values) for which the composite matrix (in Equation 3.3.1-3a) drops rank. Furthermore, the rank determination is done by eigenanaysis, which is a numerically expensive way of doing it.

Summarizing the above result, the projection method for finding a solution t, characterized by Lemma 1a, is given by Lemma 2a:

Lemma 2a.

Find matrix decompositions

$$M = M_L M_R^T$$

$$N_1 = n_L n_R^T$$

where M_L, M_R are full column rank matrices, and n_L, n_R are vectors. Then $E(t)$ $(=M-tN_1)$ loses rank for

$$t = \frac{1}{b^T a}$$

where, a and b are the unique solutions to the equations

$$n_L = M_L a$$

$$n_R = M_R b$$

[]

A generalization of the projection method to p target components, as characterized by Lemma 1b, is given by Lemma 2b.

Lemma 2b.

Find matrix decompositions

$$M = M_L M_R^T$$

$$N = N_L N_R^T$$

where M_L, M_R, N_L, N_R are full column rank matrices. Then $E(t)$ $(=M-tN)$ loses rank when t is equal to the reciprocals of the eigenvalues of the matrix

$$B^T A$$

where, A and B are the unique solutions to the equations

$$N_L = M_L A$$

$$N_R = M_R B$$

□

Comment: Unlike the rank annihilation method, the projection methodologies outlined in Lemmas 2a,b are direct "one-shot" procedures that do not require a search method. For a single-target component problem, the projection method requires at most two matrix decompositions (which at worst is equivalent to two eigenvalue problems) and two matrix inversions. In contrast, the rank annihilation method, in addition to a matrix decomposition, requires rxr eigenanalyses at each step of the search ! Even with an efficient search method (say, convergence in less than 10 steps), the potential for dramatic saving in computational requirements is apparent with the projection method. For multiple target components, the savings are even more dramatic with the projection method.

We illustrate the projection method with two examples. For the first example, we use data matrices of dimensions 3x3 that have been synthesized to simplify the numerical computations. For the second example, published results [M3] of absorbance spectra and chromatograms from a liquid chromatograph-UV spectrophotometry experiments are used to simulate absorbance profiles. Absorbances at five wavelengths and seven sampling times are used, resulting in data matrices of dimensions 5x7.

Example 1.

Given data matrices

$$N_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \text{ and}$$

$$M = \begin{bmatrix} 0 & 2 & 0 \\ 1 & 1 & 0 \end{bmatrix} \quad (=c_1 N_1 + c_2 N_2),$$

find c_1, the concentration of target component in the mixture.

Solution:

Decompositions of the data matrices N_1 and M are, for example, given by

$$N_1 = \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{n_L} \underbrace{\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}}_{n_R^T} \quad \left(= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right)$$

$$M = \underbrace{\begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}}_{M_L} \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}}_{M_R^T} \quad \left(= \begin{bmatrix} 0 & 2 & 0 \\ 1 & 1 & 0 \end{bmatrix}\right)$$

(The decompositions show that N_1 has rank 1 and M has rank 2.)

From Lemma 2a, we need to solve

$$n_L = M_L a$$

$$n_R = M_R b$$

for a and b. The unique solutions of the above equations are

$$a = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$$

$$b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

from which we have

$$t = \frac{1}{b^T a} = \frac{1}{\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}} = 2$$

Checking the solution,

$$E(2) = M - 2N_1$$

$$= \begin{bmatrix} 0 & 2 & 0 \\ 1 & 1 & 0 \end{bmatrix} - 2 \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

which has rank 1.

Note: Matrix M was indeed synthesized from pure component data matrices N_1 and N_2 as follows

$$M = 2N_1 + N_2$$

$$= 2 \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 2 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

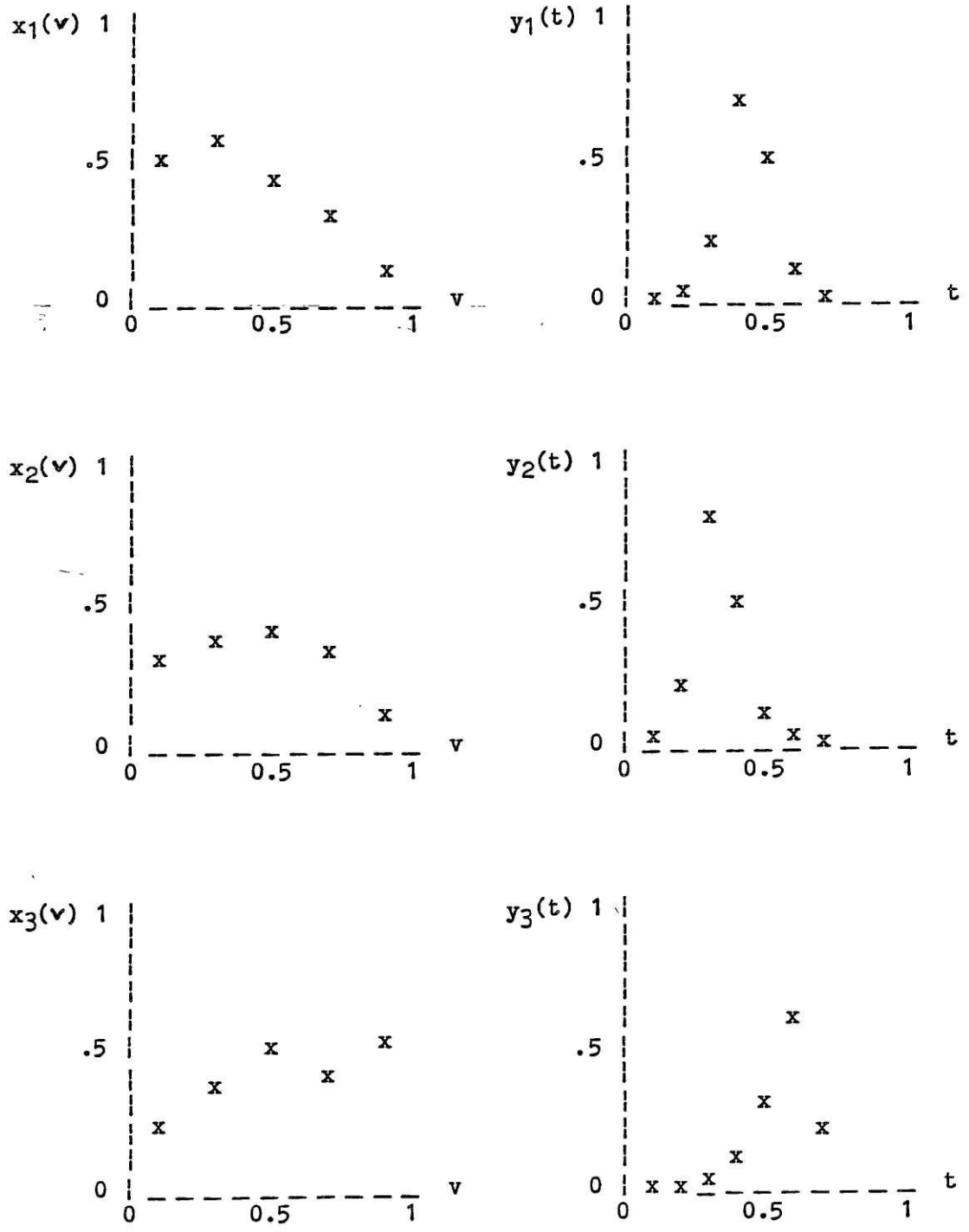
Example 2. Application of projection method to synthesized liquid chromatography data.

The absorbance spectra and concentration profiles shown in Figures 3.3.1-1a and 3.3.1-1b were obtained from Figures 1 and 3 in [M3] by selecting appropriately chosen wavelength and time samples. (The wavelength and time scales in our figures were reset to range from zero to one.) Absorbance profiles were then generated from the absorbance spectra and the concentration profiles shown in Figures 3.3.1-1a and 3.3.1-1b, respectively, as follows

$$M = c_1 x_1 y_1^T + c_2 x_2 y_2^T + c_3 x_3 y_3^T \; ; \; c_1 = c_2 = c_3 = 1$$

$$N = x_1 y_1^T$$

where x_i are the absorption spectra, y_i are the concentration profiles, and c_i are the concentrations.



(a) Absorbance spectra (b) Concentration profile

Figure 3.3.1-1. LC-UVS experimental data [M3]

The mixture data matrix M and the target component data matrix N are summarized in Figures 3.3.1-1c and 3.3.1-1d, respectively, and represent absorbances of seven sample points at five wavelengths, resulting in data matrices of dimensions 5x7.

Application of the projection method to data matrices M and N in Figures 3.3.1-1 correctly computed the concentration (=1) of the target component in the mixture to eight significant digits. McCue and Malinowski [M3] also obtained (using the method of [H4]) results that are accurate to within the experimental errors. Although direct comparison between our results and that of [M3] is difficult since we have used synthesized data matrices (see Chapter 5 for results with real data), we should note at this point that the results obtained by [M3] required (as best as can be deduced from the paper) a minimum of 30 (!) eigenanalysis. Whereas our method required 2 singular value decomposition, which at worst is equivalent to 2 eigenanalysis.

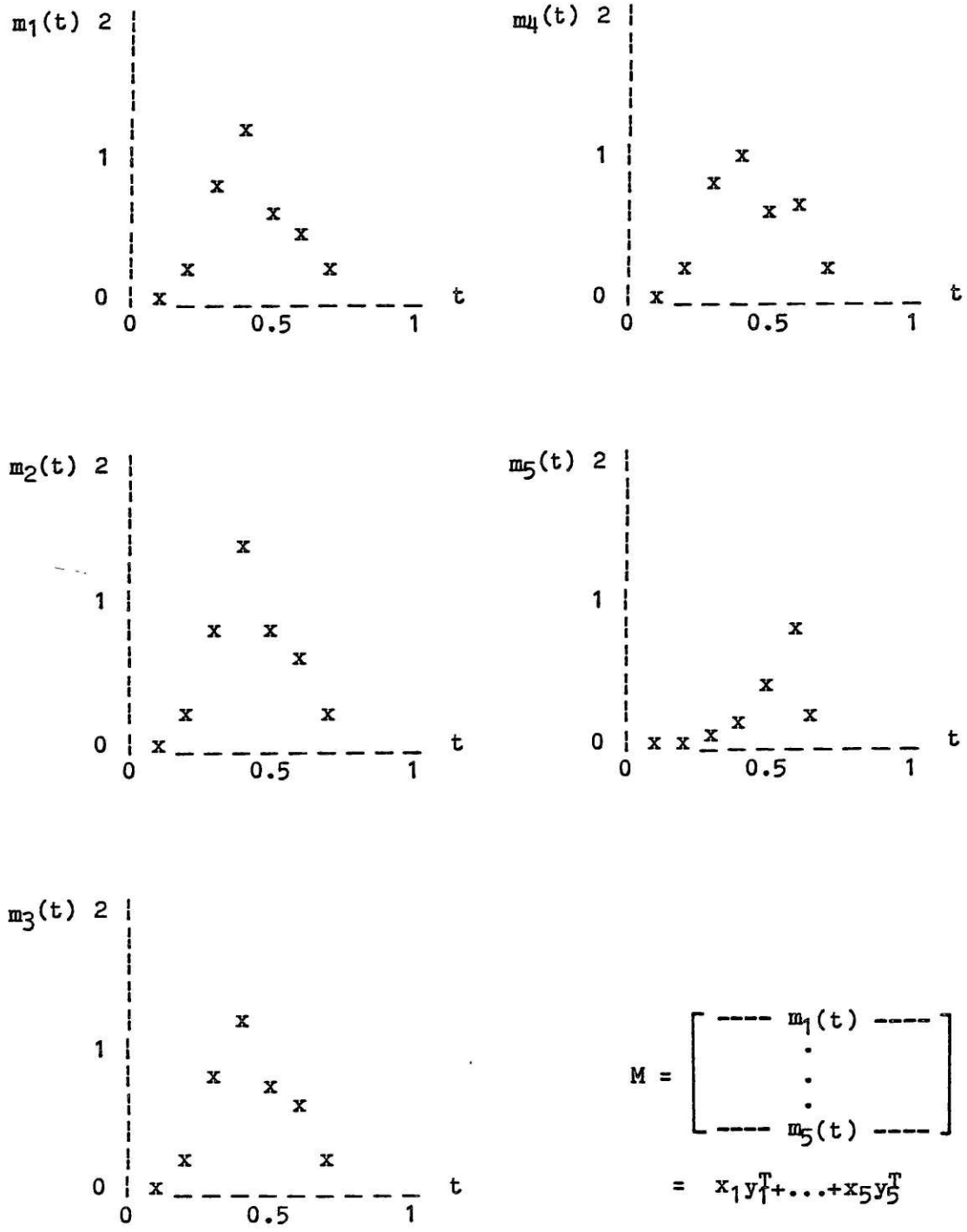


Figure 3.3.1-1c. Synthesized mixture absorbance profile.

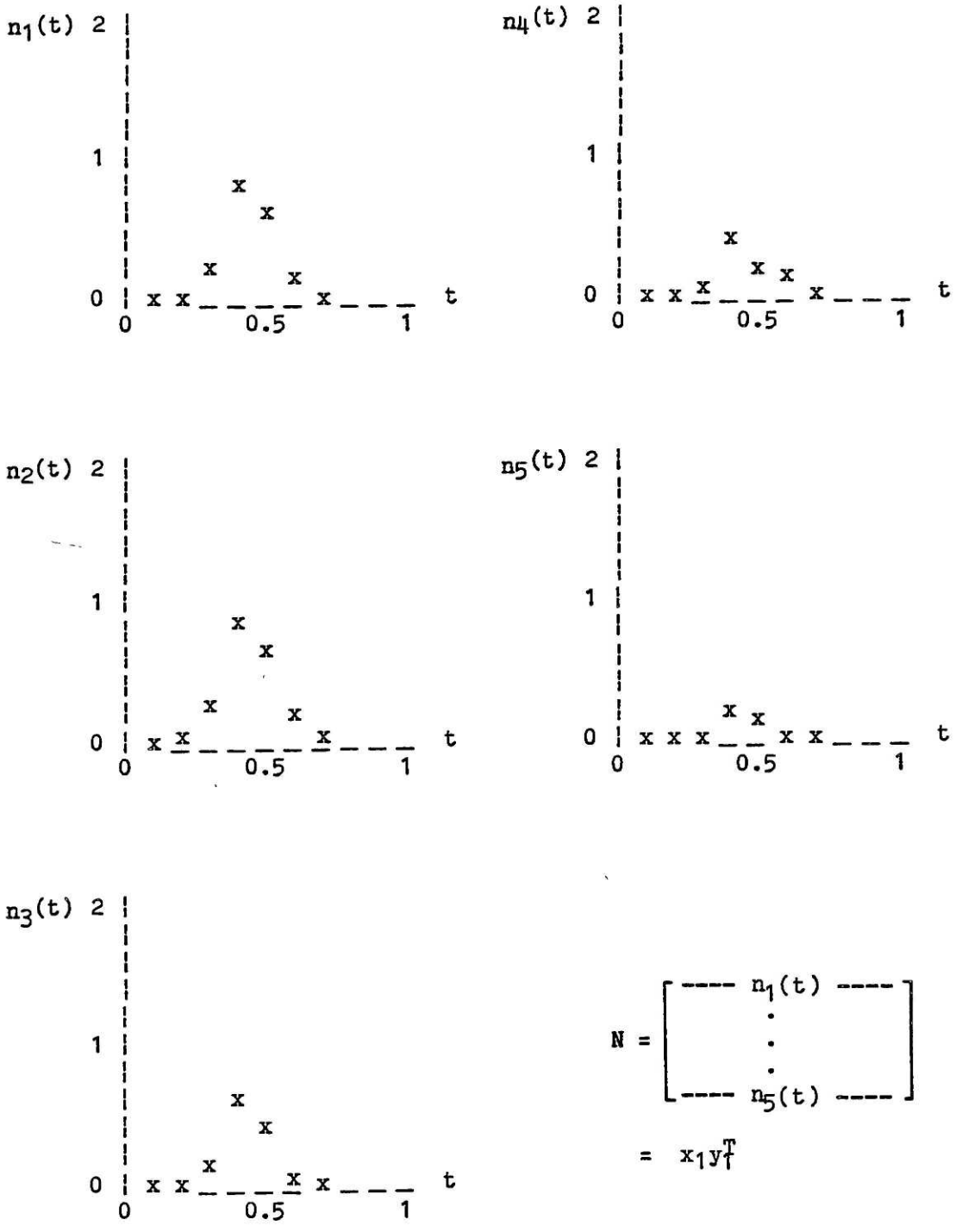


Figure 3.3.1-1d. Synthesized target component absorbance profile.

3.3.2 Noisy Data Case

The projection methodology developed in Section 3.3.1 is modified in this section to take into account the effects of experimental error/noise. In subsection (a), we address the problem of determining the number of components in a mixture by analyzing its noisy data matrix. In subsection (b), we present several methods of "filtering" the noisy data matrices to approximate the underlying noise free data matrices in a way that is conformable to the projection method. In subsection (c), we give a method to correct for data asynchronization.

(a) Number of Components: Rank Test

Determining the number of components in a mixture is the critical step underlying our solution of the target components problem. Given a noise free data matrix M whose minimum dimension is not less than the number of components in the mixture, the number of components in the mixture is given by the rank of M (under our standing independence assumptions). Several methods, such as the QR decomposition and the singular value decomposition (SVD), may be used to determine the rank of a matrix (see Appendix A, [K2,G6]). When the ideal (infinite precision) SVD is used, the rank is given by the number of nonzero singular values.

If the data matrix $\tilde{M}$ (=M+E) contains noise, however, $\tilde{M}$ and M can have different numbers of nonzero singular values. In fact $\tilde{M}$ will in general have all nonzero singular values, due to the random nature of the elements of the noise matrix E. The number of nonzero singular values of $\tilde{M}$ therefore does not in general correspond to the number of components in the mixture.

Fortunately, when the signal-to-noise ratio is large, SVD provides

an effective means for determining the rank of the underlying noise free data matrix as a consequence of the following theorem from matrix theory.

Theorem 1 (see [S10])

Let

$$\tilde{M} = M + E$$

then

$$|\tilde{s}_i - s_i| \leq ||E||_2, \text{ all } i$$

where $||.||_2$ indicates spectral norm, $|.|$ indicates absolute value, and $\tilde{s}_i$ and s_i are the singular values of $\tilde{M}$ and M, respectively. []

Hence, when one has some prior information on the magnitude of the error matrix, such as a bound d on its norm, the rank of the underlying matrix M can be taken as the number of singular values of $\tilde{M}$ greater than d (or some small multiple of d, to account for the uncertainty in the prior information on the noise). This is because, in view of Theorem 1, singular values of $\tilde{M}$ not greater than d may have arisen from adding the noise matrix E to an M that had corresponding singular values equal to zero. A complementary statement is that, even if these singular values of M were nonzero, they are indistinguishable from noise on the basis of the given data matrix $\tilde{M}$.

Similarly, in the case of finite precision computation, the zero singular values of M may be computed as being nonzero (even in the absence of non-numerical noise). In such cases, as in the case of noisy data, a threshold (conventionally denoted by) EPS is used to determine its so-called <u>numerical rank</u> [G7], where EPS=(=p.e.$||M||_2$), with p being a low-order polynomial expression in the dimensions of the data matrix

M, e the machine precision of the computer, and $||.||_2$ the spectral norm [W5].

Alternate methods for determining the number of components have been addressed in the literature. Most of the criteria used in these methods were developed within the framework of factor analysis and principal components analysis, and stress the statistical approaches to the problem. These methods generally require initial eigenanalysis of the data covariance matrix $\tilde{M}^T\tilde{M}$ (the eigenvalues of the data covariance matrix are the squares of the singular values of the data matrix $\tilde{M}$). Prior information one has on the variances of the elements of the error matrix are then used to develop criteria for determining which of the eigenvalues are significant. Thus, the final outcome of most of these alternate approaches is again essentially the selection of a threshold below which singular values of $\tilde{M}$ are neglected. An extensive review of many of these methods is given in [M1]. We list some of these methods below, under three major categories, each with different prior information on the noise matrix. In all cases, the elements of the noise matrix are assumed to be independent of one another.

(i) Variances of all the elements of the error matrices are assumed to be known and equal:

- Residual standard deviation [M1]
- Root mean square error [M1]

(ii) Variances of all the elements of the error matrices are assumed to be known, but are allowed to be unequal:

- Chi-squared method [B3]
- Standard error in the eigenvalue [H5]

(iii) Unknown variances:

- Scree test [C1]
- Imbedded error function [M1,C3]

Comment: As noted in [D3], no single method is superior or completely satisfactory on its own. As such, various methods should be examined together to provide better guidance in determining the "true" rank of a matrix (and hence, in our case, the number of components in a mixture). Also, the direct application of an SVD algorithm to $\tilde{M}$, rather than principal component analysis of $\tilde{M}^T\tilde{M}$, is often much more efficient and direct, but is hardly mentioned in the chemometrics literature; it deserves to be better known.

(b) Modes of Projections

We have shown in the preceding subsection how to determine the number of components in a mixture via a rank test on its data matrix. We give in this subsection a method for approximating (in a sense to be made clear below) the underlying noise-free data matrix from a noisy measurement data matrix. Working with the so-called reconstructed data matrix (see below), we then describe various alternate procedures for solving the target components problem with noisy data matrices.

Reconstructed Data Matrix

Given a noisy data matrix $\tilde{M}$, assume that we have determined its "true" rank r (see preceding subsection). Let $\hat{M}$ be a matrix of rank r that has been reconstructed from $\tilde{M}$ in the manner described in the following theorem:

Theorem 2 [E1,B2,G6]

Given the mxn matrix $\tilde{M}$ with SVD

$$\tilde{M} = USV^T$$

$$S = \text{diag}\,(s_1,\ldots,s_m),\ s_1 \geq s_2 \geq \ldots \geq s_m,$$

a matrix $\hat{M}$ of rank r, $r<m$, that minimizes the norm

$$||\tilde{M}-\hat{M}||_2$$

is given by

$$\hat{M} = US_oV^T$$

where

$$S_o = \text{diag}\,(s_1,\ldots,s_r,0,\ldots,0)$$

□

We shall call $\hat{M}$ the **RECONSTRUCTED DATA MATRIX.** (A similar procedure applied to a noisy target component data matrix N_1 produces a reconstructed target component data matrix.)

Comment: The reconstructed data matrix $\hat{M}$ is optimum in the sense that it minimizes the norm of the error matrix. It is in effect obtained by subtracting from the noisy data matrix $\tilde{M}$ the contributions of those nonzero singular values $\tilde{M}$ that are of the same order of magnitude as the noise.

Alternate Projections

In characterizing the target components solution (Lemma 1), it is assumed that the target component is in the mixture. In developing the projection methods (Lemma 2) for the noise-free case, this assumption translates into assuming that the row and column spaces of the target component data matrix N_1 are respectively contained in the row and

column spaces of the mixture data matrix M.

When the data matrices are noisy, however, the row and column spaces of the reconstructed target component data matrix $\hat{N}_1$ do not quite lie in the row and column spaces, respectively, of the reconstructed mixture data matrix $\hat{M}$, even if the target component is in the mixture (see Figure 3.3.2-1).

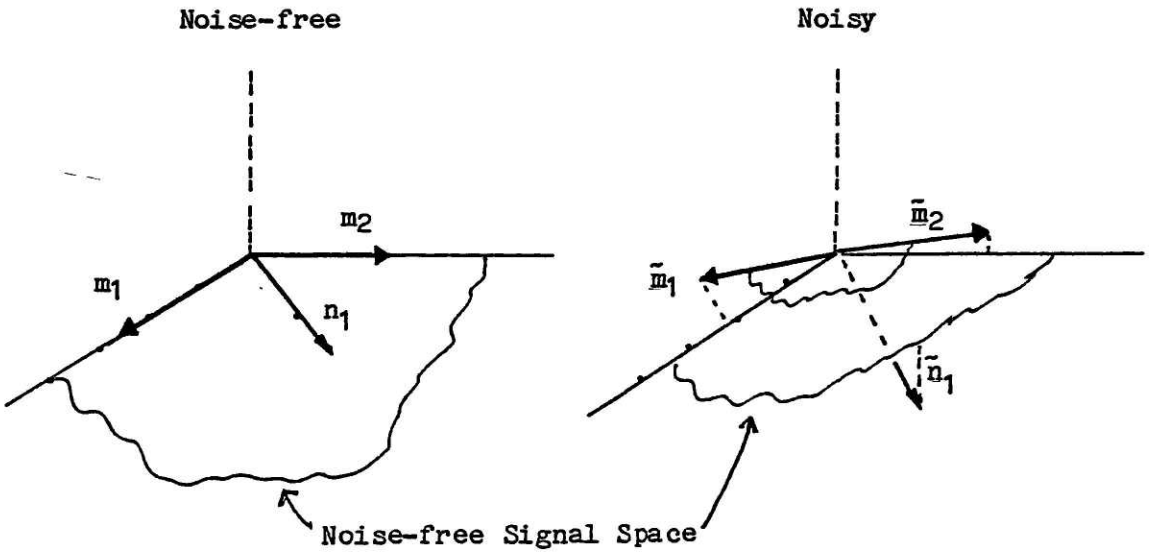


Figure 3.3.2-1. Spaces spanned by columns of noise free and noisy data matrices.

Thus, given matrix decompositions

$$\hat{\underline{M}} = M_L M_R^T \qquad (3.3.2\text{-}1a)$$

$$\hat{\underline{N}}_1 = n_L n_R^T \qquad (3.3.2\text{-}1b)$$

where $\hat{\underline{M}}$ and $\hat{\underline{N}}_1$ are reconstructed noisy data matrices, we now can only write

$$n_L+e_L = (M_L+E_L)a \qquad (3.3.2\text{-}2a)$$

$$n_R+e_R = (M_R+E_R)b \qquad (3.3.2\text{-}2b)$$

where e_L, e_R, E_L, E_R are deviations of the nominal noise-free values from the computed values. The problem is then to look for further approximations to the reconstructed data matrices $\hat{M}$ and $\hat{N}_1$ so that the row/column space of the reconstructed target component data matrix $\hat{N}_1$ is contained in the row/column space of the reconstructed mixture data matrix $\hat{M}$.

A convenient formulation of this problem is given by:

PROBLEM 3: Generalized Bilinear Perturbation Problem (GBPP)

Given data matrices $X \in R^{mxr}$, $Y \in R^{rxn}$, and $Z \in R^{mxn}$,

$$\underset{(E_X, E_y, E_Z)}{\text{Minimize:}} \quad tr(E_X^T P E_X)+tr(E_y^T Q E_y)+tr(E_Z^T R E_Z)$$

$$\text{Subject to:} \quad (Z+E_Z)=(X+E_X)(Y+E_y)$$

for some symmetric, positive semidefinite weighting matrices P, Q, and R.

□

A set of necessary conditions for a solution to Problem 3 is given in Appendix B; solutions to three special cases are given below.

(i) Setting $E_X=0$, $Y=0$, and $Q=0$, Problem 3 reduces to:

Problem 3a: Linear Least Squares (LLS) Problem

Given data matrices Z and X,

$$\underset{(E_y, E_Z)}{\text{Minimize:}} \quad tr(E_Z^T R E_Z)$$

$$\text{Subject to:} \quad Z+E_Z=XE_y$$

Solution: [S6]

When X has full column rank, and R is symmetric, positive definite,

$$E_y=(X^TRX)^{-1}X^TRZ$$

$$E_z=X(X^TRX)^{-1}X^TRZ-Z$$

More generally,

$$E_y=X^+Z$$

$$E_z=X(X^+)-Z$$

where X^+ is the R-weighted pseudoinverse of X [B5]. □

(ii) Setting Y=0 and Q=0, Problem 3 reduces to:

Problem 3b: Total Least Squares (TLS) Problem

Given data matrices X and Z,

$$\underset{(E_X, E_y, E_Z)}{\text{Minimize:}} \quad tr(E_X^TPE_X)+tr(E_Z^TRE_Z)$$

$$\text{Subject to:} \quad (Z+E_Z)=(X+E_X)(E_y)$$

Solution: (See [G4,G6] for proof)

For P=R=I, and $X^TZ \neq 0$ (where "0" is a matrix with all its elements equal to zero), let

$$C = \underset{n}{[X}, \underset{k}{Z}] \; m, \qquad m \geq n+k$$

Given the SVD of C

$$C = USV^T$$

where

$$U = \underset{n}{[U_1}, \underset{k}{U_2}] \; m \qquad V = \begin{bmatrix} V_{11} & V_{12} \\ \hline V_{21} & V_{22} \\ n & k \end{bmatrix} \begin{matrix} n \\ k \end{matrix} \qquad S = \begin{bmatrix} S_1 & 0 \\ \hline 0 & S_2 \\ n & k \end{bmatrix} \begin{matrix} n \\ k \end{matrix}$$

if $s_n > s_{n+1}$, then

$$[E_X, E_Z] = -U_2S_2[V_{12}^T, V_{22}^T] = -C \begin{bmatrix} V_{12} \\ V_{22} \end{bmatrix} [V_{12}^T, V_{22}^T]$$

$$E_y = -V_{12}V_{22}^{-1}$$

are the unique solutions to the problem.

Note: An alternative form of solution for E_y when k=1 is

$$e_y = (X^TX - s^2_{n+1}I)^{-1}X^Te_z$$

where s_{n+1} ($<s_n$) is the smallest singular value of [X,z]. []

(iii) Setting Y=0, Q=0, R=0, E_Z=0, and k=1, Problem 3 reduces to:

Problem 3c: Complementary Least Squares (CLS) problem

$$\underset{(E_X, e_y)}{\text{Minimize:}} \quad tr(E_X^T P E_X)$$

$$\text{Subject to:} \quad z=(X+E_X)e_y$$

Solution: (See Appendix B for details)

For the case when [X,z] has full column rank, P is symmetric, positive definite, and $X^Tz \neq 0$, let

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \begin{matrix} n \\ 1 \end{matrix}$$

be the eigenvector corresponding to the smallest eigenvalue of the pencil

$$\left[\begin{array}{c|c} X^TPX & X^TPz \\ \hline z^TPX & z^TPz \end{array}\right] - t\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]$$

where

$$\left[\begin{array}{c|c} X^TPX & X^TPz \\ \hline z^TPX & z^TPz \end{array}\right]$$

is a symmetric, positive definite matrix. Then the unique solutions to the problem are

$$e_y = -v_1/v_2$$

$$E_X = \frac{(z-Xe_y)e_y^T}{e_y^Te_y}$$

with $v_2 \neq 0$ and $e_y^T e_y \neq 0$. □

Any one of the three methods of projection (data fitting) detailed above (Problems 3a-c) may be used to satisfy Equation 3.3.2-2, so that the row/column space of the reconstructed target component data matrix $\hat{N}_1$ is contained in the row/column space of the reconstructed mixture data matrix $\hat{M}$. Applying LLS and TLS projection methods to the noisy single-target component problem yields:

For the linear least squares projection,

$$a_{LLS} = (M_L^T M_L)^{-1} M_L^T n_L \qquad (3.3.2\text{-}3a)$$

$$b_{LLS} = (M_R^T M_R)^{-1} M_R^T n_R \qquad (3.3.2\text{-}3b)$$

where M_L, M_R, n_L, n_R, a and b are as in Equations 3.3.2-1,2, and

for the total least squares projection,

$$a_{TLS} = (M_L^T M_L - s_a^2 I)^{-1} M_L^T n_L \qquad (3.3.2\text{-}4a)$$

$$b_{TLS} = (M_R^T M_R - s_b^2 I)^{-1} M_R^T n_R \qquad (3.3.2\text{-}4b)$$

where s_a and s_b are the smallest singular values of $[M_L, n_L]$ and $[M_R, n_R]$ respectively.

The concentrations of the target component for the noisy data case using LLS and TLS projections respectively are then given by

$$t_{LLS} = 1/(b_{LLS}^T a_{LLS}) \qquad (3.3.2\text{-}5a)$$

$$t_{TLS} = 1/(b_{TLS}^T a_{TLS}) \qquad (3.3.2\text{-}5b)$$

Note: Several efficient algorithms for implementing the above solutions are detailed in Appendix C.

Example 3. Application of the projection method to noisy synthetic data.

Noisy data matrices $\tilde{M}$ and $\tilde{N}$ were obtained from the noise free data in Example 2 as follows:

$$\tilde{M} = c_1 x_1 y_1^T + c_2 x_2 y_2^T \quad c_3 x_3 y_3^T + E_m$$

$$\tilde{N} = x_1 y_1^T + E_n$$

where the elements of the noise matrices E_m and E_n were chosen to be independent random variables drawn from a normal distribution with zero mean and standard deviation w. The results of the application of the projection method to the noisy data matrices $\tilde{M}$ and $\tilde{N}$ for $c_2=c_3=1$, and various values of c_1 and w are summarized in Table 3.3.2-1 for the linear least squares (LLS) projection and the total least squares (TLS) projection.

		LLS		TLS	
c_1	w	$\hat{c}_1^*$	s.d.	$\hat{c}_1^*$	s.d.
1	0.005	1.004	0.025	1.002	0.025
1	0.01	1.013	0.048	1.005	0.049
0.5	0.005	0.506	0.022	0.501	0.022
0.5	0.01	0.529	0.040	0.513	0.042
0.1	0.005	0.137	0.023	0.103	0.022
0.1	0.01	0.395	0.358	0.104	0.052

s.d. = Standard Deviation

* 99 sample average

Table 3.3.2-1. Projection method solution to target component problem with noisy data.

As shown in Table 3.3.2-1, the errors in the estimated values of the target component, and the associated standard deviations increase with increasing noise level. Of the two projections used (LLS and TLS), the TLS projection yielded more accurate estimates (approximately 50% less errors with TLS). The effect of decreasing the concentration of the target component in the mixture (relative to the other components in the mixture) for a fixed noise level was to increase the error in the estimated value for both the LLS and the TLS projections.

(c) Data Synchronization

We address the problem of synchronization error (of the form given by Equation 2.5-3) for the projection method:

Suppose the noise-free data matrices for a mixture and a target component are M and N_1 respectively, and assume that the target component is in the mixture, i.e., the row and column spaces of N_1 are contained in the row and column spaces of M (we have assumed here that the dimensions of the data matrices are not less than the number of components in the mixture, and that our standing independence assumptions still hold). Now suppose the actual data matrices $\tilde{M}$ and $\tilde{N}_1$ are misaligned with respect to the column index (e.g. a set of data matrices from a chromatography/spectroscopy experiment that are incorrectly registered with respect to sample eluting time: Example 4). Then, in general, the row space of $\tilde{N}_1$ is not contained in the row space of $\tilde{M}$ (except in rare cases where the row space remains unchanged with shift along the rows - e.g., a data matrix whose elements are discretized values of a periodic function would not change with a shift of one period). A data synchronization scheme is then to shift all the

rows of $\tilde{N}_1$ along the misaligned axis until the row/column space of $\tilde{N}_1$ is contained (or most nearly contained) in the row/column space of $\tilde{M}$.

An algorithm that implements this scheme is summarized in Figure 3.3.2-2, and detailed in Algorithm 1. Given misregistered data matrices

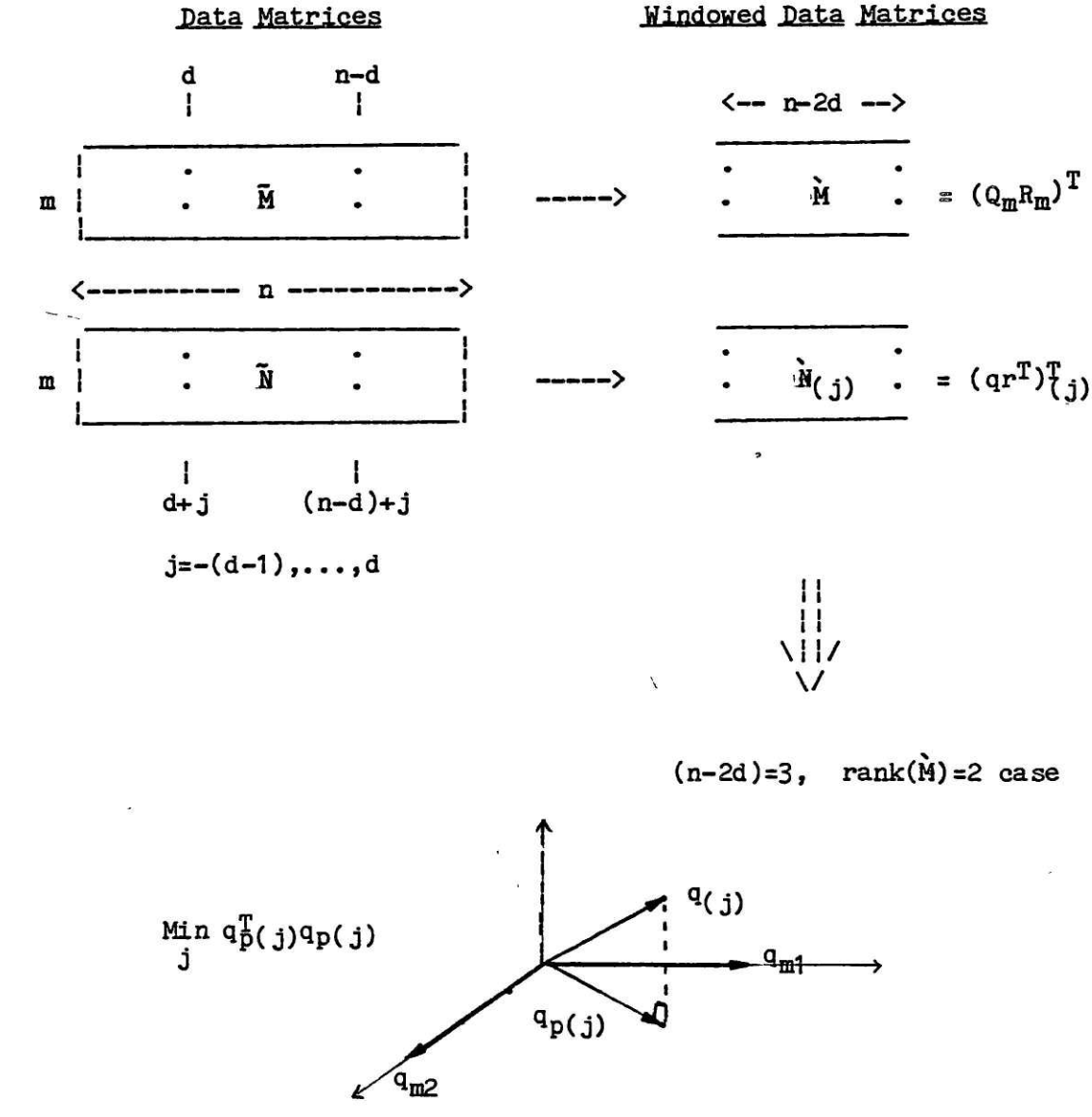


Figure 3.3.2-2. Data Synchronization Scheme.

Accept: d (starting index of window)

Comment: Window $\tilde{M}_{mxn}$ down to $\grave{M}_{mx(n-2d)}$

Do i=1,...,m

$\grave{M}_{i,j} = \tilde{M}_{i,j+d}$, j=1,...,n-2d

End Do i.

Comment: Compute QR decomposition of $\grave{M}^T$

$\grave{M}^T = Q_m R_m$

Initialize: $l_o=0$

Comment: Iterate on $\tilde{N}_1$

Do k=1,...,2d

Comment: Window $\tilde{N}_1$ down to $\grave{N}$

Do i=1,...,m

$\grave{N}_{i,j} = \tilde{N}_1 i,j+k$, j=1,...,n-2d

End Do i.

Comment: Compute QR decomposition of $\grave{N}^T$

$\grave{N}^T = qr^T$

Comment: Compute projection vector q_p

$q_p = Q_m Q_m^T q$; $l = q_p^T q_p$

Comment: Maximize (square of the) length of projection vector

If ($l > l_o$): $l_o=l$ and synch=k

End Do k.

Comment: Synchronize M with N_1

$\tilde{M}_{i,j}$ <----> $\tilde{N}_1 i,j+synch-d$

Algorithm 1: Data Synchronization Algorithm.

$\tilde{M}$ and $\tilde{N}_1$ of dimensions mxn, the first step in the algorithm is to window the data matrices down to a smaller dimension along the column index (to permit shifting of one data matrix relative to the other without "falling off the edge" of the data). The choice of d, the starting index of the window, and hence the window size (n-2d), is based on prior expectations on the degree of data misfit, and should be sufficiently large so that the windowed data can be shifted into the synchronized position. Once the data matrices have been properly windowed, the next step in the algorithm is to systematically shift the column index until the data synchronization between the two matrices is achieved. In the case of noise-free data, the data matrices M and N_1 are synchronized when l=1 in Algorithm 1 (a consequence of Bessel's inequality); with noisy data, a reasonable synchronization criterion is to maximize l ($=q_p^T q_p$), the (square of the) length of the projection vector (see Figure 3.3.2-2), and this is used in the algorithm.

Example 4. Application of the data synchronization algorithm to asynchronous synthetic data.

Data matrices in this example were generated from the absorbance and concentration profiles shown in Figures 3.3.2-3a,b,c as follows

$$\tilde{M} = c_1 x_1 y_1^T + c_2 x_2 y_2^T \quad c_3 x_3 y_3^T + E_m, \quad (c_1=c_2=c_3=1)$$

$$\tilde{N} = x_1 y_{o1}^T + E_n$$

where x_i are the component absorbance spectra, y_i are the component concentration profiles, c_i are the component concentrations in the mixture, y_{o1} is the concentration profile of the target component in a calibration run, and E_m and E_n are error matrices whose elements are

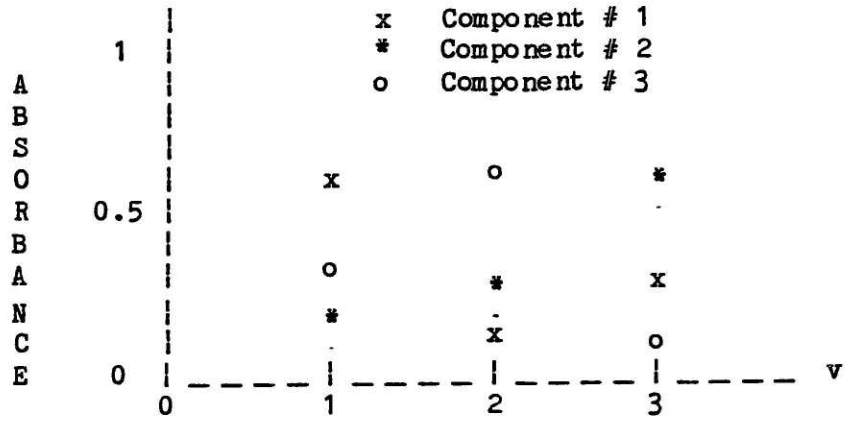


Figure 3.3.2-3a. Absorption Spectra.

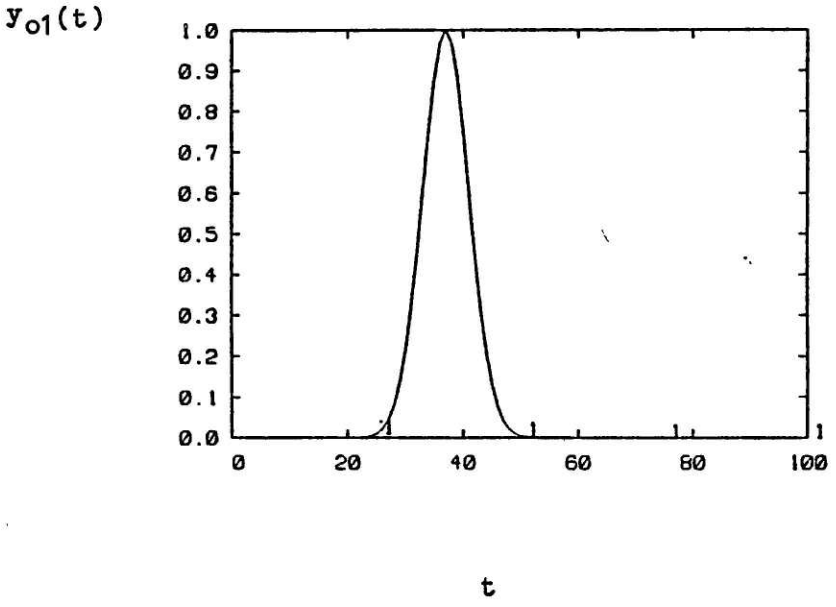


Figure 3.3.2-3b. Concentration profile of the target component in the (asynchronized) calibration run.

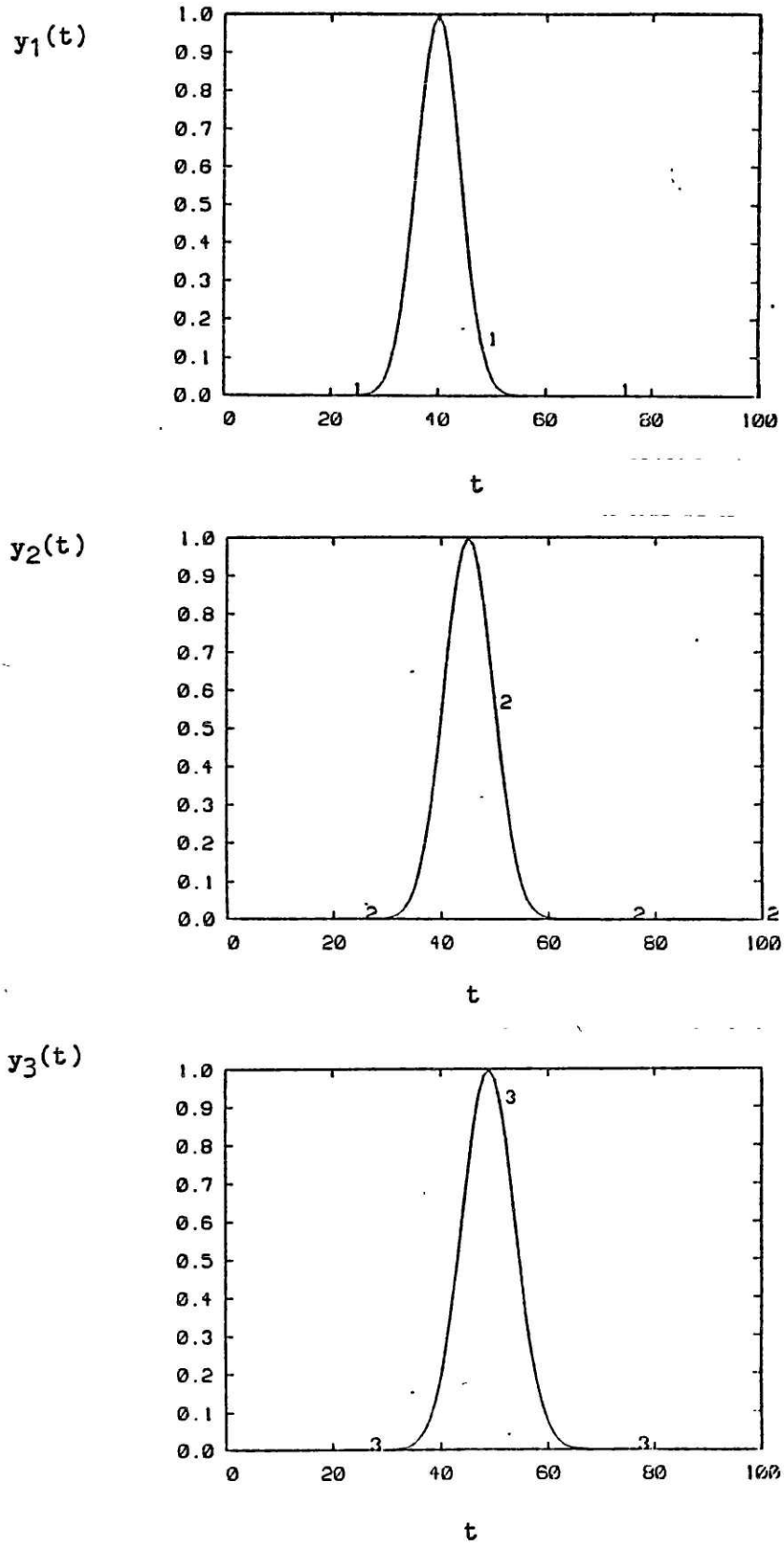


Figure 3.3.2-2c. Concentration profiles of the components in the mixture

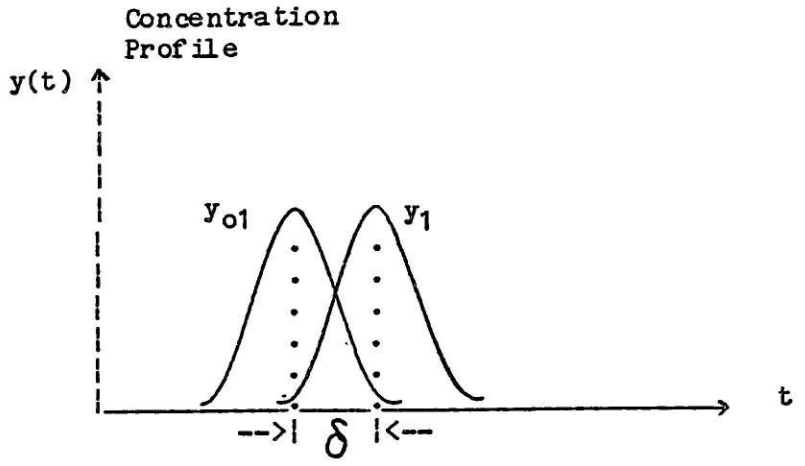
independent random variables drawn from a normal distribution with zero mean and standard deviation w.

The concentration profiles were generated as Gaussians:

$$y_i(t) = \exp(-0.5((t-t_{ri})/w_i)^2)$$

with the plate number ($=(t_{ri}/w_i)^2$) held constant at 100 to simulate, reasonably well, the concentration profiles from a real liquid chromatograph column. The starting time of the calibration run for the target component (first component) was set three time steps ahead of the mixture run to simulate asynchronization between the mixture and the calibration run (compare component #1 concentration profile in Figure 3.3.2-3c with that in Figure 3.3.2-3b).

The synchronization algorithm was then applied to the noisy asynchronous data matrices for several values of noise levels. Once the data matrices were synchronized, the LLS projection method was used to compute the concentration of the target component in the mixture. As can be seen from the Table 3.3.2-2, the synchronization algorithm provides an effective means for correcting an asynchronization problem that seriously affects the projection solution. Without application of the synchronization algorithm, there was 33% error in the projection method solution with a misregistration in the calibration run of 3 time steps (which is 75% of the standard deviation of the concentration profile of the first component). With the synchronization algorithm, the data matrices were correctly synchronized up to a standard deviation of 0.02 in the noise, and the concentration of the target component was correctly estimated to within 3% of its actual value.



(i) Without synchronization algorithm: δ =3, w=0, c_1=1.

Estimated concentration ($\hat{c}_1$) = 0.66

(ii) With synchronization algorithm: δ =3, c_1=1.

w	Synchronized* ?	$\hat{c}_1$
0	yes	1.000
0.001	yes	1.000
0.01	yes	1.009
0.02	yes	0.973
0.05	no	1.246

* "Yes" indicates the algorithm recognized that there was shift of 3 time steps, and correctly synchronized the data matrices.

Table 3.3.2-2. Simulation results from projection method with synchronization algorithm.

3.4 GENERALIZED EIGENANALYSIS METHOD

In Section 3.2 it was shown that a solution to the target components problem is given by rank tests on a composite matrix of the form

$$E(t) = M - tN \qquad (3.4\text{-}1)$$

Solving for the values of t for which the composite matrix (or the matrix pencil, as it is properly called) E(t) drops rank is, by definition, a generalized eigenvalue problem, and these values of t are called the (finite) eigenvalues of the pencil. However, what is usually discussed in the literature under the heading of generalized eigenvalue problems is the so-called regular case, where E(t) is (square and) nonsingular. In our case E(t) is in general singular, i.e., (i) the pencil is non-square, or (ii) if square, its determinant is zero for all t.

Discussions of issues involved in numerical computation of eigenvalues of singular pencils have only recently surfaced in the literature [S8,V2,V3,K5] and have not yet been fully resolved. The method we have used in solving the target components problem via generalized eigenanalysis is an adaptation of a numerical algorithm developed by Van Dooren [V2]. We have, in this thesis, amplified Van Dooren's work by developing in detail the rationale and guidelines for application of his algorithm to the noisy data matrices encountered in our work.

A description of Van Dooren's algorithm as applied to our problem is given in Section 3.4.1 for the case of noise-free data matrices. Application to noisy data, based on an error analysis (detailed in

Appendix I) of Van Dooren's algorithm, is given in Section 3.4.2.

3.4.1. Solution Method

Kronecker's theory of singular pencils, which gives a canonical decomposition of singular pencils that exposes their structure, is summarized in subsection (a). This serves as a background information on Van Dooren's algorithm, which is presented in subsection (b). Application to the noise-free target components problem is then described in subsection (c).

(a) Background: Canonical form (Kronecker structure) of a singular pencil

The results summarized here are drawn mostly from expositions on singular pencils by Gantmacher [G1] and Van Dooren [V4].

Definitions

The rank of a pencil: The rank r of a pencil is defined to be the order of the largest nonzero minor of the pencil. Hence, for a singular pencil of dimension pxq and rank r, $r \leq \min(p,q)$. Furthermore, this is the rank of E(t) for all but a finite number ($\leq \min(p,q)$) of values of t.

Strict equivalence: Two rectangular pencils M_1-tN_1 and M_2-tN_2, of the same dimensions mxn, are said to be strictly equivalent if there exist constant invertible matrices P and Q of orders m and n respectively, such that

$$P(M_1-tN_1)Q=M_2-tN_2 \qquad (3.4.1-1)$$

Basic Result

The canonical form of a singular pencil: Kronecker's theory of singular pencils [K4,G1] shows that any pencil M-tN is strictly equivalent to a unique canonical quasi-diagonal pencil of the form

$$\{0, L_{l1}, \ldots, L_{lp}, L_{r1}^T, \ldots, L_{rq}^T, I-tB, J-tI\} \qquad (3.4.1\text{-}2)$$

where

(i) L_k is the (k+1)xk bidiagonal pencil

$$L_k = \begin{bmatrix} t & & & \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ & & \ddots & t \\ & & & -1 \end{bmatrix} \quad \begin{matrix} \\ \\ \\ \\ k+1 \end{matrix} \qquad (3.4.1\text{-}3)$$

$$k$$

(ii) B is nilpotent (i.e. has all eigenvalues at 0), and both B and J are in Jordan canonical form.

The finite eigenstructure of the pencil M-tN is then defined as that of J. The eigenvalues of J are precisely the (finite) values of t for which the canonical form, and hence M-tN, drops rank. The matrix B embodies what is termed the infinite eigenstructure of the pencil. The L_{li}, L_{rj}^T give information on the structure of the left and right null spaces of the pencil, respectively; one says that the pencil has left minimal indices l_i and right minimal indices r_i. We shall not need to examine these in any further detail at present.

(b) Van Dooren's Algorithm : Noise-Free Case [V2]

Motivation

Van Dooren's algorithm, in essence, uses orthogonal transformations to alternately deflate, row- and column-wise, a singular pencil until a

regular pencil containing only the finite eigenvalues is recovered. In the process, the algorithm also reveals the canonical structure of the singular pencil. By "deflate", one means the reduction of the dimension of the problem to be solved, by removing from it whatever can be identified as not contributing to the solution. Van Dooren's method of deflating a singular pencil is a generalization of the standard method of deflating a matrix when one of its eigenvalues is known [K5,G3]. In order to motivate Van Dooren's algorithm, we present this standard method first.

Assume that an eigenvalue t_1 of an nxn matrix M is known, and suppose we are interested in computing the other eigenvalues of M. The eigenvalue problem for M is to find the values of t for which

$$| tI - M | = 0 \qquad (3.4.1\text{-}4)$$

This can be rewritten as

$$| \hat{t}I - (M - t_1 I) | = 0 \qquad (3.4.1\text{-}5)$$

where $\hat{t} = t - t_1$. In other words, the eigenvalues of M can be obtained by adding t_1 to the eigenvalues of $M - t_1 I$. Now, since t_1 is an eigenvalue of M, we already know that 0 is an eigenvalue of $M - t_1 I$. The deflation step involves displaying a smaller matrix that contains the <u>remaining</u> eigenvalues of $M - t_1 I$.

To this end, let

$$C = M - t_1 I \qquad (3.4.1\text{-}6)$$

with the SVD

$$C = [U_1, u_2] \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} [V_1, v_2]^T \qquad (3.4.1\text{-}7a)$$

$$S_1 = \text{diag}\,(s_1,\ldots,s_{n-1}) \qquad (3.4.1\text{-}7b)$$

Multiplying C by V^T and V from left and right, respectively, we have,

$$V^TCV = \left[\begin{array}{c|c} V_1^TU_1S_1 & 0 \\ \hline v_2^TU_1S_1 & 0 \end{array}\right] \begin{array}{c} n-1 \\ \hline 1 \end{array} \qquad (3.4.1\text{-}8)$$

The structure of the above matrix motivates one to label the above step as constituting "column compression" of C. Since V is orthogonal, the eigenvalues of C are the same as the eigenvalues of V^TCV, which in turn are given by the eigenvalues of the block diagonal matrices $V_1^TU_1S_1$ and 0. The remaining eigenvalues of M are thus obtained by adding t_1 to the eigenvalues of the (n-1)x(n-1) submatrix $V_1^TU_1S_1$. Hence, we have now successfully deflated the nxn eigenvalue problem to a smaller (n-1)x(n-1) eigenvalue problem.

Application to Target Component Problem

As mentioned above, Van Dooren's algorithm is essentially a sequence of orthogonal transformations that recovers a regular pencil structure with the same finite eigenvalues as the original singular pencil. This is achieved by systematically deflating the singular pencil with a series of row and column compressions that expose the Kronecker structure of the pencil, ultimately revealing a regular pencil structure.

The first phase of the algorithm deflates the pencil via column compressions on N and row compressions on M to "subtract" the infinite eigenstructure and the right null structure of the pencil. The second phase of the algorithm then further deflates the pencil by subtracting the left null structure via row compressions on N and column

compressions on M.

Van Dooren's algorithm for the rank(N)=1 (single target component case) is summarized in Figure 3.4.1-1 and outlined next in the solution to Problem 4. The underlying structures of the data matrices M and N are as stated in Section 2.3 of Chapter 2.

Problem 4.

Given

$M\ (=c_1N_1+\ldots+c_rN_r)$

$N\ (=N_j,\ 1\leq j\leq r)$

find the value of t for which the matrix pencil

$$M - tN$$

loses rank.

Solution:

Algorithm 2. Van Dooren's algorithm with infinite precision computation.

(Note: to avoid a profusion of symbols, the same "generic" notation is used for the various SVD's that occur in the algorithm, and symbols for some of the other matrices are also used repeatedly in Phases I & II. When necessary in later discussion, we shall make clear which particular instance of use is under discussion.)

Phase I. Column compression.

<Step 1.> Compress $N\ (=[u_1,U_2]\left[\begin{array}{c|c}s_1 & 0\\ \hline 0 & 0\end{array}\right]\begin{bmatrix}v_1^T\\ V_2^T\end{bmatrix})$ to full column rank:

$$NV = [Nv_1,0] = [n_1,0]$$

Update M:

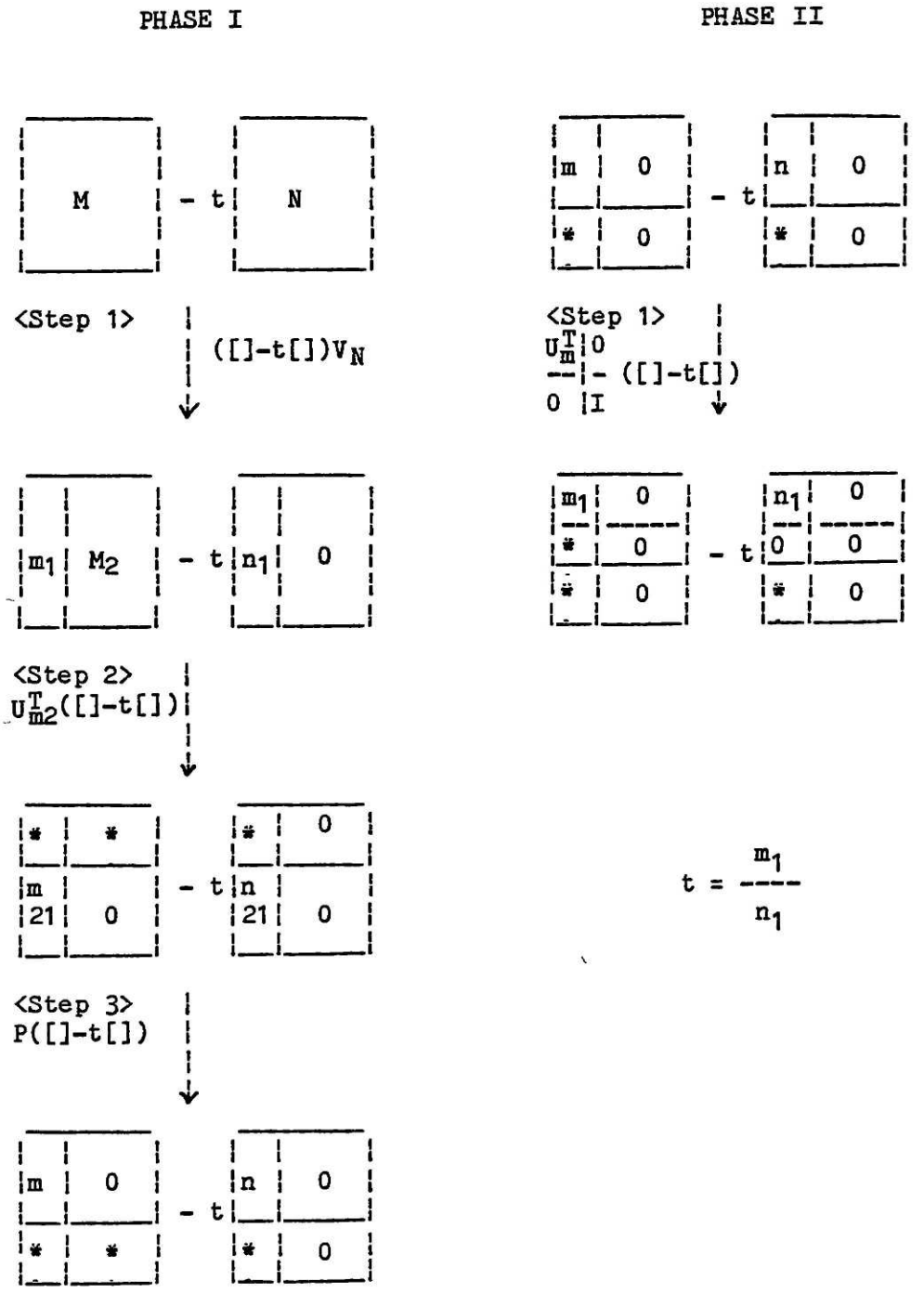


Figure 3.4.1-1. Van Dooren's algorithm for rank(N)=1: noise free case.

$$MV = [Mv_1, MV_2] = [m_1, M_2]$$

<Step 2.> Compress M_2 $(=[U_1, U_2]\left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & 0 \end{array}\right]\left[\begin{array}{c} V_1^T \\ \hline V_2^T \end{array}\right])$ to full row rank:

$$U^T[m_1, M_2] = \left[\begin{array}{c|c} U_1^T m_1 & U_1^T M_2 \\ \hline U_2^T m_1 & 0 \end{array}\right] = \left[\begin{array}{c|c} m_{11} & M_{12} \\ \hline m_{21} & 0 \end{array}\right]$$

Update n_1:

$$U^T[n_1, 0] = \left[\begin{array}{c|c} U_1^T n_1 & 0 \\ \hline U_2^T n_1 & 0 \end{array}\right] = \left[\begin{array}{c|c} n_{11} & 0 \\ \hline n_{21} & 0 \end{array}\right]$$

<Step 3.> Permute upper and lower blocks, and reinitialize:

$$P\left[\begin{array}{c|c} m_{11} & M_{12} \\ \hline m_{21} & 0 \end{array}\right] = \left[\begin{array}{c|c} m_{21} & 0 \\ \hline m_{11} & M_{12} \end{array}\right] = \left[\begin{array}{c|c} m & 0 \\ \hline * & * \end{array}\right]$$

$$P\left[\begin{array}{c|c} n_{11} & 0 \\ \hline n_{21} & 0 \end{array}\right] = \left[\begin{array}{c|c} n_{21} & 0 \\ \hline n_{11} & 0 \end{array}\right] = \left[\begin{array}{c|c} n & 0 \\ \hline * & 0 \end{array}\right]$$

Phase II. Row compression.

<Step 1.> Compress n $(= [u_1, U_2]\left[\begin{array}{c} s \\ \hline 0 \end{array}\right][1])$ to full row rank:

$$U^T n = \left[\begin{array}{c} u_1^T n \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} n_1 \\ \hline 0 \end{array}\right]$$

Update m:

$$U^T m = \left[\begin{array}{c} u_1^T m \\ \hline U_2^T m \ (=0) \end{array}\right] = \left[\begin{array}{c} m_1 \\ \hline m_2 \ (=0) \end{array}\right]$$

<End.> The regular pencil with finite eigenvalue is:

$$m_1 - t n_1$$

Solving for t,

$$t = m_1/n_1$$

Comments:

(i) It can be shown that for M and N of the form given in Problem 4, $U_2^T m=0$ in Step 1, Phase II of Algorithm 2. An application of this equality would be to the data synchronization problem (see Section 3.3.2-c); the mixture data matrix M and the target data matrix N are in synchronization when application of Van Dooren's algorithm yields $U_2^T m=0$ at this step.

(ii) The effect of Step 3 in Phase I is to simply position the regular pencil structure at the upper-left corner of the original pencil, and may be omitted without affecting the solution. The use of that step, however, simplifies the book-keeping when coding the algorithm for implementation on a computer, and is thus maintained.

3.4.2 Application of Van Dooren's Algorithm to Noisy Data Case

The critical steps in applying Van Dooren's algorithm to the noisy data case are the rank decisions made on the processed data matrices. Proper rank decisions are required to guarantee that the eigenvalues of the recovered regular pencil correspond to the eigenvalues of the underlying noise-free pencil, and that the magnitudes of the errors in the computed eigenvalues are of the order of the noise. As stated in [V4] and Appendix I, discussions on the eigenvalue sensitivity problem for singular pencils are meaningful only when the Kronecker structure of the pencil is preserved, either through restricted perturbation or via judicious application of Van Dooren's algorithm; otherwise (unlike in

the regular pencil case) the eigenvalues are generally highly discontinuous functions of the data, as will be illustrated subsquently.

In this section, we give guidelines (based on an error analysis of Van Dooren's algorithm as detailed in Appendix I) on setting the rank-determining-thresholds in a way that guarantee the recovery of the Kronecker structure of the underlying noise-free pencil from a noisy one.

To motivate the problem at hand, consider the following example:

Example 5.

Given the singular pencil

$$\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right] - t\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right]$$

whose finite eigenvalue is seen to occur at t=1, consider a pencil that is a small (<<1) perturbation of the above singular pencil,

$$\left[\begin{array}{c|c} 1+d_{11} & d_{12} \\ \hline d_{21} & d_{22} \end{array}\right] - t\left[\begin{array}{c|c} 1+e_{11} & e_{12} \\ \hline e_{21} & e_{22} \end{array}\right]$$

with $|d_{ij}|, |e_{ij}| << 1$.

Case (i) $d_{12}=d_{21}=e_{12}=e_{21}=0$.

For any arbitrarily small values of d_{22} and e_{22}, this pencil is non-singular, with eigenvalues at $(1+d_{11})/(1+e_{11})$ and d_{22}/e_{22}. Note that the second eigenvalue can be made arbitrarily large with small variations in d_{22} and e_{22}.

Case (ii) $d_{12}=d_{22}=e_{12}=e_{22}=0$, $d_{21}\neq e_{21}$.

There is no value of t for which this pencil loses rank. That is, this pencil has no regular pencil structure, and correspondingly has no

finite eigenvalue.

Case (iii) $d_{12}=d_{21}=d_{22}=e_{12}=e_{21}=e_{22}=0$.

This pencil has a single eigenvalue at $(1+d_{11})/(1+e_{11})$, and small perturbations in d_{11} and e_{11} result in small changes in the eigenvalue.

Cases (i) and (ii) are examples where the Kronecker structure (the left and right null structures in particular) of the pencil changes with arbitrarily small perturbations in its elements. This results in a change in the dimension of the regular pencil structure, and consequently, addition or subtraction of eigenvalues to and from the spectrum of the unperturbed pencil. In contrast, Case (iii) illustrates a "well-behaved" perturbation, in which the spectrum of the pencil changes continuously with continuous changes in the elements of the pencil.

[]

Hence, what is required in applying Van Dooren's algorithm to a noisy data case is a set of guidelines for making rank decisions (from prior knowledge of the noise) that in effect remove that part of the noise that results in changes in the Kronecker structure of the underlying noise-free singular pencil.

We give, in Algorithm 3, Van Dooren's algorithm (as applied to our problem) with thresholds that guarantee that, when applied to noisy data matrices, the Kronecker structure of the underlying noise free pencil is recovered. The thresholds to be used in the rank decisions made in Algorithm 3 are given in Result 1.

Result 1. (See Result I-2 in Appendix I for proof.)

Suppose we are given noisy data matrices

$\tilde{M}$ $(=M+E_m=c_1x_1y_1^T+\ldots+c_rx_ry_r^T+E_m)$

$\tilde{N}$ $(=N+E_n=x_jy_j^T+E_n,\ 1\leq j\leq r)$

such that

$$s_r(M_2) > 2||E_m||_2\left(1+\frac{2^{1/2}(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

$$s_1(N) > 2||E_n||_2$$

where $s_r(M_2)$ and $s_1(N)$ are the smallest nonzero singular values of the matrices M_2 $(= MV_2$, see Algorithm 2) and N for the noise free case.

Application of Algorithm 3 with

$$EPSN = ||E_n||_2$$

$$EPSM2 = ||E_m||_2\left(1+\frac{2^{1/2}(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

to the noisy matrix pencil $\tilde{M}-t\tilde{N}$ then recovers the Kronecker structure of the underlying noise free pencil M-tN.

□

Algorithm 3: Van Dooren's algorithm with "thresholding"

(Note: The "generic" symbols for SVD are repeated in each of the steps; the symbols for some of the other matrices are also repeated in Phases I & II)

Phase I. Column compression.

<Step 1.> Compress $\tilde{N}$ $(=[\tilde{u}_1,\tilde{U}_2]\begin{bmatrix}\tilde{s}_1 & 0\\ 0 & \tilde{S}_2\end{bmatrix}\begin{bmatrix}\tilde{v}_1^T\\ \tilde{V}_2^T\end{bmatrix},\ (\tilde{S}_2)_{1,1}\leq||E_n||_2=EPSN)$

to full numerical column rank:

$$\tilde{N}V = [\tilde{N}\tilde{v}_1,0] = [\tilde{n}_1,0]$$

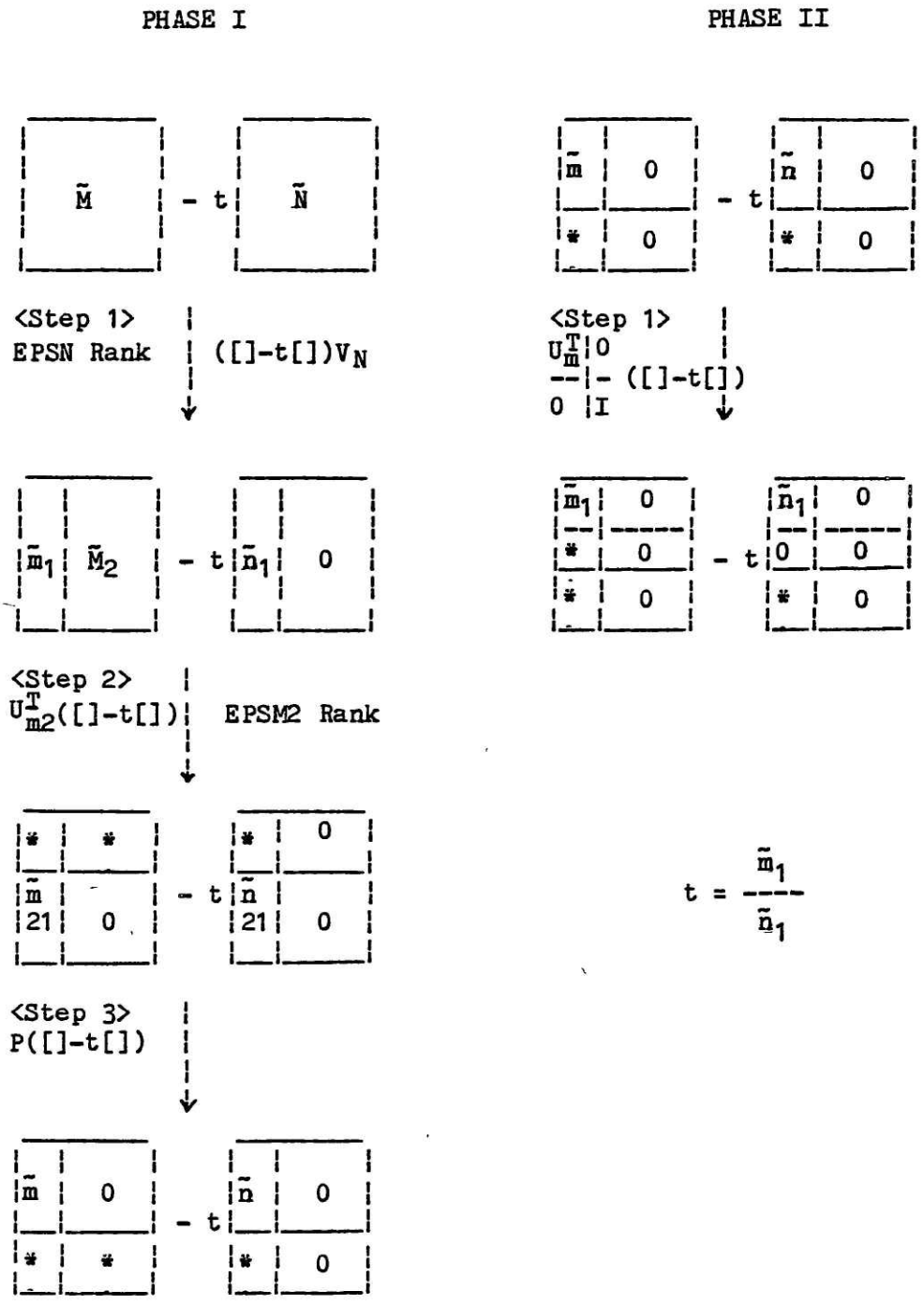


Figure 3.4.1-2. Van Dooren's algorithm for rank(N)=1: noisy data case.

Update $\tilde{M}$:

$$\tilde{M}V = [\tilde{M}\tilde{v}_1, \tilde{M}\tilde{V}_2] = [\tilde{m}_1, \tilde{M}_2]$$

<Step 2.> Compress $\tilde{M}_2$ $(=[\tilde{U}_1, \tilde{U}_2]\begin{bmatrix} \tilde{S}_1 & 0 \\ 0 & \tilde{S}_2 \end{bmatrix}\begin{bmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{bmatrix}$,

$(\tilde{S}_2)_i \leq ||E_m||_2 = EPSM2)$ to full numerical row rank:

$$\tilde{U}^T[\tilde{m}_1, \tilde{M}_2] = \begin{bmatrix} \tilde{U}_1^T\tilde{m}_1 & \tilde{U}_1^T\tilde{M}_2 \\ \tilde{U}_2^T\tilde{m}_1 & 0 \end{bmatrix} = \begin{bmatrix} \tilde{m}_{11} & \tilde{M}_{12} \\ \tilde{m}_{21} & 0 \end{bmatrix}$$

Update $\tilde{n}_1$:

$$\tilde{U}^T[\tilde{n}_1, 0] = \begin{bmatrix} \tilde{U}_1^T\tilde{n}_1 & 0 \\ \tilde{U}_2^T\tilde{n}_1 & 0 \end{bmatrix} = \begin{bmatrix} \tilde{n}_{11} & 0 \\ \tilde{n}_{21} & 0 \end{bmatrix}$$

<Step 3.> Permute upper and lower blocks, and reinitialize:

$$P\begin{bmatrix} \tilde{m}_{11} & \tilde{M}_{12} \\ \tilde{m}_{21} & 0 \end{bmatrix} = \begin{bmatrix} \tilde{m}_{21} & 0 \\ \tilde{m}_{11} & \tilde{M}_{12} \end{bmatrix} = \begin{bmatrix} \tilde{m} & 0 \\ * & * \end{bmatrix}$$

$$P\begin{bmatrix} \tilde{n}_{11} & 0 \\ \tilde{n}_{21} & 0 \end{bmatrix} = \begin{bmatrix} \tilde{n}_{21} & 0 \\ \tilde{n}_{11} & 0 \end{bmatrix} = \begin{bmatrix} \tilde{n} & 0 \\ * & 0 \end{bmatrix}$$

Phase II. Row compression.

<step 1.> Compress $\tilde{n}$ $(= [\tilde{u}_1, \tilde{U}_2]\begin{bmatrix} \tilde{s} \\ 0 \end{bmatrix}[1])$ to full row rank:

$$\tilde{U}^T\tilde{n} = \begin{bmatrix} \tilde{u}_1^T\tilde{n} \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{n}_1 \\ 0 \end{bmatrix}$$

Update $\tilde{m}$:

$$\tilde{U}^T\tilde{m} = \begin{bmatrix} \tilde{u}_1^T\tilde{m} \\ \tilde{U}_2^T\tilde{m} \end{bmatrix} = \begin{bmatrix} \tilde{m}_1 \\ \tilde{m}_2 \end{bmatrix}$$

<End.> The regular pencil with finite eigenvalue is:

$$\tilde{m}_1 - t\tilde{n}_1$$

Solving for t,

$$t = \tilde{m}_1/\tilde{n}_1$$

□

Comments:

(i) For data synchronization with noisy data via Van Dooren's algorithm, minimize $||U_2^T m||_2$ in Step 1, Phase II, Algorithm 3.

(ii) The underlying issues for the generalized eigenanalysis method via Van Dooren's algorithm and the projection method are essentially same, but addressed from slightly different perspectives and resulting in slightly different algorithms. The exact connection between the two have not been worked out, but should be explored as a future work.

Example 5. Application of Algorithm 3 to the target component problem with noisy data matrices.

The data matrices used in this example are the same as in Example 3. The threshholds used to make rank decisions are from Result 3 (Section 3.4.2c). The results of application of Algorithm 3 to the noisy target component problem for several values of noise variance are summarized in Table 3.4.1-1. It can be seen from the table that the algorithm correctly estimates the concentration of the target component to within 1% of the actual value for the cases with standard deviations of the noise less than 0.01. At higher levels of noise, due to the conservativeness of the error bounds (from which the thresholds are derived) given by Result I-3 in Appendix I, the thresholds were set too high, and resulted in incorrect rank decisions.

w	$\hat{G}_1^*$	s.d.
0.001	1.000	0.005
0.002	1.000	0.010
0.004	1.000	0.021
0.008	1.001	0.042
0.01	1.003	0.053
0.02	1.440	0.102
0.01	1.001+	0.053
0.02	1.006+	0.103

* 2000 sample average.

+ Results with prior knowledge of ranks.

Table 3.4.1-1. Application of Algorithm 3 to noisy target component problem.

CHAPTER IV: THE UNKNOWN COMPONENTS SOLUTION

CHAPTER OUTLINE

	Page
4.1 Introduction	114
4.2 Data Representation and Solution Characterization: Noise-Free Case	116
4.2.1 Measurement Space	117
(a) Data Representation and Solution Characterization for PROBLEM 2a	117
(b) Data Representation and Solution Characterization for PROBLEM 2b	124
4.2.2 Parametrized signal space	129
- Solution Space Transformation	130
- Normalized Version of Transformation	132
- Solution Characterization in the Parametrized Solution Space	135
- Normalized Version of Solution Characterization	138
4.3 Extensions to Noisy Data Case	142
4.3.1 Reconstructing the signal space	142
- Noise Model and Transformation to Isotropic Noise	142
- Approximating the Signal Space Via SVD	144
4.3.2 Representation in Parameterized Reconstructed Signal (PRS) Space	147
- Data Projection to PRS Space	147
- Error Projection to PRS Space	148
4.3.3 Representation in Normalized Parametrized Reconstructed Signal (NPRS) Space	150
- Noisy Data Representation in the NPRS Space	150
- Normalization of the Uncertainty Region in the PRS Space	152
- Bounding the Unit-Sum Normalized Uncertainty Region	157
4.4 Solution Procedures: Two- and Three-Components Cases	159
4.4.1 Two-Components Case	159
4.4.2 Three-Components Case	163
- Data Representaion	163
- Inner Bound	164
- Outer Bound	164
- Generating Disjoint Regions	166
- An Iterative Approximation	169
- Measurement Space Bound	171

Page

4.5 Examples of Data Display with Synthetic Liquid Chromatography Data..... 173

- Barycentric Coordinates..... 173
- Examples of Data Display..... 176

4.1 INTRODUCTION

The unknown components problem for the two-component case has been addressed by Lawton and Sylvestre [L1], who gave a method for estimating (via upper and lower bounds) the shapes of two overlapping functions from an observed set of an additive combination of these two functions. Applications of their method can be found in [S2], where it is used to resolve overlapping mass spectra from a gas chromatography-mass spectroscopy experiment, and in [O2] where the spectra to be resolved are absorption spectra from a liquid chromatography experiment. Spjotvoll et al. [S5] propose an alternate (restricted least squares method) approach to the two-component problem and claim that the method is applicable to an arbitrary number of components, but extensions to three or more components have not been made explicit or pursued in the literature. Extensions to the three-component case have been partially addressed by several investigators [C2,M2,O1]. Chen and Hwang's method [C2] appears to work only in special cases where the underlying spectra (functions) are as distinct as possble, and even then some trial-and-error is involved in determining the underlying spectra. (For an application of Chen and Hwang's method to liquid chromatography data, see [V1].) Meister [M2] also assumes that the spectra are as distinct as possible, and uses an iterative linear programming approach to fit a "maximum-volume" simplex, within some allowable solution region, that defines a solution set. Ohta's extension [O1] closely follows that of Lawton and Sylvestre in defining the allowable solution region, but the characterization of solution sets is unclear (it appears to be a brute force Monte Carlo method) and numerically expensive.

We have, in Chapter 2, given a concise formulation of the unknown

components problem for an arbitrary number of components. We present here a comprehensive solution approach to this problem. The general solution approach is geometric in nature, and systematically transforms the observed data matrices into representations in the solution space that reveal the internal structure of the data matrices. The outcome of this is a set of solution bands (bounds) for the unknowns. For the case of arbitrary number of components, we give a complete characterization of a solution set in a solution space that is easier to visualize than those used by [C2,M2,L1], and then show how it may be transformed into that of [L1] in a natural way so that the solution characterization in this space becomes more transparent. In addition, for the case of two and three components problems, we give comprehensive solution methods.

In Section 4.2 we address the problem of data representation and solution characterization for the noise-free data case. These concepts are extended to the noisy data case in Section 4.3. In Section 4.4, we give solution methods for the cases of two and three unknown components. Finally, in Section 4.5, examples of data display with synthetic liquid chromatography data is given for the case of three channel data from three-component mixtures, using an efficient data display scheme that does not require matrix decompositions.

Comment: We distinguish between "solution characterization" and "solution methodology" as follows: the solution characterization defines what constitutes a solution; the solution methodology, e.g. algorithms, equations, etc., provides a mean for actually obtaining that solution.

4.2 DATA REPRESENTATION AND SOLUTION CHARACTERIZATION: NOISE FREE CASE

In this section, we study in detail (with geometric interpretation) how the composition of a mixture is reflected into the structure of its data matrix, in the ideal noise-free case. With the understanding gained from this, we work backwards to characterize solution sets for the unknown components. In Section 4.2.1, the problem of data representation and solution characterization in the physically meaningful and easily visualized measurement space is addressed. In Section 4.2.2, the same problem is treated in a lower dimensional parameterized signal space, which enables a more economical description of the solution. The structure of the data matrix to be analyzed and the constraints placed on it are summarized below:

The data matrix:

We have, for PROBLEM 2a (Chapter 2), the data matrix

$$M \ (= c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T) \tag{4.2-1a}$$

with the constraints

$$c_i \geq 0, \ x_i \geq 0, \text{ and } y_i \geq 0, \ i=1,\ldots,r \tag{4.2-1b}$$

where the inequalities on the vectors x_i and y_i are component-wise inequalities, and both $\{x_i\}$ and $\{y_i\}$ are independent sets. In addition, for PROBLEM 2b, we have the unit-sum constraints on the unknown vectors

$$1^T x_i = 1 \text{ and } 1^T y_i = 1, \ i=1,\ldots,r \tag{4.2-1c}$$

where 1 ($=[1,\ldots,1]^T$) is a vector with all its elements equal to 1. As stated in Chapter 2, the problem is to estimate from M, subject to the

constraints above, the unknown vectors x_i and y_i. We shall however, limit our discussion to estimating x_i only, since the problem of estimating y_i completely parallels that of x_i; one only has to replace M by M^T and interchange x_i with y_i in what follows.

4.2.1 Measurement Space

We shall first work with the data matrix and constraints of PROBLEM 2a, and then impose the unit-sum constraints on the solution set of PROBLEM 2a to characterize a solution set for PROBLEM 2b.

(a) Data Representation and Solution Characterization for PROBLEM 2a

Given

$$M = [m_1,\ldots,m_q] \qquad (4.2.1\text{-}1)$$

where m_j $(= [m_{1j},\ldots,m_{pj}]^T)$, $j=1,\ldots,q$, are px1 vectors, we can represent each of the columns m_j of M as a vector in R^p. Hence the data matrix M is completely described by these <u>data vectors</u> or <u>data points</u> m_j in R^p. Motivated by the fact that the vectors m_j are measured physical variables, we shall call this p-dimensional space in which the m_j are represented the **MEASUREMENT SPACE**. For example, in a liquid chromatography-UV spectroscopy experiment, each of the data vectors m_j would represent an absorbance vector at a particular sampling time, and the axes of the measurement space R^p would represent absorbances at each of the p selected measurement wavelengths.

By displaying the data matrix in the measurement space, We shall now show how the composition of a mixture is reflected into the mixture data matrix.

One-component case:

In this case M is the rank-one matrix given by

$$M = [m_1, \ldots, m_q] = cx_1 y_1^T = [cy_{11}x_1, \ldots, cy_{1q}x_1] \quad (4.2.1\text{-}2)$$

so the data vectors m_j are all simply the unknown vector x_1 scaled by the weights cy_{1j}. Hence, the data points all lie on the straight line spanned by the vector x_1. In addition, the positivity constraints allow us to deduce more. Firstly, the constraint that x_1 have positive entries tells us that this vector lies in the positive (or first) orthant of RP. Secondly, the constraints on the cy_{1j} imply that the data points must lie on the same side of the origin as x_1 (on the line spanned by it).

Working backwards from the observations in the previous paragraph, we now see that, by plotting the data vectors m_i in the measurement space, we obtain the direction (but not the magnitude) of the unknown vector x_1. This is illustrated in Figure 4.2.1-1 for the case where p=3.

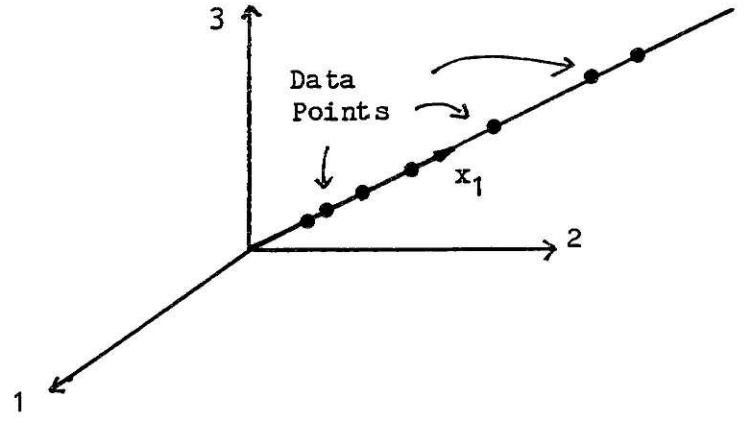


Figure 4.2.1-1. Data representation in the measurement space: one-component (rank(M)=1), p=3

Two-components case:

We now have the rank-two data matrix

$$M = [m_1,\ldots,m_q] = c_1x_1y_1^T + c_2x_2y_2^T \qquad (4.2.1\text{-}3)$$
$$= [(c_1y_{11}x_1+c_2y_{21}x_2),\ldots,(c_1y_{1q}x_1+c_2y_{2q}x_2)]$$

so that the data vectors m_j are (independent) linear combinations of the two unknown vectors x_1 and x_2. Hence, the m_j all lie on the plane spanned by x_1 and x_2 in RP. Furthermore, because of the positivity constraints on the entries of x_i and c_iy_{ij}, it is evident that (i) x_i lie in the positive (or first) orthant of RP and (ii) the points m_j lie inside the positive cone generated by x_1 and x_2 (see Figure 4.2.1-2), where the POSITIVE CONE (C^+) of any set $x_1,\ldots,x_r$ is defined to be

$$C^+ = \{x: x=a_1x_1+\ldots+a_rx_r,\ a_i\geq 0,\ i=1,\ldots,r\} \qquad (4.2.1\text{-}4)$$

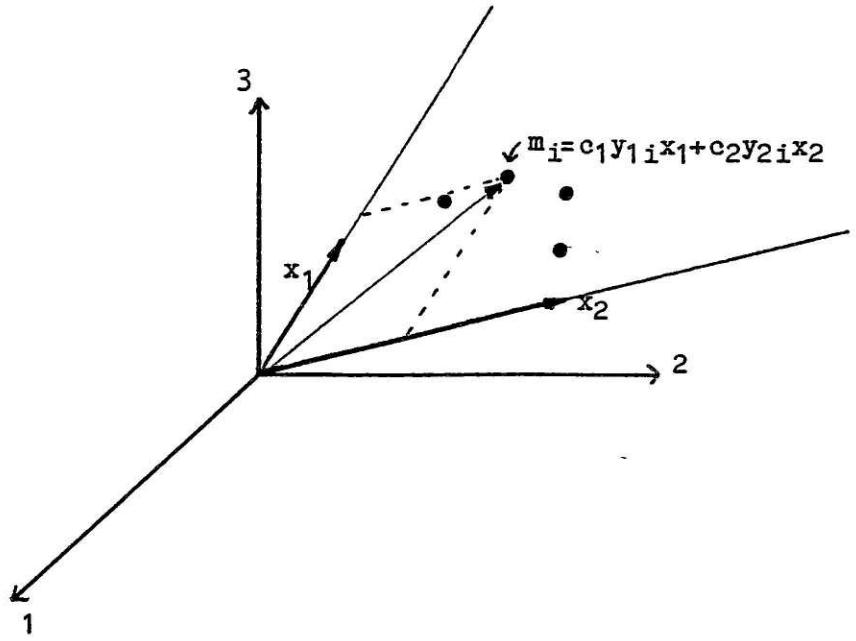


Figure 4.2.1-2.

Working backwards again, it can be seen that, by plotting the data vectors m_j in the measurement space, we obtain the plane spanned by the unknown vectors x_1 and x_2 in R^p, as well as the <u>directions</u> in which the unknown vectors x_1 and x_2 can point (but <u>not their magnitudes</u>). More specifically, our independence assumptions guarantee that the m_j generate a plane, and this is precisely the plane containing (and spanned by) x_1 and x_2. Furthermore, x_1 and x_2 must (i) lie in the positive orthant, and (ii) generate a positive cone that contains the m_j. This leads us to the possible solutions shown schematically in Figure 4.2.1-3.

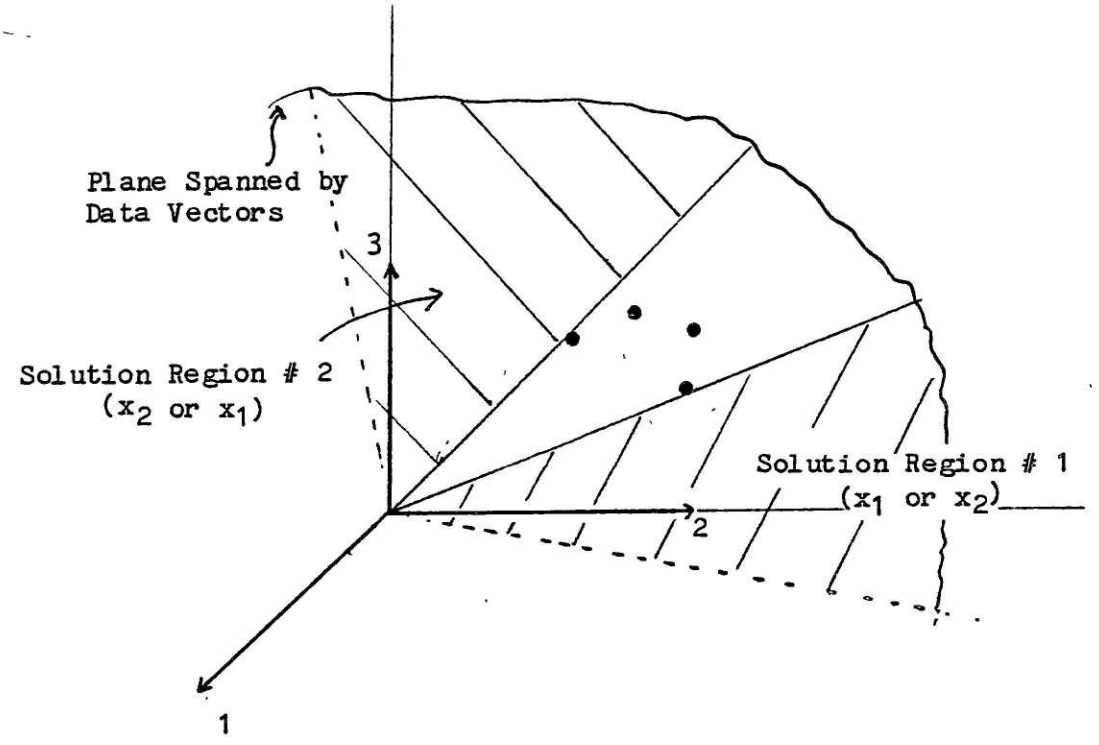


Figure 4.2.1-3. Data representation in the measurement space: Two components (rank(M)=2), p=3

Arbitrary number of components:

Generalizing the above observations to the r unknown components case, we have:

Result 1

Given a pxq data matrix M,

$$M = [m_1,\ldots,m_q] \ (=c_1x_1y_1^T+\ldots+c_rx_ry_r^T)$$

$c_i \geq 0$, $x_i \geq 0$, and $y_i \geq 0$, $i=1,\ldots,r$

(i) The (unknown) vectors x_i lie in the positive orthant of R^p;

(ii) the points m_j lie inside the positive cone generated by the vectors x_i.

□

Note that (i) follows immediately from $x \geq 0$, and (ii) follows from the definition of a positive cone, using the facts that $c_i \geq 0$ and $y_i \geq 0$. More details are given in Appendix G.

Now generalizing the process of working backwards, we can use Result 1 to deduce information about the unknown vectors x_i from the measured data vectors m_j, as summarized in Result 2 below. In stating the result, we shall make use of the following terminology: the r-dimensional subspace spanned by the data vectors m_j will be called the SIGNAL SPACE.

Result 2

Given a pxq data matrix M,

$$M = [m_1,\ldots,m_q] \ (=c_1x_1y_1^T+\ldots+c_rx_ry_r^T)$$

$c_i \geq 0$, $x_i \geq 0$, and $y_i \geq 0$, $i=1,\ldots,r$

where both $\{x_i\}$ and $\{y_i\}$ are independent sets, the unknown vectors

x_i

(i) lie in, and span, the signal space;

(ii) lie in (the portion of signal space contained in) the positive orthant of measurement space;

(iii) generate a positive cone that (lies in the portion of signal space contained in the positive orthant of measurement space and) contains m_j.

[]

In other words, the positive cone generated by the vectors $x_1,\ldots,x_r$ lies between the inner bounding cone generated by the data vectors m_j, and the outer bounding cone formed by the portion of the r-dimensional signal space that intersects the positive orthant of measurement space. The proof of (1) follows from the independence assumption, and (ii) and (iii) follow from Result 1. Details are given in Appendix G. Note that Result 2 only gives information about the directions of the x_i, and not their magnitudes.

The reader should verify that the one and two component cases are indeed accounted for by the preceding results. As a further illustration, we show in Figure 4.2.1-4 what the results imply for the three-component case.

As already noted, the unknown components solution characterized by Result 2 only gives the directions of the candidate solution vectors, not their magnitudes. This is a direct consequence of the scaling problem inherent in our bilinear estimation problem: if we find a set of c_i, x_i, and y_i that satisfy our positivity and independence constraints, and also satisfy

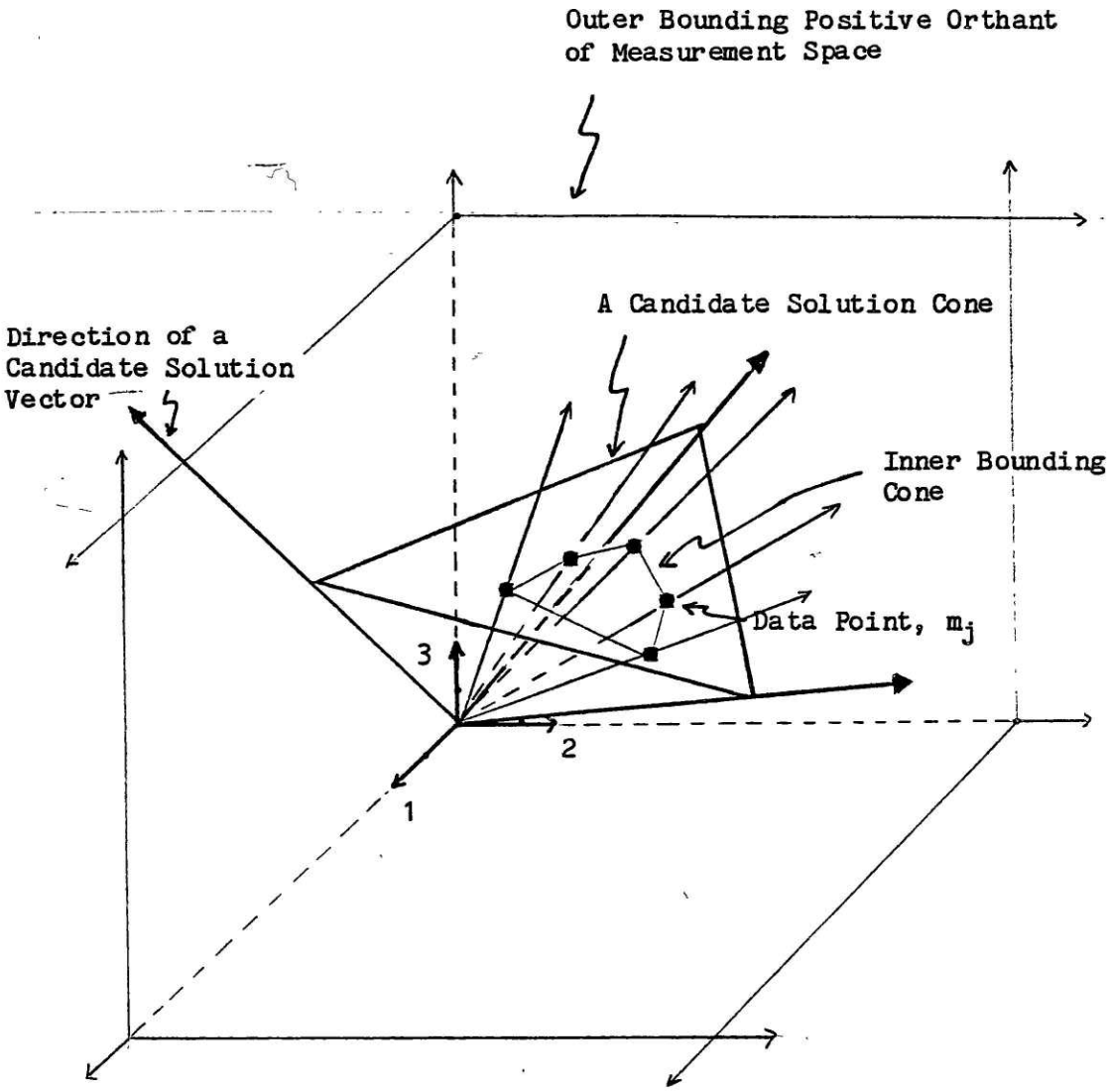

Outer Bounding Positive Orthant
of Measurement Space
A Candidate Solution Cone
Direction of a
Candidate Solution
Vector
Inner Bounding
Cone
Data Point, m_j
3
2
1

Figure 4.2.1-4.

$$M = c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T \qquad (4.2.1\text{-}5)$$

then an <u>infinite</u> number of other solutions is given by $a_i x_i$, $b_i y_i$, and $c_i/(a_i b_i)$, where a_i and b_i are arbitrary positive numbers. Thus, in the setting of our unknown components problem, we can only expect to recover x_i to within a (positive) scale factor.

This is still valuable information in many contexts, however. For example, if M is a data matrix from liquid chromatography-UV spectroscopy experiment, and x_i are the absorption spectra of the constituent components in the chemical mixture under analysis, then the knowledge of the relative shape of the spectra (hence the directions of the unknown vectors in R^p) is often sufficient to identify the unknown components.

Given the arbitrary scale factors above, we can hope to simplify our solution procedure by picking some convenient normalization for the x_i. It is to be expected that this normalization step, which essentially acknowledges that there is no information in the magnitudes of a reconstructed x_i, serves to reduce by one the dimensions of the spaces in which we work. This expectation is indeed justified. Note also that result 2 only requires the directions, and not magnitudes, of the m_j (though we shall see that in the case of noisy data the magnitudes will also be used). It will be shown next how picking a <u>unit-sum constraint</u> on the x_i enables us to realize these simplifications in computation and display results.

(b) Data Representation and Solution Characterization for PROBLEM 2b

The unit-sum constraint on the unknown vectors x_i

$$1^T x_i = 1, \quad i=1,\ldots,r \qquad (4.2.1\text{-}6)$$

restricts (the tips of) the unknown vectors x_i to lie on a (p-1)-dimensional hyperplane that passes through (the tips of) the unit coordinate vectors of R^p. We shall call this hyperplane the **NORMALIZING PLANE**. As illustrated in Figure 4.2.1-5, the normalizing plane for p=2 case corresponds to a line passing through the points (1,0) and (0,1), and for the p=3 case, to a plane passing through the points (1,0,0), (0,1,0), and (0,0,1). When restricted to the positive orthant of R^p, which is all that is of interest in view of Result 2(ii), the normalizing plane reduces to a (p-1)-dimensional regular simplex (shaded areas in Figure 4.2.1-5), with its vertices at the (tips of the) unit coordinate vectors of R^p. The unknown vectors x_i will thus correspond to points within this simplex.

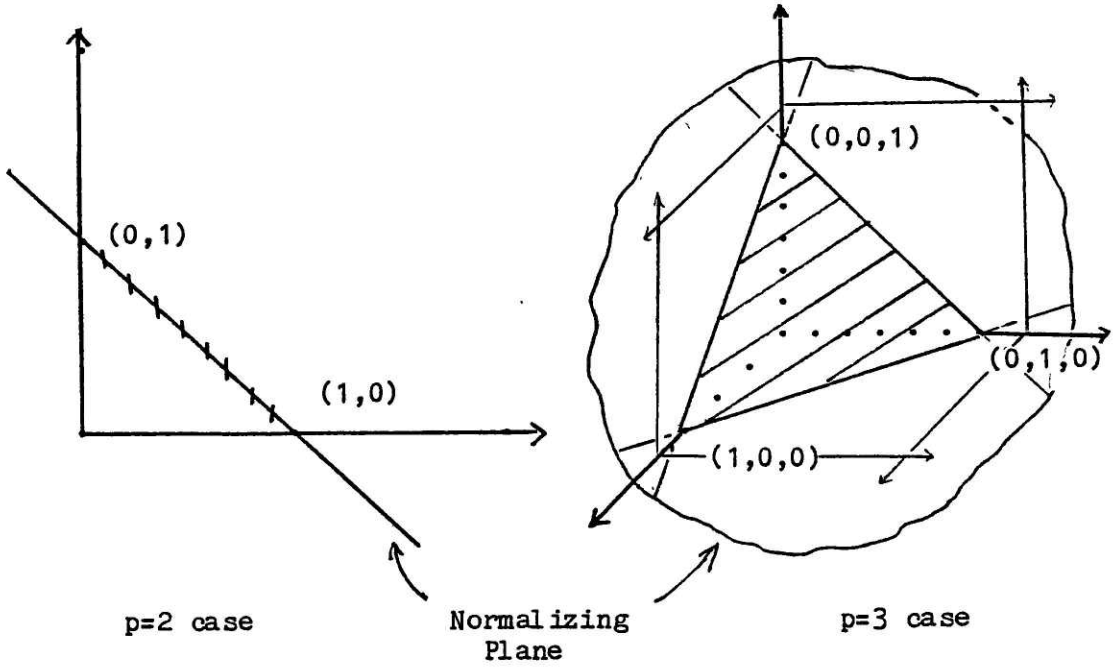


Figure 4.2.1-5.

As noted earlier, in this noise-free situation, only the directions of the data vectors m_j play a role in reconstructing the x_i, as is evident from Result 2. We can therefore work within the unit-sum normalization of these as well. In other words, we compute

$$\grave{m}_j = \frac{m_j}{1^T m_j}, \quad i=1,\ldots,q \qquad (4.2.1\text{-}7)$$

which has unit sum and (because the scale factor $1^T m_j$ is positive) points in the same direction as m_j.

Now the normalized data points $\grave{m}_j$ also appear as points lying in the portion of the normalizing plane that intersects the positive orthant of measurement space, namely in the (p-1)-simplex referred to above. These normalized data points are just the intersecting points of the normalizing plane with the rays emanating from the origin of RP in the direction of the data vectors m_i (see Figure 4.2.1-6). It should

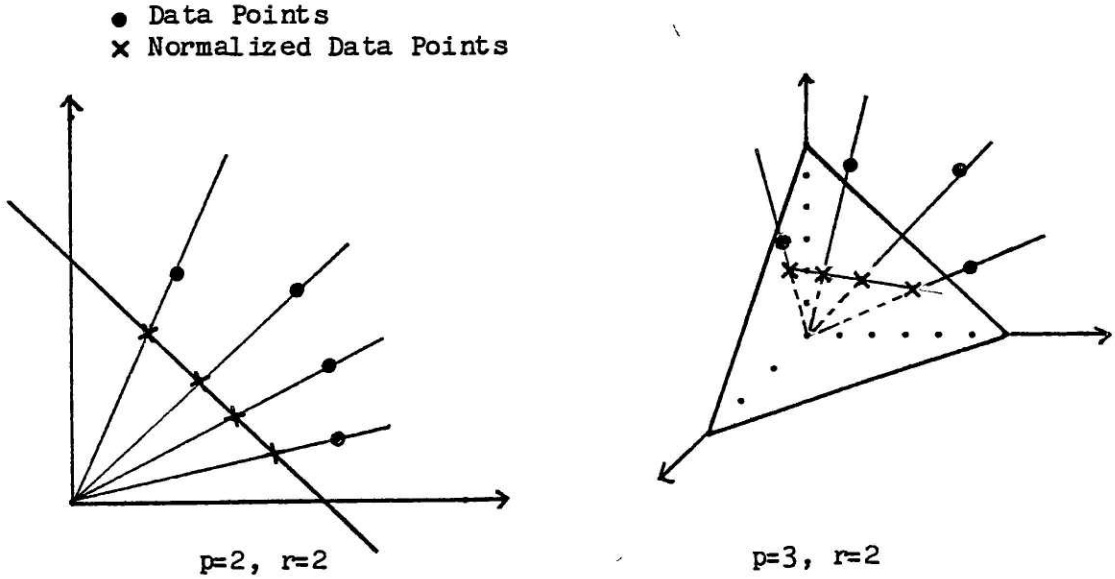


Figure 4.2.1-6.

thus be clear that these normalized data points $\grave{m}_j$ (as well as the normalized unknown vectors $\grave{x}_i$) all lie in the intersection of the signal space with the normalizing plane, and we shall term this intersection the NORMALIZED SIGNAL SPACE. For an r-component mixture, the normalized signal space is (r-1)-dimensional, and all the normalized data points fall within this space.

After the above normalization, Result 2 now takes the form of Result 3 as summarized below.

Result 3

Given a unit-sum normalized data matrix $\grave{M}$,

$$\grave{M} = [\grave{m}_1,\ldots,\grave{m}_q] \; (=\grave{c}_1\grave{x}_1\grave{y}_1^T+\ldots+\grave{c}_r\grave{x}_r\grave{y}_r^T)$$

$$\grave{c}_i\geq 0,\ \grave{x}_i\geq 0,\ \text{and}\ \grave{y}_i\geq 0,\ i=1,\ldots,r$$

$$1^T\grave{m}_j=1,\ j=1,\ldots,q$$

$$1^T\grave{x}_i=1,\ i=1,\ldots,r$$

where both $\{\grave{x}_i\}$ and $\{\grave{y}_i\}$ are independent sets, the (tips of) normalized unknown vectors $\grave{x}_i$

(i) lie in, and span, the normalized signal space;

(ii) lie in (the portion of normalized signal space contained in) the positive orthant of measurement space;

and the convex hull of (tips of) $\grave{x}_i$

(iii) contains (the tips of) the normalized data vectors $\grave{m}_j$.

[]

With our standard independence assumptions, the convex hull of (tips of) $\grave{x}_i$ form a (r-1)-dimensional simplex. Hence, from Result 3, we have that the solution points (the tips of $\grave{x}_i$) are the vertices of a (r-1)-dimensional simplex that lies between the inner bound generated by

(the tips of) the normalized data vectors $\grave{m}_j$, and the <u>outer bound</u> formed by the intersection of the normalized signal space with the normalizing plane (a (p-1)-dimensional regular simplex) - see Figures 4.2.1-7a, b. A proof of Result 3 is given in Appendix G.

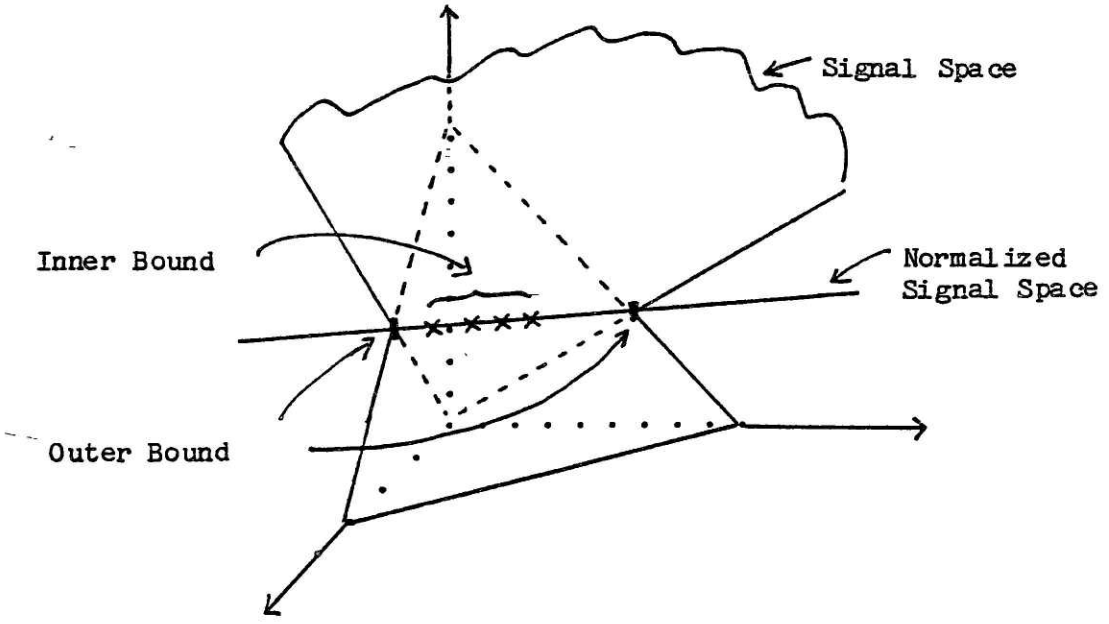


Figure 4.2.1-7a. (p=3, r=2)

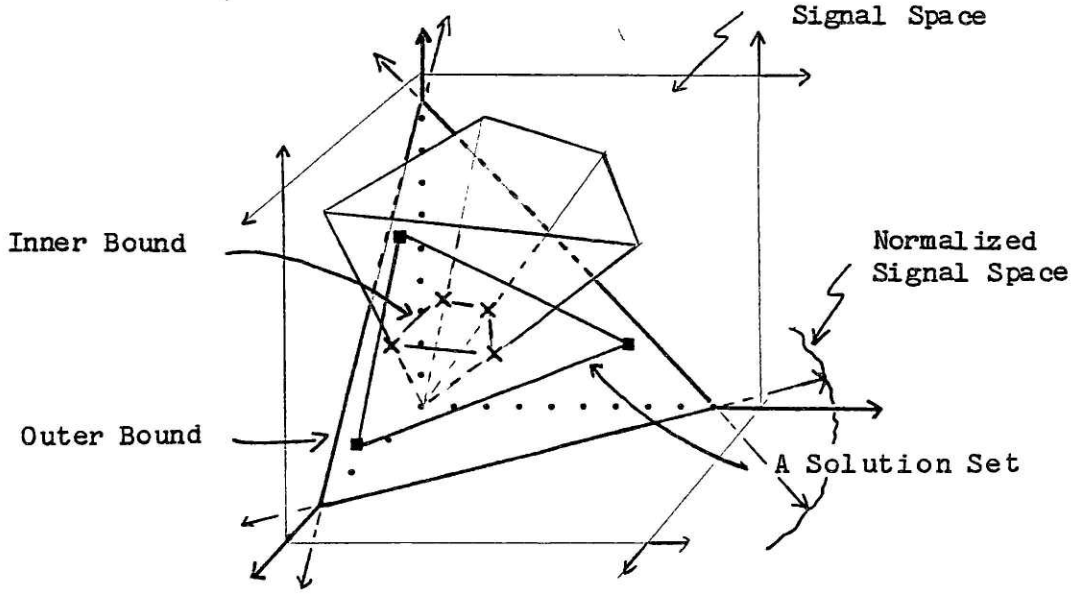


Figure 4.2.1-7b. (p=3, r=3)

4.2.2 Parametrized Signal Space

In Section 4.2.1 it was shown that the unknown vectors x_i (and hence the data vectors m_i) lie in an r-dimensional subspace (the signal space) of R^p (the measurement space), where r is the number of components in the mixture and p is the dimension of the measurement space. In this section, we first give a method for transforming the unknown components solution from the measurement space to the lower (if $r<p$) dimensional signal space. Geometrically, this transformation is equivalent to performing rotations on the coordinate axes of R^p until r of the axes align with the column space of M (or equivalently, p-r of the axes are orthogonal to the column space of M). Then, in the rotated coordinate system, only r coordinates are needed to completely characterize any point in the subspace spanned by the columns of M, thereby reducing the number of parameters to be estimated/displayed.

The above transformation procedure is actually a very effective way to determine r from M in the first place, and is the procedure that is extended in Section 4.3 to the case of noisy data. Thus even if it happens that r=p, we would still carry out the transformation, because it is through this that we in fact establish that r=p.

We go on in this section to show how the unit-sum normalization discussed in Section 4.2.1 may be efficiently carried out in conjunction with the preceding transformation. The normalization enables us to further reduce the dimension of the space within which we have to work, from r to r-1. The advantages of working in a lower dimensional space are:

(i) data display in a lower dimensional space (particularly advantageous for $r\leq 3$), and

(ii) simplified solution methodologies.

We outline the details of the transformation below.

Solution space transformation

Given the pxq data matrix

$$M \; (= c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T) \qquad (4.2.2\text{-}1)$$

let the SVD of M be given by,

$$M = [U_1, U_2] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} V_1^T \\ \hline V_2^T \end{array}\right] = U_1 S_1 V_1^T \qquad (4.2.2\text{-}2)$$

$$S_1 = \mathrm{diag}(s_1, \ldots, s_r)$$

Then the columns of U_1 form a set of orthonormal basis vectors for the column space of M [K2] (i.e., the signal space). Since

$$M = U_1 Z \qquad (4.2.2\text{-}3a)$$

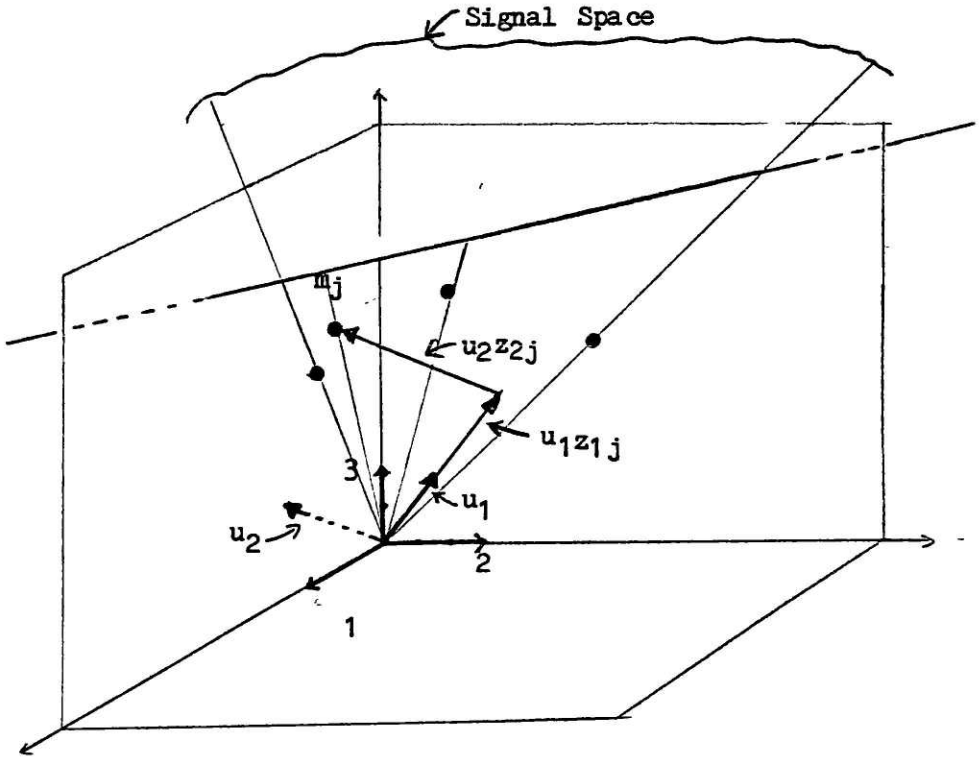
where

$$Z = S_1 V_1^T \qquad (4.2.2\text{-}3b)$$

is an rxq matrix of full column rank, we can express the data vectors m_j as linear combinations of the columns of U_1

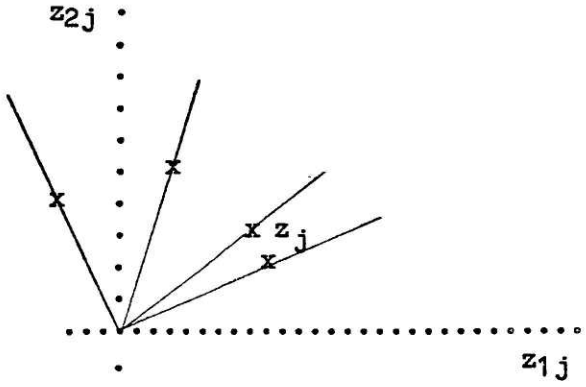
$$m_j = U_1 z_j = u_1 z_{1j} + \ldots + u_r z_{rj}, \qquad j=1,\ldots,q \qquad (4.2.2\text{-}3c)$$

where z_j is the j-th column of Z. Thus, we can represent the vectors mj in terms of r-dimensional vectors z_j, which are of lower dimension. We shall thus call Z the PARAMETRIZED DATA MATRIX, and its column space will be termed the PARAMETRIZED SIGNAL SPACE (see Figure 4.2.2-1).



Measurement Space

Parametrization



Parametrized Signal Space

Figure 4.2.2-1.

Furthermore, since the unknown vectors x_i, $i=1,\ldots,r$, also span the column space of M, it is the case that the columns of U_1 and the unknown vectors x_i span the same space. Hence, we can express the x_i as linear combinations of the columns of U_1. That is,

$$x_i = U_1 w_i = u_1 w_{1i} + \ldots + u_r w_{ri} \qquad (4.2.2\text{-}4a)$$

or

$$X = U_1 W \qquad (4.2.2\text{-}4b)$$

where X is a pxr matrix of unknown vectors and W is a rxq matrix of unknown weighting (parametrized) variables. The problem of estimating the matrix X is now reduced to that of estimating W, a lower dimensional matrix.

Applying Equations 4.2.2-1,3,4 to PROBLEM 2a (Chapter 2), the unknown components problem in the parametrized signal space is given by:

PROBLEM 2c.

Given	$Z\ (=c_1 w_1 y_1^T + \ldots + c_r w_r y_r^T)$
Estimate	w_i, $i=1,\ldots,r$
Subject to	$c_i \geq 0$, $y_i \geq 0$, and $U_1 w_i \geq 0$

□

Normalized Version of Transformation

As in PROBLEM 2b, a unit-sum constrained (normalized) version of PROBLEM 2c is given by:

PROBLEM 2d.

Given	$Z\ (=c_1 w_1 y_1^T + \ldots + c_r w_r y_r^T)$
Estimate	w_i, $i=1,\ldots,r$

Subject to (i) $c_i \geq 0$, $y_i \geq 0$, and $U_1 w_i \geq 0$

(ii) $1^T w_i = 1$, $i=1,\ldots,r$

□

With the above unit-sum constraints on the parametrized unknown vectors w_i the solution is further restricted to the normalizing plane in this space (see Figure 4.2.2-2). Also, as suggested earlier, it suffices in this noise-free case to work with unit-sum normalizations of the parametrized data as well. The effect of this is to map the data from the p-dimensional measurement space to the (r-1)-dimensional normalizing plane in the parametrized signal space.

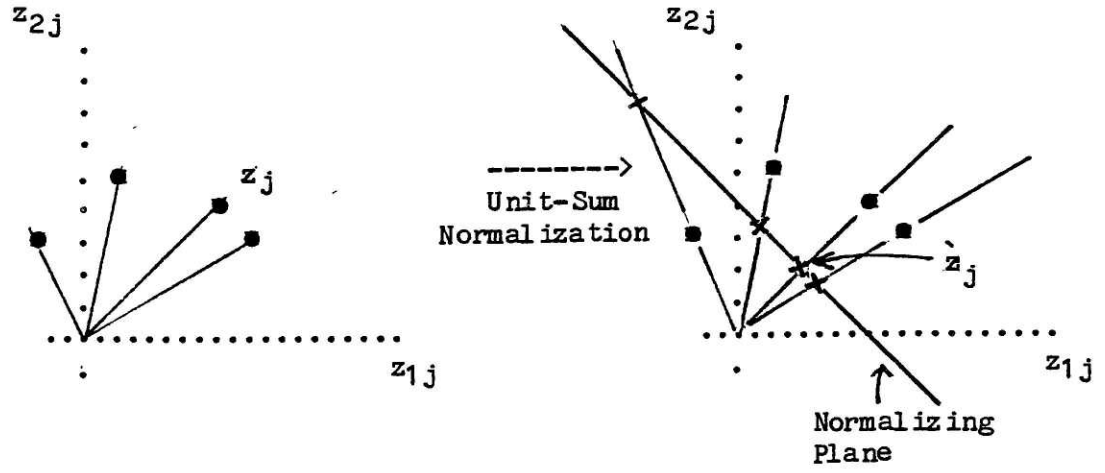


Figure 4.2.2-2.

To carry out the normalizations in a manner that allows us to map back easily to the original normalized problem, using Equations 4.2.2-3c and 4.2.2-4a, we proceed as follows. First normalize the columns m_j of M and u_i of U_1 to unit sum by computing, for $1^T m_j \neq 0$ and $1^T u_i \neq 0$

$$\grave{m}_j = \frac{m_j}{1^T m_j} \qquad (4.2.2\text{-}5a)$$

$$\grave{u}_i = \frac{u_i}{1^T u_i} \qquad (4.2.2\text{-}5b)$$

Then instead of Equations 4.2.2-3c and 4.2.2-4a, we can write

$$\grave{m}_j = \grave{U}_1 \grave{z}_j \qquad (4.2.2\text{-}6a)$$

and

$$x_i = \grave{U}_1 \grave{w}_i \qquad (4.2.2\text{-}6b)$$

where $\grave{U}_1$ is the matrix whose columns are the $\grave{u}_i$ (which are mutually orthogonal but no longer of unit length), and $\grave{z}_j$ and $\grave{w}_i$ are appropriate vectors. Now using the fact that

$$1^T \grave{U}_1 = 1^T \qquad (4.2.2\text{-}6c)$$

we find, on premultiplying Equation 4.2.2-6a,b by 1^T, that

$$1^T \grave{z}_j = 1 \qquad (4.2.2\text{-}7a)$$

and

$$1^T x_i = 1^T \grave{w}_i \qquad (4.2.2\text{-}7a)$$

The first of these equations show that $\grave{z}_j$ is already the unit-sum parametrized data vector we seek, and examining the step that led to

Equation 4.2.2-6a gives an explicit formula for $\grave{z}_j$

$$\grave{z}_j = ([z_{1j}(1^T u_1),\ldots,z_{rj}(1^T u_r)]^T)/(1^T m_j) \qquad (4.2.2\text{-}8)$$

The second equation, Equation 4.2.2-7b, show that if $\grave{w}_i$ is picked to have unit-sum normalization (so $1^T\grave{w}_i=1$), then x_i, computed according to Equation 4.2.2-6b, will also have unit sum.

The unit-sum normalization of the $\grave{z}_j$ allows us to represent them with r-1 instead of r parameters, for example by their last r-1 components: $\grave{z}_{2j},\ldots,\grave{z}_{rj}$. The normalized original data vector $\grave{m}_j$ (see Equation 4.2.2-6a) can be written in terms of these as follows (see Figure 4.2.2-3):

$$\grave{m}_j = \grave{u}_1 + \grave{z}_{2j}(\grave{u}_2-\grave{u}_1) + \ldots + \grave{z}_{rj}(\grave{u}_r-\grave{u}_1) \qquad (4.2.2\text{-}9)$$

Solution Characterization in the Parametrized Solution Space

We have shown in the preceding section that the data vectors m_j in R^p are related to the parametrized vectors z_j in R^r by a linear transformation

$$m_j = U_1 z_j \qquad (4.2.2\text{-}10)$$

Since the columns of U_1 are orthonormal vectors that form a set of basis vectors for the column space of $\{m_j\}$, the transformation is one-to-one, and its inverse transformation exist, and is given by

$$z_j = U_1^T m_j \qquad (4.2.2\text{-}11)$$

Therefore, every vector in the space spanned by the data vectors m_j maps one-to-one, through the transformation U_1^T, to a corresponding vector in the space spanned by z_j.

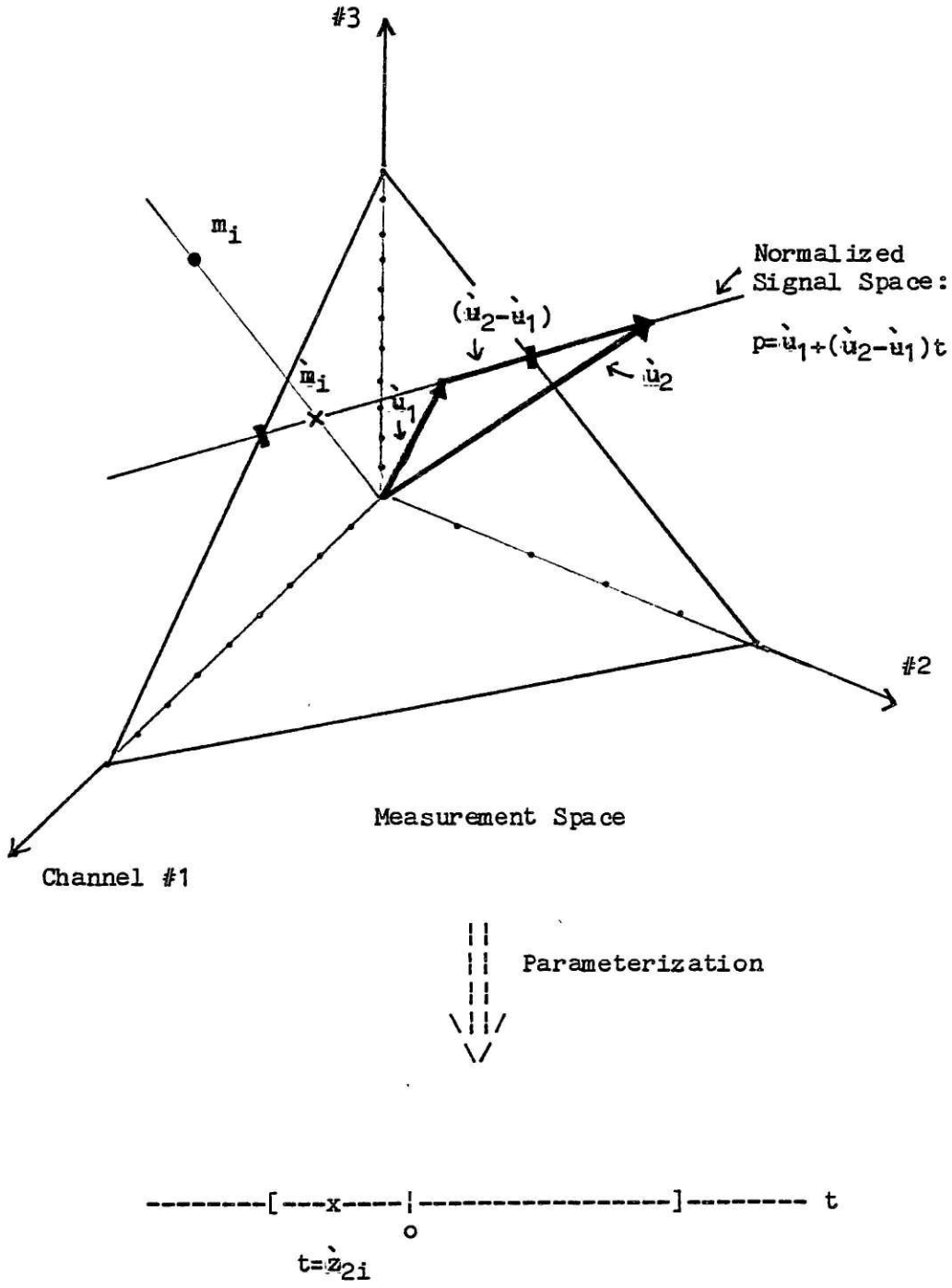
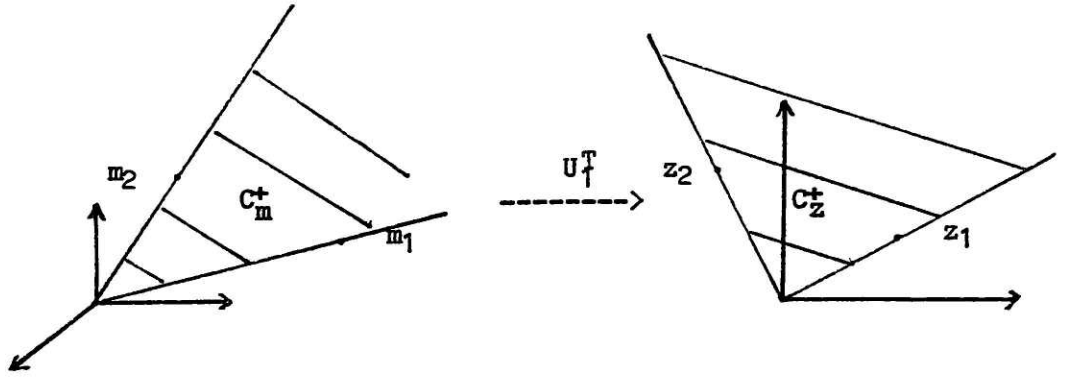

#3
m_i
Normalized
Signal Space:
$p=\grave{u}_1+(\grave{u}_2-\grave{u}_1)t$
$(\grave{u}_2-\grave{u}_1)$
$\grave{m}_i$
$\grave{u}_1$
$\grave{u}_2$
#2
Measurement Space
Channel #1
Parameterization
--------[---x----|----------]-------- t
o
$t=\grave{z}_{2i}$
Parameterized-Normalized-Signal-Space

Figure 4.2.2-3.

Now, the positive cone C_m^+ of any set $c_1,\ldots,c_s$ (see Equation 4.2.1-4 for definition) in the space spanned by the m_j maps, by the above transformation, into the positive cone C_z^+ of the corresponding set $U_1^T c_1,\ldots,U_1^T c_s$ in the space spanned by the z_j (see Figure 4.2.2-4): we have,

$$\begin{aligned} C_z^+ &= U_1^T C_m^+ \\ &= \{z: z=U_1^T(a_1c_1+\ldots+a_sc_s),\ a_i\geq 0,\ i=1,\ldots,s\} \\ &= \{z: z=a_1(U_1^Tc_1)+\ldots+a_s(U_1^Tc_s),\ a_i\geq 0,\ i=1,\ldots,s\} \end{aligned} \qquad (4.2.2\text{-}12)$$



3-Dimensional

2-Dimensional

$C_m^+=\{m: m=a_1m_1+a_2m_2,\ a_i\geq 0\}$

$C_z^+=\{z: z=U_1^T(a_1m_1+a_2m_2),\ a_i\geq 0\}$

$=\{z: z=a_1U_1^Tm_1+a_2U_1^Tm_2,\ a_i\geq 0\}$

$=\{z: z=a_1z_1+a_2z_2,\ a_i\geq 0\}$

Figure 4.2.2-4. (p=3, r=2)

Hence a characterization of the solution set to Problem 2c (unknown components problem in the parametrized space) is given by (see Figure 4.2.2-5):

Result 4

Given an rxq parametrized data matrix Z

$$Z = [z_1,\ldots,z_q] \ (=c_1w_1y_1^T+\ldots+c_rw_ry_r^T)$$

$$c_i\geq 0,\ y_i\geq 0,\ U_1w_i\geq 0,\ i=1,\ldots,r,\ U_1^TU_1=I$$

where both $\{w_i\}$ and $\{y_i\}$ are independent sets, the parametrized unknown vectors w_i

(i) lie in, and span, the parametrized signal space;

(ii) lie inside a cone (an <u>outer bounding cone</u>) in R^r given by

$$\bigcap_{k=1}^{p} \{a: u_{k1}a_1+\ldots+u_{kr}a_r<0,\ a=[a_1,\ldots,a_r]\in R^r\} \qquad (*)$$

where $\cap$ denotes "intersection of";

(iii) generate a positive cone that (lie inside the cone given by (*) and) contains z_j.

□

Result (i) follows from the independence assumption, (ii) essentially gives the transformed outer bounding cone of Result 3, and (iii) follows from the definition of positive cone and the positivity constraints $c_i\geq 0, y_i\geq 0$. A detailed proof of above result is given in Appendix G.

Normalized Version of Solution Characterization

With the unit-sum normalization scheme detailed in subsection "Normalized Version of Transformation", the normalized version of Result 4 (and a solution charactrization of Problem 2d) is given by Result 5

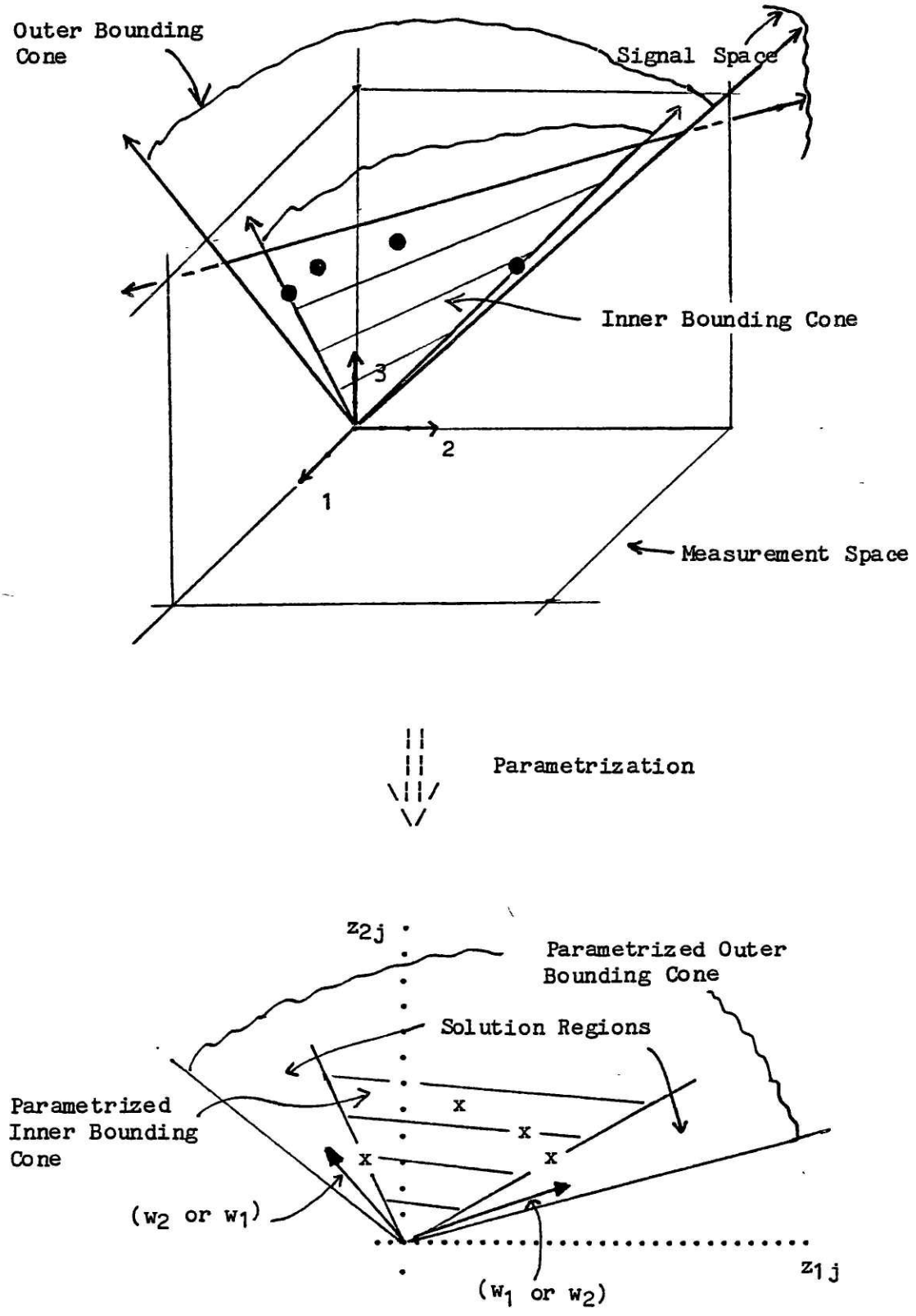

Outer Bounding Cone
Signal Space
Inner Bounding Cone
3
2
1
Measurement Space
Parametrization
z_{2j}
Parametrized Outer Bounding Cone
Solution Regions
Parametrized Inner Bounding Cone
x
x
x
x
$(w_2$ or $w_1)$
z_{1j}
$(w_1$ or $w_2)$

Figure 4.2.2-5.

(see Figure 4.2.2-6).

Result 5

Given a unit-sum normalized parametrized data matrix $\grave{Z}$

$$\grave{Z} = [\grave{z}_1,\ldots,\grave{z}_q] \ (=\grave{c}_1\grave{w}_1\grave{y}_1^T+\ldots+\grave{c}_r\grave{w}_r\grave{y}_r^T)$$

$$\grave{c}_i\geq 0,\ \grave{y}_i\geq 0,\ \grave{U}_1\grave{w}_i\geq 0,\ i=1,\ldots,r$$

$$1^T\grave{z}_j=1,\ j=1,\ldots,q$$

$$1^T\grave{w}_i=1,\ i=1,\ldots,r$$

$$1^T\grave{U}_1=1^T$$

where both $\{\grave{w}_i\}$ and $\{\grave{y}_i\}$ are independent sets, the (tips of) normalized parametrized unknown vectors $\grave{w}_i$

(i) lie in, and span, the normalized parametrized signal space;

(ii) lie inside a convex polygon given by the intersection of the normalizing plane (in the parametrized signal space) with the outer bounding cone in Result 4(ii);

and the convex hull of (tips of) $\grave{w}_i$

(iii) form an (r-1)-dimensional simplex that (lie inside the convex polygon in (ii) and) contains the (tips of) $\grave{z}_j$.

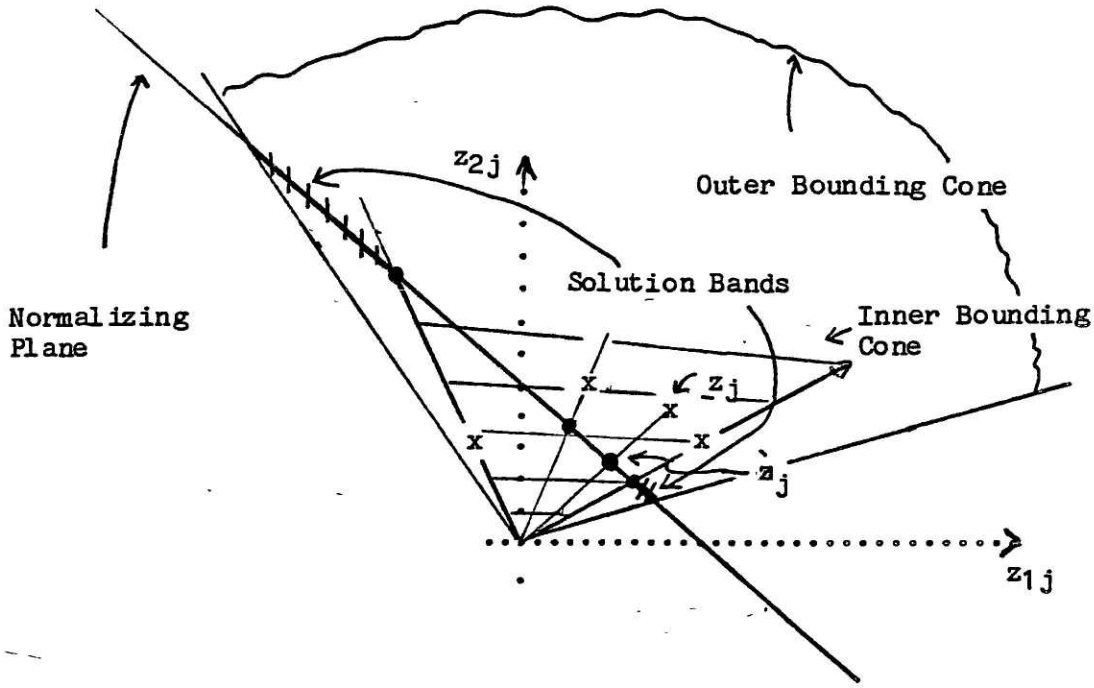

z_{2j}
Outer Bounding Cone
Solution Bands
Normalizing Plane
Inner Bounding Cone
z_j
z_j
z_{1j}
Parametrized Signal Space

Figure 4.2.2-6.

4.3 EXTENSIONS TO NOISY DATA CASE

Ideas and results developed in Section 4.2 for the noise free case are extended and modified in this section to address the noisy data case. In Section 4.3.1 we give methods of approximating the underlying signal space from a noisy data matrix. We then address in Section 4.3.2 the problem of projecting data and uncertainty regions onto this approximated signal space. Finally, in Section 4.3.3, we discuss the normalized version of the preceding section.

4.3.1. Reconstructing the signal space

Noise Model and Transformation to Isotropic Noise

We have shown in Section 4.2 that, given the noise-free data matrix M of an r-component mixture, the columns of M span an r-dimensional subspace (the signal space) in R^p (the measurement space). Now, consider a perturbation of M by an error matrix E (see Section 2.5, Chapter 2)

$$\tilde{M} = M + E \qquad (4.3.1\text{-}1)$$

so that the data vectors m_j are modified by the error vectors e_j

$$\tilde{m}_j = m_j + e_j \qquad (4.3.1\text{-}2)$$

If we assume that the primary source of error is the noise in the absorbance detectors, we may (see Section 2.5, Chapter 2) model the i-th entries of the error vectors e_j (i.e., the measurements from the detector at i-th wavelength) to be random variables from independent symmetric distributions of mean 0 and standard deviation w_i. We can then model the error vectors e_j as independent random vectors, whose mean is 0 and variance is a diagonal matrix with entries w_i^2:

$$\text{Variance } (e_j) = W^2, \quad W = \text{diag } \{w_i\} \qquad (4.3.1\text{-}3)$$

It is convenient, before proceeding further, to scale the absorbance measurements relative to each other, so as to reflect the different levels of noise at the different detectors. Note that Equation (4.3.1-3) constitutes a set of equations, one for absorbance at each of the chosen wavelengths. Multiplying both sides of any of these constituent equations by any nonzero scalar does not destroy or add to the information in that equation. If we now multiply the i-th equation by bw_i^{-1}, for each i and some fixed b>0, or equivalently multiply both sides of Equation 4.3.1-2 by bW^{-1}, we get a set of equations of the same form as Equation 4.3.1-2, except that the noise sources in each equation now all have standard deviation b (see Figure 4.3.1-1).

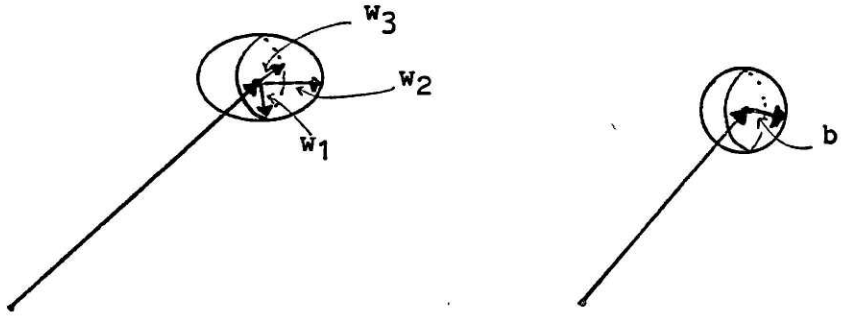


Figure 4.3.1-1. Transformation to Isotropic Noise.

We shall from now on, when referring to Equation 4.3.1-2, assume that this transformation has been carried out, so that $\tilde{m}_j$ in Equation 4.3.1-2 will henceforth be taken as an absorbance vector whose entries

have been scaled by some fixed multiple b of the reciprocals of the standard deviations of the associated detectors, and similarly for the m_j and e_j. The entries of e_j are thus to be taken from now on as (independent, symmetric, zero mean) variables of standard deviation b. The actual value of b is chosen by considerations of convenience in handling the resulting numbers in Equation 4.3.1-2.

Approximating the Signal Space Via SVD

Now given noisy data vectors $\tilde{m}_j$, unless the error vectors e_j lie in the same space as the data vectors m_j, the $\tilde{m}_j$ do not lie in the signal space of the underlying noise-free data matrix M (see Figure 4.3.1-2).

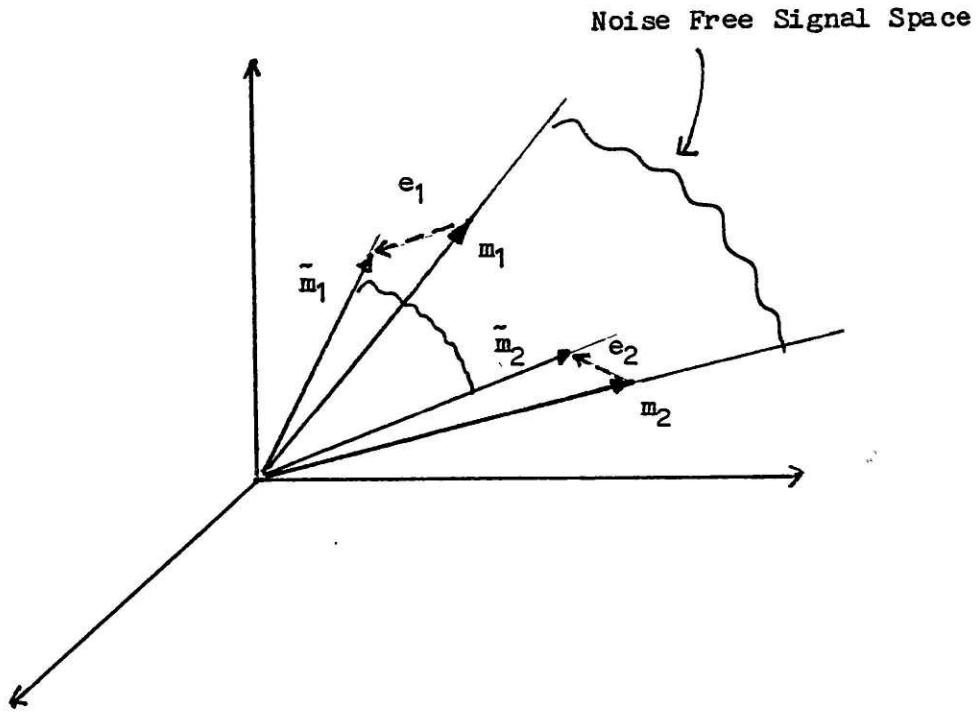


Figure 4.3.1-1. Effect of noise on signal space.

In fact, because of the random nature of the elements of e_j, we would expect the error vectors e_j to span the measurement space R^p when $q \geq p$

(which is typically the case in our application), and consequently the noisy data matrix $\tilde{M}$ would have full-rank. Clearly, the degree to which the noisy data vectors $\tilde{m}_j$ deviate from the underlying signal space depends on the magnitudes of the error vectors e_j, and these magnitudes consequently determine the fidelity of the reconstructed signal space. The question then is: how do we deduce, from examination of the full-rank matrix $\tilde{M}$, that its columns might have lain in a lower dimensional signal space, if not for the noise, and how do we estimate this underlying signal space ?

We have already addressed the first part of this question in Section 3.3.2a. There it was shown that the singular value decomposition (SVD) provides an effective mean for determining the "true" rank r of a noisy data matrix. The SVD also serves to compute a subspace of dimension r that most nearly (in a natural sense to be made clear subsequently) contains the columns of $\tilde{M}$, and that therefore serves us as an estimate of the signal space: specifically, the space spanned by the first r left singular vectors of $\tilde{M}$, which we shall term the RECONSTRUCTED SIGNAL SPACE.

A geometric interpretation of the above reconstruction is as follows. Of all the subspaces of dimension r, the reconstructed signal space minimizes the sum of squares of the distances from the data vectors $\tilde{m}_j$ to itself. Since we have assumed isotropic noise structure, this is an appropriate criterion to be used in approximating the underlying signal space. We shall now detail this construction below:

Given noisy data matrix $\tilde{M}$ with the singular value decomposition (see Appendix A)

$$\tilde{M} = [U_1, U_2] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & S_2 \end{array}\right] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} \tag{4.3.1-4}$$

where S_1 is an rxr diagonal matrix that contains the "significant" singular values of M, and hence determines the "true" rank of the matrix (see Section 3.3.2a for a discussion on the number of components), we have shown in Section 3.2.2 that the reconstructed data matrix (see Chapter 3)

$$\hat{M} = [U_1, U_2] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 S_1 V_1^T \tag{4.3.1-5}$$

is optimum in the sense that it minimizes the error

$$||\tilde{M}-\hat{M}||_2 \tag{4.3.1-6}$$

under the constraint that $\hat{M}$ has rank r (Theorem 3, Chapter 3). The space spanned by the columns of $\hat{M}$ (or equivalently by the columns of U_1) is, as defined above, the reconstructed signal space. The geometric interpretation of the "optimal" property of the reconstructed signal space is easily shown by noting that

$$\begin{aligned} ||\tilde{M}-\hat{M}||_2 &= ||\tilde{M}-U_1 S_1 V_1^T||_2 \\ &= ||\tilde{M}-U_1 U_1^T \tilde{M}||_2 \\ &= ||(I-U_1 U_1^T)\tilde{M}||_2 \\ &= ((||(I-U_1 U_1^T)\tilde{m}_1||_2)^2+\ldots+(||(I-U_1 U_1^T)\tilde{m}_q||_2)^2)^{1/2} \end{aligned} \tag{4.3.1-7}$$

That is, as stated above, the reconstructed signal space minimizes the sum of squares of the distances from the data vectors $\tilde{m}_j$ to itself.

Furthermore, since

$$||\tilde{M}-\hat{M}||_2 = ||U_2 S_2 V_2^T||_2 \qquad (4.3.1\text{-}8)$$
$$= s_{r+1}$$
$$\leq ||E||_2 \qquad \text{(from Theorem 1, Chapter 3)}$$

the distances of the tips of the noisy data vectors $\tilde{m}_j$ from the reconstructed signal space are assured to be less than the spectral norm of the error matrix. In addition, since

$$||\tilde{m}_j-\hat{m}_j||_2 \leq ||E||_2 \qquad (4.3.1\text{-}9a)$$
$$||\tilde{m}_j-m_j||_2 \leq ||E||_2 \qquad (4.3.1\text{-}9b)$$

the distances between the tips of the reconstructed (projected) data vectors and the underlying noise free data vectors are guaranteed to be less than twice the spectral norm of the error matrix, i.e.

$$||m_j-\hat{m}_j||_2 \leq 2.||E||_2 \qquad (4.3.1\text{-}10)$$

4.3.2 Representation in Parametrized Reconstructed Signal (PRS) Space

Data Projection to PRS Space

Since the $\hat{m}_j$ lie in the r-dimensional span(U_1), we can parametrize them with lower dimensional <u>parametered reconstructed data vectors</u> $\hat{z}_j$ as follows (cf. Section 4.2.2)

$$\hat{m}_j = U_1 \hat{z}_j \qquad (4.3.2\text{-}1)$$

where $\hat{z}_j$ is the j-th column of

$$\hat{Z} = S_1 V_1^T \qquad (4.3.2\text{-}2)$$

Error Projection to PRS Space

Given error vectors e_j in the measurement space R^p, consider their projections $\hat{e}_j$ to the reconstructed signal space span(U_1), given by

$$\hat{e}_j = U_1 U_1^T e_j \tag{4.3.2-3}$$

Since projected error vectors $\hat{e}_j$ lie in span(U_1), we can parametrize $\hat{e}_j$ with lower dimensional parametred error vectors $\hat{a}_j$ as follows (see Section 4.2.2)

$$\hat{e}_j = U_1 \hat{a}_j \tag{4.3.2-4}$$

where

$$\hat{a}_j = U_1^T e_j \tag{4.3.2-5}$$

If the elements of e_j are independent, normally distributed, zero mean random variables with variance w^2, then the elements of $\hat{a}_j$ are also independent, normally distributed, zero mean random variables with variances w^2:

$$\begin{aligned} E(\hat{a}_j) &= E(U_1^T e_j) \\ &= U_1^T E(e_j) \\ &= 0 \end{aligned} \tag{4.3.2-6a}$$

$$\begin{aligned} E(\hat{a}_j \hat{a}_j^T) &= E(U_1^T e_j e_j^T U_1) \\ &= U_1^T E(e_j e_j^T) U_1 \\ &= U_1^T w^2 I U_1 \\ &= w^2 U_1^T U_1 \\ &= w^2 I \end{aligned} \tag{4.3.2-6b}$$

Note that we are able to show Equation 4.3.2-6b precisely because of the transformation to isotropic noise.

Alternatively, if the elements of e_j are bounded in the following way ("box" bound)

$$-w1 \leq e_j \leq w1 \qquad (4.3.2\text{-}7)$$

where the inequalities are componentwise, then the elements of the parametrized error vectors a_j are (conservatively) bounded by

$$-w|U_1|^T 1 \leq \hat{a}_j \leq w|U_1|^T 1 \qquad (4.3.2\text{-}8)$$

where $|.|$ is the componentwise absolute value operator, and the inequalities are componentwise.

Note: The region defined by Equation 4.3.2-8 (r-dimensional cube) is a (conservative) bound on the projection of the region defined by Equation 4.3.2-7 (p-dimensional cube) onto the r-dimensional subspace span(U_1), not the projection itself. However, it does provide an efficient means of bounding the projected error. The actual projection of the error bound given by Equation 4.3.2-9 onto the span(U_1) is given by the convex hull of the projections of the 2^p vertices of the p-dimensional cube defined by Equation 4.3.2-9 onto span(U_1) - see Figure 3.3.2-1.

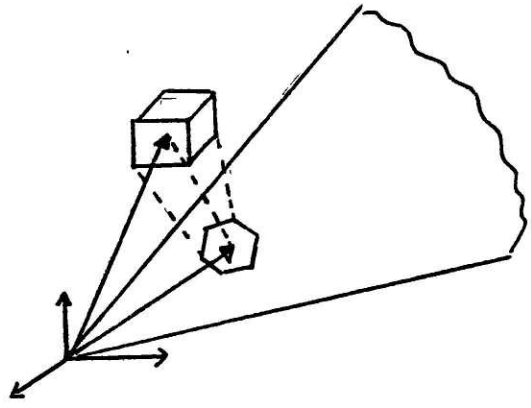


Figure 4.3.2-1. Uncertainty region projection.

4.3.3 Representation in Normalized Parameterized Reconstructed Signal (NPRS) Space

Noisy Data Representation in the NPRS Space

Given noisy data matrix $\tilde{M}$, the problem of data representaion in the NPRS space completely parallels that of the noise-free case (cf. Section 4.2.2) if we work with the reconstructed data matrix $\hat{M}$. Hence given the SVD of $\hat{M}$

$$\hat{M} = [U_1, U_2] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 S_1 V_1^T \qquad (4.3.3\text{-}1)$$

the <u>normalized parametrized reconstructed</u> data vectors $\grave{z}_j$ (note that we have used the same notation as the noise-free case - cf. Equation 4.2.2-8) are given by

$$\grave{z}_j = ([\hat{z}_{1j}(1^T u_1), \ldots, \hat{z}_{rj}(1^T u_r)]^T)/(1^T \hat{m}_j) \qquad (4.3.3\text{-}2)$$

where $\hat{z}_j$ $(=[\hat{z}_{1j}, \ldots, \hat{z}_{rj}]^T)$ is the j-th column of the matrix

$$\hat{Z} = S_1 V_1^T \qquad (4.3.3\text{-}3)$$

The overall procedure is schematically outlined in Figure 4.3.3-1, and detailed in the algorithm below.

Algorithm

<Step 1> Compute SVD of $\tilde{M}$

$$\tilde{M} = [U_1, U_2] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & S_2 \end{array}\right] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 S_1 V_1^T$$

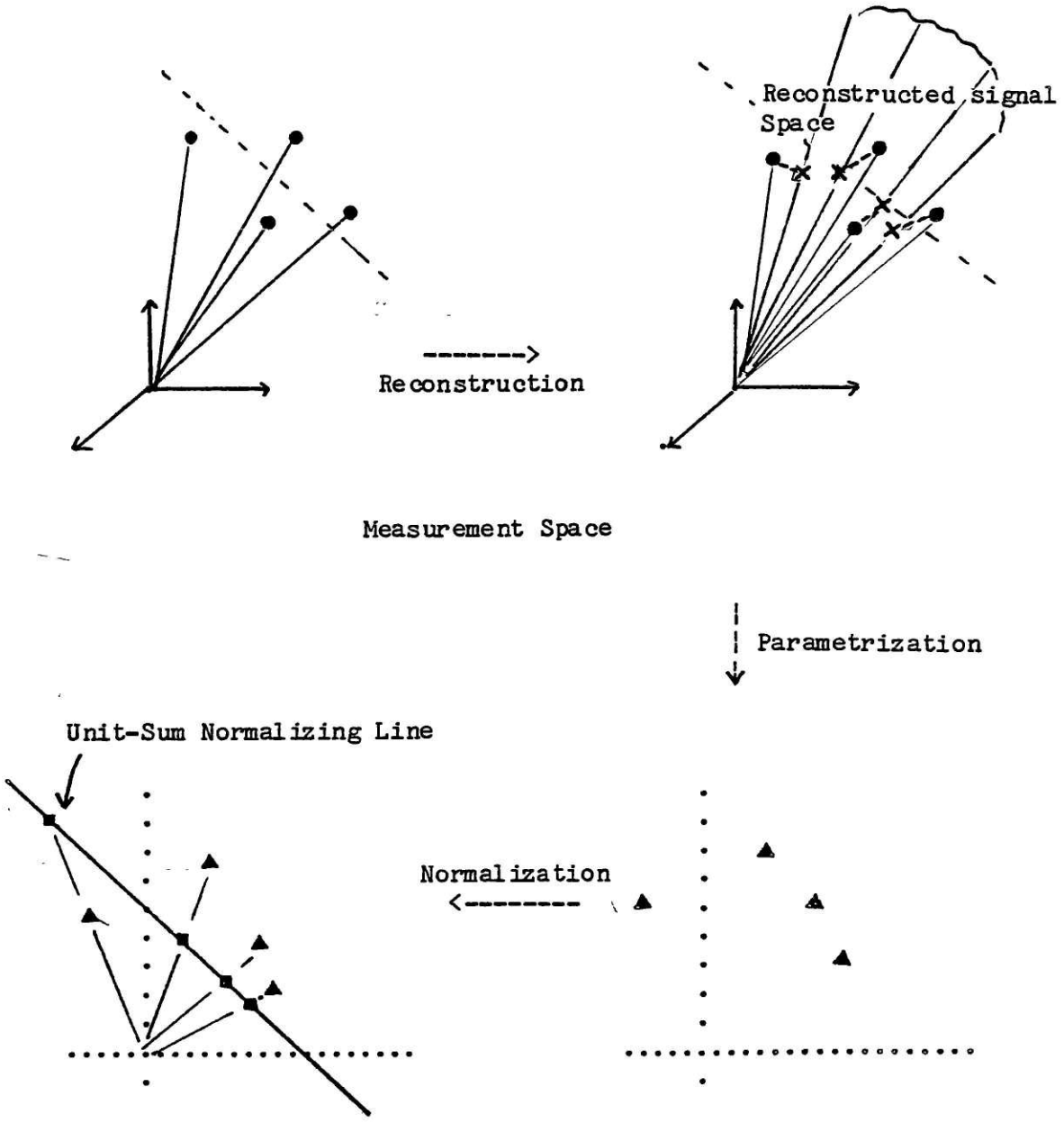

Reconstructed signal Space
Reconstruction
Measurement Space
Parametrization
Unit-Sum Normalizing Line
Normalization
Parametrized Space

Figure 4.3.3-1.

- Dimension of the signal space:

= Numerical rank of $\tilde{M}$, $(S_2)_i \leq EPS = ||E||_2$

- Reconstructed data matrix $\hat{M}$:

$$= [U_1, U_2] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 S_1 V_1^T$$

- Optimally reconstructed signal space:

= span (U_1)

<Step 2> Normalize signal space basis vectors u_j to unit sum:

$$\grave{u}_j = \frac{u_j}{1^T u_j}, \quad i=1,\ldots,r$$

Update S_1:

$$\grave{s}_j = s_j 1^T u_j, \quad i=1,\ldots,r$$

<Step 3> Normalize columns of $\hat{M}$ $(= \grave{U}_1 \grave{S}_1 V_1^T)$ to unit sum:

$$\grave{M} = \hat{M}Q \; (= \grave{U}_1 \grave{S}_1 V_1^T Q)$$

$$Q = \text{diag}\,(q_1,\ldots,q_r)$$

$$q_j = (1^T \tilde{m}_j)^{-1}$$

<Step 4> Find unit-sum normalized parameterizing variables:

$$\grave{Z} = \grave{S}_1 V_1{}^T Q$$

[]

Normalization of the Uncertainty Region in the PRS Space

Given noisy parametrized reconstructed data vectors $\hat{z}_j$ (Equation 4.3.2-2), we shall define the uncertainty regions for these vectors to

be the regions $\{z\}_j$ obtained by letting the corresponding error terms $\hat{a}_j$ (Equation 4.3.2-5) vary over all their uncertainty region $\{\hat{a}\}_j$ (see Figure 4.3.3-2). That is

$$\{z\}_j = \{z: z=\hat{z}_j+\hat{a}_j,\ \hat{a}_j \in \{\hat{a}\}_j\} \qquad (4.3.3\text{-}1)$$

where the uncertainty region $\{\hat{a}\}_j$ is given by the box bound

$$\{\hat{a}\}_j = \{\hat{a}: -w \leq \hat{a} \leq w,\ w>0\} \qquad (4.3.3\text{-}2)$$

Note that the region $\{z\}_j$ is (approximately) the region in which the noise free parametrizing data vectors could lie - approximately because the reconstructed signal space is only an estimate of the underlying noise-free signal space, hence the uncertainty region $\{z\}_j$ is only an approximation to the "true" uncertainty region. In this section, we give a method of representing the uncertainty regions $\{z\}_j$ when they are transformed to be unit-sum normalized.

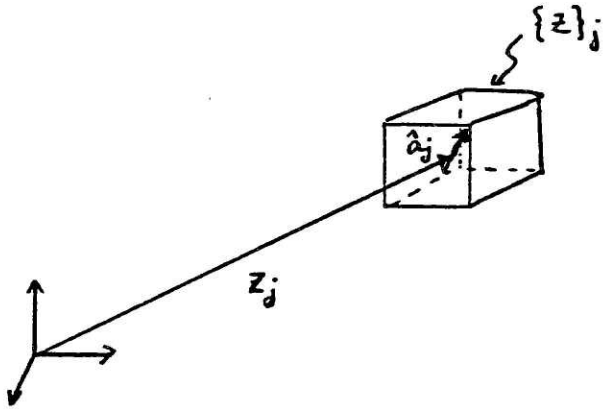


Figure 4.3.3-2.

The uncertainty region around the unit-sum normalized parameterized data vector $\grave{z}_j$,

$$\grave{z}_j = \frac{\hat{z}_j}{1^T \hat{z}_j} \tag{4.3.3-3}$$

is the unit-sum normalization of the uncertainty box $\{z\}_j$, and is given by the convex hull of the unit-sum normalized vertices of the uncertainty box (see Figure 4.3.3-3).

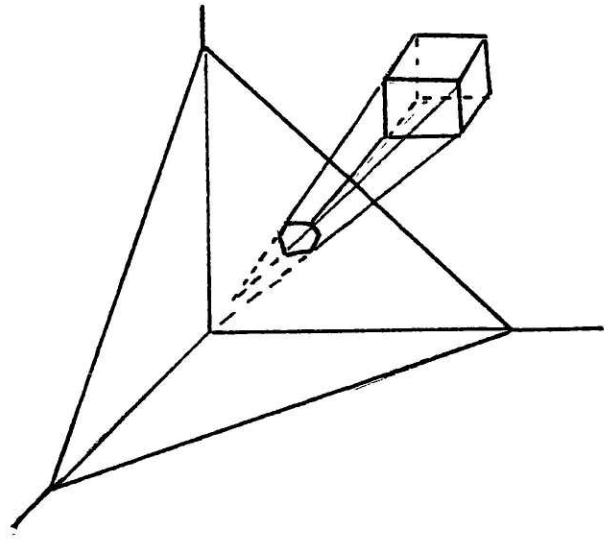


Figure 4.3.3-3.

This can be seen by noting that every point p that is an element of the uncertainty box $\{z\}_j$ is a convex combination of the 2^r vertices of the box (where r is the dimension of the parameterized space), i.e.

$$\begin{aligned} p &= c_1 z_{v1} + \ldots + c_k z_{vk}, \quad k=2^r \\ c_1 &+ \ldots + c_k = 1 \\ c_i &\geq 0 \end{aligned} \tag{4.3.3-4}$$

where z_{vi} are the 2^r vertices of the uncertainty box $\{z\}_j$. The unit-sum

normalization of p is then given by

$$\frac{p}{1^T p} = \frac{c_1 z_{v1} + \ldots + c_k z_{vk}}{c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk}} \qquad (4.3.3\text{-}5)$$

$$= \frac{c_1 1^T z_{v1}}{c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk}} \cdot \frac{z_{v1}}{1^T z_{v1}} + \ldots$$

$$\ldots + \frac{c_k 1^T z_{vk}}{c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk}} \cdot \frac{z_{vk}}{1^T z_{vk}}$$

If we assume that $1^T p$ $(=c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk})$ is not zero for any p in z_j, that is

$$c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk} \neq 0 \qquad (4.3.3\text{-}6)$$

for any (subject to the constraints in Equation 4.3.3-3) c_i, then the unit sum normalizing factor $1^T p$ $(=c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk})$ does not change its sign for any allowable values of c_i (since its value must be zero before it changes its sign). Furthermore, by looking at all the limiting cases where each of the c_i is equal to 1, i=1,...,k, we find that all the unit-sum normalized vertices $(1^T z_{vi}$, i=1,...,k) must have the same sign, otherwise $1^T p$ would change its sign. Hence

$$\frac{c_i 1^T z_{vi}}{c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk}} \geq 0, \quad i=1,\ldots,k \qquad (4.3.3\text{-}7)$$

Furthermore, since

$$\frac{c_i 1^T z_{v1}}{c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk}} + \ldots + \frac{c_i 1^T z_{vk}}{c_1 1^T z_{v1} + \ldots + c_k 1^T z_{vk}} = 1 \qquad (4.3.3\text{-}8)$$

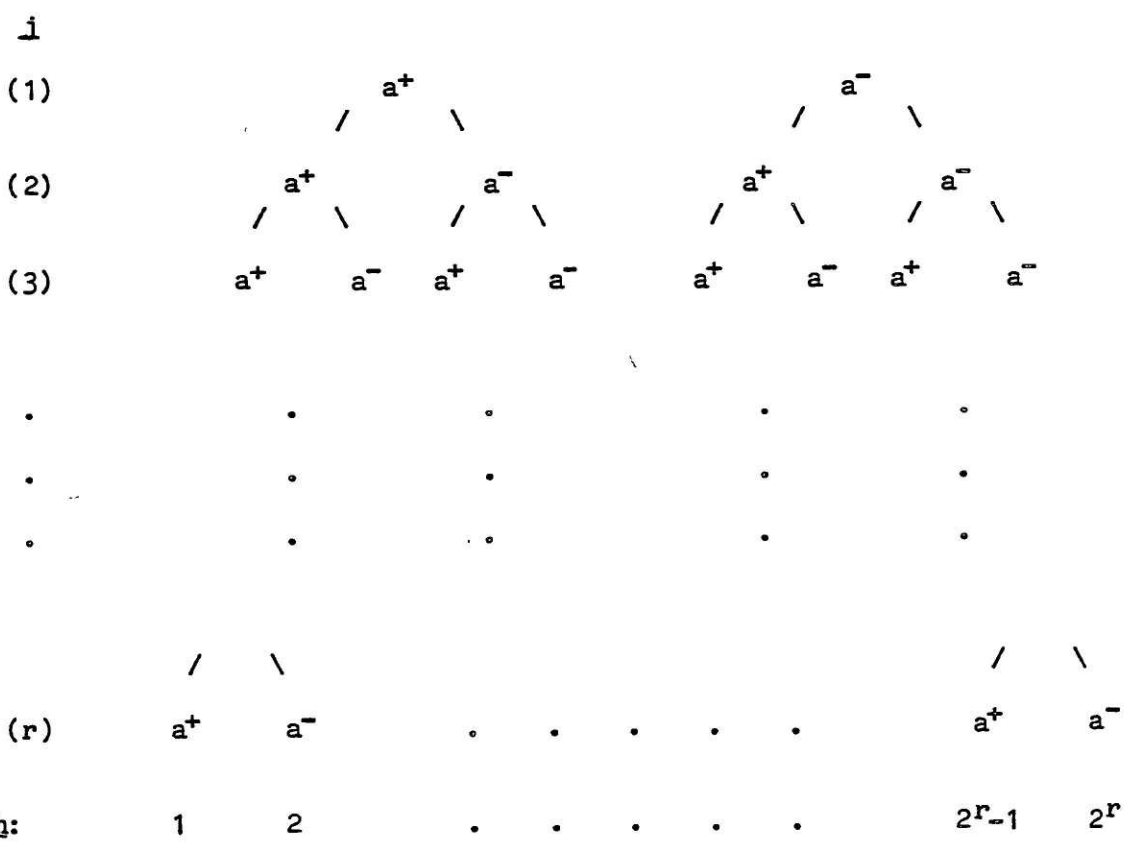
we have that the unit-sum normalized p is a convex combination of the unit-sum normalized vertices of the uncertainty box.

Therefore, the uncertainty region around the unit-sum normalized vector $\grave{z}_j$ is given by the convex hull of the unit-sum normalized vertices of the uncertainty box.

The vertices z_{vi} of the uncertainty box are given by

$$z_{vi} = \hat{z} + e_1 a_{1i} + \ldots + e_r a_{ri} \qquad (4.3.3\text{-}9)$$

where e_j ($=[0,..,0,1,0,..,0]^T$) is the j-th unit vector and a_{ji} is the j-th element of the i-th path in the "tree" shown in Figure 4.4.3-4.



(Superscripts "+" and "−" indicate upper and lower bounds)

Figure 4.4.3-3. Vertex tree for the uncertainty region.

Bounding the Unit-Sum Normalized Uncertainty Region

We give here a (conservative) bound on the transformation of the uncertainty region $\{z\}_j$ to the NPRS space when the errors $\hat{a}_i$ are given by a box bound (Equation 4.3.3-2). For convenience, we shall drop the set notation $\{.\}$ when describing the uncertainty regions in this subsection.

Unit-sum normalization of Equation 4.3.3-1 yields

$$\frac{z_i}{1^T z_i} = \frac{\hat{z}_i + \hat{a}_i}{1^T(\hat{z}_i + \hat{a}_i)} \qquad (4.3.3\text{-}10)$$

An upper bound on the each of the elements of the normalized uncertainty regions is then given by

$$\max_{\hat{a}_i} \frac{\hat{z}_{ji} + \hat{a}_{ji}}{1^T(\hat{z}_i + \hat{a}_i)}, \quad j=1,\ldots,r \qquad (4.3.3\text{-}11)$$

and a lower bound by

$$\min_{\hat{a}_i} \frac{\hat{z}_{ji} + \hat{a}_{ji}}{1^T(\hat{z}_i + \hat{a}_i)}, \quad j=1,\ldots,r \qquad (4.3.3\text{-}12)$$

In addition, a tighter bound on the uncertainty region can be obtained by considering the bounds on the relative slopes of the elements of the uncertainty region vector

$$\frac{\hat{z}_{ji} + \hat{a}_{ji}}{\hat{z}_{ki} + \hat{a}_{ki}}, \quad j,k=1,\ldots,r \; ; \; j \neq k \qquad (4.3.3\text{-}13)$$

An upper bound is given by

$$\max_{\hat{a}_{ji},\hat{a}_{ki}} \frac{\hat{z}_{ji} + \hat{a}_{ji}}{\hat{z}_{ki} + \hat{a}_{ki}}, \quad j,k=1,\ldots,r \; ; \; j \neq k \tag{4.3.3-14}$$

and an lower bound by

$$\min_{\hat{a}_{ji},\hat{a}_{ki}} \frac{\hat{z}_{ji} + \hat{a}_{ji}}{\hat{z}_{ki} + \hat{a}_{ki}}, \quad j,k=1,\ldots,r \; ; \; j \neq k \tag{4.3.3-15}$$

4.4 SOLUTION PROCEDURES: TWO- AND THREE-COMPONENTS CASES

We present in this section, detailed procedures for obtaining the solution sets characterized in the previous sections. We shall limit our discussion to the two- and three-components cases, and to solutions with unit-sum constraints on the unknown vectors (PROBLEM 2b). The procedures will be developed in the parametrized domains, and the solutions obtained will then be transformed back to the measurement space. For the case of noisy data matrices, we shall work with the reconstructed data matrices. Hence, from now on, we shall not (with repect to the notations used) distinguish between the noise-free and noisy cases. The solution sets will be in the form of bounds on the elements of the unknown vectors. The two-components case is discussed in Section 4.4.1, and the three-components case in Section 4.4.2.

4.4.1 Two-Components Case

Consider data representation on the normalized parametrized solution space. We have, from Equation (4.2.2-6a), a parametrization of the normalized data vectors $\grave{m}_j$ given by

$$\grave{m}_j = \grave{U}_1 \grave{z}_j = \grave{u}_1 \grave{z}_{1j} + \grave{u}_2 \grave{z}_{2j}, \quad j=1,\ldots,q \tag{4.4.1-1}$$

Furthermore, since the $\grave{z}_j$ are unit-sum normalized, using the substitution

$$\grave{z}_{1j} = 1 - \grave{z}_{2j} \tag{4.4.1-2}$$

we have a complete <u>data representation</u> on a 1-dimensional space by the parametrization (see Figure 4.4.1-1)

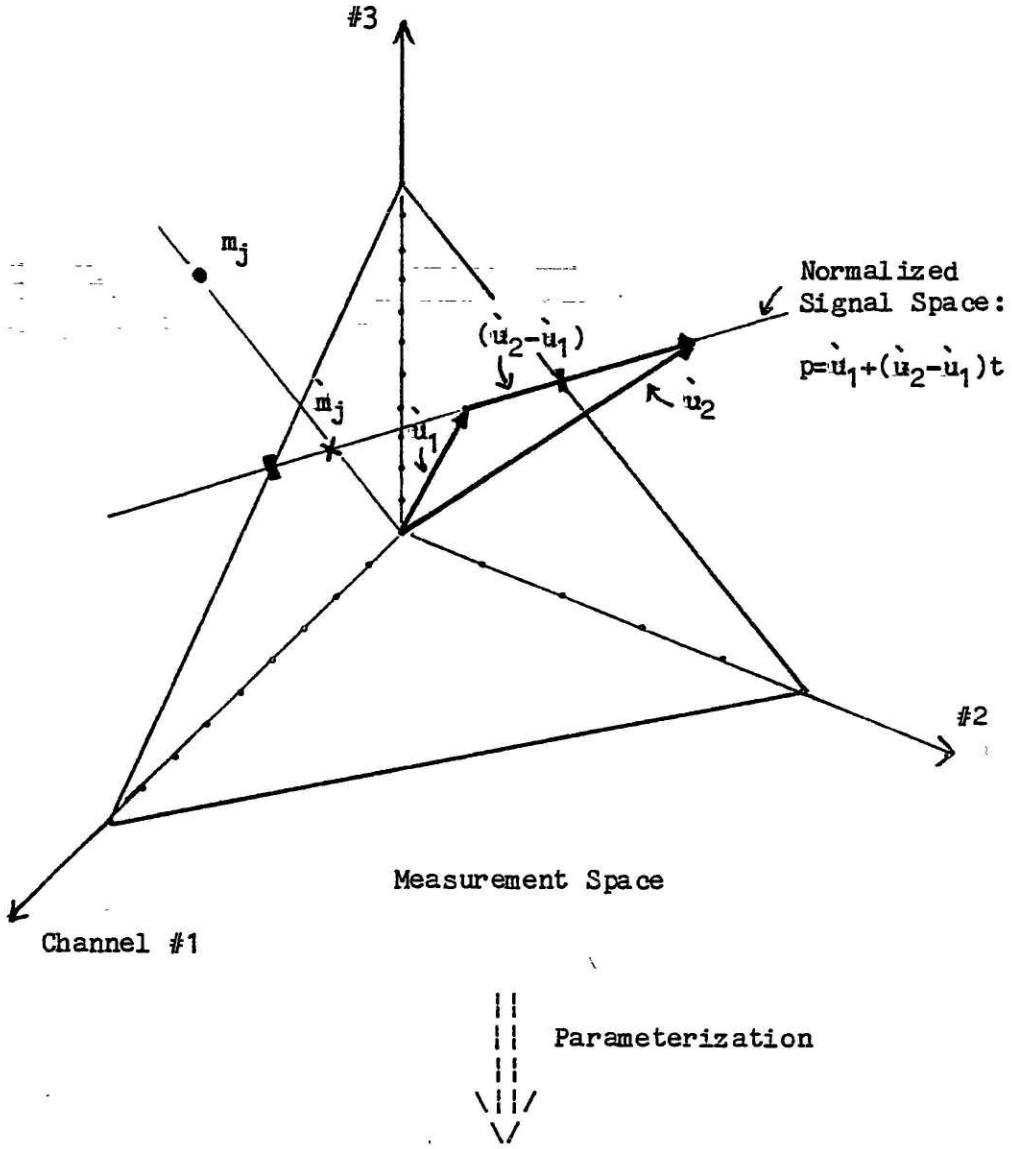

#3
m_j
Normalized
Signal Space:
$p=\dot{u}_1+(\dot{u}_2-\dot{u}_1)t$
$(\dot{u}_2-\dot{u}_1)$
$\dot{m}_j$
$\dot{u}_1$
$\dot{u}_2$
#2
Measurement Space
Channel #1
Parameterization
--------[---x-----|----------]-------- t
o
$t=\dot{z}_{2j}$
Parameterized-Normalized-Signal-Space

Figure 4.4.1-1.

$$\grave{m}_j = \grave{u}_1 + (\grave{u}_2 - \grave{u}_1)\grave{z}_{2j}, \quad j=1,\ldots,q \tag{4.4.1-3}$$

The normalized unknown vectors $\grave{x}_j$ can correspondingly be written as

$$\grave{x}_j = \grave{u}_1 + (\grave{u}_2 - \grave{u}_1)\grave{w}_{2j}, \quad j=1,\ldots,q \tag{4.4.1-4a}$$

and, more generally, an arbitrary point p in the normalized parametrized signal space can be written as

$$p = \grave{u}_1 + (\grave{u}_2 - \grave{u}_1)t \tag{4.4.1-4b}$$

The inner bound on normalized-parametrized solutions $\grave{w}_{2j}$, j=1,2, is given by the convex hull of the parametrized data vectors $\grave{z}_{2j}$ (Result 5, Section 4.2.2), which for this case is given by the line segment joining the points

$$\max_j \grave{z}_{2j} \tag{4.4.1-5a}$$

and

$$\min_j \grave{z}_{2j} \tag{4.4.1-5b}$$

The outer bound is imposed by the componentwise positivity constraint on the normalized unknown vectors $\grave{x}_j$, and can thus be shown from Equation 4.4.1-4 to be given by the line segment joining the points

$$\min_i \left(\frac{\grave{u}_{i1}}{\grave{u}_{i1}-\grave{u}_{i2}}\right) \tag{4.4.1-6a}$$

and

$$\max_i \left(\frac{\grave{u}_{i1}}{\grave{u}_{i1}-\grave{u}_{i2}}\right) \tag{4.4.1-6b}$$

where the subscripts "+" and "-" on the minimizing/maximizing index i indicate, respectively, minimization/maximization over positive and negative values of the ratio

$$\left(\frac{\grave{u}_{i1}}{\grave{u}_{i1}-\grave{u}_{i2}}\right) \qquad (4.4.1\text{-}7)$$

A geometric interpretation of the outer bound given by Equation 4.4.1-6 is as follows (see Figure 4.4.1-1):

The two points that describe the outer bound are precisely the intersections of the normalized signal space (a straight line in this case) with the boundaries of the positive orthant of the measurement space (see Section 4.2.1). The points p on this normalized signal space are parametrized as in Equation 4.4.1-4b. As mentioned in Appendix A (under Frobenius-Perron Theorem), the vector $\grave{u}_1$ has all its elements positive. Thus, when t=0, the vector p also has all its elements positive, that is, it has its tip in the positive orthant of measurement space, and consequently inside the outer bound. Now, as we change t from t=0, the vector p moves towards the boundary of the positive orthant. The values of t that describe the outer bound are those values closest to t=0 for which some component of p reaches 0.

Because the normalized signal space is a line, the inner and outer bounds on the solution sets generate two disjoint regions (see Figure 4.4.1-2) in each of which one of the two unknown parameterizing variables lies.

Once the two solution regions for the parameterizing variables $\grave{w}_{21}$ and $\grave{w}_{22}$ are found, bounds on the unknown vectors $\grave{x}_1$ and $\grave{x}_2$ are given by:

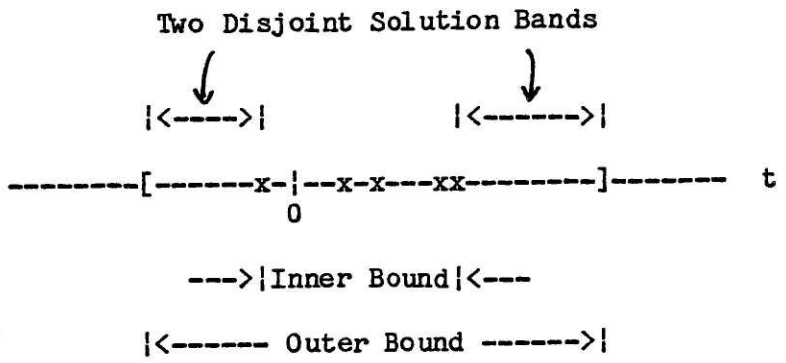


Figure 4.4.1-2.

(i) for $(\grave{u}_{i2}-\grave{u}_{i1})$ positive

$$\grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{-}_{21} \leq \grave{x}_{i1} \leq \grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{+}_{21} \qquad (4.4.1\text{-}9a)$$

$$\grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{-}_{22} \leq \grave{x}_{i2} \leq \grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{+}_{22} \qquad (4.4.1\text{-}9b)$$

(ii) for $(\grave{u}_{i2}-\grave{u}_{i1})$ negative

$$\grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{+}_{21} \leq \grave{x}_{i1} \leq \grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{-}_{21} \qquad (4.4.1\text{-}9c)$$

$$\grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{+}_{22} \leq \grave{x}_{i2} \leq \grave{u}_{i1} + (\grave{u}_{i2}-\grave{u}_{i1})\grave{w}^{-}_{22} \qquad (4.4.1\text{-}9d)$$

where the superscripts "+" and "-" indicate upper and lower bounds.

4.4.2 Three-Components Case

Data Representation

Consider data representation on the normalized parametrized solution space. We have, from Equation (4.2.2-6a), a parameterization of the normalized data vectors $\grave{m}_j$ given by

$$\grave{m}_j = \grave{U}_1\grave{z}_j = \grave{u}_1\grave{z}_{1j} + \grave{u}_2\grave{z}_{2j} + \grave{u}_3\grave{z}_{3j}, \quad j=1,\ldots,q \qquad (4.4.2\text{-}1)$$

Furthermore, since the $\dot{z}_j$ are unit-sum normalized, using the substitution

$$\dot{z}_{1j} = 1 - \dot{z}_{2j} - \dot{z}_{3j} \qquad (4.4.2\text{-}2)$$

we have a complete <u>data representation</u> on a 2-dimensional space by the parametrization (see Figure 4.4.2-1)

$$\dot{m}_j = \dot{u}_1 + (\dot{u}_2 - \dot{u}_1)\dot{z}_{2j} + (\dot{u}_3 - \dot{u}_1)\dot{z}_{3j}, \quad j=1,\ldots,q \qquad (4.4.2\text{-}3)$$

The normalized unknown vectors $\dot{x}_j$ can correspondingly be written as

$$\dot{x}_j = \dot{u}_1 + (\dot{u}_2 - \dot{u}_1)\dot{w}_{2j} + (\dot{u}_3 - \dot{u}_1)\dot{w}_{3j}, \quad j=1,\ldots,q \qquad (4.4.2\text{-}4a)$$

and, more generally, an arbitrary point p in the normalized parametrized space can be written as

$$p = \dot{u}_1 + (\dot{u}_2 - \dot{u}_1)t_1 + (\dot{u}_3 - \dot{u}_1)t_2 \qquad (4.4.2\text{-}4a)$$

Inner Bound

The <u>Inner bound</u> (I) on the normalized parametrized solutions $\dot{w}_{2j}$ and $\dot{w}_{3j}$, j=1,...,3, is given by the convex hull of the parametrized data points with the coordinates $(\dot{z}_{2j},\dot{z}_{3j})$ in the 2-dimensional parametrized space (Result 5, Section 4.2.2), as illustrated in Figure 4.4.2-2. Several efficient algorithms and references for constructing a convex hull of points in a 2-dimensional space can be found in [S1].

Outer Bound

The <u>ouTer bound</u> (T) on the solution sets is generated by the componentwise positivity constraint on the normalized unknown vectors $\dot{x}_j$

$$\dot{u}_1 + (\dot{u}_2 - \dot{u}_1)\dot{w}_{2j} + (\dot{u}_3 - \dot{u}_1)\dot{w}_{3j}\ (=\dot{x}_j) \geq 0, \quad j=1,\ldots,3 \qquad (4.4.2\text{-}5)$$

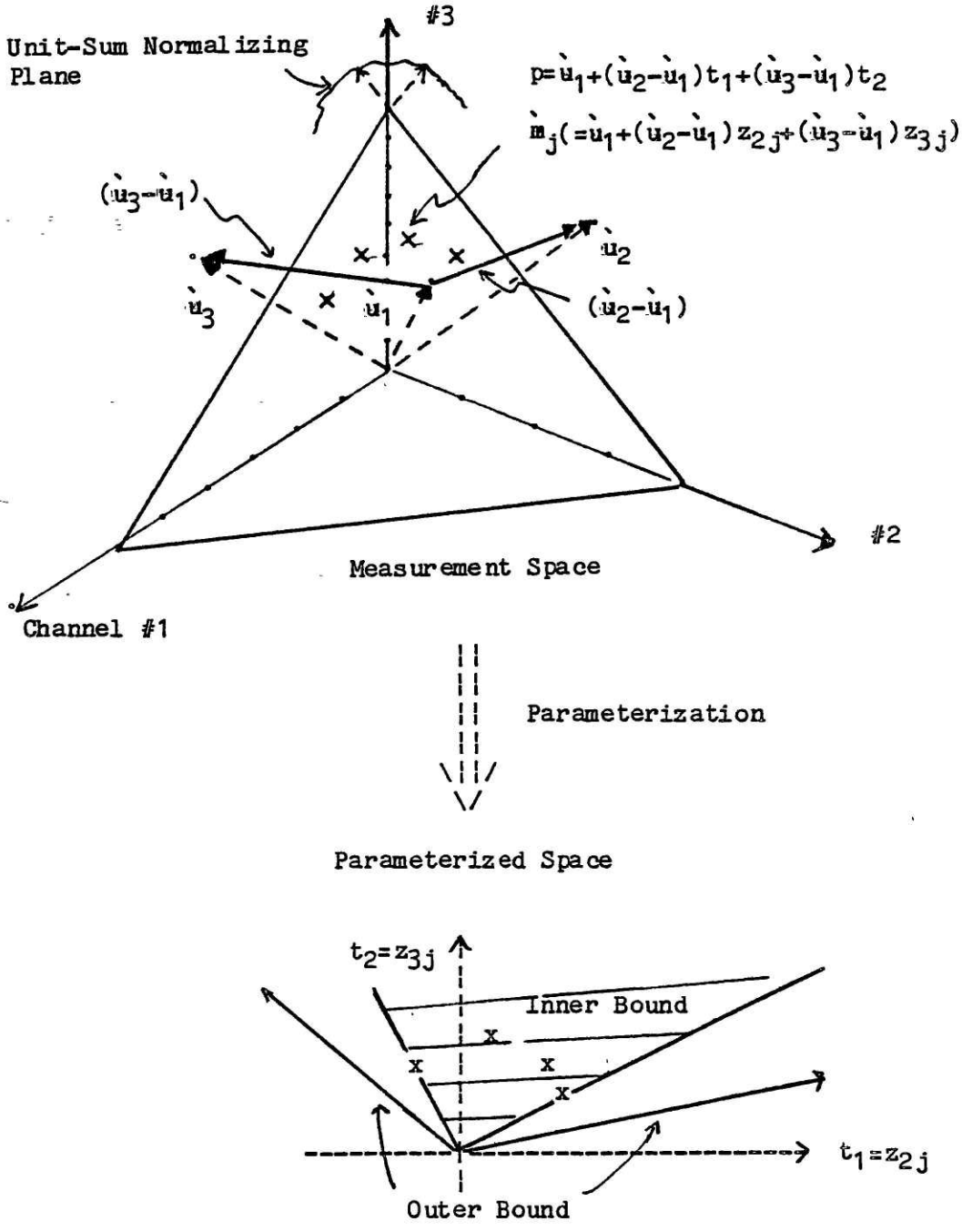


Figure 4.4.2-1. (p=3, r=3 case)

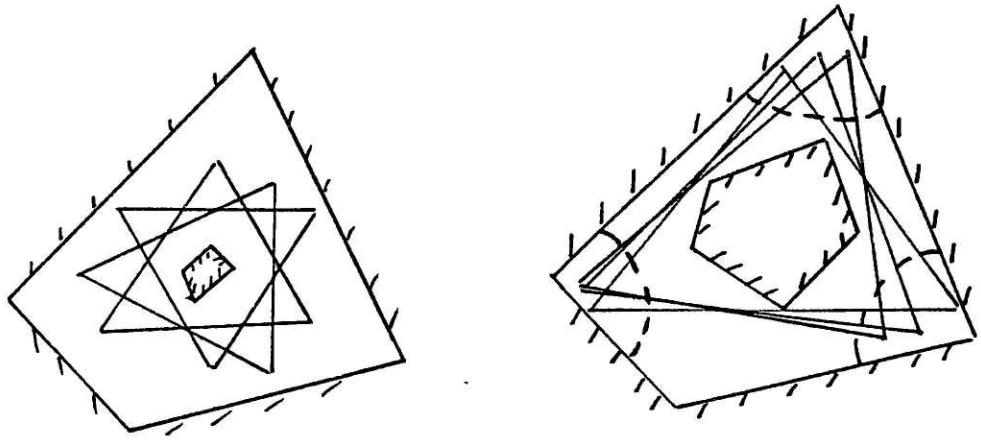
where the inequalities are componentwise inequalities. Each of the elements of the inequality in Equation 4.4.2-5

$$\grave{u}_{i1} + (\grave{u}_{i2} - \grave{u}_{i1})\grave{w}_{2j} + (\grave{u}_{i3} - \grave{u}_{i1})\grave{w}_{3j} \; (=\grave{x}_j) \geq 0, \quad j=1,\ldots,p \tag{4.4.2-6}$$

restricts the solution space (the region in which the solutions $\grave{w}_{2j}$ and $\grave{w}_{3j}$ can lie) to a half-plane in R^2. Consequently, the intersection of all the allowable half-planes as defined by the inequalities in Equation 4.4.2-6 form the outer bound.

Generating Disjoint Regions

Given the inner and outer bounds specified above, the allowable solution sets correspond to the vertices of all triangles that enclose the inner bound while lying inside the outer bound (Result 5, Section 4.2.2). Unlike the two-component case, however, the bounds in the three-component case may <u>not</u> divide the allowable solution region into three disjoint sets, each of which contains one of the unknown components.



(a) No disjoint solution regions. (b) Disjoint solution regions.

Figure 4.4.2-2.

This is the case in Figure 4.4.2-2a, where the inner bound is much smaller than the outer one. In Figure 4.4.2-2b, the bounds are closer together, and the vertices of triangles that lie in T and enclose I form three disjoint regions. We should also note at this point that there may be more than one set of three disjoint regions.

For the cases where we do not have disjoint regions, additional data samples (which may be added in the form of more time samples and/or wavelengths at which the absorbances are detected) may help enlarge the inner bound and/or shrink the outer bound so that disjoint regions are generated. For example, in the case of three wavelengths and three components, it is clear that additional time samples can expand (but not shrink) the inner bound. The exact effect of additional data samples on the bounds requires further investigation, which we shall not pursue much further in this thesis.

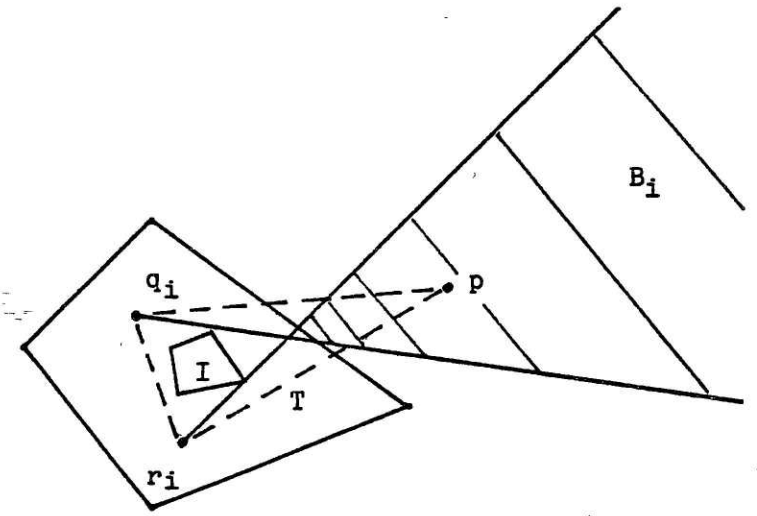
We give in this section a method [G9] for describing the boundaries of the solution sets for the three-component case, generating three disjoint solution regions when the bounds are sufficiently close to each other. In addition, we give an iterative algorithm that quickly approximates the three disjoint regions when they exist.

In constructing the solution sets, we shall use an "extended" solution set B defined as follows (see Figure 4.4.2-3)

$$B = \{p : \text{there are points } q \text{ and } r \text{ in } T \text{ for which the triangle } pqr \text{ contains } I\}$$

We shall call the intersection of B and T the VERTEX DOMAIN V. It is clear that each point in the vertex domain contributes one vertex of a triangle that corresponds to an allowable solution triple, and

conversely.



B={Union of all B_i}

Figure 4.4.2-3.

We now describe a geometric construction that generates the boundary of the set B [G9]:

> The boundary of B is generated by the point p of a rotating triangle pqr that circumscribes I, with points q and r lying on the boundary of T.

As the point r in Figure 4.4.2-4 moves along the segment r_1r_2 and the point q moves along the segment q_1q_2, the point p traces a portion of the boundary of B, and when the point r completely traverses the boundary of T, the path outlined by point p defines the boundary of B. In [G9] it is shown that the path traced by p is composed of several conic arcs, with new conics being generated when the points q or r cross to a different line segment or when the contact points of the triangle pqr with I changes.

Once the boundary of B is completely traced out, bounds on the unknown components can be isolated if the boundary of B separates the solution region into three disjoint sets (see Figure 4.4.2-5).

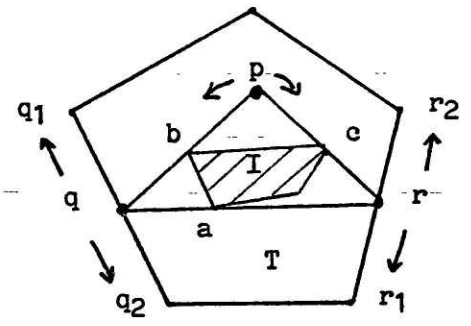


Figure 4.4.2-4 [G9]

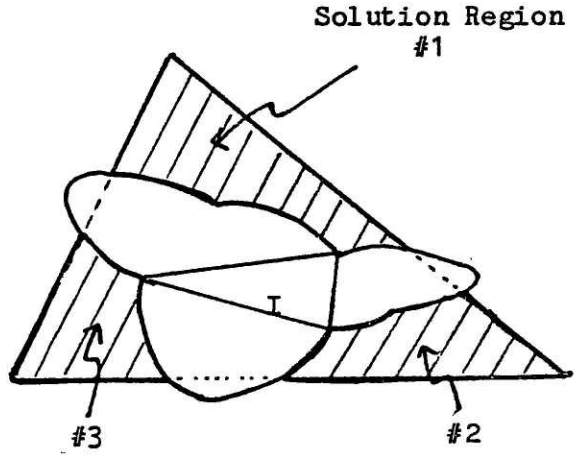


Figure 4.4.2-5 [G9]

An Iterative Approximation

If the vertex domain is composed of three disjoint sets, then an iterative geometric construction (see Figure 4.4.2-6) that efficiently approximates the vertex domain is as follows (see [G9] for proof of existence, convergence, etc.):

Algorithm

Step 1: Starting from any point (p_o in Figure 4.4.2-6a) along the boundary of T, construct the two tangent lines to L

Step 2: For each tangent line, determine the other point of intersection (p_1^+ or p_1^- in Figure 4.4.2-6a)of

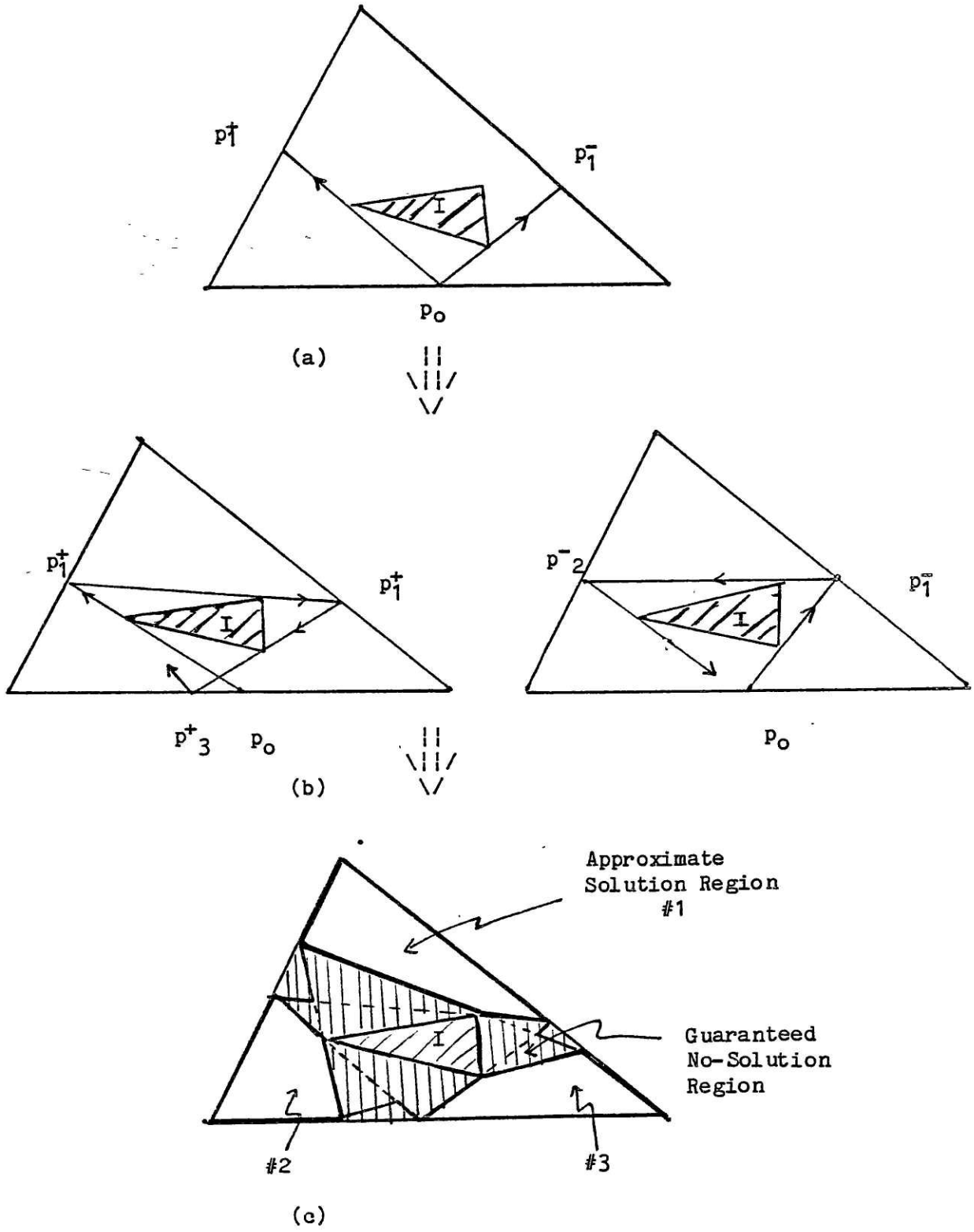

p_1^+
p_1^-
I
p_o
(a)
p_1^+
p_1^+
I
p^+_3
p_o
(b)
p^-_2
p_1^-
I
p_o
Approximate
Solution Region
#1
Guaranteed
No-Solution
Region
I
#2
#3
(c)

Figure 4.4.2-6.

the line with the boundary of T.

Step 3: Using these two points (p_1^+ or p_1^-) as new starting positions, draw "new" tangent lines to I (note that each of these two points now has only one "new" tangent line).

Step 4: Repeat the process from Step (2); stop when the path traversed by each of the two sequence of tangent lines converges to a triangle (within some predetermined tolerance).

The two constructed triangles at the end of Step 4 have vertices in T and sides that touch I, and are referred to as "circumscribing-inscribing" (or CI) triangles.

The interior of the CI triangles can immediately be guaranteed to not contain the vertices of the triangle triples. Beyond this, the intersection of the exteriors of the two CI triangles and the outer bound T forms six disjoint regions, of which only three contain the vertex domain. A simple test for containment is the ability or inability to construct a solution triangle with a vertex in one of the regions. For the regions that do not contain the vertex domain, no point in them can be a vertex of a solution triangle.

The construction above then serves to rule out components that map to points guaranteed to not lie in the vertex domain, and hence reduces the number of components that one may wish to test from some database.

Measurement Space Bounds

Assuming that the vertex domains are three disjoint regions, so

that each of the unknown components is isolated, we can transform the solution regions in the parametrized space to the measurement space as follows:

We have, for each of the three unknown components,

$$\dot{x}_j = \dot{u}_1 + (\dot{u}_2 - \dot{u}_1)\dot{w}_{2j} + (\dot{u}_3 - \dot{u}_1)\dot{w}_{3j} \qquad j=1,\ldots,3 \tag{4.4.2-7}$$

where $(\dot{w}_{2j},\dot{w}_{3j})$, $j=1,\ldots3$, are the coordinates of the points in the three solution regions. For each component j, let (v_{k2j},v_{k3j}) be the coordinates of the k-th vertex of the convex polygon that bounds the j-th solution region. Then an upper and a lower bound on the i-th elements of the unknown vectors $\dot{x}_j$, $j=1,\ldots,3$, are given by

$$\dot{x}^+_{ij} = \max_k \{\dot{u}_{i1} + (\dot{u}_{i2} - \dot{u}_{i1})\dot{v}_{k2j} + (\dot{u}_{i3} - \dot{u}_{i1})\dot{v}_{k3j}\}$$

$$\dot{x}^-_{ij} = \min_k \{\dot{u}_{i1} + (\dot{u}_{i2} - \dot{u}_{i1})\dot{v}_{k2j} + (\dot{u}_{i3} - \dot{u}_{ji})\dot{v}_{k3j}\} \tag{4.4.2-8}$$

where the superscripts "+" and "-" indicate upper and lower bounds.

Note that, if the solution regions are generated by the iterative algorithm outlined above, then they are convex polygons; if not, form convex polygons by constructing the convex hull of the regions (recall that the boundaries of the vertex domains are either segments of the outer bound T, or arcs of conics that are concave inward, so that the convex hulls of these regions result in convex polygons).

4.5 EXAMPLES OF DATA DISPLAY WITH SYNTHETIC LIQUID CHROMATOGRAPHY DATA

We shall in this section give examples of data display with synthetic liquid chromatography data that illustrate the utility of our methods. The examples will be limited to three-components mixtures with absorbance measurements at three wavelengths, so that the measurement space is 3-dimensional. The 3-dimensional plot will be then transformed to a 2-dimensional display by the use of barycentric coordinates (discussed below). The 3-component data matrix M used here is generated as in Example 4 of Chapter 3. Two different sets of resolutions are used to study the effect of peak separations on the data display.

Barycentric Coordinates

When the measurement space is restricted to three dimensions, data display on a 2-dimensional space is possible with a parameterization that does not require matrix decompositions. The first step in this approach is the unit-sum normalization of the data vectors m_j

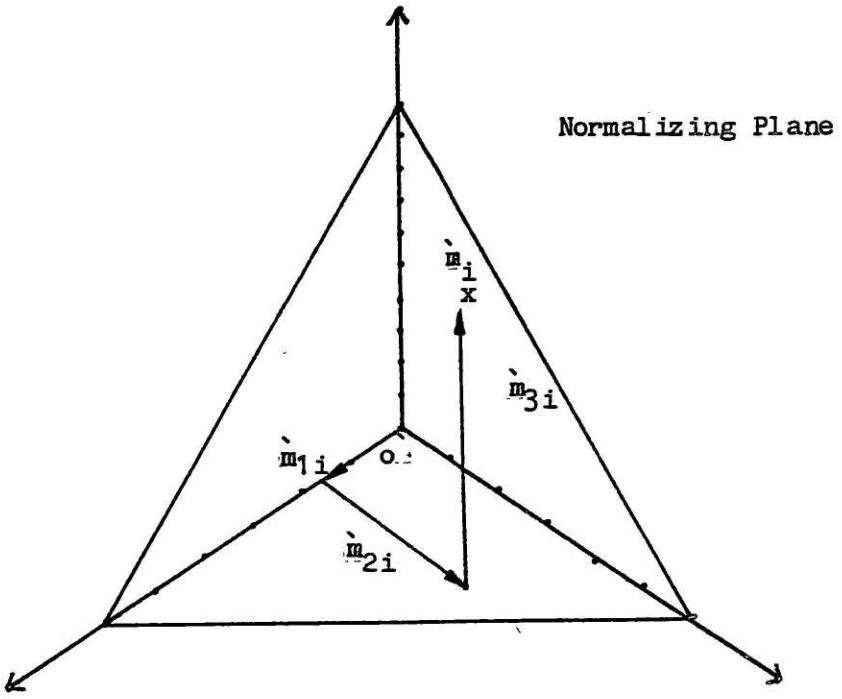
$$\grave{m}_j = \frac{m_j}{1^T m_j} \tag{4.5-1}$$

resulting in a constraint on the elements of $\grave{m}_i$ given by

$$1^T \grave{m}_j = \grave{m}_{1j} + \grave{m}_{2j} + \grave{m}_{3j} = 1 \tag{4.5-2}$$

We can then set up a coordinate system (the so-called barycentric coordinate system [C4]) for the portion of the normalizing plane that lies in the positive orthant and that forms an equilateral triangle as follows. Let the centroid of the equilateral triangle be the origin, and the line segments from this origin to the vertices of the triangle be

the three unit-length axes of the coordinate system (see Figure 4.5-1). The normalized data vectors then map to unique points on the triangle by marking off their components $\grave{m}_{1j}$, $\grave{m}_{2j}$, and $\grave{m}_{3j}$, along the axes as shown in Figure 4.5-1. (The use of such coordinates for the study of 3-phase mixtures is standard in physical chemistry [V1, S4].)



Barycentric Coordinate System

Figure. 4.5-1

Given these coordinates, we can transform them into a rectangular coordinate system by the easily derived transformation (see Figure 4.5-2)

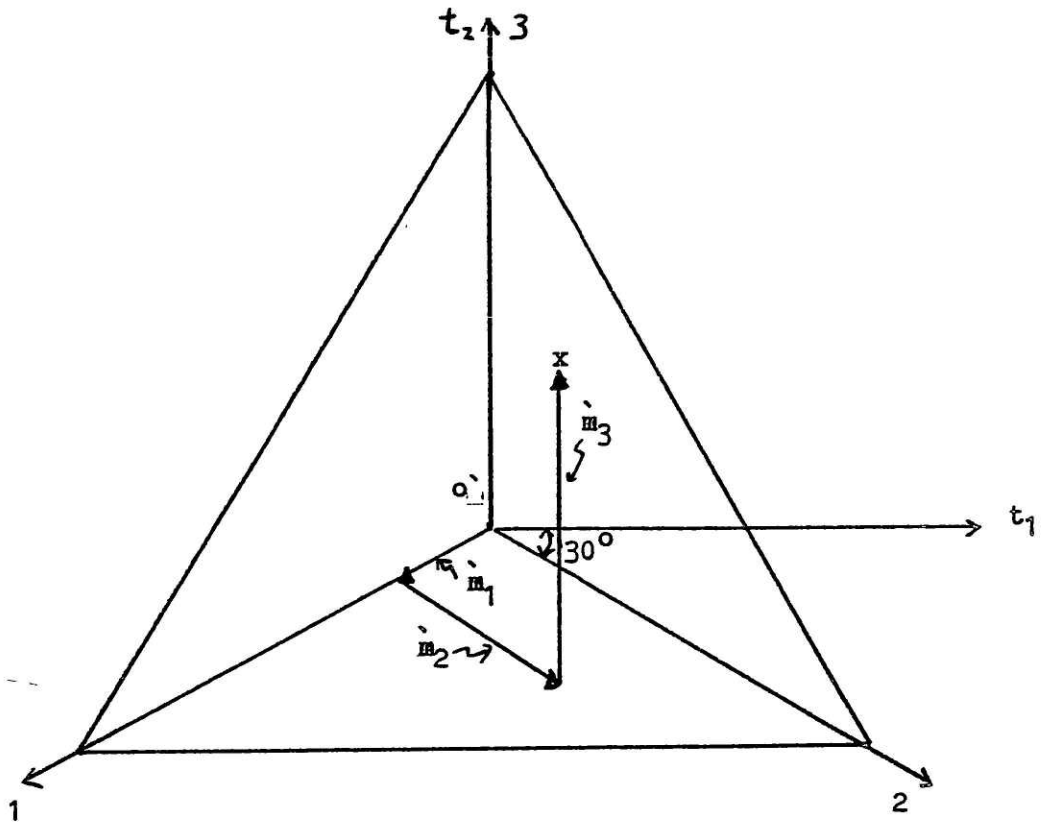


Figure 4.5-2.

$$\begin{bmatrix} t_1 \\ t_2 \end{bmatrix} = \left[\begin{array}{c|c|c} -\cos 30 & \cos 30 & 0 \\ \hline -\sin 30 & -\sin 30 & 1 \end{array}\right] \begin{bmatrix} \dot{m}_1 \\ \dot{m}_2 \\ \dot{m}_3 \end{bmatrix} \tag{4.5-3}$$

The inverse transformation from the parametrized space to the normalized measurement space is then given by solving Equations 4.5-2,3 to obtain

$$\begin{bmatrix} \dot{m}_1 \\ \dot{m}_2 \\ \dot{m}_3 \end{bmatrix} = \left[\begin{array}{c|c} -1/(2\cos 30) & -1/(2(1+\sin 30)) \\ \hline 1/(2\cos 30) & -1/(2(1+\sin 30)) \\ \hline -1/\cos 30 & 0 \end{array}\right] \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} + \begin{bmatrix} 1/(2(1+\sin 30)) \\ 1/(2(1+\sin 30)) \\ 1-1/(1+\sin 30) \end{bmatrix}$$

(4.5-4)

Examples of Data Display

Figure 4.5-3 illustrates the data display for a noise free case with closely eluting components in the mixture: (a) shows the concentration profiles; (b), the absorbance profiles at three measurement channels; (c), the 2-dimensional data display of the measurement matrix on the parametrized t_1, t_2 space (given by Equation 4.5-3). The three pure components in the mixture are depicted in this space as the vertices of the solid (pure-component) triangle; the outer dotted triangle represents the positive orthant portion of the unit-sum normalized plane, and hence the outer bound (see Section 4.2.1). The data is represented by x's with sampling time of 0.1 units.

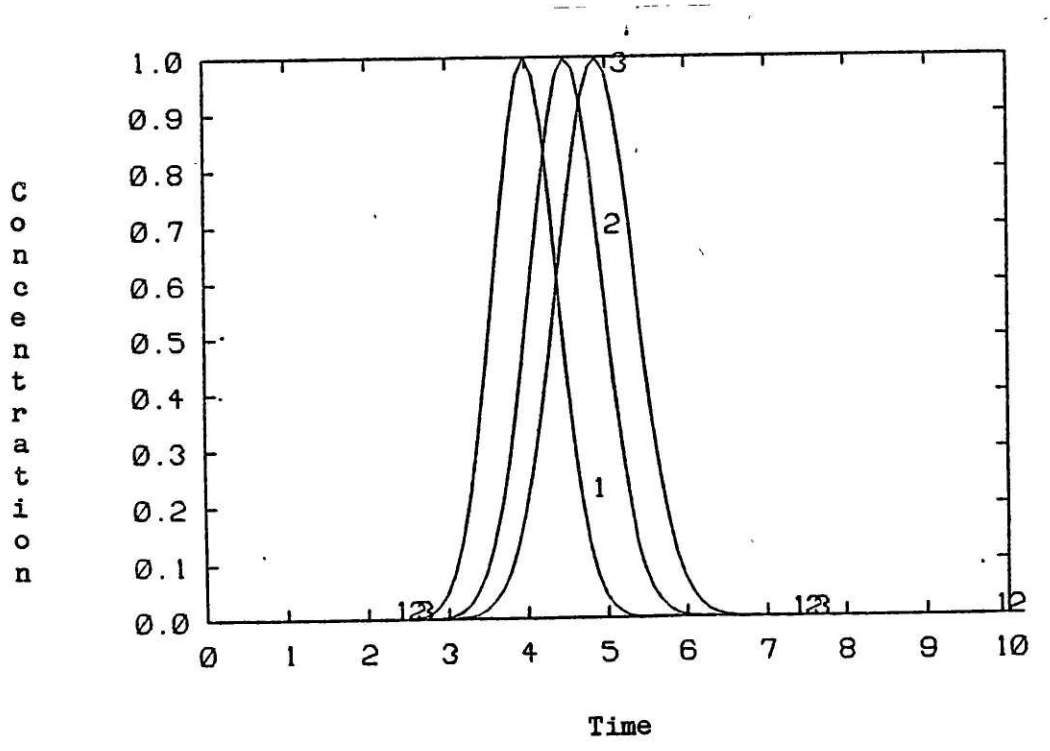


Figure 4.5-3a. Concentration Profiles.

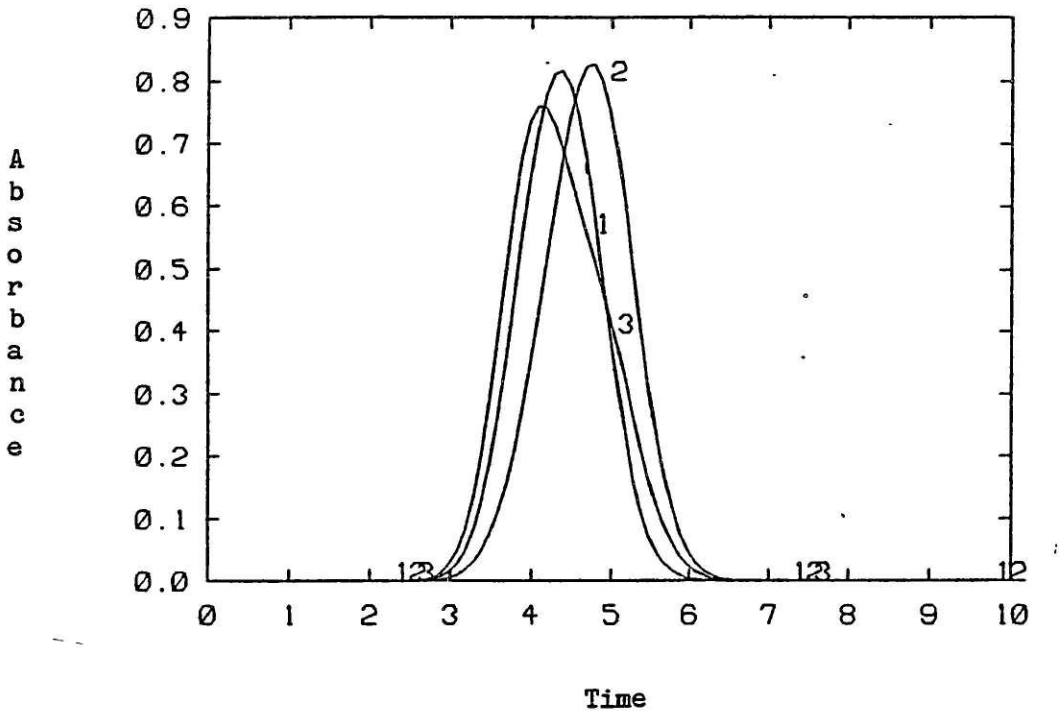


Figure 4.5-3b. Absorbance Profiles: Noise-Free case.

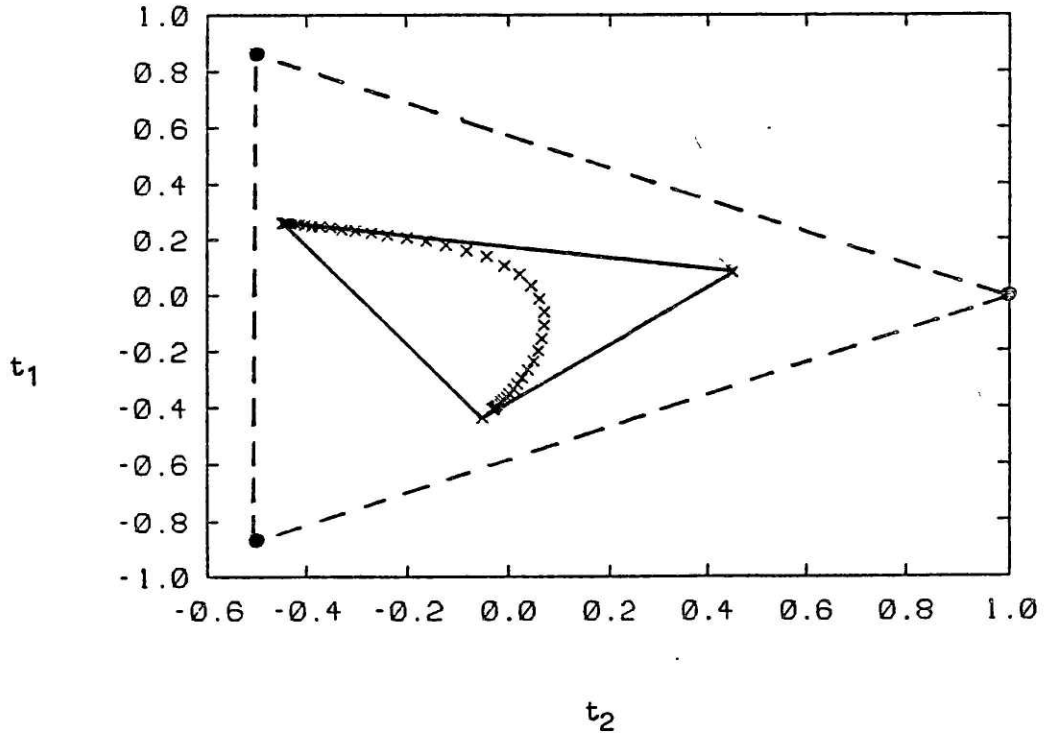


Figure 4.5-3c. 2-dimensional Data Display: Noise-Free Case.

As can be seen from Figure 4.5-3c, for this example, two of the pure components are easily identified by the cluster of points at the beginning and the end of the path traversed by the normalized data points. This is, of course, because the initial element is essentially pure component #1, and the final element is essentially pure component #3. Also note that, as expected, the pure-component triangle encloses the data points.

The effect of noisy data is shown in Figure 4.5-4. The standard deviation of the noise was 0.01, which is approximately 1% of the peak absorbance. Some of the data with low signal-to-noise ratio (100:3) were "filtered" out by discarding measurement vectors below a set threshold. Note that now some of the data points no longer lie inside the pure-component triangle. Nevertheless, two of the components are once again well approximated by the two clusters of points at the beginning and the end of the path traversed by the normalized data points.

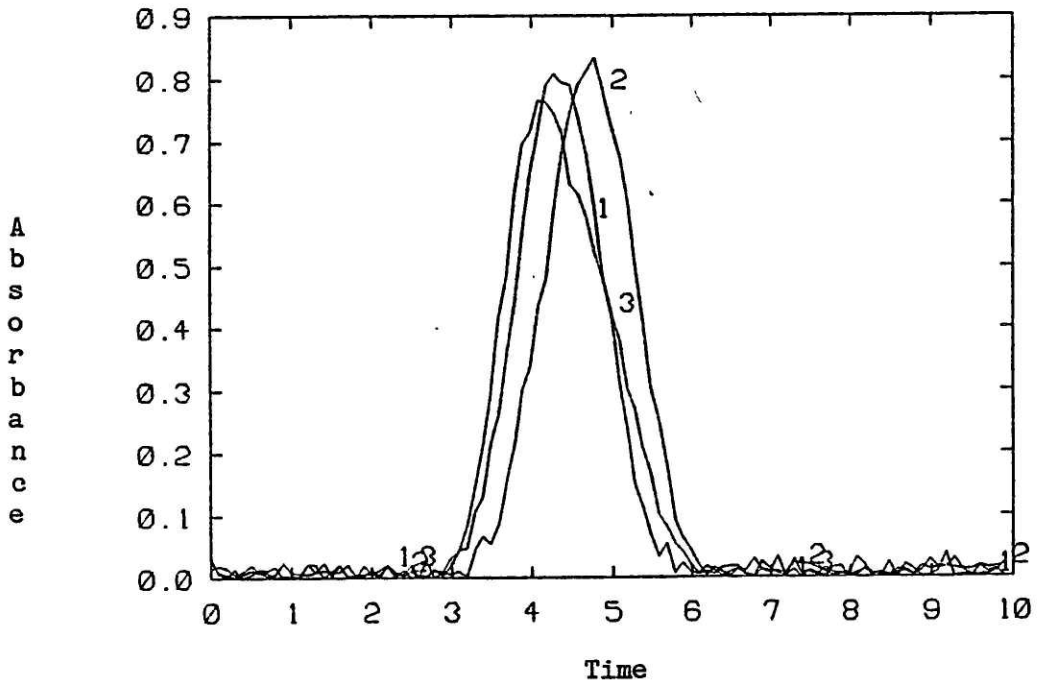


Figure 4.5-4a. Absorbance Profile: Noisy Data Case (s.d.=0.01).

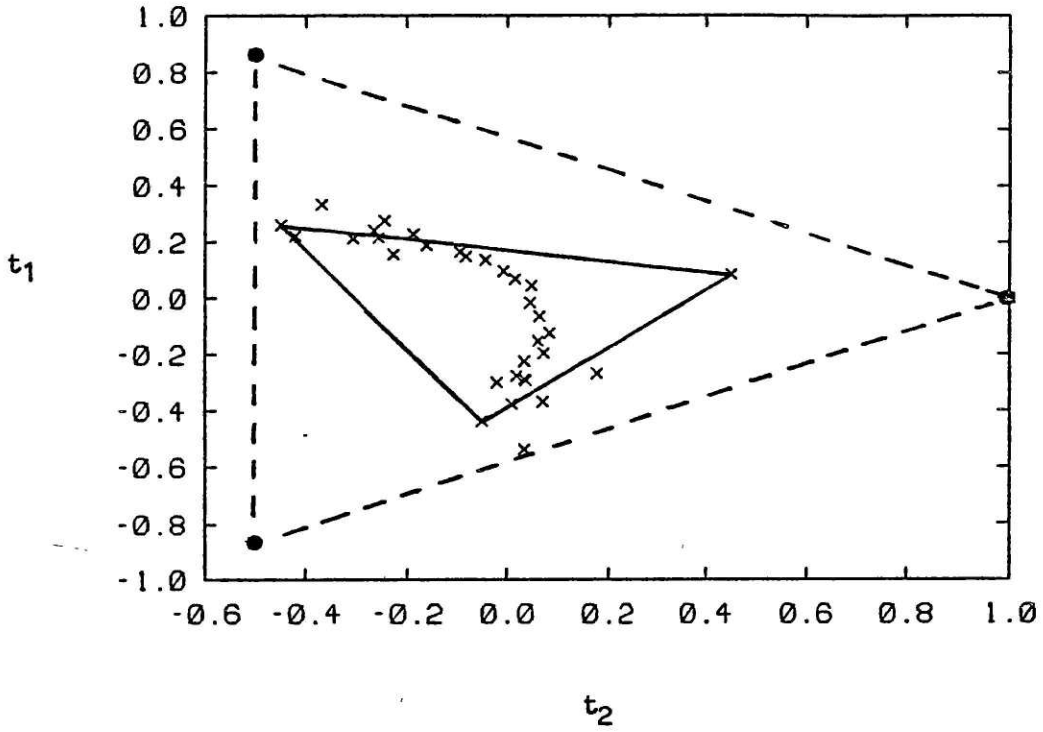


Figure 4.5-4b. 2-dimensional Data Display: Noisy Data Case (s.d.=0.01)

The effect of improved component resolutions is shown in Figures 4.5-5 and 6 for noise free and noisy data cases, respectively. As can be seen from Figures 4.5-5c and 4.5-6b, the data points appoximate the pure-components triangle more accurately with wider peak separations.

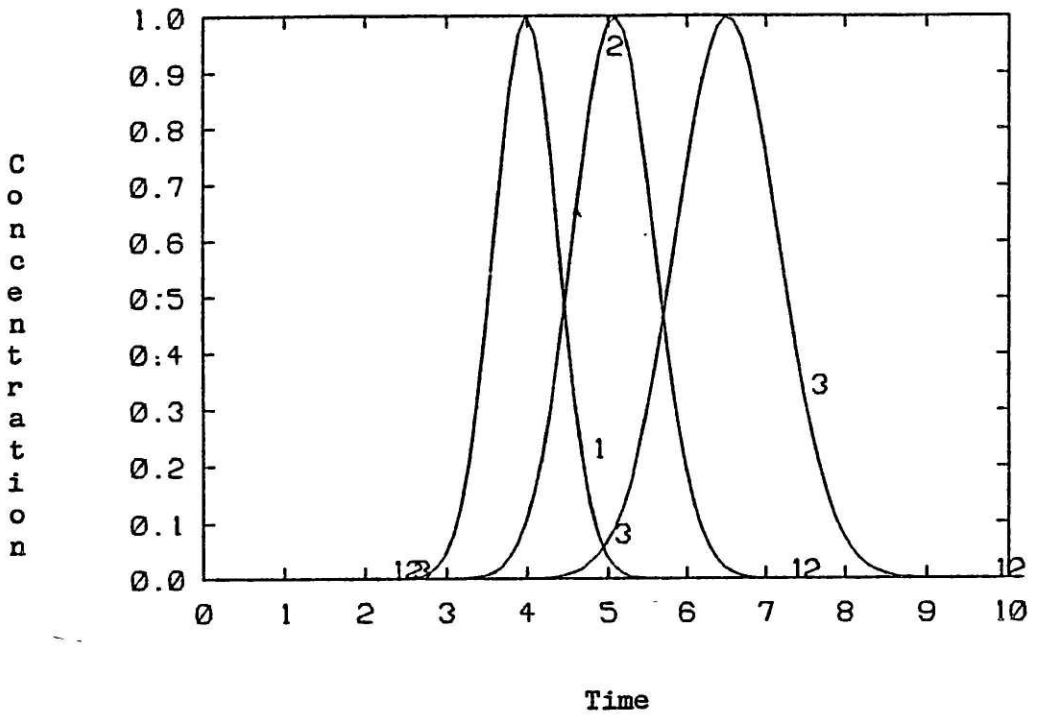


Figure 4.5-5a. Concentration Profiles.

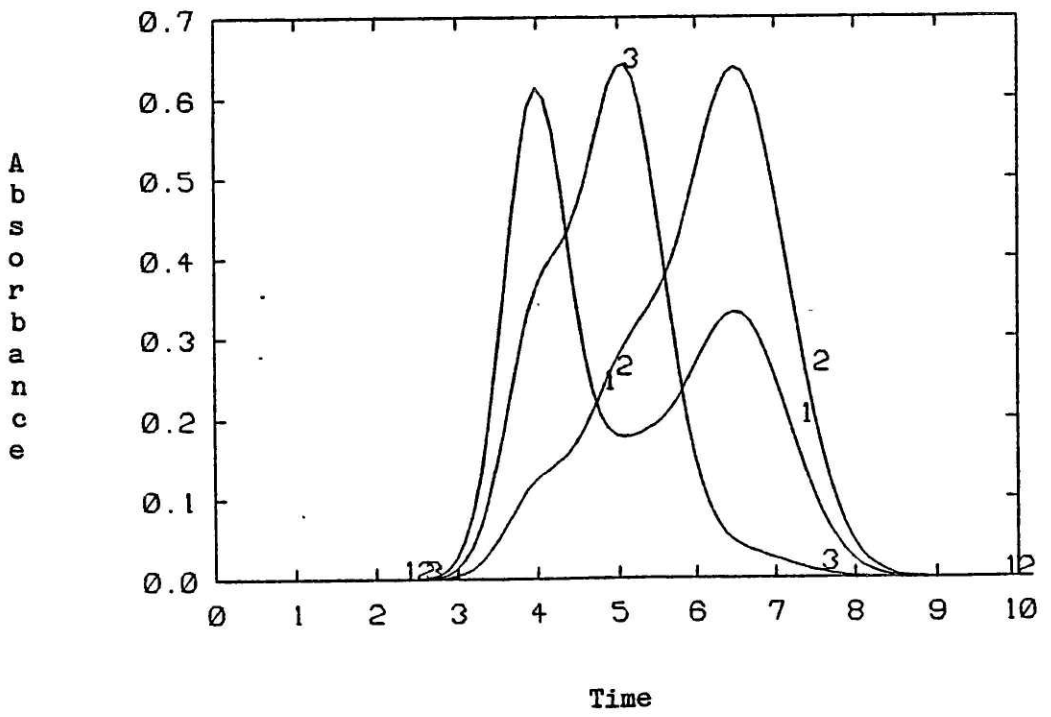


Figure 4.5-5b. Absorbance Profiles: Noise Free case.

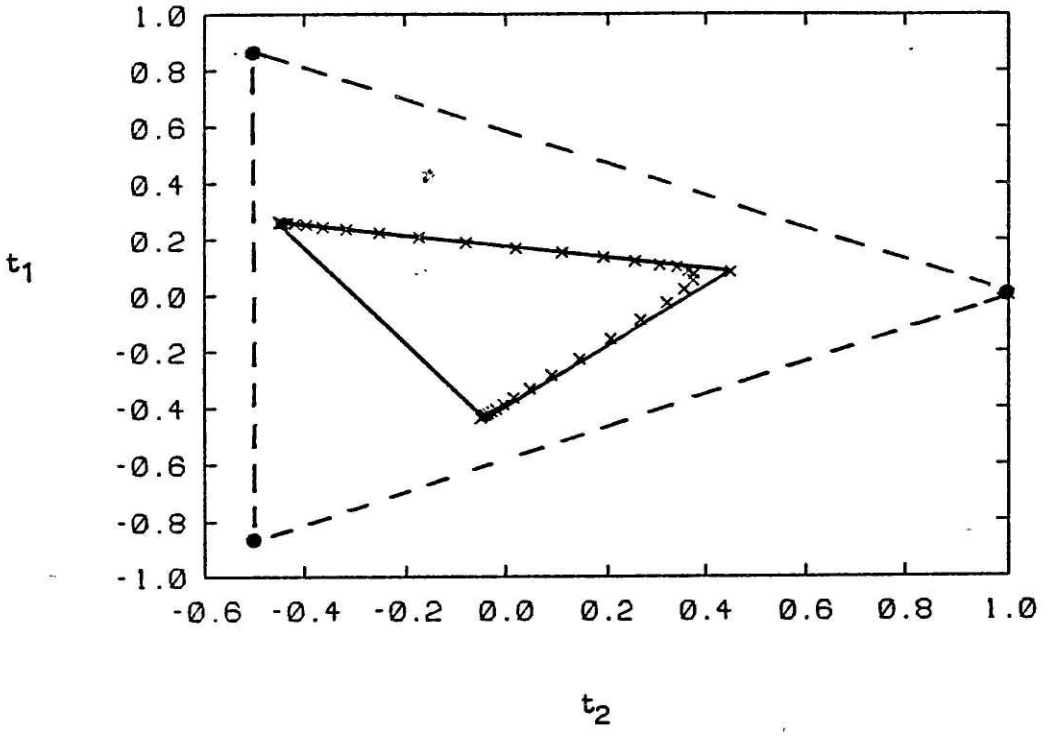


Figure 4.5-5c. 2-dimensional Data Display: Noise Free Case.

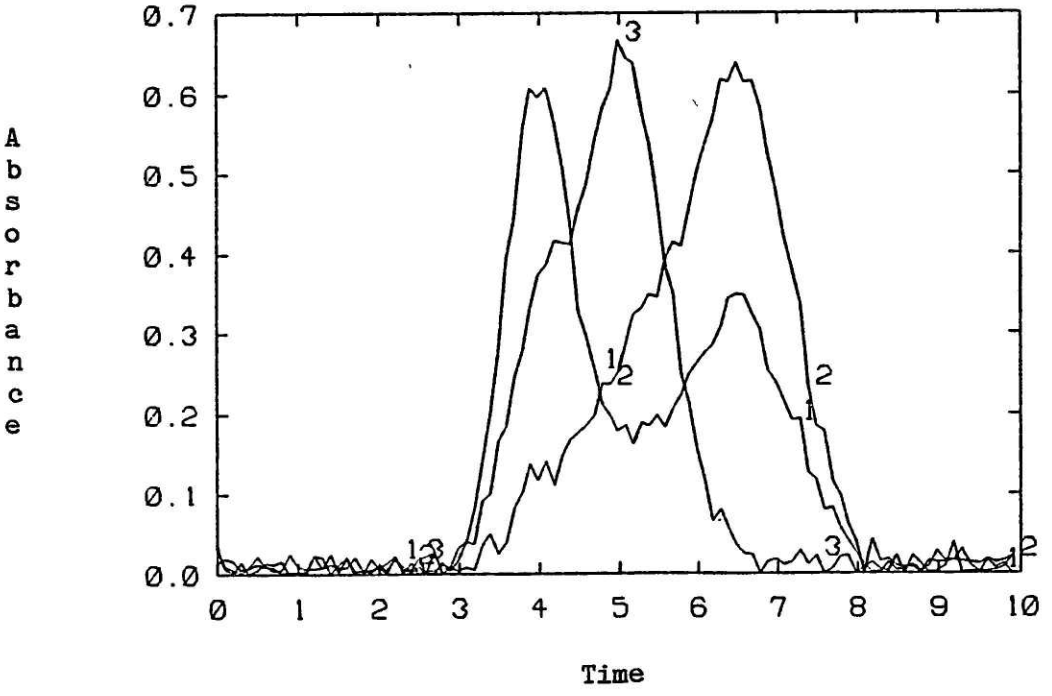


Figure 4.5-6a. Absorbance Profiles: Noisy Data case.

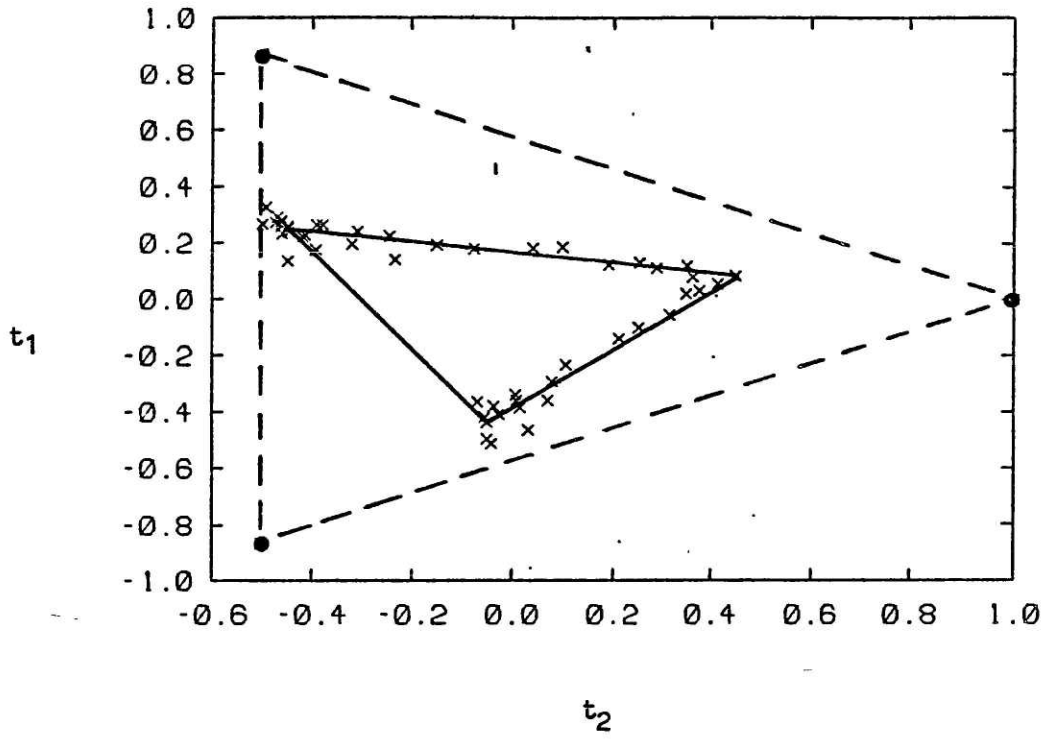


Figure 4.5-6b. 2-dimensional Data Display: Noisy Data Case.

Effect of Measurement Asynchronization in the Channels

We illustrate here the sensitivity of the data display in detecting the structural details of the measurement matrix. For this example, the absorbance profile of a pure component at Channel #2 was offset by one-half the sampling time of 0.1 (see Figure 4.5-7a). This discrepancy is not easily recognizable by standard methods of analyzing the absorbance profiles at the three channels. However, as shown in Figure 4.5-7b, the data display alerts one to suspect possible experimental errors that result in aberrations from the expected pure component behavior (recall that for a pure component, the data vectors plot at a point).

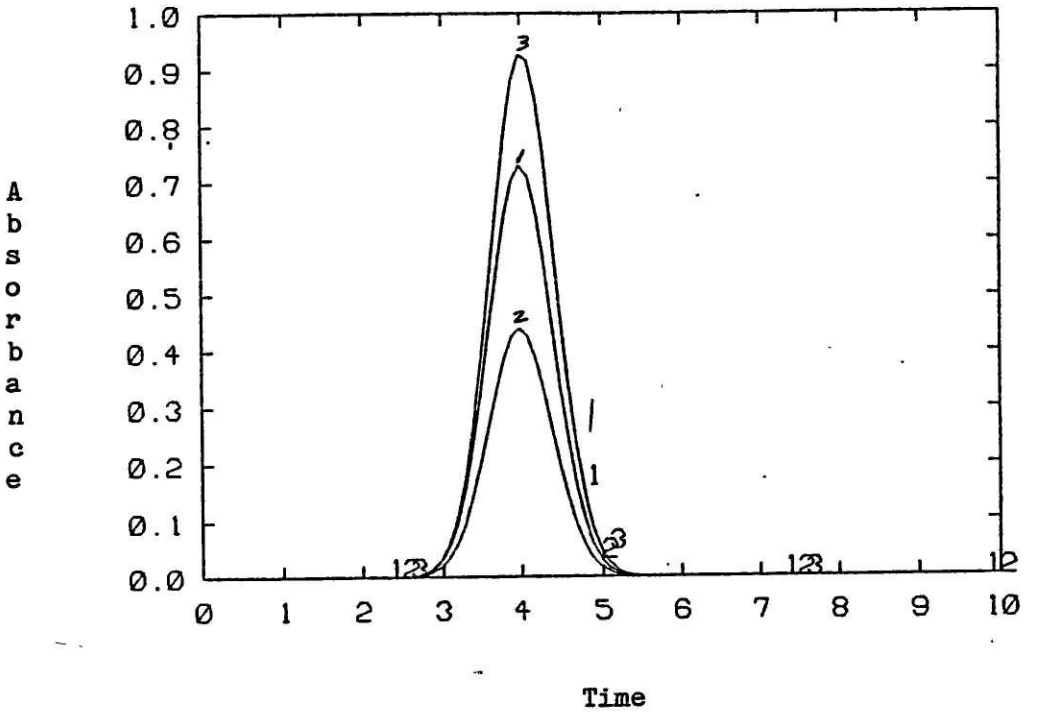


Figure 4.5-7a. Absorbance Profiles: Asynchronized Channel 2.

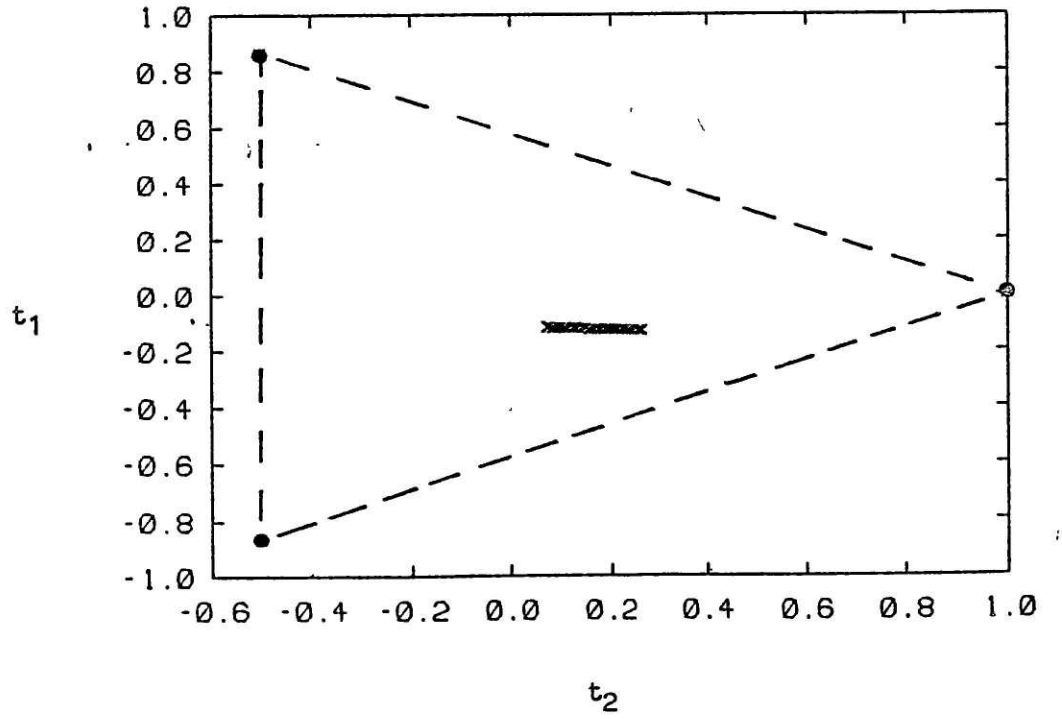


Figure 4.5-7b. 2-dimensional Data Display: Asynchronized Channel 2.

CHAPTER V: NUMERICAL RESULTS: REAL DATA

CHAPTER OUTLINE

	Page
5.1 Description of Experiment.....	186
5.2 Target Components Problem.....	191
5.3 Unknown Components Problem.....	192

5.1 DESCRIPTION OF EXPERIMENT

The experimental data used in this chapter are absorbance profiles from a liquid chromatography experiment carried out by the Waters Chromatography Division of Millipore Corporation.

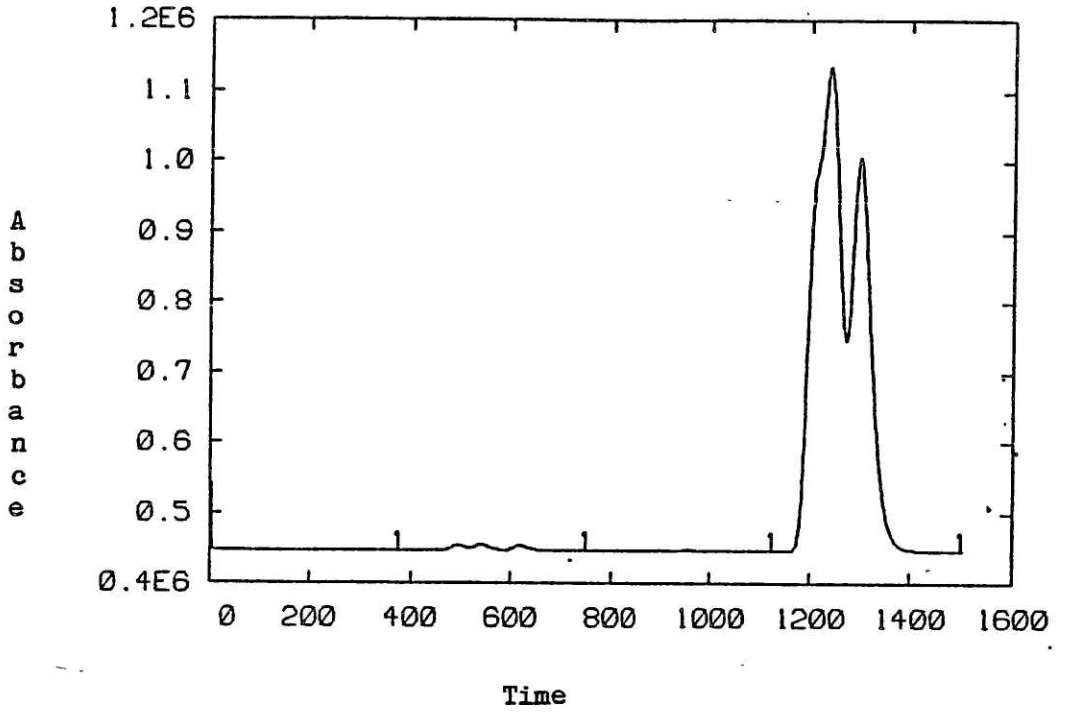
The chemical mixture used in the experiment is a three-component mixture consisting of 5:4:4 ratio (by weight) of the following compounds

- Ethylbenzene,
- o-xylene, and
- p-xylene

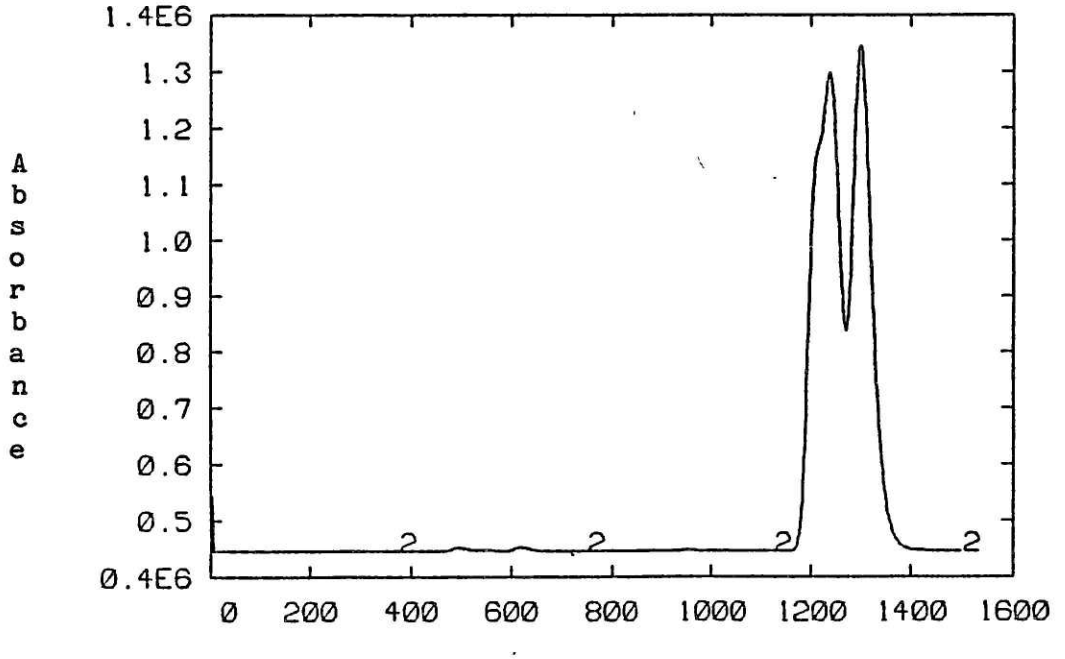
in a 64% CH_3CN mobile phase solvent. The absorbances were measured at wavelengths of 255, 260, 265, and 270 nm. The separation was carried out on a NOVA-PAK^{TM} brand C18 column. The experiment was performed at a flow rate of 1 ml/min, and the temperature was stabilized at 25°C. The sampling rate was 5 samples per second. The single component used for the calibration run in the target component problem is Ethylbenzene, and the absorbance measurements were made under similar conditions.

The absorbance profiles of the mixture and the single component for the case of 3000:1 signal-to-noise (S/N) ratio (peak-to-peak) are summarized in Figures 5.1-1 and 5.1-2, respectively, at the four specified wavelengths. Data was also provided for S/N ratios of 1000:1 and 100:1, though these are not shown here.

NOVA-PAK is a trademark of Waters Chromatography Division of Millipore Corporation.

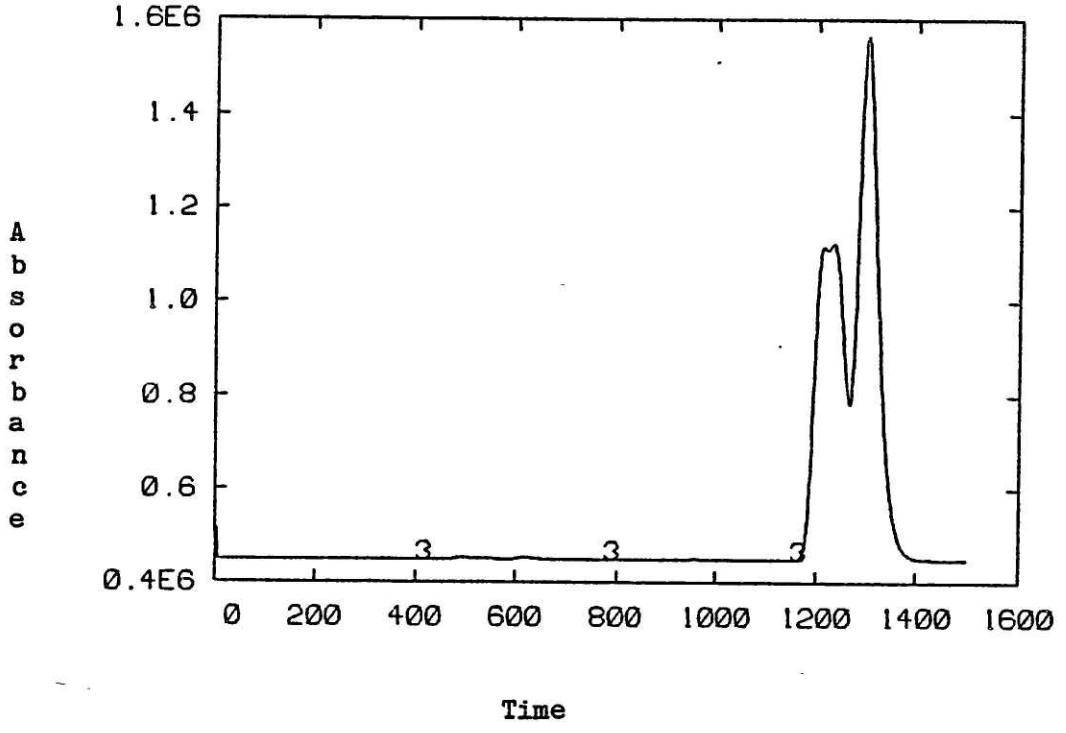


(a) Absorbance at wavelength of 255 nm.

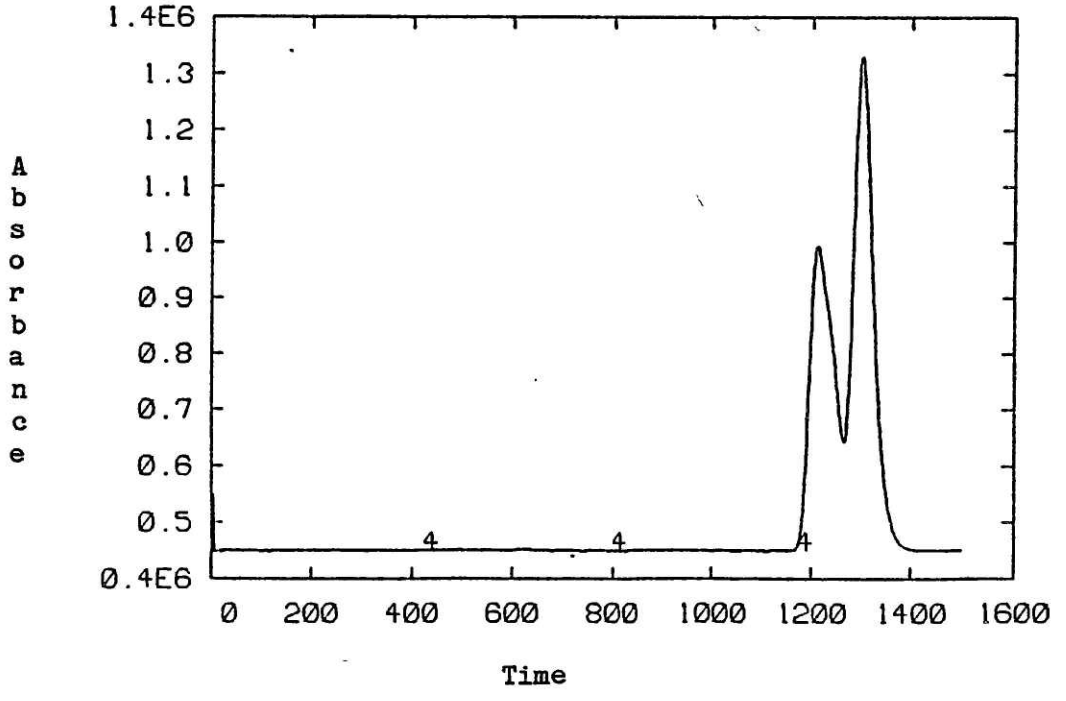


(b) Absorbance at wavelength of 260 nm.

Figure 5.1-1. Absorbance profiles of the mixture.

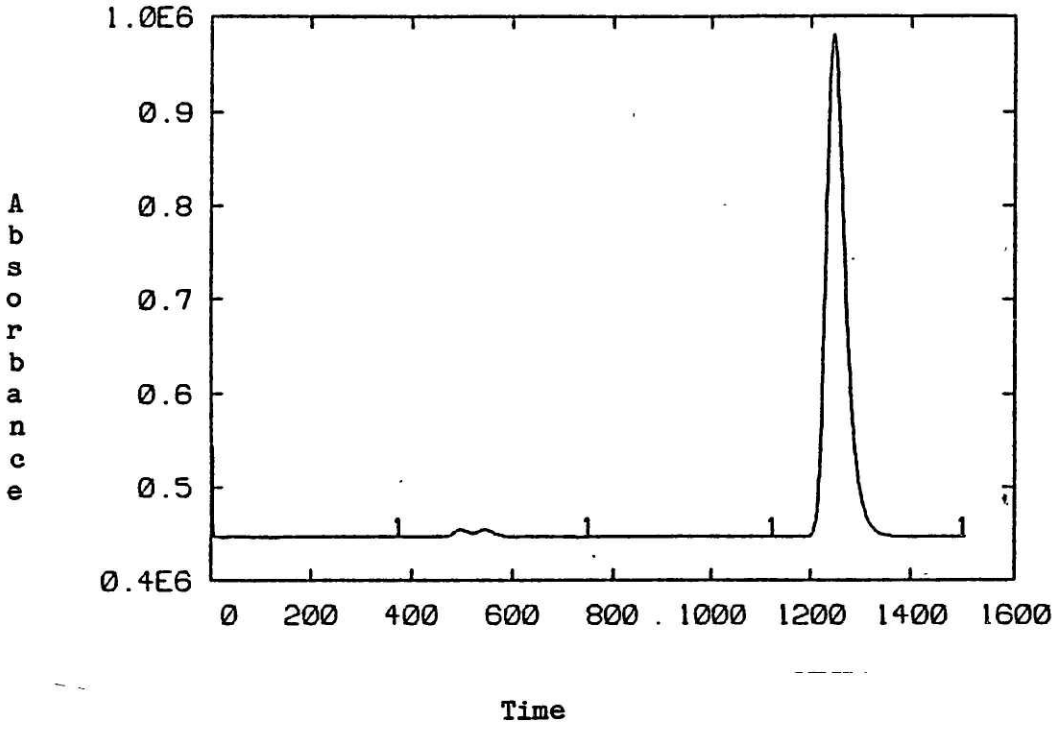


(c) Absorbance at wavelength of 265 nm.

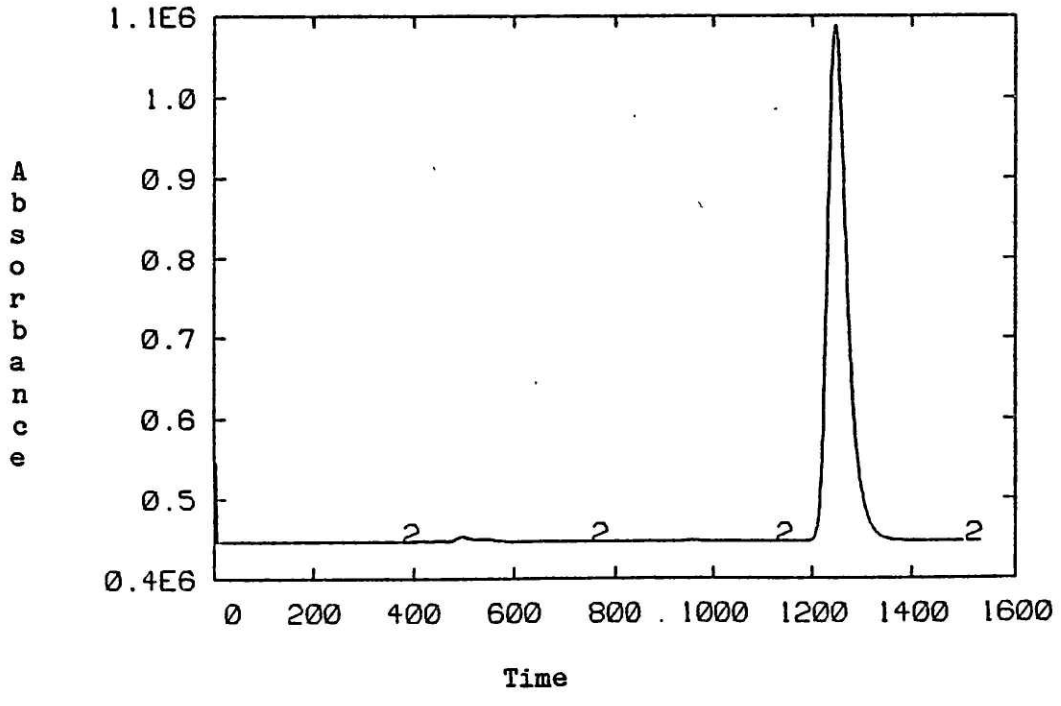


(d) Absorbance at wavelength of 270 nm.

Figure 5.1-1. Absorbance profiles of the mixture.

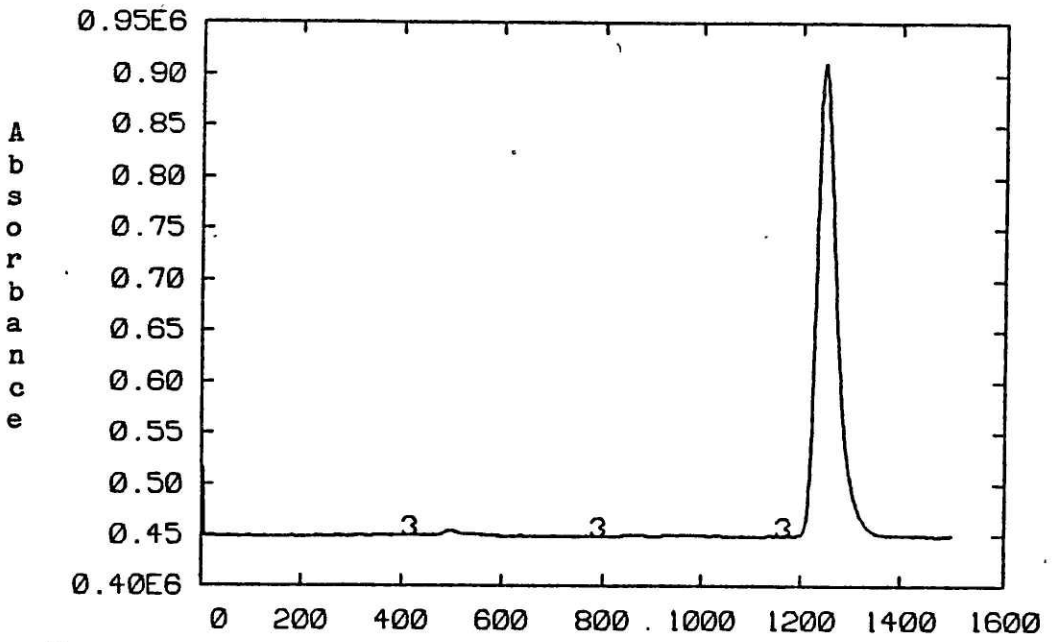


(c) Absorbance at wavelength of 255 nm.

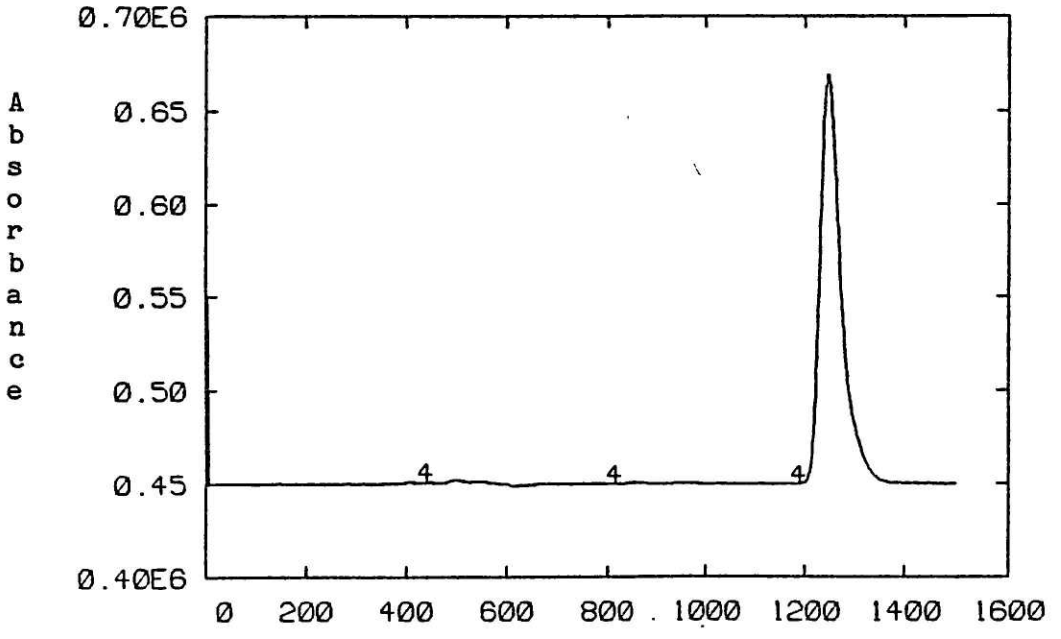


(b) Absorbance at wavelength of 260 nm.

Figure 5.1-2. Absorbance profiles of the single component.



(c) Absorbance at wavelength of 265 nm.



(d) Absorbance at wavelength of 270 nm.

Figure 5.1-2. Absorbance profiles of the single component.

5.2 TARGET COMPONENTS PROBLEM

The results of application of our target components methods to the preceding experimental data are summarized in Table 5.2-1. The methods used are the projection method via linear least squares projection (Section 3.3.2b), with and without the data synchronization algorithm (Section 3.3.2c). The results tabulated are for 6-sample averages at various signal-to-noise ratios and target component concentrations, with $\bar{c}$ denoting the averaged values and s.d. the standard deviations of the computed values of c.

(S/N)	Target Conc. (c)	With Syn. Alg.		Without Syn. Alg.	
		$\bar{c}$	s.d.	$\bar{c}$	s.d.
3000:1	1	1.0101	0.0200	1.0301	0.0187
1000:1	0.25	0.2587	0.0055	0.2640	0.0072
100:1	0.025	0.0259	0.0022	0.0263	0.0024

Table 5.2-1. Results of the application of projection method to experimental data.

As can be seen from the results shown in Table 5.2-1, the projection method successfully estimates the concentrations of the target component in the mixture with, at worst, less than 4% error. Also, the synchronization algorithm is seen to improve the accuracy of the estimated concentrations.

5.3 UNKNOWN COMPONENTS PROBLEM

The absorbance profiles shown in Figure 5.1-1 were displayed on the normalizing plane of two 3-dimensional measurement spaces. The first measurement space used was a 3-dimensional absorbance space with its coordinates representing absorbances at the first three wavelengths (255, 260, and 265 nm); the second measurement space used the first, third and fourth wavelengths (255, 265, and 270 nm). The data representation on a two-dimensional space was obtained with the parameterization of the normalized data vectors as follows (see Section 4.5):

$$t_{1i} = \cos 30(\grave{m}_{2j} - \grave{m}_{1j}) \qquad (5.3\text{-}1a)$$

$$t_{2i} = \grave{m}_{3j} - \sin 30(\grave{m}_{1j} + \grave{m}_{2j}) \qquad (5.3\text{-}1b)$$

The data displays for the mixture are shown in Figures 5.3-1 and 5.2-2 for the first and the second measurement spaces, respectively. In Figure 5.3.1, the normalized data points for the three-component mixture are seen to lie on a straight line, which one would not expect for a three-component mixture. It turns out that the absorbances of the constituent components at these three wavelengths are very close to being linearly dependent, thus resulting in a display that mimics a two-component mixture.

It is possible, however, to deduce that there are indeed three components in the mixture by looking at this display if one is aware of the underlying mechanisms of a liquid chromatography experiment (that is, by using prior information on the structure of the concentration profile). More specifically, with the Gaussian concentration profiles of the constituent components, for reasonably well separated peaks, there

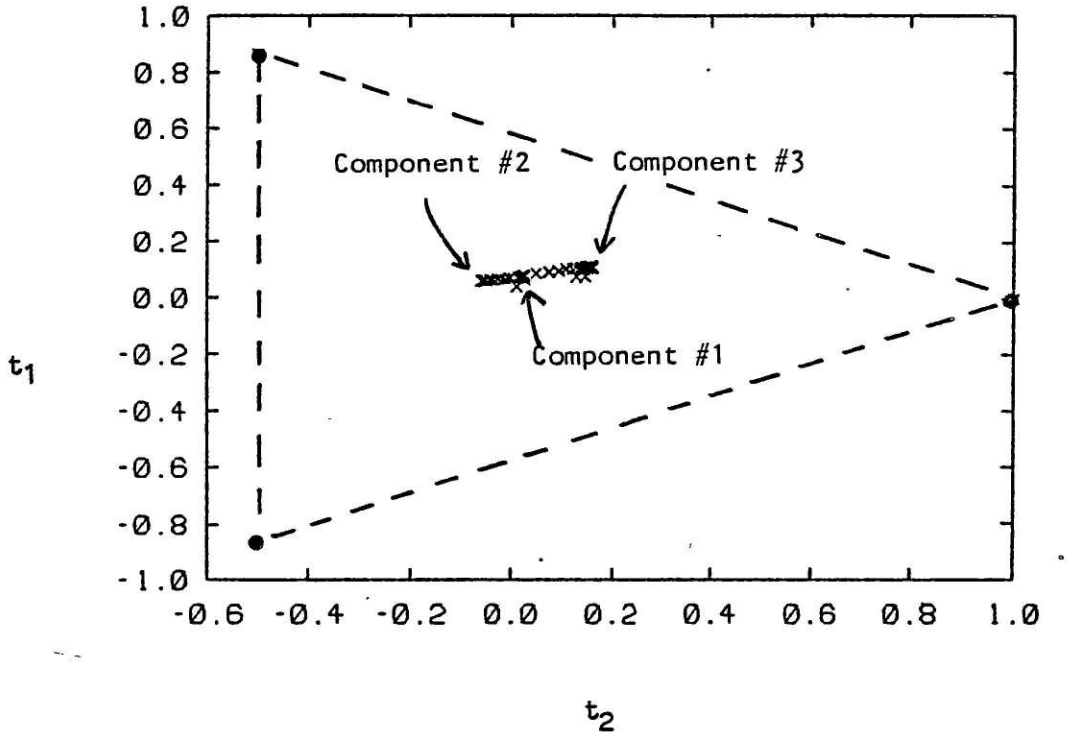


Figure 5.3-1. Data representation on a parametrized space of the normalizing plane for the first measurement space

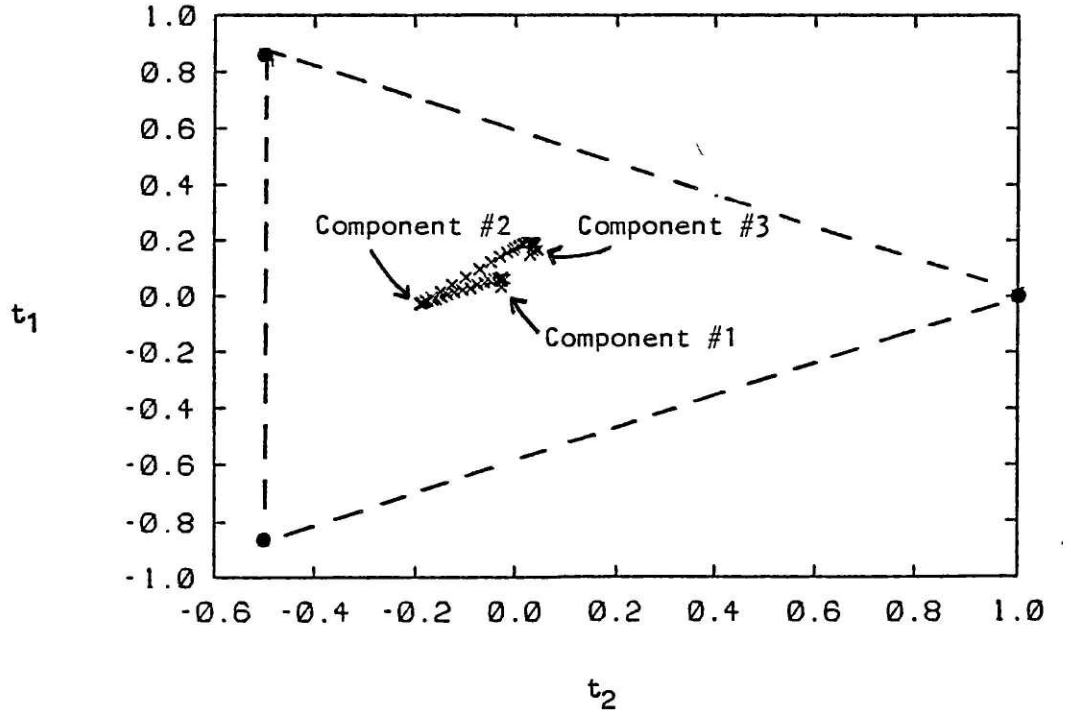


Figure 5.3-2. Data representation on a parametrized space of the normalizing plane for the second measurement space.

are regions in the absorbance profile where we have (or are very close to having) only pure component absorbances. In these regions we would observe pure component behavior in our display, that is, data points plotting to a point (or to a cluster of points), as seen in Figure 5.3.1 at three distinct locations.

The data display in the second measurement space is seen to be more like what we would expect from a three-component mixture. The three corner points of the path described by the data points essentially represent the unit-sum normalized absorbance spectra of the constituent components at the three wavelengths that characterize the measurement space.

An interesting fact emerged from the data representation on the normalizing plane for the supposedly pure component (component #2) shown in Figure 5.3-3. It is evident that there are actually two components present, a fact that would have been difficult to detect from the absorbance profiles shown in Figure 5.1-2, even under visual examination by a well trained chromatographer. When we compare Figures 5.3-3 and 5.3-1, it is seen that the normalized absorbance spectra of the impurity closely matches that of component #3 in the mixture. One possible explanation for this is that when doing the calibration run, some traces of the component #3 were introduced into the pure component run by residues left over from previous runs on the mixture; or perhaps the supposedly pure component was not so pure after all! (Although this contaminated "pure" component was used in the calibration run for the target component problem in Section 5.2, due to the low concentration of impurity, its effect was negligible.)

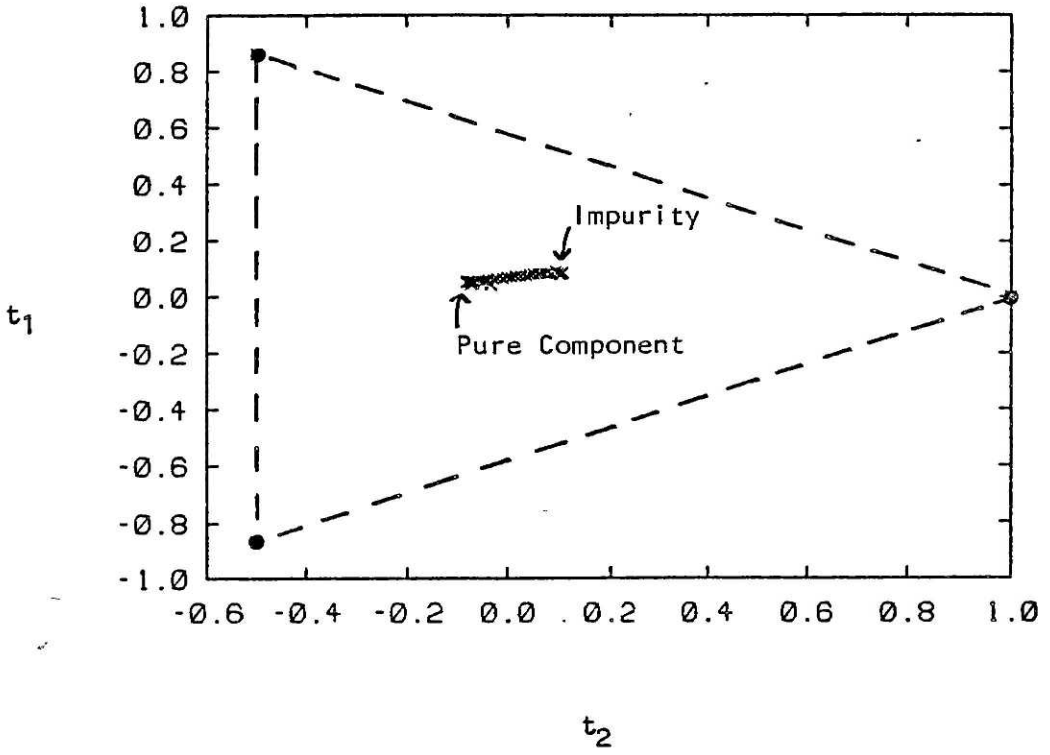


Figure 5.3-3. Data representation on a parametrized space of the normalizing plane for the first measurement space: "Pure" component.

CHAPTER VI: CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

CONCLUSIONS

Summary of Contributions

We have in this thesis addressed two important categories of chemical identification problems: the target components and the unknown components problems. We have developed new analytical methods and algorithms, which have been successfully applied to both synthetic and real data from liquid chromatography experiments. The results have shown that these methods represent a significant advance over existing methods of analysis.

For the target components problem, we have obtained two non-iterative and numerically stable solution methods that, in comparison to existing methods, dramatically reduce computational requirements. In addition, our algorithms compensate for data synchronization errors in measurements from liquid chromatography that adversely affect existing methods. The two methods have been applied to both synthetic and experimental data from liquid chromatography, and are found to be robust in the presence of sensor noise and other experimental errors such as the synchronization error.

For the unknown components problem, we have extended earlier work on the case of two and three unknown components, giving a concise and physically/geometrically motivated formulation of the problem and a characterization of all possible solution sets for an arbitrary number of unknown components. We have also devised a numerically sound reduction and representation scheme for noisy data that delineates fine structural details of such chemical measurements, details that typically elude existing methods of analysis. The representation scheme reduces to a two-dimensional display for the practically important case of three

components, thus facilitating interactive data analysis. Our method has been shown to be particularly effective in analyzing absorbance profiles of chemical mixtures from liquid chromatography.

Directions for Further Work

We outline below some related directions that may be fruitfully pursued in future work.

Exploiting prior knowledge of the structure of concentration profiles in liquid chromatography

In liquid chromatography, the concentration profile of a pure component can be approximated by a Gaussian function, which is completely described by two parameters: the retention time (time at which the concentration profile peaks), and the standard deviation of the distribution. (Other more refined *a priori* models do exist, e.g., the Gaussian-plus-exponential-tail model [BO.5] that takes into account the effect of tailing in liquid chromatograms, though they generally require more parameters to describe.) Such *a priori* knowledge may be exploited in several ways.

Given the absorbance profile of an unknown mixture, one possibility is to search over the two (or more) parameters that uniquely define the concentration profile of a pure component, with appropriately chosen criteria for deciding whether a particular pure component profile could have contributed to generation of the mixture absorbance profile. One such criterion is to test for the containment of a chosen pure component profile in the space spanned by the rows of the mixture data matrix (recall that we have chosen the rows of the data matrices to represent absorbances at a particular wavelength). When all sets of parameters

that satisfy this test are found, we have solved the identification problem.

An additional possible use for the prior knowledge of the Gaussian structure of the concentration profiles is in the analysis of, for example, a three-component mixture. Because the constituent components of a mixture elute at different rates, and the concentration profiles are Gaussian, there are regions at the beginning and the end of a complete mixture absorbance profile where one has essentially single-component profiles. Hence the components that elute first and last in a three-component mixture are respectively identifiable, in the data representation plane discussed in Chapter 4, as the initial and final clusters in the path traversed by the data points when plotted sequentially on this plane.

Once the two end components have been identified, several possibilities exist for identifying the middle component. For example, we may use the two end points as two of the vertices for all sets of triangles that enclose the data points and are enclosed by the outer bound, and then find the region in which the third vertex (and hence the third component) can lie. A second possibility is to find the triangle with the minimum area that encloses the data points, with two of its vertices fixed at the two known, pure component points. Alternatively, in the case of noisy data, one can carry out a two-line least squares fit of the data [L3], with one end of each of the two line segments fixed at the points that represent the two known pure components, and the point of intersection taken to represent the third component.

Updating the data representation with additional data samples

Additional data can be added to the existing set of measurements by

either detecting at new wavelengths or by taking more time samples. In either case, efficient numerical algorithms for updating the solutions without starting from scratch are desirable.

Representation of uncertainty regions

In Chapter 4, we have addressed the problem of error represention (projections to lower dimensional subspaces and hyperplanes) when the uncertainty regions are given by box-bounds; similar problems exist when the uncertainty regions are given by probability density functions, norm bounds, or others.

Hybrid target components-unknown components approach

For the three unknown components case, if we can isolate one or two of the unknown components, we can reduce the dimension of the problem by using the target components method to subtract from the mixture data matrix the known components. As mentioned earlier, such sequential analysis is a very real possibility in liquid chromatography, because one can often identify the first and last elements in a group.

Long term noise

In a liquid chromatography experiment, long-term noise (or baseline drift) is characterized by fluctuations in the baseline signal with periods in the range of 6 to 60 cycles per hour, and generally originate from the seperation system through either chemical or fluidic effects. For example, variations in column temperature, changes in flow rate of the mobile phase, disequilibrium in the chemical composition of the sample, and so on, all cause large scale fluctuations that are approximately linearly varying within the range of most peak detections. In the application of our methods to real data, the long term noise was not a significant factor. Nevertheless, it's effect should be studied

since significant drifts do occur in some liquid chromatography experiments.

Other applications involving bilinear problem

Although we have in this thesis emphasized applications to data from liquid chromatography-UV spectroscopy experiments, the methodologies developed in this thesis are applicable to data from other chemical (e.g. excitation-emission spectra from fluorescence spectroscopy experiments) and non-chemical experiments that have bilinear structure, and applications to these areas should be pursued.

REFERENCES

A1 Anderson, T.W., Introduction to Multivariate Analysis, John Wiley and Sons, N.Y.

A2 Athans, M. and Schweppe, F.C., "Gradient Matrices and Matrix Calculations," MIT Lincoln Lab Tech. Note 1965-53 (Unpublished), Lexington, Massachusetts.

B1 Balke, S.T., Quantitative Column Liquid Chromatography, Elsevier, 1984.

B2 Barnes, E.R., "Some Matrix Minimization Problems," IEEE Conference on Decision and Control, 1980, Vol. 1, 244-246.

B3 Bartlett, M.S., Br. J. Psychol. Stat., Sect. 3, 77(1950)

B4 Bellman, R., Introduction to Matrix Analysis, McGraw-Hill, 1960.

B5 Ben-Israel, A. and Greville, T.N.E., Generalized Inverses: Theory and Applications, Krieger, N.Y., 1980.

C1 Cattell, R.B., "Statistical Weighted Principal Component Analysis of Rapid Scanning Wavelength Kinetics Experiment," Multivariate Behav. Res., 1, 245 (1966).

C2 Chen, J.-H. and Hwang, L.-P., "Reconstruction of mass spectra of components of unknown mixtures based on factor analysis," Analytica Chemica Acta, 133 (1981) 272-281.

C3 Cochran, R.N. and Horne, F.H., Anal. Chem., 46, 846 (1977).

C4 Coxeter, H.S.M., Introduction to Geometry, 2nd ed., Wiley, 1969.

D1 Davis, C. and Kahan, W.M., "The rotation of eigenvectors by a perturbation. III," SIAM J. Numer. Anal. 7 (1970), pp.1-46.

D2 Dongarra, J.J., et al., LINPACK User's Guide, SIAM, 1979.

D3 Duewer, D.L., Kowalski, B.R., and Fasching, J.L., "Improving the Reliability of Factor Analysis of Chemical Data by Utilizing the Measured Analytical Uncertainty," Anal. Chem., 48, 2002 (1976).

E1 Eckart, C. and Young, G., "The approximation of one matrix by another of lower rank," Psychometrika, Vol. 1, No. 3, 1936, pp. 211-218.

E2 Emami-Naeini, A. and Van Dooren, P., "Computation of Zeros of Linear Multivariate Systems," Automatica, Vol. 18, No. 4, pp. 415-430, 1982.

F1 Forney, G., "Minimal bases of rational vector spaces with applications to multivariable linear systems," SIAM J. Contr., Vol. 13, pp. 493-520, 1975.

F2 Frank, I.E. and Kowalski, B.R., "Chemometrics", Anal. Chem., 1982, 54, 232R-243R.

G1 Gantmacher, F.R., Theory of Matrices I & II, Chelsea, N.Y., 1959.

G2 Giddings, J.C., Dynamics of Chromatography, Vol. 1 of Chromatographic Science Series (J.C.Giddings, ed.), Dekker, N.Y.,1965.

G3 Golub, G.H. and Wilkinson, J.H., "Ill-conditioned eigensystems and the computation of the Jordan Canonical form," SIAM Review, Vol.18, No.4, October 1976, pp. 578-618.

G4 Golub, G.H. and VanLoan, C.F., "An Analysis of the Total Least Squares Problem," SIAM J. Numerical Anal. Vol.17, No.6, Dec. 1980.

G5 Golub, G.H. and Reinch, C., "Singular Value Decomposition and Least Squares Solutions," Numer. Math. 14, 1970, 403-420.

G6 Golub, G.H. and VanLoan, C.F., "Matrix Computations," The Johns Hopkins University Press, Baltimore, 1983.

G7 Golub, G.H., Klema, V., and Stewart, G., "Rank degeneracy and least squares problem," Report STAN-CS-76-559, Computer Science Department, Stanford University, Stanford, California.

G8 Graham, A., Kronecker Products and Matrix Calculus: with Application, Ellis Horwood Limited, 1981.

G7 Greschak, J.P., "Reconstructing Convex Sets," Ph.D. Thesis, Dept. of Electrical Engineering, M.I.T., Cambridge, Ma., February, 1985.

H1 Hamilton, R.J., Sewell, P.A., Introduction to high performance liquid chromatography, 2nd Edition, Chapman and Hall, 1982.

H2 Ho, C.-N., Christian, G.D., and Davidson, E.R., "Simultaneous multicomponent rank annihilation and applications to multicomponent fluorescent data recquired by the video fluorometer," Anal. Chem., 53, 92-98,1981.

H3 Ho, C.-N. et al., Anal. Chem., "Application of the Method of Rank Annihilation to Fluorescent Multicomponent Mixtures of Polynuclear Aromatic Hydrocarbons," Vol. 52, No. 7, June 1980, 1071-1079.

H4 Ho, C.-N. et al., Anal. Chem., "Application of the Method of Rank Annihilation to Quantitative Analysis of Multicomponent Fluorescense Data from the Video Fluorometer," Vol. 50, No. 8, July 1978, 1108-1113.

H5 Hugus, Z.Z. and El-Awady, A.A., "The Determination of the Number of Species Present in a System: a New Matrix Rank Treatment of Spectrophotometric Data," J. Phys. Chem., 75, 2954 (1971).

J1 Johnson, D.W., Callis, J.B., and Christian, G.D., "Rapid Scanning Flourescence Spectroscopy," Analytical Chemistry, Vol.49, No.8, July 1977, 747A-757A.

K1 Kailath, T., Linear Systems, Pretice-Hall, Inc., 1980.

K2 Klema, V.C. and Laub, A.J., "The Singular Value Decomposition: Its computation and Some Applications," IEEE Trans. on Automatic Cont., Vol.AC-25, No.2, April 1980.

K3 Kormos, D.W. and Waugh, J.S., "Abstract Factor Analysis of Solid-State Nuclear Magnetic Resonance Spectra," Anal. Chem., Vol.55, No.4, April 1983, 633-638.

K4 Kronecker, L., "Algebraische Reduction der Schaaren bilinear Formen," S.-B. Akad. Berlin 1890, pp 763-76.

K5 Kublanovskaya, V.N., "An approach to solving the spectral problem of A-sB," in Matrix Pencils, Ed. A.Ruhe & B.Kagstrom, Lecture Notes in Mathematics, Vol. 973, Springer, N.Y., 1982.

L1 Lawton, W.H. and Sylvestre, E.A., "Self modeling curve resolution," Technometrics, 13, 3, 617-633, Aug. 1971.

L2 Ledermann, W., "Reduction of singular pencils of matrices," Proc. Edinburg Math. Soc., Vol.4, pp.92-105 (1935)

L3 Lenstra, A.K. et al., "Two lines least squares," Annals of Discrete Mathematics, 16, 1982, 201-211.

L4 Luenberger, D.G., Introduction to Dynamic Systems: Theory, Models, and Applications, John Wiley & Sons, 1979.

M1 Malinowski, E.R. and Howery, D.G., Factor Analysis in Chemistry, Wiley, N.Y., 1980.

M2 Meister, a., "Estimation of component spectra by principal components method," Analytica Chemica Acta, 161 (1984) 149-161.

M3 McCue, M. and Malinowski, E.R., "Rank annihilation factor analysis of unresolved LC peaks," J. Chrom. Sci., 21, 229-234, May 1983.

O1 Ohta, N., "Estimating absorption bands of component dyes by means of principal component analysis," Anal. Chem., 45, 3, 553-557, March, 1973.

O2 Osten, D.W. and Kowalski, B.R., "Multivariate Curve Resolution in Liquid Chromatography," Anal. Chem., Vol. 56, No.6, pp. 991-995, May 1984.

P1 Paraskevopoulos, P.N. et al., "Eigenvalue-Eigenvector Sensitivity Analysis of Linear Time-Invariant Singular System," IEEE Trans. Aut. Cont., Vol. AC-29, No.4, April 1984.

P2 Parris, N.A., Instrumental Liquid Chromatography, 2nd Edition, Elsevier, 1984.

P3 Porter, B. and Crossley, R., Modal Control: Theory and applications, Taylor & Francis Ltd., London, 1972.

R1 Rao, C.R., "Matrix approximations and reduction of dimensionality in multivariate statistical analysis," P.R. Khrishnaiah, ed., Multivariate Analysis - V, North-Holland Publishing Company, 1980

R2 Ritter, G.L., et.al., "Factor Analysis of the Mass Spectra of Mixtures," Anal. Chem., Vol.48, No.3, March 1976, 591-595.

S1 Sedgewick, R., Algorithms, Addison-Wesley, 1893.

S2 Sharaf, M.A. and Kowalski, B.R., "Quantitative resolution of fused chromatographic peaks in gas chromatograph / mass spectrometry," Anal. Chem., 54, 1291-1296, 1982.

S3 Snyder, L.R., "A Rapid Approach to Selecting the Best Experimental Conditions for High-Speed Liquid Column Chromatography. Part I - Estimating Initial Sample Resolution and the Final Resolution Required by a Given Problem," J. of Chromatographic Science, Vol. 10, April 1972, pp. 200-212

S4 Snyder, L.R. and Kirkland, J.J., Introduction to Modern Liquid Chromatography, 2nd Edition, Wiley, 1979.

S5 Spjotvoll, E., Marten, H., Volden, R., "Restricted Least Squares Estimation of the Spectra and Concentration of Two Unknown Constituents Availlable in Mixtures," Technometrics, Vol. 24, No.3, pp. 173-180, August, 1982.

S6 Strang, G., Linear Algebra and its Applications, 2nd Edition, Academic Press, 1980.

S7 Stewart, F.M., Introduction to Linear Algebra, D. Von Nostrand Company, Inc.

S8 Stewart, G.W., "On the sensitivity of the eigenvalue problem Ax=sBx, " SIAM J. Numer. Anal., Vol.9, No.4, Dec. 1972.

S9 Stewart, G.W., "Error and perturbation bounds for subspaces associated with certain eigenvalue problems," SIAM Review, Vol.15, No.4, October 1973.

S10 Stewart, G.W., Introduction to Matrix Computations, Academic Press, 1973.

T1 Turnbull, H.W., "On the reduction of singular matrix pencils," Proc. Edinburgh Math. Soc., Vol.4, pp. 67-76 (1935)

V1 Vandeginste, B.G.M., "Application of Cheometrics," Pure and Appl. Chem., Vol. 55, No. 12, pp. 2007-2016, 1983.

V2 Van Dooren, P., "The Computation of Kronecker's Canonacal Form of a Singular Pencils," Lin. Alg. & Appl., Vol.27, pp. 103-141, 1979.

V3 Van Dooren, P., "The generalized eigenstructure problem in linear system theory," IEEE Trans. Automatic Control, AC-26, 1, 111-129, Feb., 1981.

V4 Van Dooren, P., "Reducing subspaces: definitions, properties and algorithms," in Matrix Pencils, Ed. A.Ruhe & B.Kagstrom, Lecture Notes in Mathematics, Vol. 973, Springer, N.Y., 1982.

V5 Van Dooren, P., "Reducing subspaces: computational aspects and applications in linear systems theory," Internal Memo, Philips Resarch Laboratory, Brussels, Belgium, (1984).

V6 Vickrey, T.M. (Ed.), Liquid Chromatography Detectors, Chromatographic Science Series, Vol. 23, Marcel Dekker, Inc., 1983.

W1 Wax, M., Shan, T.-J., and Kailath, T., "Location and the Spectral Density estimation of Multiple Source," Asolimar Conference, 1982.

W2 Wedin, P.A., "Perturbation bounds in connection with singular value decomposition," BIT, 12(1972), pp.99-111

W3 Wedin, P.A., "On Angles between Subspaces of a Finite Dimensional Inner Product Space," in Matrix Pencils, ED. A. Ruhe and B. Kagstrom, Lecture Notes in Mathematics, Springer, N.Y., 1982.

W4 Wilkinson, C.R. and Reinch C., Linear Algebra, Springer-Verlag, N.Y., 1971.

W5 Wilkinson, J.H., The Algebraic Eigenvalue Problems, Clarendon Press, Oxford (1965).

APPENDIX A. MODES OF DECOMPOSITION

Matrix decompositions, in theory and in computation, play a central role in the projection method: an in-depth understanding of the nature of matrix decompositions is essential in formulating the projection method, and a good working knowledge of the numerical properties of the various matrix decompositions is critical in assuring numerical stability of the projection algorithms in practice. The purpose of this section is to provide a brief survey of three typical decompositions,

- LU decomposition,
- QR decomposition, and
- singular value decomposition (SVD),

that can be used in the projection method. For in-depth treatment on the subject, several excellent references on matrix decompositions are available in the literature (see [S6,S7] for theoretical flavor, and [G6] for numerical treatment.)

Matrix decompositions are the results of a sequence of matrix operations (e.g. rotations, elementary row/column operations) performed on a matrix that reveal "hidden" structural information about the matrix. One such piece of information that can be obtained from a matrix decomposition is the rank of a matrix. The rank of a matrix is the number of independent columns/rows in a matrix (it can be shown that the number of independent columns in a matrix is equal to the number of independent rows in a matrix), and hence determines the dimension of the column/row space of a matrix. More specifically for the problems considered in this thesis, the rank of a mixture data matrix M gives the number of independent components in that mixture.

Matrix decompositions typically factor a matrix into three matrices (see Equation A-1):

$$M = LDR \tag{A-1}$$

one (L) that contains a set of basis vectors for the column space, another (R) that contains a set of basis vectors for the row space, and a third (D), generally a diagonal matrix, that provides additional freedom in normalizing the lengths of the basis vectors that span the row and column spaces. Alternatively, the "normalizing" matrix can be absorbed by the two matrices that contain the basis vectors, in which case the decomposition factors a matrix into two matrices.

The three decompositions to be discussed below differ from each other in that different operations, with different numerical properties and varying degrees of computational requirements, are performed on the given matrix to yield a set of factored matrices of different form and structure. The decompositions are described in order of increasing computational requirements and increasing numerical accuracy as well.

LU decomposition [S6,G6]

For the case when M is a general matrix of dimension pxq, the LU decomposition with "pivoting" decomposes the matrix into

$$PM = LU \tag{A-2}$$

where P is some permutation matrix, L is a square (pxp) lower triangular matrix with 1's on the diagonal, and U is a pxq upper triangular matrix with its first r rows independent and its remaining (p-r) rows containing only zeros.

A decomposition of M

$$M = M_L M_R^T \tag{A-3}$$

with full column rank matrices M_L and M_R is then obtained by setting the first r columns of the matrix product $P^{-1}L$ to be equal to M_L, and the first r rows of L to be equal to M_R^T. Since the last (n-r) rows of U are all zeros, we have

$$M_L M_R^T = P^{-1}LU = M \tag{A-4}$$

QR decomposition [S6,G6]

It can be shown that any pxq matrix M can be factored into

$$MP = QR \tag{A-5a}$$

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix} \begin{matrix} r \\ p-r \end{matrix} \qquad \begin{matrix} r & q-r \end{matrix} \tag{A-5b}$$

where r=rank(M), Q is orthogonal, and R_{11} is upper triangular, and P is a permutation matrix [G6].

A decomposition of M

$$M = M_L M_R^T \tag{A-6}$$

with full column rank matrices M_L and M_R is then obtained by setting the first r columns of Q to be equal to M_L, and the first r rows of the matrix product RP^{-1} to be equal to M_R^T.

The main advantage of this decomposition over the LU decomposition is the exclusive use of orthogonal transformations that results in a numerically stable factorization with almost all matrices (except in few

isolated cases). A Fortran coding of an algorithm that implements this decomposition can be found in [D2], and is the one we recommend.

Singular Value Decomposition (SVD) [K2,G6]

The singular value decomposition (SVD) is one of the most important decompositions in matrix analysis and computation. As such, it is used in wide-ranging areas of applications in linear algebra and linear systems:

- determining the rank of a matrix,
- solving linear least squares problems,
- determining dependencies or near-dependencies among columns or rows of a matrix, etc.

An excellent overview of its concepts, computation, and applications, along with numerous references on related topics, is given by [K2]. For numerical computation of SVD, an efficient and stable algorithm can be found in [G5]. Of the several different versions of this algorithm that has been implemented, the one in LINPACK [D2] is recommended.

The existenceof SVD is stated in the following theorem:

THEOREM A1. Singular Value Decomposition

If $M \in R^{pxq}$, where R^{pxq} denotes the vector space of all mxn real matrices, then there exist orthogonal matrices

$$U = [u_1,\ldots,u_p] \in R^{pxp}$$

and

$$V = [v_1,\ldots,v_q] \in R^{qxq}$$

such that

$$U^T M V = S$$

where

$$S = \text{diag}(s_1, \ldots, s_k), \qquad k = \min(p, q)$$

with

$$s_1 \geq s_2 \geq \ldots \geq s_k \geq 0$$

Proof: see [G6]

□

From Theorem A1, since U and V are orthogonal, we have

$$M = USV^T \tag{A-7}$$

as desired. Furthermore, if the SVD of M is given by Theorem 1, and

$$s_1 \geq \ldots \geq s_r > s_{r+1} = \ldots = s_p = 0 \tag{A-8}$$

then

(i) rank(M) = r,

(ii) the first r columns of U span the column space of M,

and (iii) the first r columns of V span the row space of M.

The numbers s_i are called the singular values of M and they are the positive square roots of the eigenvalues of $M^T M$, the columns of U are called the left singular vectors of M and they are the orthonormal eigenvectors of MM^T, and the columns of V are called the right singular vectors of M and they are the orthonormal eigenvectors of $M^T M$.

Frobenius-Perron Theorem [L4, G1]

Some interesting results on the first (left and right) singular vectors of a positive (or a non-negative) matrix M are given by the Frobenius-Perron Theorem. We shall first define a positive matrix, and

then state the results.

If $A = [a_{ij}]$ is a matrix, we write:

(i) $A > 0$ if $a_{ij} > 0$ for all i,j

(ii) $A \gneq 0$ if $a_{ij} \geq$ for all i,j, and $a_{ij} > 0$ for at least one element

(iii) $A \geq 0$ if $a_{ij} \geq 0$ for all i,j.

THEOREM A2

If $A>0$, then there exist $s_o>0$ and $x_o>0$ such that

(i) $Ax_o=s_ox_o$;

(ii) if $s\neq s_o$ is any other eigenvalue of A, then $|s|<s_o$;

(iii) s_o is an eigenvalue of geometric and algebraic multiplicity 1.

THEOREM A3

Let $A\geq 0$ and suppose $A^m>0$ for some positive integer m. Then conclusions (i), (ii), and (iii) of Theorem A-2 apply to A.

THEOREM A4

Let $A\geq 0$. Then there exists $s_o\geq 0$, and $x_o\geq$ such that

(i) $Ax_o=s_ox_o$;

(ii) if $s\neq s_o$ is any other eigenvalue of A, then $|s|\leq s_o$.

□

Since the left and right singular vectors of a matrix M are respectively the eigenvectors of the matrices MM^T and M^TM, we can substitute MM^T and M^TM in place of A in Theorems A2-A4 and obtain similar results for the singular vectors of M.

APPENDIX B. GENERALIZED BILINEAR PERTURBATION PROBLEM.

The Problem

The problem to be addressed in this appendix is:

Given data matrices $X \in R^{mxr}$, $Y \in R^{rxn}$, and $Z \in R^{mxn}$,

$$\underset{(E_x, E_y, E_z)}{\text{Minimize:}} \quad tr[E_x^T P E_x]+tr[E_y^T Q E_y]+tr[E_z^T R E_z] \tag{B-1a}$$

$$\text{Subject to:} \quad [Z+E_z]=[X+E_x][Y+E_y] \tag{B-1b}$$

for some symmetric, positive semidefinite weighting matrices P, Q, and R.

We shall first transform the above problem to an unconstrained minimization problem via the Lagrange multiplier technique for handling equality constraints, and then derive a set of necessary conditions for a minimum. Using these conditions, we then give a solution to a special case.

Necessary Conditions for a Minimum

By introducing the Lagrange multiplier matrix $L \quad R^{mxn}$, the problem can be recast as an unconstrained minimization problem involving the criterion

$$\begin{aligned} J = tr([E_x^T P E_x]) + tr([E_y^T Q E_y]) + tr([E_z^T R E_z]) \\ + tr(L^T([Z+E_z]-[X+E_x][Y+E_y])) \end{aligned} \tag{B-2}$$

Necessary conditions for a minimum are then given by setting the first derivatives of J with respect to the minimizing variables and L to zero (see [A2,G8] for the derivatives of a function with respect to matrices):

$$\frac{\partial J}{\partial E_x} = 2PE_x - L[Y+E_y]^T = 0 \qquad \text{(B-3a)}$$

$$\frac{\partial J}{\partial E_y} = 2QE_y - [X+E_x]^T L = 0 \qquad \text{(B-3b)}$$

$$\frac{\partial J}{\partial E_z} = 2RE_z + L = 0 \qquad \text{(B-3c)}$$

$$\frac{\partial J}{\partial L} = [Z+E_z]-[X+E_x][Y+E_y] = 0 \qquad \text{(B-3d)}$$

where, "0" is a matrix with all it's elements equal to zero.

The case where E_z=0, Y=0, Q=0, and z, e_y are vectors: Complementary Least Squares Problem

The problem in this case is

Given X and z: $X^T z \neq 0$, and [X, z] with full column rank,

$$\underset{(E_x, e_y)}{\text{Minimize:}} \quad tr[E_x^T P E_x] \qquad \text{(B-4a)}$$

$$\text{Subject to:} \quad z = [X+E_x]e_y \qquad \text{(B-4b)}$$

Using Lagrange multiplier vector l, the criterion J of Equation (B-2) is

$$J = tr[E_x^T P E_x] + l^T(z-[X+E_x]e_y) \qquad \text{(B-5)}$$

From Equation (B-3), the necessary conditions for a minimum are

$$\frac{\partial J}{\partial E_x} = 2PE_x - 1e_y^T = 0 \qquad \text{(B-6a)}$$

$$\frac{\partial J}{\partial e_y} = 2Qe_y - [X+E_x]^T 1 = 0 \qquad \text{(B-6b)}$$

$$\frac{\partial J}{\partial 1} = z-[X+E_x][Y+E_y] = 0 \qquad \text{(B-6c)}$$

Solving for E_x in Equation (B-6a),

$$E_x = 0.5P^{-1}1e_y^T \qquad \text{(B-7)}$$

Substituting for E_x in Equation (B-6c)

$$z = [X+0.5P^{-1}1e_y^T]e_y \qquad \text{(B-8)}$$

Solving for the Lagrange multiplier 1 in Equation (B-8)

$$1 = \frac{2P[z-Xe_y]}{e_y^T e_y} \qquad \text{(B-9)}$$

Eliminating 1 from Equation (B-7) by substituting Equation (B-9)

$$E_x = \frac{[z-Xe_y]e_y^T}{e_y^T e_y} \qquad \text{(B-10)}$$

Using Equations (B-6c) and (B-10), we can replace the cost-function J in Equation (B-4a) with a cost-function $\tilde{J}$ to be minimize with respect to a reduce set of variables that satify the necessary conditions for a minimum (Equations B-6a,b,c)

$$\tilde{J} = \operatorname{tr}[E_x^T E_x]$$

$$= \operatorname{tr} \frac{e_y[z-Xe_y]^T P[z-Xe_y]e_y^T}{(e_y^T e_y)^2}$$

$$= \frac{[z-Xe_y]^T P[z-Xe_y]e_y^T e_y}{(e_y^T e_y)^2}$$

$$= \frac{[z-Xe_y]^T P[z-Xe_y]}{(e_y^T e_y)}$$

$$= \frac{[-e_y^T,1]\begin{bmatrix} X^T \\ z^T \end{bmatrix} P\,[X,z] \begin{bmatrix} -e_y \\ 1 \end{bmatrix}}{[-e_y^T,1]\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]\begin{bmatrix} -e_y \\ 1 \end{bmatrix}}$$

$$= \frac{[-e_y^T,1]\left[\begin{array}{c|c} X^TPX & X^TPz \\ \hline z^TPX & z^TPz \end{array}\right]\begin{bmatrix} -e_y \\ 1 \end{bmatrix}}{[-e_y^T,1]\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]\begin{bmatrix} -e_y \\ 1 \end{bmatrix}} \qquad \text{(B-11)}$$

to be minimized with respect to e_y.

Clearly $\tilde{J}$ is minimized when $[-e_y^T,1]$ is the eigenvector (normalized so that the last element is "1") corresponding to the smallest eigenvalue of the pencil

$$\left[\begin{array}{c|c} X^TPX & X^TPz \\ \hline z^TPX & z^TPz \end{array}\right] - t\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \qquad \text{(B-12)}$$

The uniqueness of the solution is given by Lemma B1; the existence of a

solution by Lemma B2.

Lemma B1

Let $t_1 \geq t_2 \geq \ldots \geq t_{n+1}$ denote the eigenvalues of the matrix pencil

$$\left[\begin{array}{c|c} X^TX & X^Tz \\ \hline z^TX & z^Tz \end{array}\right] - t \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \begin{matrix} n \\ 1 \end{matrix}$$

$$\begin{matrix} n & 1 \end{matrix}$$

where

$$\left[\begin{array}{c|c} X^TX & X^Tz \\ \hline z^TX & z^Tz \end{array}\right]$$

is a symmetric, positive definite matrix. If $X^Tz \neq 0$, then $t_n > t_{n+1}$.

Proof:

Let the SVD of X be given by

$$X = [U_1, U_2] \begin{bmatrix} S^{1/2} \\ 0 \end{bmatrix} V^T$$

We have, as detailed in the proof of Lemma B3, $t_n = 1/\tilde{s}_2$ and $t_{n+1} = 1/\tilde{s}_1$, where $\tilde{s}_1 \geq \tilde{s}_2 \geq \ldots \geq \tilde{s}_n$ denote the eigenvalues of the matrix

$$S^{-1} + \frac{S^{-1/2} U_1^T z z^T U_1 S^{-1/2}}{z^T (I - U_1 U_1^T) z}$$

with $z^T(I - U_1 U_1^T)z > 0$.

Let $\grave{s}_1 \geq \ldots \geq \grave{s}_n$ denote the eigenvalues of S^{-1}, then it can be shown that [W5]

$$\grave{s}_i \leq \tilde{s}_i \leq \grave{s}_{i+1}, \qquad i=2,\ldots,n-1$$

$$\grave{s}_1 \leq \tilde{s}_1 \leq \grave{s}_1+d$$

where

$$d = \frac{z^T U_1 S^{-1} U_1^T z}{z^T (I-U_1 U_1^T) z}$$

Also, from Lemma B3,

$$\grave{s}_1 < \tilde{s}_1$$

Hence,

$$\tilde{s}_2 < \tilde{s}_1$$

and

$$t_n > t_{n+1}$$

Q.E.D.

Lemma B2

Let $v=[v_1^T, v_2]^T$, $v^Tv=1$, denote the eigenvector corresponding to the smallest eigenvalue (of multiplicity 1) of the matrix pencil

$$\left[\begin{array}{c|c} X^TX & X^Tz \\ \hline z^TX & z^Tz \end{array}\right] - t\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \begin{matrix} n \\ 1 \end{matrix}$$

$$\begin{matrix} \\ n \end{matrix} \quad \begin{matrix} \\ 1 \end{matrix}$$

where

$$\left[\begin{array}{c|c} X^TX & X^Tz \\ \hline z^TX & z^Tz \end{array}\right]$$

is a symmetric, positive definite matrix. Then, a sufficient

condition for $v_2 \neq 0$ is that $X^T z \neq 0$.

Proof:

We have that v is the vector that minimizes

$$\min_{v,\, v^T v=1} J = \frac{[v_1^T, v_2] \left[\begin{array}{c|c} X^T X & X^T z \\ \hline z^T X & z^T z \end{array}\right] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}}{[v_1^T, v_2] \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}}$$

$$= \frac{v_1^T X^T X v_1 + v_1^T X^T z v_2 + v_2^T z^T X v_1 + v_2 z^T z v_2}{v_1^T v_1}$$

If $X^T z \neq 0$, then

$$\min_v J < \min_{v_1} \frac{v_1^T X^T X v_1}{v_1^T v_1} \qquad \text{(from Lemma B3)}$$

Assume $v_2 = 0$, then

$$\min_v J = \min_{[v_1^T, v_2 = 0]} J = \min_{v_1} \frac{v_1^T X^T X v_1}{v_1^T v_1}$$

which contradicts the preceding ineqality, hence $v_2 \neq 0$.

Q.E.D.

Lemma B3

Given a symmetric, positive definite matrix

$$\left[\begin{array}{c|c} X^T X & X^T z \\ \hline z^T X & z^T z \end{array}\right] \qquad (\#1)$$

a sufficient condition for the smallest eigenvalue of the matrix pencil

$$\left[\begin{array}{c|c} X^TX & X^Tz \\ \hline z^TX & z^Tz \end{array}\right] - t\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \qquad (*2)$$

to be less than the smallest eigenvalue of X^TX is that $X^Tz \neq 0$.

Proof:

The eigenvalues of the pencil in Equation (*2) are equal to the reciprocals of the eigenvalues of the matrix

$$\left[\begin{array}{c|c} X^TX & X^Tz \\ \hline z^TX & z^Tz \end{array}\right]^{-1} \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] \qquad (*3)$$

Using the identity for the inverse of a block matrix [K1] (for A invertible)

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]^{-1} = \left[\begin{array}{c|c} A^{-1}+EK^{-1}F & -EK^{-1} \\ \hline -K^{-1}F & K^{-1} \end{array}\right]$$

where $K=B-CA^{-1}D$, $E=A^{-1}D$, and $F=CA^{-1}$, matrix in (*3) can be rewritten as

$$\left[\begin{array}{c|c} (X^TX)^{-1} + \dfrac{(X^TX)^{-1}X^Tzz^TX^T(X^TX)^{-1}}{z^T(I-X(X^TX)^{-1}X^T)z} & 0 \\ \hline \dfrac{-z^T(X^TX)^{-1}}{z^T(I-X(X^TX)^{-1}X^T)z} & 0 \end{array}\right] \qquad (*4)$$

(it can easily be shown that with the stated assumptions on (*1), $z^T(I-X(X^TX)^{-1}X^T)z>0$). Thus, the eigenvalues of (*3) are "0" and the

eigenvalues of the matrix

$$(X^TX)^{-1} + \frac{(X^TX)^{-1}X^Tzz^TX^T(X^TX)^{-1}}{z^T(I-X(X^TX)^{-1}X^T)z} \qquad (*5)$$

Let the SVD of X be given by

$$X = [U_1, U_2]\begin{bmatrix} S^{1/2} \\ \hline 0 \end{bmatrix} V^T \qquad (*6)$$

then

$$X^TX = VSV^T \qquad (*7)$$

and

$$(X^TX)^{-1} = VS^{-1}V^T \qquad (*8)$$

Consider the similarity transformation of (*5) given by

$$V^T \left[(X^TX)^{-1} + \frac{(X^TX)^{-1}X^Tzz^TX^T(X^TX)^{-1}}{z^T(I-X(X^TX)^{-1}X^T)z} \right] V$$

$$= S^{-1} + \frac{S^{-1/2}U_1^Tzz^TU_1S^{-1/2}}{z^T(I-U_1U_1^T)z} \qquad (*9)$$

Now, the largest eigenvalue of (*9) is given by

$$\max_{a,\, a^Ta=1} a^T\left(S^{-1} + \frac{S^{-1/2}U^Tzz^TUS^{-1/2}}{z^T(I-X(X^TX)^{-1}X^T)z}\right)a$$

$$= \max_{a,\, a^Ta=1} a^TS^{-1}a + a^T\left(\frac{S^{-1/2}U^Tzz^TUS^{-1/2}}{z^T(I-X(X^TX)^{-1}X^T)z}\right)a \qquad (*10)$$

$$> \max_{a,\, a^T a = 1} a^T S^{-1} a = \text{largest eigenvalue of } (X^T X)^{-1}$$

since

$$a^T \left(\frac{S^{-1/2} U^T z z^T U S^{-1/2}}{z^T (I - X(X^T X)^{-1} X^T) z} \right) a > 0$$

Hence, we have the result:

Smallest eigenvalue of (*2) < smallest eigenvalue of $X^T X$.

Q.E.D.

APPENDIX C. EFFICIENT ALGORITHMS FOR THE PROJECTION METHOD

In implementing the projection method, savings in numerical computations can be realized by the exploiting special structure of the method. Some of these features are summarized below.

SVD approach.

Although SVD is expensive (numerically), it is the most reliable way to decompose a matrix, and yields much useful structural information about that matrix. For example, the left and right singular vectors corresponding to the nonzero singular values form two sets of orthonormal basis vectors for the row and column spaces, respectively. As will be shown below, matrices with orthonormal rows/columns greatly simplify matrix operations, and hence simplify numerical solution via the projection method. We present below the projection method solution via SVD:

Given SVD of M and N_1

$$M = U_m S_m V_m^T, \qquad S_m = \text{diag}\,(s_{m1},\ldots,s_{mr}) \tag{C-1a}$$

$$N_1 = u_n s_{n1} v_n^T \tag{C-1b}$$

(we have assumed here that M and N_1 are reconstructed data matrices-see Chapter 3), let

$$M_L = U_m S_m^{1/2} \tag{C-2a}$$

$$M_R = V_m S_m^{1/2} \tag{C-2b}$$

and

$$n_L = u_n s_{n1}^{1/2} \tag{C-3a}$$

$$n_R = v_n s_{n1}^{1/2} \tag{C-3b}$$

Substituting Equations (C-2) and (C-3) into Equations (3.3.2-3) and (3.3.2-4), and simplifying, we have:

for linear least squares projection,

$$a_{LLS} = S_m^{-1/2} U_m^T u_n s_{n1}^{1/2} \tag{C-4a}$$

$$b_{LLS} = S_m^{-1/2} V_m^T v_n s_{n1}^{1/2} \tag{C-4b}$$

and

for total least squares projection,

$$a_{TLS} = (S_m - s_a^2 I)^{-1} S_m^{-1/2} U_m^T u_n s_{n1}^{1/2} \tag{C-5a}$$

$$b_{TLS} = (S_m - s_b^2 I)^{-1} S_m^{-1/2} V_m^T v_n s_{n1}^{1/2} \tag{C-5b}$$

where,

s_a = min. singular value of $[S_m^{1/2} U_m, s_{n1}^{1/2} u_n]$, and

s_b = min. singular value of $[S_m^{1/2} V_m, s_{n1}^{1/2} u_n]$

Substituting Equations (C-4) and (C-5) into Equation (3.3.2-5), and simplifying, yields the projection solutions:

$$t_{LLS} = \frac{1}{u_n^T U_m \operatorname{diag}\left(\frac{s_{n1}}{s_{mi}}\right) V_m^T v_n} \tag{C-6a}$$

and

$$t_{TLS} = \frac{1}{u_n^T U_m \operatorname{diag}\left(\frac{s_{n1} s_{mi}}{(s_{mi} - s_a^2)(s_{mi} - s_a^2)}\right) V_m^T v_n} \tag{C-6b}$$

for the linear least squares and the total least squares projections, respectively.

Comment: Note that, in contrast to the solutions given in Equations

(3.3.2-3), (3.3.2-4), and (3.3.2-5), the solutions given in Equation (C-6) do not involve matrix inversions.

For the total least squares solution, the singular values s_a and s_b can be computed efficiently by exploiting special structures of the matrices $[S_m^{1/2}U_m, s_{n1}^{1/2}u_n]$ and $[S_m^{1/2}V_m, s_{n1}^{1/2}u_n]$. Since singular values of a matrix M equal the square roots of the eigenvalues of M^TM, we have:

$$s_a^2 = \text{min. eigenvalue of } \begin{bmatrix} S_m, & S_m^{1/2}U_m^T u_n s_{n1}^{1/2} \\ s_{n1}^{1/2}u_n^T U_m S_m^{1/2}, & s_{n1} \end{bmatrix} \tag{C-7a}$$

and

$$s_b^2 = \text{min. eigenvalue of } \begin{bmatrix} S_m, & S_m^{1/2}V_m^T v_n s_{n1}^{1/2} \\ s_{n1}^{1/2}v_n^T V_m S_m^{1/2}, & s_{n1} \end{bmatrix} \tag{C-7b}$$

Thus, computation of s_a and s_b is reduced to computation of minimum eigenvalues of matrices of the form

$$M^TM = \left[\begin{array}{ccc|c} d_1 & & & a_1 \\ & \ddots & 0 & \vdots \\ & 0 & & \vdots \\ & & d_r & a_r \\ \hline a_1 & \cdots & a_r & a_{r+1} \end{array}\right] \tag{C-8}$$

i.e a symmetric matrix that is formed by adding a row and a column to a diagonal matrix $D=\text{diag}(d_1,\ldots,d_r)$, $d_1 \geq \ldots \geq d_r$. The eigenvalues of this matrix are the roots of the polynomial equation,

$$\begin{aligned} P(s) &= \det(sI-M) \\ &= (s-a_{r+1})\prod_{i=1}^{r}(s-d_i) - \sum_{i=1}^{r}\left(a_i^2 \sum_{\substack{j=1 \\ j\neq i}}^{r}(s-d_j)\right) \end{aligned} \tag{C-9}$$

and satisfy the Sturmian Separation Theorem [B4], which states that

$$s_{j+1} \leq d_j \leq s_j, \quad j=1,\ldots,r \tag{C-10}$$

where s_i are the roots of P(s), and hence eigenvalues of M^TM. Since M^TM is a symmetric positive definite matrix, we have, for the smallest eigenvalue of M^TM, a bound

$$0 < s_{r+1} \leq d_r, \tag{C-11}$$

which can be used to advantage in computing s_{r+1} by standard root finding methods such as the bisection method, Newton-Raphson iteration method, etc. Furthermore, computation of P(s), given by Equation (C-9), at each iteration of a root finding algorithm can be done efficiently by a recursion formula described in Algorithm 1.

Algorithm 1.

Comment: Initialize

$Z = 1$

$P(s) = s - a_{r+1}$

Comment: Start recursion

Do $j=1,\ldots,r$

$P(s) = (s-d_j)P(s) - a_j^2 Z$

$Z = (s-d_j)Z$

End do j.

[]

<u>QR decomposition approach.</u>

An alternative to SVD that yields reliable decomposition at reduced computational cost is the QR decomposition (see Appendix A). We describe

here projection method via QR decomposition:

Given data matrices M and N_1 of mixture and target component, respectively, QR decomposition yields

$$M = Q_m R_m^T \tag{C-12a}$$

$$N_1 = q_{n1} r_{n1}^T \tag{C-12b}$$

Let,

$$M_L = Q_m \tag{C-13a}$$

$$M_R = R_m \tag{C-13b}$$

and

$$n_L = q_{n1} \tag{C-14a}$$

$$n_R = r_{n1} \tag{C-14b}$$

Substituting Equations (C-13) and (C-14) into Equations (3.3.2-3) and (3.3.2-4), and simplifying, we have:

for linear least squares projection,

$$a_{LLS} = Q_m^T q_{n1} \tag{C-15a}$$

$$b_{LLS} = (R_m^T R_m)^{-1} R_m^T r_{n1} \tag{C-15b}$$

and

for total least squares projection,

$$a_{TLS} = \frac{Q_m^T q_{n1}}{1-s_a^2} \tag{C-16a}$$

$$b_{TLS} = (R_m^T R_m - s_b^2 I)^{-1} R_m^T r_{n1} \tag{C-16b}$$

where,

s_a = min. singular value of $[Q_m, q_{n1}]$

s_b = min. singular value of $[R_m, r_{n1}]$

The projection solutions for the two projection methods are,

$$t_{LLS} = b_{LLS}^T a_{LLS} \qquad \text{(C-17a)}$$

$$t_{TLS} = b_{TLS}^T a_{TLS} \qquad \text{(C-17b)}$$

where a_{LLS}, b_{LLS}, a_{TLS}, and b_{TLS} are given by Equations (C-15) and (C-16).

Comment: In comparison to the SVD approach, the QR decomposition approach requires additional matrix operations in obtaining the final solution (consequence of working with a matrix, R, that does not have orthonormal columns). With the LLS projection, QR decomposition approach requires a matrix inverse. Nevertheless, computational savings in subsituting SVD with QR decomposition may offset this difference. With the TLS projection method, however, Equation (C-16b) requires computation of a singular value, s_b, for which, unlike the computation of s_a, there are no special structures to exploit. Hence, unless methods such as inverse iteration method [G6] for computing minimum eigenvalue of positive definite symmetric matrix can be exploited efficiently, the QR decomposition approach may not offer much advantage over the SVD approach for the TLS projection method.

Sequential data acquisition: Solution updating for QR decomposition approach.

We address the problem of updating the projection method solution when an additional row of data is observed or appended (such a problem may arise in a Liquid Chromatography-Spectroscopy experiment where additional data points are observed as the mixture elutes from the

column, or in an Excitation -Emission Spectroscopy experiment where the mixture is excited or detected at additional wavelengths). The objective here is to efficiently process the additional data without having to resolve the problem completly with new data. We give update schemes for QR decomposition and LLS projection.

Given data matrices M and N_1, and their QR decompositions

$$M = Q_m R_m^T \tag{C-18a}$$

$$N_1 = q_{n1} r_{n1}^T \tag{C-18b}$$

consider the problem of updating the decomposition when an additional row of data points are appended as shown (see [G6])

$$\bar{M} = \begin{bmatrix} M \\ x^T \end{bmatrix} \tag{C-19a}$$

$$\bar{N}_1 = \begin{bmatrix} N_1 \\ y^T \end{bmatrix} \tag{C-19b}$$

We have,

$$\bar{M} = \begin{bmatrix} M \\ x^T \end{bmatrix} = \left[\begin{array}{c|c} Q_m & 0 \\ \hline 0 & 1 \end{array}\right] \begin{bmatrix} R_m^T \\ x^T \end{bmatrix} \tag{C-20a}$$

$$\bar{N}_1 = \begin{bmatrix} N_1 \\ y^T \end{bmatrix} = \left[\begin{array}{c|c} q_{n1} & 0 \\ \hline 0 & 1 \end{array}\right] \begin{bmatrix} r_{n1}^T \\ y^T \end{bmatrix} \tag{C-20b}$$

or

$$\left[\begin{array}{c|c} Q_m^T & 0 \\ \hline 0 & 1 \end{array}\right] \bar{M} = \begin{bmatrix} R_m^T \\ x^T \end{bmatrix} \tag{C-21a}$$

$$\left[\begin{array}{c|c} q_{n1}^T & 0 \\ \hline 0 & 1 \end{array}\right] \tilde{N}_1 = \begin{bmatrix} r_{n1} \\ y^T \end{bmatrix} \qquad \text{(C-21b)}$$

Use Givens' transformations J_m^T and J_n^T to zero out x and y, respectively,

$$[J_m^T]\left[\begin{array}{c|c} Q_m^T & 0 \\ \hline 0 & 1 \end{array}\right] \tilde{M} = [J_m^T]\begin{bmatrix} R_m^T \\ x^T \end{bmatrix} = \begin{bmatrix} \tilde{R}_m^T \\ 0 \end{bmatrix} \qquad \text{(C-22a)}$$

$\hat{Q}_m^T$

$$[J_n^T]\left[\begin{array}{c|c} q_{n1}^T & 0 \\ \hline 0 & 1 \end{array}\right] \tilde{N}_1 = [J_n^T]\begin{bmatrix} r_{n1} \\ y^T \end{bmatrix} = \begin{bmatrix} \tilde{r}_{n1}^T \\ 0 \end{bmatrix} \qquad \text{(C-22b)}$$

$\hat{Q}_n^T$

We can now express $\tilde{M}$ and $\tilde{N}_1$ as

$$\tilde{M} = \hat{Q}_m \begin{bmatrix} \tilde{R}_m^T \\ 0 \end{bmatrix} = \tilde{Q}_m \tilde{R}_m^T \qquad \text{(C-23a)}$$

$$\tilde{N}_1 = \hat{Q}_n \begin{bmatrix} \tilde{r}_{n1}^T \\ 0 \end{bmatrix} = \tilde{q}_{n1} \tilde{r}_{n1}^T \qquad \text{(C-23b)}$$

Let J_m and J_n be partitioned as follow,

$$J_m = [J_{m1}, j_{m2}]$$

$$J_n = [j_{n1}, J_{n2}]$$

where j_{m2} and j_{n1} are vectors. Partial update equations for the LLS projections are then given by:

$$\begin{aligned} \tilde{a}_{LLS} &= (\tilde{Q}_m^T \tilde{Q}_m)^{-1} \tilde{Q}_m^T \tilde{q}_{n1} = \tilde{Q}_m^T \tilde{q}_{n1} \\ &= J_{m1}^T \left[\begin{array}{c|c} Q_m^T & 0 \\ \hline 0 & 1 \end{array}\right]\left[\begin{array}{c|c} q_{n1} & 0 \\ \hline 0 & 1 \end{array}\right] j_{n1} \end{aligned} \qquad \text{(C-24a)}$$

$$= J_{m1}^T \left[\begin{array}{c|c} Q_m^T q_{n1} & 0 \\ \hline 0 & 1 \end{array}\right] j_{n1}$$

$$= J_{m1}^T \left[\begin{array}{c|c} a_{LLS} & 0 \\ \hline 0 & 1 \end{array}\right] j_{n1}$$

and,

$$\bar{b}_{LLS} = (\bar{R}_m^T \bar{R}_m)^{-1} \bar{R}_m^T \bar{r}_{n1}$$

$$= (J_{m1}^T \; R_m^T [R_m, x] J_{m1})^{-1} J_{m1}^T \; R_m^T [r_{n1}, y] j_{n1} \qquad \text{(C-24b)}$$
$$\qquad\quad x^T \qquad\qquad\qquad\qquad x^T$$

$$= (J_{m1}^T \left[\begin{array}{c|c} R_m^T R_m & R_m^T x \\ \hline x^T R_m & x^T x \end{array}\right] J_{m1})^{-1} J_{m1}^T \left[\begin{array}{c|c} R_m^T r_{n1} & R_m^T y \\ \hline x^T r_{n1} & x^T y \end{array}\right] j_{n1}$$

APPENDIX D. A QUALITATIVE ERROR ANALYSIS OF THE PROJECTION METHOD

In this section we discuss the effect of noisy data on the projection method solution. A general approach will be taken in the sense that we will be addressing qualitative aspects of the expected values of bias and error variance. We begin by assuming that the noisy data matrices are of the form given by Equation (2.5-1), and the errors are random in nature with

$$\begin{aligned} e_{i,j} &\sim \text{Normal}(0,w^2) \\ E(e_{i,j}e_{k,l}) &= 0 \text{ for } i\neq k, j\neq l \end{aligned} \tag{D-1}$$

where $e_{i,j}$ is the i,j-th element of the error matrices. It is also assumed that the errors are independent from sample to sample.

Define,

$$d = t - t_o \tag{D-2}$$

where t_o and t are the projection method solutions for the noise free data matrix and noisy data matrix, respectively, and let,

$$d = f(e) \tag{D-3}$$

where e is a vector composed of all the elements of the error matrices (e.g. by stacking columns of error matrices in one vector). That is, assume that the error in projection method solution is some nonlinear function of the elements of the error matrices. Taylor series expansion of f(e) about e=0 yields,

$$\begin{aligned} d &= f(e) \\ &= f(0) + g^Te + 0.5e^THe + \cdots \end{aligned} \tag{D-4}$$

where

$$g_i = \left.\frac{\partial f}{\partial e_i}\right|_{e=0} \tag{D-5a}$$

$$H_{i,j} = \left.\frac{\partial^2 f}{\partial e_i \partial e_j}\right|_{e=0} \tag{D-5b}$$

Since f(0)=0, to a second order approximation, we have,

$$d = g^T e + 0.5 e^T H e \tag{D-6}$$

Then, the solution bias and variance are given by,

(i) bias

$$\begin{aligned} \text{Bias}(d) &= E(d) \\ &= \underbrace{g^T E(e)}_{0} + 0.5E(e^T H e) \\ &= 0.5E\Big(\sum_i e_i^2 H_{i,i} + \sum_i \Big(e_i \sum_{j\neq i} e_j H_{i,j}\Big)\Big) \\ &= 0.5\sum_i E(e_i^2) H_{i,i} + 0.5\sum_i \sum_{j\neq i} \underbrace{E(e_i e_j)}_{0} H_{i,j} \\ &= 0.5 w^2 \text{tr}(H) \end{aligned} \tag{D-7}$$

(ii) variance

$$\begin{aligned} \text{Var}(d) &= E(d^2) - (E(d))^2 \\ &= E(g^T e e^T g) + \underbrace{E(g^T e e^T H e)}_{0} + 0.25E(e^T H e e^T H e) - (E(d))^2 \\ &= w^2 g^T g + 0.25 w^4 \Big(3\sum_i H_{i,i}^2 + \sum_i \Big(H_{i,i} \sum_{j\neq i} H_{j,j}\Big) + 2\sum_i \sum_{j\neq i} H_{i,j}^2\Big) \\ &\quad - 0.25 w^4 \Big(\sum_i H_{i,i}^2 + \sum_i \Big(H_{i,i} \sum_{j\neq i} H_{j,j}\Big)\Big) \\ &= w^2 g^T g + 0.5 w^4 \sum_i \sum_j H_{i,j}^2 \\ &= w^2 g^T g + 0.5\, w^4 \text{tr}(H^T H) \end{aligned} \tag{D-8}$$

where E(.) indicates "expected value of", and tr(.) indicates the "trace of".

APPENDIX E. A FIRST ORDER EIGENVALUE SENSITIVITY FORMULA FOR REGULAR PENCILS

Given eigenvalues t_i, $i=1,\ldots,n$, $t_i \neq t_j$ for $i \neq j$, and their associated left and right eigenvectors w_i and v_i of a regular pencil $M-tN$, consider the eigenvalue problem of a perturbed pencil $\tilde{M}-t\tilde{N}$,

$$\tilde{M} - t\tilde{N} = (M+E_m) - t(N+E_n).$$

In what follows, we work under the constraint (or assumption) that E_m, E_n are structured such that the eigenvalues of the perturbed pencil are continuous functions of the perturbations.

Let,

$$\begin{aligned} \tilde{t}_i &= t_i + e_{ti} \\ \tilde{w}_i &= w_i + e_{wi} \qquad i=1,\ldots,n \\ \tilde{v}_i &= v_i + e_{vi} \end{aligned}$$

be the eigenvalues, and left and right eigenvectors of the perturbed pencil. We shall assume here that $||E_m||_2$ and $||E_n||_2$ are such that

$$\tilde{t}_i(M+E_m, NE_n) \neq \tilde{t}_j(M+E_m, N+E_n), \text{ for } i \neq j.$$

Then, from the definitions of eigenvalues and eigenvectors:

$$\{[M+E_m]-(t_i+e_{ti})[N+E_n]\}(v_i+e_{vi}) = 0 \qquad i=1,\ldots,n$$

and multiplying to the left by the $(w_i+e_{wi})^T$,

$$(w_i+e_{wi})^T\{[M+E_m]-(t_i+e_{ti})[N+E_n]\}(v_i+e_{vi}) = 0 \qquad i=1,\ldots,n \qquad (E-1)$$

For small E_m, E_n, Equation (E-1) can be simplified by ignoring second- and higher-order error terms, and noting that for the

unperturbed pencil

$$w_i^T[M-t_iN]v_i = 0 \qquad i=1,\ldots,n$$

We then have

$$w_i^T[E_m - t_iE_n - e_{ti}N]v_i = 0 \qquad i=1,\ldots,n$$

or solving for e_{ti},

$$e_{ti} = \frac{w_i^T[E_m - t_iE_n]v_i}{w_i^TNv_i}, \qquad i=1,\ldots,n \tag{E-2}$$

to a first-order approximation.

APPENDIX F. PROOFS OF LEMMAS IN CHAPTER 3

Lemma 2a.

Given matrix decompositions

$$M = M_L M_R^T$$

$$N_1 = n_L n_R^T$$

where M_L, M_R are full column rank matrices, and n_L, n_R are vectors. Then $E(t)(=M-tN_1)$ loses rank for

$$t = \frac{1}{b^T a}$$

where, a and b are the unique solutions to the equations

$$n_L = M_L a$$

$$n_R = M_R b$$

Proof:

Since $M = c_1 N_1 + \ldots + c_r N_r$, we have

$R(n_L) = R(N_1) \subset R(M) = R(M_L)$, and

$R(n_R) = R(N_1^T) \subset R(M^T) = R(M_R)$

where R(.) indicates "range of", and $\subset$ indicates "contained in". Hence we can express n_L and n_R as linear combinations of columns of M_L and M_R, respectively, i.e.

$$n_L = M_L a$$

$$n_R = M_R b$$

where the vectors a,b are the unique weightings on the columns of M_L, M_R so that the above equalities hold. We can now rewrite E(t) as

$$E(t) = M - tN_1$$

$$= M_L M_R^T - t n_L n_R^T$$

$$= M_L M_R^T - t M_L a b^T M_R^T$$

$$= M_L (I - t a b^T) M_R^T$$

Since M_L, M_R have full column ranks, matrix E(t) loses rank when

$$I - t a b^T$$

loses rank. That is, when

$$|I - t a b^T| = 0$$

where $|.|$ indicates "determinant of".

Using the identity,

$$|I - t a b^T| = |1 - t b^T a|$$

and solving for t, we have

$$t = \frac{1}{b^T a}$$

Q.E.D.

Lemma 2b.

Find matrix decompositions

$$M = M_L M_R^T$$

$$N = N_L N_R^T$$

where M_L, M_R, N_L, N_R are full column rank matrices. Then E(t)(=M-tN) loses rank when t is equal to the reciprocals of the eigenvalues of the matrix

$$B^T A$$

where, A and B are the unique solutions to the equations

$$N_L = M_L A$$

$$N_R = M_R B$$

Proof:

Same as in Lemma 2a after replacing the vectors a, b, n_L, n_R with the matrices A, B, N_L, N_R, respectively, except now we have to solve for the values of t for which

$$|I - tB^TA| = 0,$$

i.e., the reciprocals of the eigenvalues of B^TA.

Q.E.D.

APPENDIX G. PROOFS OF RESULTS IN CHAPTER 4

Result 1

Given a pxq data matrix M,

$$M = [m_1,\ldots,m_q] \ (=c_1x_1y_1^T+\ldots+c_rx_ry_r^T)$$

$c_i \geq 0$, $x_i \geq 0$, and $y_i \geq 0$, $i=1,\ldots,r$

(i) The (unknown) vectors x_i lie in the positive orthant of R^p;

(ii) the points m_j lie inside the positive cone generated by the vectors x_i.

Proof:

(i) From the positivity constraints on x_i;

(ii) we have,

$$\begin{aligned} M &= [m_1,\ldots,m_q] \\ &= [c_1x_1y_1^T+\ldots+c_rx_ry_r^T] \qquad (*) \\ &= [(c_1y_{11}x_1+\ldots+c_ry_{r1}x_r),\ldots,(c_1y_{1q}x_1+\ldots+c_ry_{rq}x_r)] \end{aligned}$$

hence the result from the definition of positive cone, and the positivity constraints on c_i, y_i.

Q.E.D.

Result 2

Given a pxq data matrix M,

$$M = [m_1,\ldots,m_q] \ (=c_1x_1y_1^T+\ldots+c_rx_ry_r^T)$$

$c_i \geq 0$, $x_i \geq 0$, and $y_i \geq 0$, $i=1,\ldots,r$

where both $\{x_i\}$ and $\{y_i\}$ are independent sets, the unknown vectors x_i

(i) lie in, and span, the signal space;

(ii) lie in (the portion of signal space contained in) the

positive orthant of measurement space;

(iii) generate a positive cone that (lies in the portion of signal space contained in the positive orthant of measurement space and) contains m_j.

Proof:

(i) From Equation (*) in Result 1, and the independence assumptions on $\{x_i\},\{y_i\}$;

(ii) from the positivity constraints on x_i;

(iii) from Equation (*) in Result 1, the definition of positive cone, and the positivity constraints on c_i, y_i. Q.E.D.

Result 3

Given a unit-sum normalized data matrix $\grave{M}$,

$$\grave{M} = [\grave{m}_1,\ldots,\grave{m}_q] \ (=\grave{c}_1\grave{x}_1\grave{y}_1^T+\ldots+\grave{c}_r\grave{x}_r\grave{y}_r^T)$$

$$\grave{c}_i\geq 0,\ \grave{x}_i\geq 0,\ \text{and}\ \grave{y}_i\geq 0,\ i=1,\ldots,r$$

$$1^T\grave{m}_j=1,\ j=1,\ldots,q$$

$$1^T\grave{x}_i=1,\ i=1,\ldots,r$$

where both $\{\grave{x}_i\}$ and $\{\grave{y}_i\}$ are independent sets, the (tips of) normalized unknown vectors $\grave{x}_i$

(i) lie in, and span, the normalized signal space;

(ii) lie in (the portion of normalized signal space contained in) the positive orthant of measurement space;

and the convex hull of (tips of) $\grave{x}_i$

(iii) contains (the tips of) the normalized data vectors $\grave{m}_j$.

Proof:

(i) From the independence assumptions on $\{\grave{x}_i\},\{\grave{y}_i\}$;

(ii) from the positivity and unit-sum constraints on $\grave{x}_i$;

(iii) we have,

$$\hat{m}_j = \hat{c}_1\hat{y}_{1j}\hat{x}_1+\ldots+\hat{c}_r\hat{y}_{rj}\hat{x}_r, \quad j=1,\ldots,q$$

$$\hat{c}_1\hat{y}_{1j}+\ldots+\hat{c}_r\hat{y}_{rj} = 1, \quad \hat{c}_i\hat{y}_{ij}\geq 0, \quad i=1,\ldots,r$$

that is, $\hat{m}_j$ are convex combinations of $\hat{x}_1,\ldots,\hat{x}_r$, hence the result.

Q.E.D.

Result 4

Given an rxq parametrized data matrix Z

$$Z = [z_1,\ldots,z_q] \ (=c_1w_1y_1^T+\ldots+c_rw_ry_r^T)$$

$$c_i\geq 0, \ y_i\geq 0, \ U_1w_i\geq 0, \ i=1,\ldots,r, \ U_1^TU_1=I$$

where both $\{w_i\}$ and $\{y_i\}$ are independent sets, the parametrized unknown vectors w_i

(i) lie in, and span, the parametrized signal space;

(ii) lie inside a cone (an outer bounding cone) in R^r given by

$$\bigcap_{k=1}^{p}\{a: u_{k1}a_1+\ldots+u_{kr}a_r\geq 0, \ a=[a_1,\ldots,a_r]\in R^r\} \qquad (*)$$

where $\bigcap$ denotes "intersection of";

(iii) generate a positive cone that (lie inside the cone given by (*) and) contains z_j.

Proof:

We have,

$$Z = [z_1,\ldots,z_q] = [c_1w_1y_1^T+\ldots+c_rw_ry_r^T]$$

$$= [(c_1y_{11}w_1+\ldots+c_ry_{r1}w_r),\ldots,(c_1y_{1q}w_1+\ldots+c_ry_{rq}w_r)]$$

Hence,

(i) from the independence assumptions on w_i, y_i;

(ii) from the constraint $U_1w_i\geq 0$: we have,

$$u_{j1}w_{1i}+\ldots+u_{jr}w_{ri} \geq 0, \quad j=1,\ldots p \qquad (**)$$

for each j, j=1,...,p, Equation (**) restricts the solution space (the region in which the solutions w_i can lie) to a half-space in R^r. Thus, the intersection of all the allowable half-space as defined by the inequalities in Equation (**) form the outer bound;

(iii) from the definition of positive cone, and the positivity constraints on c_i, y_i.

Q.E.D.

Result 5

Given a unit-sum normalized parametrized data matrix $\grave{Z}$

$$\grave{Z} = [\grave{z}_1,\ldots,\grave{z}_q] \ (=\grave{c}_1\grave{w}_1\grave{y}_1^T+\ldots+\grave{c}_r\grave{w}_r\grave{y}_r^T)$$

$$\grave{c}_i \geq 0,\ \grave{y}_i \geq 0,\ \grave{U}_1\grave{w}_i \geq 0,\ i=1,\ldots,r$$

$$1^T\grave{z}_j=1,\ j=1,\ldots,q$$

$$1^T\grave{w}_i=1,\ i=1,\ldots,r$$

$$1^T\grave{U}_1=1^T$$

where both $\{\grave{w}_i\}$ and $\{\grave{y}_i\}$ are independent sets, the (tips of) normalized parametrized unknown vectors $\grave{w}_i$

(i) lie in, and span, the normalized parametrized signal space;

(ii) lie inside a convex polygon given by the intersection of the normalizing plane (in the parametrized signal space) with the outer bounding cone in Result 4(ii);

and the convex hull of (tips of) $\grave{w}_i$

(iii) form an (r-1)-dimensional simplex that (lie inside the convex polygon in (ii) and) contains the (tips of) $\grave{z}_j$.

Proof:

(i) From the independence assumptions on $\grave{w}_i, \grave{y}_i$;

(ii) from Result 4(ii) and the unit-sum constraint on $\grave{w}_i$;

(iii) we have

$$\grave{z}_j = \grave{c}_1\grave{y}_{1j}\grave{w}_1+\ldots+\grave{c}_r\grave{y}_{rj}\grave{w}_r, \quad j=1,\ldots,q$$

$$\grave{c}_1\grave{y}_{1j}+\ldots+\grave{c}_r\grave{y}_{rj}=1, \quad \grave{c}_i\grave{y}_{ij}\geq 0$$

that is, $\grave{z}_j$ are convex combinations of $\grave{w}_1,\ldots,\grave{w}_r$, hence the result.

Q.E.D.

APPENDIX H. A MAXIMUM LIKELIHOOD APPROACH TO THE TARGET COMPONENT PROBLEM: A PARTIAL SOLUTION

When the data matrices contain experimental noise/errors, an alternate approach to solving the target components problem is the maximum likelihood estimation method. In this approach, the inherent structures of the data matrices are hypothesized, and the parameters of the hypothesized matrices are estimated under a maximum likelihood criterion. Afer which a single eigenvalue problem is solved, using the parameters estimated in the first step, to yield the desired concentration of the target component.

Given a set of coupled independent observations $\{M_i, N_i\}$, $i=1,\ldots,d$, the hypothesized structures of the observed data matrices are of the form:

$$\underbrace{\begin{bmatrix} M_i \\ N_i \end{bmatrix}}_{Z_i} = \underbrace{\begin{bmatrix} USV^T \\ Upsq^T V^T \end{bmatrix}}_{Z_o} + \underbrace{\begin{bmatrix} E_{1i} \\ E_{2i} \end{bmatrix}}_{E_i} \quad i=1,\ldots,d \qquad (H\text{-}1)$$

where

$$\{E_i\}_{kl} \sim \text{Normal}(0, w^2), \quad k=1,\ldots,m, \; l=1,\ldots n$$

$$E(Z)=Z_o$$

$$E(Z-Z_o)^T(Z-Z_o)=w^2 I$$

$$U^T U=I, \; V^T V=I, \; p^T p=1, \; q^T q=1$$

S diagonal, positive definite; $s>0$

$m \leq n$ (we assume without loss of generality)

where the operator E(.) indicates "expected value of". The joint density function (the likelihood function) of the sampled data set is then given

by [A1]

$$L(Z_o, w) = \frac{\text{Exp}\left(\sum_{i=1}^{d} \text{tr}(Z_i - Z_o)^T (Z_i - Z_o)/2mw^2\right)}{((2\pi)^{mn}(mw^2)^n)^{d/2}} \qquad \text{(H-2)}$$

The problem to be solved is to find U, V, S, p, q, s, and w that maximize the likelihood function. Once the maximizing variables are found, the value of t for which the matrix

$$S - t.psq^T \qquad \text{(H-3)}$$

drops rank is solved to find the concentration of the target component.

We shall formulate this approach in a sequence of three problems, each of which addresses a maximum likelihood estimation of a "structured" matrix that builds upon (and is progressively more complex than) the preceding one, with the last being the one stated above.

Maximum Likelihood Estimation of Structured Matrices

Given an observation matrix $Z \in R^{mxn}$

$$Z = Z_o + E$$
$$E_{i,j} \sim N(0, w^2) \qquad \text{(H-4a)}$$

the probability density function of Z is given by

$$p(Z) = \frac{\text{Exp}(-0.5\,\text{tr}(Z - Z_o)^T R^{-1}(Z - Z_o))}{((2\pi)^{mn}|R|)^{1/2}} \qquad \text{(H-4b)}$$

where

$$E(Z) = Z_o$$

$$E(Z-Z_o)^T(Z-Z_o) = R = mw^2 I$$

$|.|$ indicates the determinant operator

Now, consider a sampled data set consisting of d independent observations $Z_1,\ldots,Z_d$. The joint density function (the likelihood function) is then given by

$$L(Z_o,w) = \frac{\text{Exp}(-0.5 \sum_{i=1}^{d} \text{tr}(Z_i-Z_o)^T R^{-1}(Z_i-Z_o))}{((2\pi)^{mn}|R|)^{1/2}}$$

$$= \frac{\text{Exp}(-\sum_{i=1}^{d} \text{tr}(Z_i-Z_o)^T(Z_i-Z_o)/(2mw^2))}{((2\pi)^{mn}(mw^2)^n)^{d/2}} \qquad \text{(H-5)}$$

and its log-likelihood function by

$$\text{Log } L(Z_o,w) = -d(\log((2\pi)^{mn}(mw^2)^n))/2 - 0.5\text{tr}\sum_{i=1}^{d}(Z_i-Z_o)^T(Z_i-Z_o)/(mw^2)$$

$$= -d(\log((2\pi)^{mn}(mw^2)^n))/2 - 0.5\text{tr}(\grave{R}-2\grave{Z}_o^T Z_o+Z_o^T Z_o)/(mw^2) \qquad \text{(H-6)}$$

where

$$\grave{R} = \sum_{i=1}^{d} Z_i^T Z_i/d \qquad \text{(H-7a)}$$

$$\grave{Z}_o = \sum_{i=1}^{d} Z_i/d \qquad \text{(H-7b)}$$

Problem A

Given a set of independent observations Z_i, i=1,...,d, where each of the Z_j is described by Equation (H-4), find the maximum likelihood (ML) estimate of Z_o and w^2.

Solution:

The ML estimate of Z_o and w^2 are given by maximizing the log-likelihood function (Equation H-6)

$$\max_{Z_o, w^2} \log L(Z_o, w^2) = -0.5d\text{Log}((2\ )^{mn}(mw^2)^n) - 0.5\text{tr}(\grave{R}-2\grave{Z}_o^T Z_o + Z_o^T Z_o)/(mw^2) \qquad \text{(H-8)}$$

The necessary conditions for maximum are

$$\frac{\partial \log L}{\partial Z_o} = \frac{-d}{2mw^2}(-2\grave{Z}_o + 2Z_o) = 0 \qquad \text{(H-9a)}$$

$$\frac{\partial \log L}{\partial w^2} = \frac{-dnm}{2mw^2} + \frac{-dm}{2mw^4}\text{tr}(\grave{R}-2\grave{Z}_o^T Z_o + Z_o^T Z_o) = 0 \qquad \text{(H-9b)}$$

Solving for the maximizing Z_o and w^2 in Equation (H-9)

$$\hat{Z}_o = \grave{Z}_o \qquad \text{(H-10a)}$$

$$\hat{w}^2 = \text{tr}\left(\frac{\grave{R}-2\grave{Z}_o^T Z_o + Z_o^T Z_o}{n}\right) \qquad \text{(H-10b)}$$

Where "$\hat{\ }$" indicates maximum likelihood estimate.

Problem B

Given a set of independent observations Z_i, i=1,...,d, where each Z_j is described by Equation (H-4), find the ML estimates of Z_o and w^2 subject to the constraint rank(Z_o)=r, r<min(m,n)

Solution:

The ML estimates of Z_o and w^2 are given by maximizing the log-likelihood function subject to the given constraint

$$\max_{\substack{Z_o, w^2 \\ \text{rank}(Z_o)=r}} \log L(Z_o, w^2) = -0.5d\text{Log}((2\pi)^{mn}(mw^2)^n) - 0.5\text{tr}(\grave{R}-2\grave{Z}_o^T Z_o+Z_o^T Z_o)/(mw^2) \quad \text{(H-11)}$$

Maximizing with respect to w^2 yields (as in Problem A)

$$\hat{w}^2(Z_o) = \text{tr}\left(\frac{\grave{R}-2\grave{Z}_o^T Z_o+Z_o^T Z_o}{n}\right) \quad \text{(H-12)}$$

Substituting ML estimate of w^2 given by Equation (H-12) into Equation (H-11) yields

$$\log L(Z_o, \hat{w}^2) = \log L_1(Z_o) = -0.5d(\log\{(2\pi)^{mn}((m/n)\text{tr}(\grave{R}-2\grave{Z}_o^T Z_o+Z_o^T Z_o))^n\}) - 0.5d(n/m) \quad \text{(H-13)}$$

to be maximized with respect to Z_o (subject to rank(Z_o)=r).

Since logarithm is a strictly monotonic function, and by dropping the constant term that does not affect the maximization, Z_o that maximizes log $L_1(Z_o)$ can be found by maximizing the following function

in Z_o

$$L_2(Z_o) = \mathrm{tr}(2\grave{Z}_o^T Z_o - Z_o^T Z_o) \tag{H-14}$$

The constraint on Z_o (i.e., rank(Z_o)=r) can be incorporated into the maximization process by recasting the problem as

$$\max_{U_r, V_r, S_r} \tilde{L}_2(U_r, V_r, S_r) = \mathrm{tr}\Big(2\grave{Z}_o^T [U_r, U_{m-r}] \left[\begin{array}{c|c} S_r & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix}$$

$$- [V_r, V_{n-r}] \left[\begin{array}{c|c} S_r^2 & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} \Big)$$

$$= \mathrm{tr}\Big(2\grave{Z}_o^T [U_r, U_{m-r}] \left[\begin{array}{c|c} S_r & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} \Big)$$

$$- \mathrm{tr}(S_r^2)$$

(H-15)

where U_r, V_r, and S_r are the appropriate components of the SVD of Z_o given by

$$Z_o = [U_r, U_{m-r}] \left[\begin{array}{c|c} S_r & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix}$$

$$\begin{aligned} U^T U &= I_{nxn} \\ V^T V &= I_{mxm} \\ S_r &= \mathrm{diag}\,(s_1, \ldots, s_r) \\ s_i &> 0, \quad i=1, \ldots, r \end{aligned} \tag{H-16}$$

From Lemma H1 (see end of this appendix), U_r and V_r that maximize

$$\mathrm{tr}\left(2\grave{Z}_o^T [U_r, U_{m-r}] \left[\begin{array}{c|c} S_r & 0 \\ \hline 0 & 0 \end{array}\right] \begin{bmatrix} V_r^T \\ V_{n-r}^T \end{bmatrix} \right) \tag{H-17}$$

for all possible values of S_r are given by

$$\hat{U}_r = \grave{U}_r \tag{H-18a}$$

$$\hat{V}_r = \grave{V}_r \tag{18b}$$

where $\grave{U}_r$ and $\grave{V}_r$ are the appropriate components of the SVD of $\grave{Z}_o$ given by

$$\grave{Z}_o = [\grave{U}_r, \grave{U}_{m-r}] \left[\begin{array}{c|c} \grave{S}_r & 0 \\ \hline 0 & \grave{S}_{m-r} \end{array}\right] \begin{bmatrix} \grave{V}_r^T \\ \grave{V}_{n-r}^T \end{bmatrix}$$

$$\begin{aligned} &\grave{U}^T\grave{U} = I_{nxn} \\ &\grave{V}^T\grave{V} = I_{mxm} \\ &\grave{S}_r = \mathrm{diag}\,(\grave{s}_1, \ldots, \grave{s}_r) \\ &\grave{S}_{m-r} = \begin{bmatrix} \grave{s}_{r+1} & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & \grave{s}_m & 0 \end{bmatrix} \quad \begin{matrix} \\ \\ m-r \end{matrix} \\ &\qquad\qquad\quad m-r \qquad\quad n-m \\ &\grave{s}_i \geq 0, \quad i=1,\ldots,m \end{aligned} \tag{H-19}$$

Clearly, U_r and V_r that maximize Equation (H-17) also maximize $\tilde{L}_2$ (Equation H-15). Substituting the ML estimates of U_r and V_r (Equation H-18) into Equation (H-15) yields

$$\tilde{L}_2(\hat{U}_r, \hat{V}_r, S_r) = 2\mathrm{tr}(\grave{V}_r\grave{S}_r\grave{V}^T_r) - \mathrm{tr}S_r^2 \tag{H-20}$$

Maximizing Equation (H-20) with respect to S_r, we have

$$\frac{\partial \tilde{L}_2}{\partial S_r} = 2\grave{S}_r - 2S_r = 0 \qquad \text{(H-21)}$$

yielding the ML estimate of S_r

$$\hat{S}_r = \grave{S}_r \qquad \text{(H-22)}$$

Thus, the ML estimate of Z_o is

$$\hat{Z}_o = \grave{U}_r \grave{S}_r \grave{V}_r \qquad \text{(H-23)}$$

Substituting Equation (H-23) into Equation (H-12)

$$\begin{aligned} \hat{w}^2(\hat{Z}_o) &= tr(\grave{R} - 2\grave{Z}_o^T\hat{Z}_o + \hat{Z}_o^T\hat{Z}_o)/n \\ &= tr(\grave{R} - \grave{V}^T\grave{S}_r^2\grave{V}_r^T)/n \\ &= (tr\grave{R} - tr\grave{S}_r^2)/n \end{aligned} \qquad \text{(H-24)}$$

Problem C

Given a set of independent observations Z_i, i=1,...,d, where each Z_i is of the form

$$\underbrace{\begin{bmatrix} Z_{1i} \\ \cdots \\ Z_{2i} \end{bmatrix}}_{Z_i} = \underbrace{\begin{bmatrix} USV^T \\ \cdots\cdots \\ Upsq^TV^T \end{bmatrix}}_{Z_o} + \underbrace{\begin{bmatrix} E_{1i} \\ \cdots \\ E_{2i} \end{bmatrix}}_{E_i} \qquad \text{(H-25)}$$

where

$$E(Z_i)=Z_o$$

$$E(Z_i-Z_o)^T(Z_i-Z_o)=w^2I$$

$$U^TU=I,\ V^TV=I,\ p^Tp=1,\ q^Tq=1$$

S diagonal, positive definite; $s>0$

Find ML estimates of Z_o (i.e., U, V, S, p, q, and s) and w^2.

Solution:

The ML estimates of Z_o and w^2 are given by maximizing the log-likelihood function (Equation H-6) subject to constraints (structural) on Z_o (Equation H-25).

The ML estimate of w^2 for this problem is also given by Equation (H-12) of Problem B

$$\hat{w}^2(Z_o) = \mathrm{tr}\left(\frac{\grave{R}-2\grave{Z}_o^TZ_o+Z_o^TZ_o}{n}\right) \qquad \text{(H-26)}$$

where

$$\grave{Z}_o = \sum_{i=1}^{d} Z_i/d = \frac{1}{d}\sum_{i=1}^{d}\begin{bmatrix} Z_{1i} \\ \cdots \\ Z_{2i}\end{bmatrix} = \begin{bmatrix}\grave{Z}_1 \\ \vdots \\ \grave{Z}_2\end{bmatrix} \qquad \text{(H-27)}$$

Substituting Equation (H-26) into the log-likelihood function (Equation H-6) and simplifying (as in Problem B), the function to be maximized with respect to Z_o is

$$L(Z_o) = \mathrm{tr}(2\grave{Z}^T_oZ_o - Z_o^TZ_o) \qquad \text{(H-28)}$$

Incorporating the constraints on Z_o into Equation (H-29) yields

$$L(Z_o) = \tilde{L}(U,V,S,p,q,s)$$

$$= \mathrm{tr}\left(2\grave{Z}_o^T\begin{bmatrix} USV^T \\ \cdots\cdots \\ Upsq^TV^T\end{bmatrix} - [USV^T, Vqsp^TU^T]\begin{bmatrix} USV^T \\ \cdots\cdots \\ Upsq^TV^T\end{bmatrix}\right)$$

$$= \mathrm{tr}(\grave{Z}_1^T U S V^T + \grave{Z}_2^T U p s q^T V^T - V S^2 V^T - s^2 q q^T)$$

$$= 2\mathrm{tr}(\grave{Z}_1^T U S V^T) + 2s\mathrm{tr}(\grave{Z}_2^T U p q^T V^T) - \mathrm{tr}S^2 - s^2 \qquad \text{(H-29)}$$

ML estimate of s

Maximizing $\grave{L}$ with respect to s in Equation (29), we have

$$\frac{\partial \grave{L}}{\partial s} = 2\mathrm{tr}(\grave{Z}_2^T U p q^T V^T) - 2s = 0 \qquad \text{(H-30a)}$$

yielding the ML estimate of $\hat{s}$

$$\hat{s}(U, V, p, q) = \mathrm{tr}(\grave{Z}_2^T U p q^T V^T) \qquad \text{(H-30b)}$$

Substituting $\hat{s}$ into Equation (H-29)

$$\grave{L}(U, V, S, p, q, \hat{s}) = 2\mathrm{tr}(\grave{Z}_1^T U S V^T) + (\mathrm{tr}(\grave{Z}^T U p q^T V^T))^2 - \mathrm{tr}S^2 \qquad \text{(H-31)}$$

ML estimates of p and q

Let $\hat{U}$ and and $\hat{V}$ be the U and V that maximize $\tilde{L}$ (Equation 31). Then, clearly, p and q that maximize

$$(\mathrm{tr}(\grave{Z}_2^T \hat{U} p q^T \hat{V}^T))^2 \qquad \text{(H-32)}$$

also maximize $\tilde{L}$.

From Lemma H2 and the identity

$$\mathrm{tr}(\grave{Z}_2^T \hat{U} p q^T \hat{V}^T) = \mathrm{tr}(q^T (\hat{V}^T \grave{Z}_2^T \hat{U}) p) \qquad \text{(H-33)}$$

$p(\hat{U}, \hat{V})$ and $q(\hat{U}, \hat{V})$ that maximize Equation (H-32) are given by

$$\hat{p}(\hat{U}, \hat{V}) = \text{First left singular vector of } \hat{U}^T \grave{Z}_2 \hat{V} \qquad \text{(H-34a)}$$

$$\hat{q}(\hat{U},\hat{V}) = \text{First right singular vector of } \hat{U}^T\grave{Z}_2\hat{V} \qquad \text{(H-34b)}$$

Substituting ML estimates of p and q into Equation (H-31)

$$\grave{L}(U,V,S,\hat{p},\hat{q},\hat{s}) = 2tr(\grave{Z}_1^T USV^T) + s_*^2 - trS^2 \qquad \text{(H-35)}$$

where

s_* = The largest singular value of $(V^T\grave{Z}_2^T U)$

ML estimate of S

We have from Equation (H-35)

$$\begin{aligned}\grave{L}(U,V,S,\hat{p},\hat{q},\hat{s}) &= 2tr(\grave{Z}_1^T USV^T) + s_*^2 - trS^2 \\ &= 2tr(\grave{Z}_1^T(s_1u_1v_1^T+\ldots+s_ru_rv_r^T)) + s_*^2 \\ &\quad - (s_1^2+\ldots+s_r^2)\end{aligned} \qquad \text{(H-36)}$$

The necessary conditions for s_i, i=1,...,r to be maximizing are

$$\frac{\partial \grave{L}}{\partial s_i} = 2v_i^T\grave{Z}_1^Tu_i - 2s_i = 0, \quad i=1,\ldots,r \qquad \text{(H-37)}$$

and the ML estimates of S is

$$\begin{aligned}\hat{S} &= \text{diag}(\hat{s}_1,\ldots,\hat{s}_r) \\ \hat{s}_i &= v_i^T\grave{Z}_1^Tu_i\end{aligned} \qquad \text{(H-38)}$$

Substituting the ML estimate of S into Equation (H-36)

$$\begin{aligned}\grave{L}(U,V,\hat{S},\hat{p},\hat{q},\hat{s}) &= 2tr(\grave{Z}_1^T(\hat{s}_1u_1v_1^T+\ldots+\hat{s}_ru_rv_r^T)) + s_*^2 - (\hat{s}_1^2+\ldots+\hat{s}_r^2) \\ &= 2\hat{s}_1v_1^T\grave{Z}^T{}_1u_1 + \ldots + 2\hat{s}_rv_r^T\grave{Z}^T{}_ru_r + s_*^2 - (\hat{s}_1^2+\ldots+\hat{s}_r^2) \\ &= (v_1^T\grave{Z}^T{}_1u_1)^2 + \ldots + (v_r^T\grave{Z}^T{}_ru_r)^2 + s_*^2\end{aligned} \qquad \text{(H-39)}$$

ML estimates of U and V

We have, so far, maximized the orignal likelihood function (Equation H-29) with respect to the variables S, p, q, and s, and we need now only to maximize it with respect to U and V. When the maximizing variables $\hat{S}$, $\hat{p}$, $\hat{q}$, and $\hat{s}$ are substituted into Equation (H-29), The resulting likelihood function is

$$\tilde{L}(U,V) = (v_1^T \grave{Z}^T{}_1 u_1)^2 + \ldots + (v_r^T \grave{Z}^T{}_r u_r)^2 + s_*^2$$
$$s_* = \text{The largest singular value of } (V^T \grave{Z}_2^T U) \qquad \text{(H-40)}$$

to be maximized with respect to U and V.

We give below "closed" form solutions for the maximizing U and V when USV^T is a full ranked square matrix. That is, U and V are orthogonal matrices. In this case,

$$\begin{aligned} s_* &= \text{The largest singular value of } (V^T \grave{Z}_2^T U) \\ &= \text{The largest singular value of } \grave{Z}_2 \end{aligned} \qquad \text{(H-41)}$$

and U and V that maximizes

$$(v_1^T \grave{Z}^T{}_1 u_1)^2 + \ldots + (v_r^T \grave{Z}^T{}_r u_r)^2 \qquad \text{(H-42)}$$

also maximizes $\tilde{L}$ given by Equation (H-40).

The U and V that maximize Equation (H-42) are given by an application of Lemma H2, and the fact that the colums of U,V are orthonormal

$$\begin{aligned} \hat{U} &= \text{First r left singular vectors of } \grave{Z}_1 \\ \hat{V} &= \text{First r right singular vectors of } \grave{Z}_1 \end{aligned} \qquad \text{(H-43)}$$

A quick sketch of a proof of this result is as follows:

Cleary, from Lemma H2, the maximizing u_1 and v_1 are the first left and right singular vectors of $\grave{Z}_1$. For u_2 and v_2, we have the constraints that they be orthogonal to u_1 and v_1, respectively; the maximizing u_2 and v_2 are then simply the second left and right singular vectors of $\grave{Z}_1$. Rest of the proof follows similarly: that each of the subsquent sets of vectors be orthogonal to spaces spanned by previous sets of vectors, etc.

An alternative way of deriving this result is to apply to Equation (H-35) the fact that U and V are orthogonal matrices. We then need to maximize

$$2\mathrm{tr}(\grave{Z}_1 USV^T) - \mathrm{tr}S^2 \qquad \text{(H-44)}$$

with respect to U, V, and S. This, we have done in Problem B (see Equations H-15, H-18, and H-22).

Summarizing, if U and V are orthogonal matrices, then:

$\hat{U}$ = First r left singular vectors of $\grave{Z}_1$

$\hat{V}$ = First r right singular vectors of $\grave{Z}_1$

$\hat{S}$ = Diag(first r singular values of $\grave{Z}_1$)

$\hat{p}$ = First left singular vector of $\hat{U}^T\grave{Z}_2\hat{V}$

$\hat{q}$ = First right singular vector of $\hat{U}^T\grave{Z}_2\hat{V}$

$\hat{s}$ = Largest singular value of $\grave{Z}_2$

The concentration of the target component is then given by solving for the value of c for which the expression in Equation (H-3) drops rank. That is, the value of c for which

$$|\hat{S} - c\hat{p}\hat{s}\hat{q}^T| = 0 \qquad \text{(H-45)}$$

Using identities for the determinants

$$|\hat{S}-c\hat{p}\hat{s}\hat{q}^T| = |\hat{S}(I-c\hat{S}^{-1}\hat{p}\hat{s}\hat{q}^T)|$$
$$= |\hat{S}|.|I-c\hat{S}^{-1}\hat{p}\hat{s}\hat{q}^T|$$
$$= |\hat{S}|.|1-c\hat{s}\hat{q}^T\hat{S}^{-1}\hat{p}|$$

c is given by

$$c = \frac{1}{\hat{s}\hat{q}^T\hat{S}^{-1}\hat{p}} \qquad \text{(H-46)}$$

It can be shown that this solution is identical to the projection method with the linear least squares projection.

An Extension to Multiple Target Components Case

In this case, the structure of the data matrix is

$$\underbrace{\begin{bmatrix} Z_{1i} \\ \cdots \\ Z_{2i} \end{bmatrix}}_{Z_i} = \underbrace{\begin{bmatrix} US_1V^T \\ \cdots\cdots\cdots \\ UPS_2Q^TV^T \end{bmatrix}}_{Z_o} + \underbrace{\begin{bmatrix} E_{1i} \\ \cdots \\ E_{2i} \end{bmatrix}}_{E_i} \qquad \text{(H-47)}$$

where,

$$S_1 = \text{Diag}(s_{11},\ldots,s_{1r}),\ s_{1i} > 0$$
$$S_2 = \text{Diag}(s_{21},\ldots,s_{2t}),\ s_{2i} > 0,\ t < r$$
$$P^TP = Q^TQ = I_{txt}$$

and other variables are as in Equation (H-25).

The solution procedure parallels that of single target component, and requires maximizing the function

$$\mathcal{L} = \text{tr}(2(\dot{Z}_1^TUS_1V^T+\dot{Z}_2^TUPS_2Q^TV^T) - VS_1^2V^T - QS_2^2Q^T) \qquad \text{(H-48)}$$

with respect to U, V, P, Q, S_1, and S_2.

It can be easily shown that the maximizing varibles are

$$\begin{aligned}
&\hat{s}_{1i} = \hat{v}_i^T \grave{Z}_1 \hat{u}_i, \quad i=1,\ldots,r \\
&\hat{P} = \text{First t left singular vectors of } \hat{U}^T \grave{Z}_2 \hat{V} \\
&\hat{Q} = \text{First t right singular vectors of } \hat{U}^T \grave{Z}_2 \hat{V} \\
&\hat{S}_2 = \text{Diag(first t singular values of } \hat{U}^T \grave{Z}_2 \hat{V})
\end{aligned} \tag{H-49}$$

where $\hat{U}$ and $\hat{V}$ are U and V that maximize

$$\tilde{L} = \sum_{i=1}^{r} (v_i^T \grave{Z}_1^T u_i)^2 + \sum_{i=1}^{t} s_{*i} \tag{H-50}$$

$$s_{*i} = \text{i-th singular value of } V^T \grave{Z}_2^T U$$

Lemma H1 [R1]

Let $A, B \in R^{mxn}$ have sigular value decompositions

$$A = U_a S_a V_a^T$$

$$B = U_b S_b V_b^T$$

Then,

$$|tr(APB^TQ)| \leq tr(S_a S_b)$$

with equality when $P=V_a V_b^T$ and $Q=U_b U_a^T$. (Where P and Q are orthogonal matrices, $|.|$ indicates abosolute value operator, and "tr" indicates the trace operator.)

[]

Lemma H2

Let $A \in R^{mxn}$ have SVD

$$A = USV^T$$

Then,

$$|q^T A p| \leq s_1$$

with equality when $q=u_1$ and $p=v_1$, for all q and p such that $q^T q = p^T p = 1$.

Proof:

We have,

$$q^T A p = q^T U S V^T \leq s_1 q^T U V^T p \leq s_1$$

1st inequality by using $s_1 \geq \ldots \geq s_m$

2nd inequality by using $q^T U V^T p \leq 1$

Clearly, equality is attained when $q=u_1$ and $p=v_1$. []

APPENDIX I. AN ERROR ANALYSIS OF VAN DOOREN'S ALGORITHM

Many basic problems in linear systems theory can be viewed as different manifestations of the generalized eigenstructure problem [V3]. The governing operator for this class of problems is a matrix pencil of the form

$$M - tN \qquad \text{(I-1)}$$

of some dimensions mxn. When $m \neq n$, or $m=n$ but the determinant of the pencil is zero, the pencil is said to be singular; otherwise it is said to be regular. The algebraic theory of the generalized eigenstructure problem was first addressed by Kronecker [K4], in 1890, with his work on strict equivalence of singular pencils and their canonical representations, and formed the basis for subsequent investigations on the subject [G1,T1,L2]. More recent papers [V2,V4,S8] on the subject have emphasized the numerical aspects of the problem: mainly that of formulating mumerical algorithms for the canonical decomposition and the computation of the eigenvalues of the singular pencils.

An important issue in interpreting the numerical results of these algorithms is the sensitivities of the computed canonical structure and eigenvalues to perturbations in the associated matrix pencil. The sensitivity analysis of the eigenvalues of the singular pencils have just begun to surface in the literature [V3,V4] and have not yet been fully understood. The purpose of this appendix is to consolidate existing results on the subject and to provide a cohesive and unified exposition on the issues involved in the sensitivity analysis of singular pencils.

In addition, an in-depth error analysis of Van Dooren's algorithm

as applied to our problem is presented. From this analysis, a set of guidelines for setting thresholds (used in rank decisions required by the algorithm when applied to noisy data matrices) is given that guarantees that the Kronecker structure of the underlying data matrix is recovered.

The issues to be addressed are motivated, defined, and formulated in subsection (a); sensitivity results for singular pencils, and results on error propagation in Van Dooren's algorithm are presented in subsection (b).

(a) Motivation, definition, and formulation

Given the singular pencil

$$\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right] - t\left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 0 \end{array}\right] \qquad \text{(I-2)}$$

whose finite eigenvalue is seen to occur at t=1, consider a pencil that is a small (<<1) perturbation of the pencil in (I-2),

$$\left[\begin{array}{c|c} 1+d_{11} & d_{12} \\ \hline d_{21} & d_{22} \end{array}\right] - t\left[\begin{array}{c|c} 1+e_{11} & e_{12} \\ \hline e_{21} & e_{22} \end{array}\right] \qquad \text{(I-3)}$$

with $|d_{ij}|, |e_{ij}| << 1$.

Case (i) $d_{12}=d_{21}=e_{12}=e_{21}=0$.

For any arbitrarily small values of d_{22} and e_{22}, this pencil is non-singular, with eigenvalues at $(1+d_{11})/(1+e_{11})$ and d_{22}/e_{22}. Note that the second eigenvalue can be made arbitrarily large with small variations in d_{22} and e_{22}.

Case (ii) $d_{12}=d_{22}=e_{12}=e_{22}=0$, $d_{21}\neq e_{21}$.

There is no value of t for which this pencil losses rank. That is, this pencil has no regular pencil structure, and correspondingly has no finite eigenvalue.

Case (iii) $d_{12}=d_{21}=d_{22}=e_{12}=e_{21}=e_{22}=0$.

This pencil has a single eigenvalue at $(1+d_{11})/(1+e_{11})$, and small perturbations in d_{11} and e_{11} result in small changes in the eigenvalue.

Cases (i) and (ii) are examples where the Kronecker structure, the left and right null structures in particular, of the pencil changes with arbitrarily small perturbations in its elements. This results in a change in the dimension of the regular pencil structure, and consequently, addition or subtraction of eigenvalues to and from the spectrum of the unperturbed pencil. In contrast, Case (iii) illustrates a "well-behaved" perturbation, in which the spectrum of the pencil changes continuously with continuous changes in the elements of the pencil.

With regular pencils, for sufficiently small perturbations, there is, except for the cases of repeated eigenvalues, a one-to-one correspondence between eigenvalues of the original and the perturbed systems, and the sensitivity issue revolves around addressing the following questions:

Given the eigenvalues of the unperturbed system,

(i) what are the first order approximations of the eigenvalues of the perturbed systems [P1,P3]?, i.e.,

$$t_i(M+E_m) - t_i(M) \gtrless ? \qquad \text{(I-4a)}$$

$$t_i(M+E_m, N+E_n) - t_i(M, N) \gtrless ? \qquad \text{(I-4b)}$$

or (ii) what are the bounds on the eigenvalues of the perturbed system [S8]?, i.e.,

$$|t_i(M+E_m) - t_i(M)| \leq \ ? \tag{I-5a}$$

$$|t_i(M+E_m, N+E_n) - t_i(M,N)| \leq \ ? \tag{I-5b}$$

Similarly, in addressing sensitivity of the eigenvalues of singular pencils to perturbations, one could extend ideas developed for matrices and regular pencils, and look for first order approximations or perturbation bounds on the perturbed eigenvalues. As was shown in cases (i) and (ii) above, however, some perturbations on singular pencils (independent of the magnitudes of the perturbations) do not preserve the one-to-one correspondence between the unperturbed and the perturbed eigenvalues. That is, there may be no eigenvalues in the perturbed pencil with which to associate the unperturbed eigenvalues in computing bounds for, or approximating, the eigenvalues of the perturbed pencil. Hence, for singular pencils, the sensitivity analysis of the form discussed above are retricted to perturbations that preserve this one-to-one correspondence.

Alternatively, with unrestricted perturbations, one can address the sensitivity problem as follows. Given a pencil $\tilde{M}-t\tilde{N}$ that is known to be composed of some underlying nominal pencil $M-tN$ and a perturbation pencil E_m-tE_n, i.e.

$$\begin{aligned} \tilde{M}-t\tilde{N} &= (M-tN) + (E_m-tE_n) \\ &= (M+E_m) - t(N+E_n) \end{aligned} \tag{I-6}$$

where $||M|| >> ||E_m||$ and $||N|| >> ||E_n||$,

(i) are there ways (algorithms) to recover some nearby matrix

pencil $\hat{M}-t\hat{N}$ with the same kronecker structure as the nominal pencil?

and (ii) how are the eigenvalues of the nominal pencil related to this nearby pencil?

These questions and other related issues are addressed in the subsequent subsections.

(b) Restricted perturbations: Sensitivity results for regular pencils.

We have seen in the previous section that certain perturbations in the elements of a singular pencil lead to ill-posed eigenvalue sensitivity problems in the sense that arbitrarily small perturbations in the elements of a singular pencil yield discontinuous changes in its eigenvalues. In [V3,V4,V5], Van Dooren introduces the concept of restricted perturbations, and addresses eigenvalue sensitivity issues within that framework. Essentially, the idea is to restrict perturbations to cases where the eigenvalues of the singular pencil vary continuously with small variations (in a restricted way) in the elements of the pencil. In [V4,V5], this restriction is shown to be equivalent to allowing only those perturbations that preserve the dimension and order (see [F1] for definitions) of the left and right null spaces, that is, perturbations that preserve the Kronecker structure of the singular pencil. Imposing such restrictions is justifiable only in special cases ([V3], for example); nevertheless, with appropriate use of Van Dooren's algorithm (see [E2],subsection (c)), arbitrary perturbations can be made to be conformable to restricted perturbations on the "processed" singular pencil.

Once the perturbations are restricted as above, the eigenvalue sensitivity problem for singular pencils reduces to that of regular pencils, which has been addressed in the literature [S9,P1]: Stewart [S9] addresses the problem of obtaining bounds on the eigenvalues, Paraskevorous et al. [P1] give a first order sensitivity formula for the eigenvalues of regular pencils. In Appendix E, we give an alternative first order sensitivity formula for regular pencils that is more concise and direct than [P1].

We give below examples of perturbations that preserve Kronecker structure:

Example I-1.

Given a decomposition of an arbitrary pencil M-tN by orthogonal transformations P and Q:

$$P^T(M-tN)Q = \left[\begin{array}{c|c|c} M_r-tN_r & * & * \\ \hline 0 & M_{reg}-tN_{reg} & * \\ \hline 0 & 0 & M_l-tN_l \end{array}\right] \qquad (\#)$$

where

(i) M_r-tN_r contains all the right null structure,

(ii) $M_{reg}-tN_{reg}$ contains the regular pencil structure,

and (iii) M_l-tN_l contains all the left null structure,

consider a perturbed pencil of the form

$$\bar{M}-t\bar{N} = [P_1,P_2,P_3]\left[\begin{array}{c|c|c} M_r-tN_r & * & * \\ \hline 0 & \bar{M}_{reg}-t\bar{N}_{reg} & * \\ \hline 0 & 0 & M_l-tN_l \end{array}\right]\left[\begin{array}{c} Q_1^T \\ \hline Q_2^T \\ \hline Q_3^T \end{array}\right]$$

where

$$\tilde{M}-t\tilde{N} = (M+E_m)-t(N+E_n),$$

$$\tilde{M}_{reg}-t\tilde{N}_{reg} = (M_{reg}+E_{mreg})-t(N_{reg}+E_{nreg}),$$

and $||E_{mreg}||_2$ and $||E_{nreg}||_2$ are sufficiently small so that $\tilde{M}_{reg}-t\tilde{N}_{reg}$ remains a regular pencil (clearly, the Kronecker structure is unaltered). Then, we have,

$$E_m-tE_n = [P_1,P_2,P_3]\left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & E_{mreg}-tE_{nreg} & 0 \\ \hline 0 & 0 & 0 \end{array}\right]\left[\begin{array}{c} Q_1^T \\ \hline Q_2^T \\ \hline Q_3^T \end{array}\right]$$

$$= P_2[E_{mreg}-tE_{nreg}]Q_2^T$$

or

$$\left[\begin{array}{c} P_1^T \\ \hline P_2^T \\ \hline P_3^T \end{array}\right][E_m-tE_n][Q_1,Q_2,Q_3] = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & E_{mreg}-tE_{nreg} & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$$

That is, in order to preserve the Kronecker structure of the singular pencil, it is sufficient that E_m and E_n satisfy

$$\text{span}(E_m) = \text{span}(P_2)$$

$$\text{span}(E_m^T) = \text{span}(Q_2)$$

$$\text{span}(E_n) = \text{span}(P_2)$$

$$\text{span}(E_n^T) = \text{span}(Q_2)$$

[]

Example I-2.

A less restrictive condition than that given in Example I-1 can

be obtained by noting that we can allow perturbations in blocks marked "*" in Equation (#) in Example I-1 without altering the Kronecker structure of the pencil. Thus, we now have:

$$\tilde{M}-t\tilde{N} = (M-tN) + (E_m-tE_n)$$

$$= [P_1,P_2,P_3]\left[\begin{array}{c|c|c} M_r-tN_r & * & * \\ \hline 0 & M_{reg}-tN_{reg} & * \\ \hline 0 & 0 & M_l-tN_l \end{array}\right]\left[\begin{array}{c} Q_1^T \\ Q_2^T \\ Q_3^T \end{array}\right]$$

$$+$$

$$[P_1,P_2,P_3]\left[\begin{array}{c|c|c} 0 & E_{m12}-tE_{n12} & E_{m13}-tE_{n13} \\ \hline 0 & E_{mreg}-tE_{nreg} & E_{m23}-tE_{n23} \\ \hline 0 & 0 & 0 \end{array}\right]\left[\begin{array}{c} Q_1^T \\ Q_2^T \\ Q_3^T \end{array}\right]$$

and,

$$\left[\begin{array}{c} P_1^T \\ P_2^T \\ P_3^T \end{array}\right][E_m-tE_n][Q_1,Q_2,Q_3] = \left[\begin{array}{c|c|c} 0 & E_{m12}-tE_{n12} & E_{m13}-tE_{n13} \\ \hline 0 & E_{mreg}-tE_{nreg} & E_{m23}-tE_{n23} \\ \hline 0 & 0 & 0 \end{array}\right]$$

Thus, in order to preserve the Kronecker structure of the singular pencil, it is sufficient that E_m and E_n satisfy

$$\text{span}(E_m) \subset \text{span}([P_2,P_3])$$

$$\text{span}(E_m^T) \subset \text{span}([Q_1,Q_2])$$

$$\text{span}(E_n) \subset \text{span}([P_2,P_3])$$

$$\text{span}(E_n^T) \subset \text{span}([Q_1,Q_2])$$

[]

Note: Orthogonal transformations P and Q that yield decomposition in Equation (#) in Example I-1 are given by [V4,V5]:

$$P = [P_1, P_2, P_3] \quad ; \quad Q = [Q_1, Q_2, Q_3]$$

$$\text{span}(Q_1) = X_{min}$$
$$\text{span}(Q_1, Q_2) = X_{max}$$
$$\text{span}(P_1) = Y_{min}$$
$$\text{span}(P_1, P_2) = Y_{max}$$

where X_{min}, Y_{min} and X_{max}, Y_{max} are minimal and maximal pairs of right reducing subspaces in the sense that for any pair X,Y of right reducing subspaces, the following holds

$$\{0\} \subsetneq X_{min} \subset X \subset X_{max} \subset R^n$$
$$\{0\} \subset Y_{min} \subset Y \subset Y_{max} \subset R^m$$

[]

(c) Arbitrary perturbations: Derivation and application of error bounds for Van Dooren's algorithm.

It is shown in [V4] that for noisy data cases, application of Van Dooren's algorithm with numerical rank yields a regular pencil that is guaranteed to be that of some "neighboring" singular pencil that is EPS "away" from the noisy one, where EPS is the threshold used in the rank decisions. However, the algorithm does not guarantee that the computed Kronecker structure is that of the underlying noise free pencil. (Although Van Dooren's algorithm does appear to preserve Kronecker structure for some singular pencils, to the best of our knowledge, the conditions under which this holds true have not been explicitly stated

or proven in the literature.) As was stated in the previous section, discussions on eigenvalue sensitivity problems are meaningful only when the Kronecker structure of the pencil is preserved, either through restricted perturbation or via judicious application of Van Dooren's algorithm.

In this section, we give a modified version (Algorithm I-1) of Van Dooren's algorithm that guarantees that the Kronecker structure of the underlying noise free pencil is preserved (see Result I-1). We shall also show that by using error bounds generated from Algorithm I-1 to set rank decisions on Algorithm 3 (Chapter 3), Algorithm 3 can be made to preserve the Kronecker structure of the underlying noise free pencil (see Result I-2).

Algorithm I-1 is essentially Van Dooren's algorithm with rank decisions based on the norm of the error matrices (deviations from noise free case) that propagate along the algorithm. By providing bounds on the error matrices at each step of the algorithm (see Result I-3), correct rank decisions are made - correct in the sense that the Kronecker structure of the underlying noise free pencil is preserved.

An important feature of the algorithm is that at each step of the algorithm, the deflated pencil is subject to an additional transformation that "rotates" the pencil, without altering its pencil structure, so that the norms of the propagated error matrices are minimized, resulting in a tighter error bound. Unfortunately, however, this rotation requires the knowledge of the underlying noise free pencil (which is unknown to us !), and hence the algorithm can not be implemented in practice.

Nevertheless, understanding of error propagation in Algorithm I-1

provides valuable insight into error propagation in Van Dooren's algorithm , and since, as shown in Result I-3, the computation of error bounds requires only the knowledge of the norms of the initial error matrices, Algorithm I-1 becomes an important tool in implementing Van Dooren's algorithm to noisy pencils: e.g. error bounds in Algorithm I-1 allow us to make "correct" rank decisions in Van Dooren's algorithm as applied to our problem (see Result I-2).

The noisy data matrices are assumed to be of the form

$$\tilde{M} = M + E_m = c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T + E_m \qquad \text{(I-7a)}$$

$$\tilde{N} = N + E_n = x_j y_j^T + E_n, \quad 1 \leq j \leq r \qquad \text{(I-7b)}$$

where the structure of M and N are as stated in Chapter 2. Furthermore, we shall assume that $s_r(M) < ||E_m||_2$ and $s_1(N) < ||E_n||_2$, where $s_r(M)$ and $s_1(N)$ are the smallest nonzero singular values of the underlying noise free data matrices M and N.

The algorithm is summarized in Figure I-1, and detailed in Algorithm I-1:

Algorithm I-1. Van Dooren's algorithm with minimum error transformation. (Note: vectors and matrices without "~" correspond to noise free case, Algorithm 2, Chapter 3)

(Note: Symbols for the SVD are repeated in each of the steps.)

Phase I. Column compression.

<Step 1.> Compress $\tilde{N}$ $\left(=[\tilde{u}_1, \tilde{U}_2]\left[\begin{array}{c|c} \tilde{s}_1 & 0 \\ \hline 0 & \tilde{S}_2 \end{array}\right]\begin{bmatrix} \tilde{v}_1^T \\ \tilde{V}_2^T \end{bmatrix}\right)$, $(\tilde{S}_2)_i \leq ||E_n||_2 = EPSN$

to full numerical column rank:

$$\tilde{N}\tilde{V} = [\tilde{N}\tilde{v}_1, 0] = [\tilde{n}_1, 0]$$

Update $\tilde{M}$:

$$\tilde{M}\tilde{V} = [\tilde{M}\tilde{v}_1, \tilde{M}\tilde{V}_2] = [\tilde{m}_1, \tilde{M}_2]$$

Minimum error transformation:

$$([\tilde{m}_1, \tilde{M}_2] - t[\tilde{n}_1, 0]) \left[\begin{array}{c|c} q_1 & 0 \\ \hline 0 & Q_2 \end{array}\right] = [\tilde{m}_1 q_1, \tilde{M}_2 Q_2] - t[\tilde{n}_1 q_1, 0]$$

$$q_1: \quad \min_{q_1} \|\tilde{V}_1 q_1 - v_1\|_2$$

$$Q_2: \quad \text{(i)} \quad \min_{Q_2} \|\tilde{V}_2 Q_2 - V_2\|_2$$

$$\text{(ii)} \quad \min_{Q_2^T Q_2 = I} \|\tilde{V}_2 Q_2 - V_2\|_2$$

Define:

$$e_{n1} = \tilde{n}_1 q_1 - n_1$$

$$e_{m1} = \tilde{m}_1 q_1 - m_1$$

$$E_{m2} = \tilde{M}_2 Q_2 - M_2$$

<Step 2.> Compress $\tilde{M}_2 Q_2$ $\left(=[\tilde{U}_1, \tilde{U}_2] \left[\begin{array}{c|c} \tilde{S}_1 & 0 \\ \hline 0 & \tilde{S}_2 \end{array}\right] \left[\begin{array}{c} \tilde{V}_1^T \\ \hline \tilde{V}_2^T \end{array}\right]\right.$,

$(\tilde{S}_2)_i \leq \|E_{m2}\|_2 = \text{EPSM2})$ to full <u>numerical</u> row rank:

$$\tilde{U}^T[\tilde{m}_1 q_1, \tilde{M}_2 Q_2] = \left[\begin{array}{c|c} \tilde{U}_1^T \tilde{m}_1 q_1 & \tilde{U}_1^T \tilde{M}_2 Q_2 \\ \hline \tilde{U}_2^T \tilde{m}_1 q_1 & 0 \end{array}\right] = \left[\begin{array}{c|c} \tilde{m}_{11} & \tilde{M}_{12} \\ \hline \tilde{m}_{21} & 0 \end{array}\right]$$

Update $\tilde{n}_1$:

$$\tilde{U}^T[\tilde{n}_1 q_1, 0] = \left[\begin{array}{c|c} \tilde{U}_1^T \tilde{n}_1 q_1 & 0 \\ \hline \tilde{U}_2^T \tilde{n}_1 q_1 & 0 \end{array}\right] = \left[\begin{array}{c|c} \tilde{n}_{11} & 0 \\ \hline \tilde{n}_{21} & 0 \end{array}\right]$$

Minimum error transformation:

$$\left[\begin{array}{c|c} I & 0 \\ \hline 0 & R^T \end{array}\right]\left[\begin{array}{c|c} \tilde{m}_{11} & \tilde{M}_{12} \\ \hline \tilde{m}_{21} & 0 \end{array}\right] - t\left[\begin{array}{c|c} \tilde{n}_{11} & 0 \\ \hline \tilde{n}_{21} & 0 \end{array}\right]$$

$$= \left[\begin{array}{c|c} \tilde{m}_{11} & \tilde{M}_{12} \\ \hline R^T\tilde{m}_{21} & 0 \end{array}\right] - t\left[\begin{array}{c|c} \tilde{n}_{11} & 0 \\ \hline R^T\tilde{n}_{21} & 0 \end{array}\right]$$

$$R: \quad \min_R \|R^T\tilde{U}_2^T - U_2^T\|_2$$

Define:

$$e_{m21} = R^T\tilde{m}_{21} - m_{21}$$

$$e_{n21} = R^T\tilde{n}_{21} - n_{21}$$

<Step 3.> Permute upper and lower blocks, and reinitialize:

$$P\left[\begin{array}{c|c} \tilde{m}_{11} & \tilde{M}_{12} \\ \hline R^T\tilde{m}_{21} & 0 \end{array}\right] = \left[\begin{array}{c|c} R^T\tilde{m}_{21} & 0 \\ \hline \tilde{m}_{11} & \tilde{M}_{12} \end{array}\right] = \left[\begin{array}{c|c} \tilde{m} & 0 \\ \hline * & * \end{array}\right]$$

$$P\left[\begin{array}{c|c} \tilde{n}_{11} & 0 \\ \hline R^T\tilde{n}_{21} & 0 \end{array}\right] = \left[\begin{array}{c|c} R^T\tilde{n}_{21} & 0 \\ \hline \tilde{n}_{11} & 0 \end{array}\right] = \left[\begin{array}{c|c} \tilde{n} & 0 \\ \hline * & 0 \end{array}\right]$$

Phase II. Row compression.

<step 1.> Compress $\tilde{n}$ $(= [\tilde{u}_1, \tilde{U}_2]\begin{bmatrix} \tilde{s} \\ 0 \end{bmatrix}[1])$ to full row rank:

$$\tilde{U}^T\tilde{n} = \begin{bmatrix} \tilde{u}_1^T\tilde{n} \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{n}_1 \\ 0 \end{bmatrix}$$

Update $\tilde{m}$:

$$\tilde{U}^T\tilde{m} = \begin{bmatrix} \tilde{u}_1^T\tilde{m} \\ \tilde{U}_2^T\tilde{m} \end{bmatrix} = \begin{bmatrix} \tilde{m}_1 \\ \tilde{m}_2 \end{bmatrix}$$

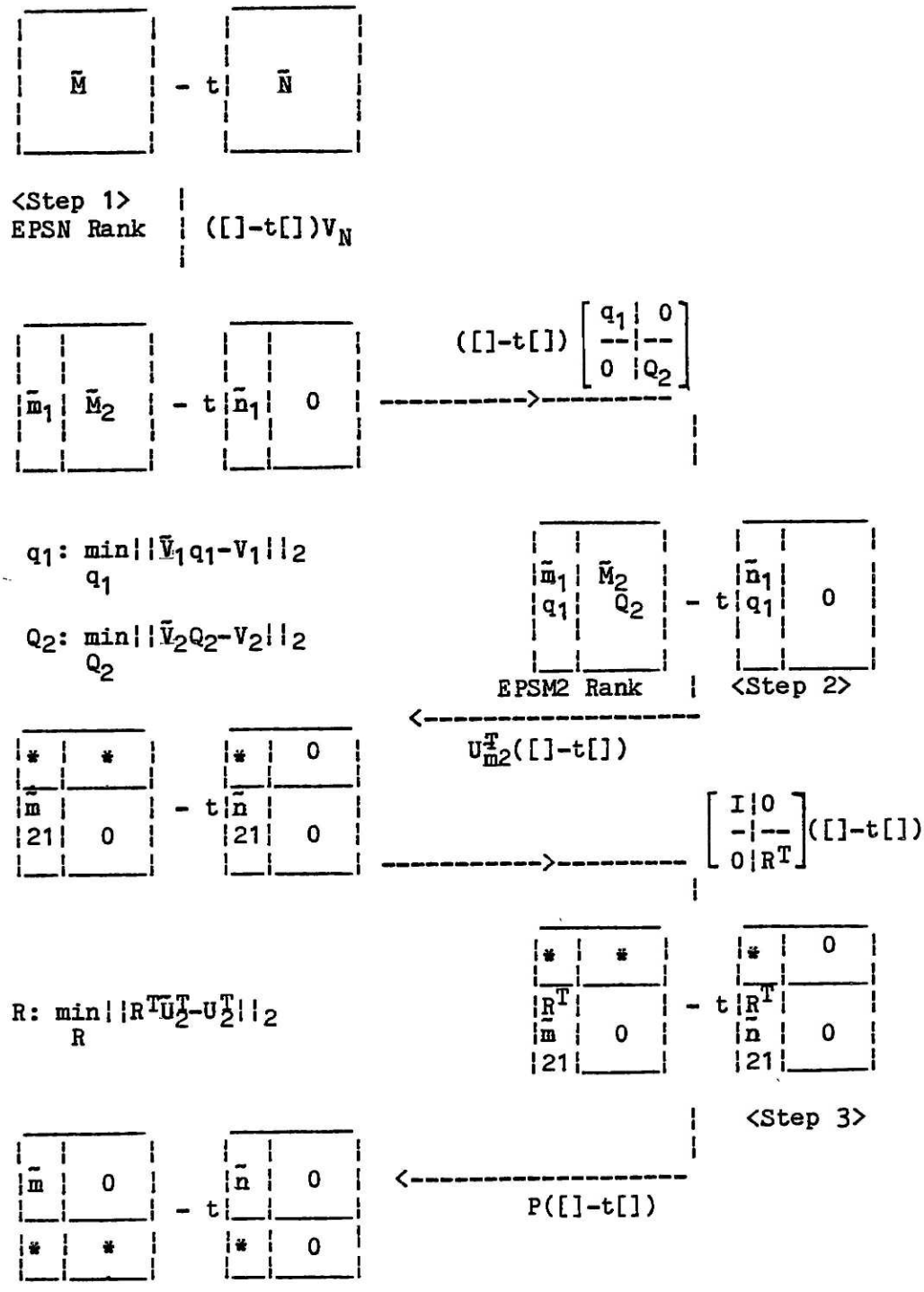


Figure I-1a. Modified Van Dooren's algorithm: Phase I.

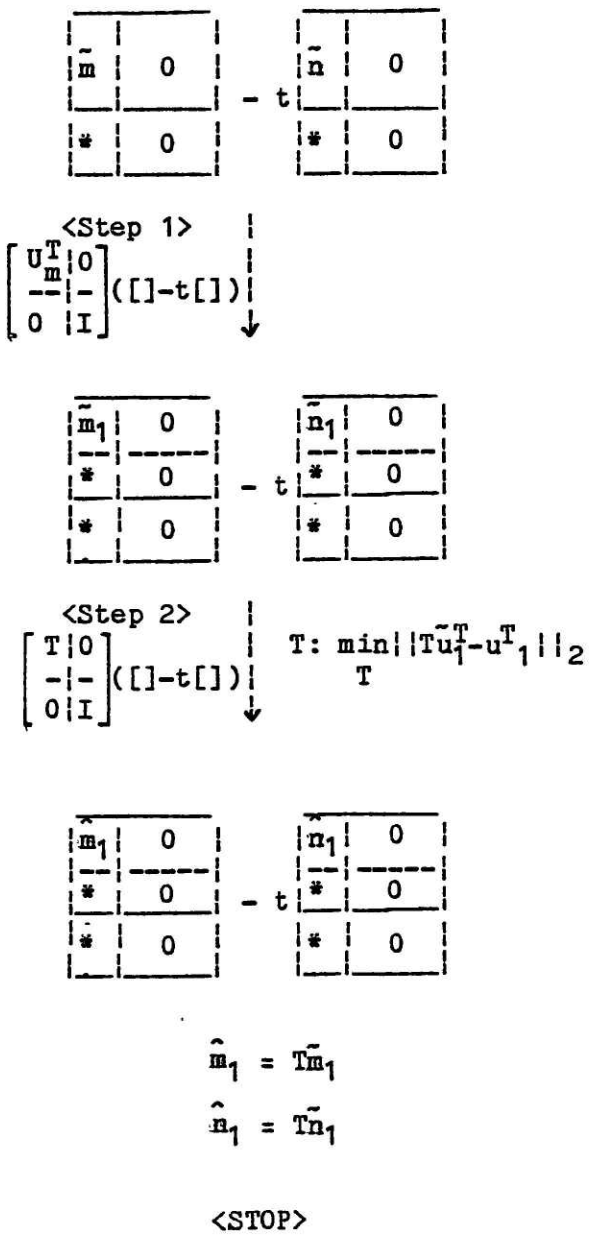


Figure I-1b. Modified Van Dooren's algorithm: Phase II.

Minimum error transformation:

$$\left[\begin{array}{c|c} T^T & 0 \\ \hline 0 & I \end{array}\right]\begin{bmatrix}\tilde{m}_1 \\ \tilde{m}_2\end{bmatrix} - t\begin{bmatrix}\tilde{n}_1 \\ 0\end{bmatrix} = \begin{bmatrix}T^T\tilde{m}_1 \\ \tilde{m}_2\end{bmatrix} - t\begin{bmatrix}T\tilde{n}_1 \\ 0\end{bmatrix}$$

$$T: \quad \min_T \|T^T\tilde{u}_1^T - u_1^T\|_2$$

Define:

$$e_n = T^T\tilde{n}_1 - n_1$$

$$e_m = T^T\tilde{m}_1 - m_1$$

<End.> The regular pencil with finite eigenvalue is:

$$\tilde{m}_1 - t\tilde{n}_1$$

Solving for t,

$$t = \tilde{m}_1/\tilde{n}_1$$

□

Thresholds to be used in the rank decisions made in Algorithm I-1 to guarantee preservation of the Kronecker structure of the underlying noise free pencil are given in Result I-1.

Result I-1.

Given noisy data matrices

$$\tilde{M} \; (= M + E_m = c_1x_1y_1^T + \ldots + c_rx_ry_r^T + E_m)$$

$$\tilde{N} \; (= N + E_n = x_jy_j^T + E_n, \quad 1 \le j \le r)$$

such that

$$s_r(M_2) > 2\|E_m\|_2 \left(1 + \frac{(\tilde{s}_1(\tilde{M})/\|E_m\|_2)+1}{(\tilde{s}_1(\tilde{N})/\|E_n\|_2)-2}\right)$$

$$s_1(N) > 2\|E_n\|_2$$

where $s_r(M_2)$ and $s_1(N)$ are the smallest nonzero singular values of

the matrices M_2 (= MV_2, see Algorithm 2, Chapter 3) and N for the noise free case.

Application of Algorithm I-1 with

$$EPSN = ||E_n||_2$$

$$EPSM2 = ||E_m||_2 \left(1 + \frac{(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

to the noisy matrix pencil $\tilde{M}-t\tilde{N}$:

(i) preserves the Kronecker structure of the underlying noise free pencil M-tN,

and (ii) yields a bound

$$\frac{T^T\tilde{m}_1-b_m}{T^T\tilde{n}_1+b_n} \leq t \leq \frac{T^T\tilde{m}_1+b_m}{T^T\tilde{n}_1-b_n}$$

where t is the finite eigenvalue of the noise free pencil M-tN, and b_m and b_n are upper bounds on $||T^T\tilde{m}_1-m_1||_2$ and $||T^T\tilde{n}_1-n_1||_2$ given by Result I-3.

Proof:

(i) In [V2] it is shown that the Kronecker structure of a singular pencil is uniquely determined by the rank decisions made in the algorithm (Van Dooren's). For rank(N)=1 case (Algorithm 3, Chapter 3),

the algorithm requires two rank decisions:

- rank of N, and
- rank of M_2 (=MV_2),

where V_2 is the matrix containing the right singular vectors corresponding to the zero singular values of N. Thus, Kronecker structure of the perturbed pencil $\tilde{M}-t\tilde{N}$, given by Algorithm 4, is

identical to that of the unperturbed pencil M-tN if

(a) rank(N) = numerical rank($\tilde{N}$)

and (b) rank(MV_2) = numerical rank($\tilde{M}\tilde{V}_2Q_2$).

Condition (a) is satisfied with EPSN=$||E_n||_2$, since

$$\tilde{N} = N + E_n \text{ and } s_1(N) > 2||E_n||_2$$

Condition (b) is also satisfied with

$$\text{EPSM2} = ||E_m||_2 \left(1 + \frac{(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

since

$$\tilde{M}\tilde{V}_2Q_2 = MV_2 + E_{m2}$$

and

$$s_r(MV_2) > 2||E_m||_2 \left(1 + \frac{(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right) \geq 2||E_{m2}||_2$$

where the last inequality is given by Result 4.

(ii) We have (see definition of e_m and e_n in Result I-3)

$$T^T\tilde{m}_1 - m_1 = e_m$$

$$T^T\tilde{n}_1 - n_1 = e_n$$

or solving for $t=m_1/n_1$,

$$t=m_1/n_1 = (T^T\tilde{m}_1 - e_m)/(T^T\tilde{n}_1 - e_n)$$

Thus,

$$\frac{T^T\tilde{m}_1 - b_m}{T^T\tilde{n}_1 + b_n} \leq t \leq \frac{T^T\tilde{m}_1 + b_m}{T^T\tilde{n}_1 - b_n}$$

where b_m and b_n are upper bounds on $||e_m||_2$ and $||e_n||_2$, respectively.

Q.E.D.

Thresholds to be used in the rank decisions made in Algorithm 3 (Chapter3) to guarantee preservation of the Kronecker structure of the underlying noise free pencil are given in Result I-2.

Result I-2.

Given noisy data matrices

$$\tilde{M} \; (= M + E_m = c_1 x_1 y_1^T + \ldots + c_r x_r y_r^T + E_m)$$

$$\tilde{N} \; (= N + E_n = x_j y_j^T + E_n, \quad 1 \leq j \leq r)$$

such that

$$s_r(M_2) > 2||E_m||_2 \left(1 + \frac{2^{1/2}(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

$$s_1(N) > 2||E_n||_2$$

where $s_r(M_2)$ and $s_1(N)$ are the smallest nonzero singular values of the matrices M_2 (= MV_2, see Algorithm 2, Chapter 3) and N for the noise-free case.

Application of Algorithm 3 (Chapter 3) with

$$\text{EPSN} = ||E_n||_2$$

$$\text{EPSM2} = ||E_m||_2 \left(1 + \frac{2^{1/2}(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

to the noisy matrix pencil $\tilde{M}-t\tilde{N}$ preserves the Kronecker structure of the underlying noise free pencil M-tN.

Proof: Consider Algorithm 4 with $Q_2^T Q_2 = I$. Since Q_2 is orthogonal, $s_r(\tilde{M}\tilde{V}_2 Q_2)$ of Algorithm I-1 is equal to $s_r(\tilde{M}\tilde{V}_2)$ of Algorithm 3 (Chapter 3). Thus, with

$$EPSM2 = ||E_m||_2 \left(1 + \frac{2^{1/2}(\tilde{s}_1(\tilde{M})/||E_m||_2)+1}{(\tilde{s}_1(\tilde{N})/||E_n||_2)-2}\right)$$

we have,

$$\text{numerical rank } (\tilde{M}\tilde{V}_2) = \text{numerical rank } (\tilde{M}\tilde{V}_2 Q_2)$$
$$= \text{rank } (MV_2)$$

Rest of the proof follows from proof in Result I-1. Q.E.D.

A set of bounds on the propagation of error terms (deviations from noise free case) for Algorithm I-1 is given in Result I-3.

Result I-3. Error bounds for Algorithm I-1. (Note: all norms $||.||$ in this result are spectral norms $||.||_2$.)

Phase I. Column compression.

<step 1.>

(i)

$$||e_{n1}|| = ||\tilde{n}_1 q_1 - n_1||$$
$$= ||(N+E_n)\tilde{v}_1 q_1 - Nv_1||$$
$$= ||N(\tilde{v}_1 q_1 - v_1) + E_n \tilde{v}_1 q_1||$$
$$= ||N(\tilde{v}_1 \tilde{v}_1^T v_1 - v_1) + E_n \tilde{v}_1 \tilde{v}_1^T v_1|| \qquad \text{(Th. I-1)}$$
$$\leq ||N||.||\tilde{v}_1 \tilde{v}_1^T v_1 - v_1|| + ||E_n||$$
$$= ||N||\sin\theta(\tilde{v}_1, v_1) + ||E_n|| \qquad \text{(Co. I-1)}$$
$$\leq ||N|| \left(\frac{||E_n||}{s_1(N) - ||E_n||}\right) + ||E_n|| \qquad \text{(Th. I-3)}$$

$$\leq (||\tilde{N}||+||E_n||)\left(\frac{||E_n||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right)+||E_n||$$

$$= ||E_n||\left(1+\frac{\tilde{s}_1(\tilde{N})+||E_n||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right)$$

$$= ||E_n||\left(1+\frac{(\tilde{s}_1(\tilde{N})/||E_n||)+1}{(\tilde{s}_1(\tilde{N})/||E_n||)-2}\right)$$

(ii)

$$||e_{m1}|| = ||\tilde{m}_1 q_1 - m_1||$$

$$= ||(M+E_m)\tilde{v}_1 q_1 - Mv_1||$$

$$= ||M(\tilde{v}_1 q_1 - v_1)+E_m\tilde{v}_1 q_1||$$

$$= ||M(\tilde{v}_1\tilde{v}_1^T v_1 - v_1)+E_m\tilde{v}_1\tilde{v}_1^T v_1|| \qquad \text{(Th. I-1)}$$

$$\leq ||M||.||\tilde{v}_1\tilde{v}_1^T v_1 - v_1||+||E_m||$$

$$= ||M||\sin\theta(\tilde{v}_1, v_1)+||E_m|| \qquad \text{(Co. I-1)}$$

$$\leq ||M||\left(\frac{||E_n||}{s_1(N)-||E_n||}\right)+||E_m|| \qquad \text{(Th. I-3)}$$

$$\leq (||\tilde{M}||+||E_m||)\left(\frac{||E_n||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right)+||E_m||$$

$$= ||E_m||\left(1+\frac{\tilde{s}_1(\tilde{M})+||E_m||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right)\cdot\frac{||E_n||}{||E_m||}$$

$$= ||E_m||\left(1+\frac{(\tilde{s}_1(\tilde{M})/||E_m||)+1}{(\tilde{s}_1(\tilde{N})/||E_n||)-2}\right)$$

(iii)

(a) Invertible Q_2

$$||E_{m2}|| = ||\tilde{M}_2 Q_2 - M_2||$$

$$= ||(M+E_m)\tilde{V}_2Q_2-MV_2||$$

$$= ||M(\tilde{V}_2Q_2-V_2)+E_m\tilde{V}_2Q_2||$$

$$= ||M(\tilde{V}_2\tilde{V}_2^T V_2-V_2)+E_m\tilde{V}_2\tilde{V}_2^T V_2|| \qquad \text{(Th. I-1)}$$

$$\leq ||M||.||\tilde{V}_2\tilde{V}_2^T V_2-V_2||+||E_m||$$

$$= ||M||\sin\theta(\tilde{V}_2,V_2)+||E_m|| \qquad \text{(Co. I-1)}$$

$$\leq ||M||\left(\frac{||E_n||}{s_1(N)-||E_n||}\right)+||E_m|| \qquad \text{(Th. I-3)}$$

$$\leq (||\tilde{M}||+||E_m||)\left(\frac{||E_n||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right)+||E_m||$$

$$= ||E_m||\left(1+\frac{\tilde{s}_1(\tilde{M})+||E_m||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right)\cdot\frac{||E_n||}{||E_m||}$$

$$= ||E_m||\left(1+\frac{(\tilde{s}_1(\tilde{M})/||E_m||)+1}{(\tilde{s}_1(\tilde{N})/||E_n||)-2}\right)$$

(b) Orthogonal Q_2.

$$||E_{m2}|| = ||\tilde{M}_2Q_2-M_2||$$

$$= ||(M+E_m)\tilde{V}_2Q_2-MV_2||$$

$$= ||M(\tilde{V}_2Q_2-V_2)+E_m\tilde{V}_2Q_2||$$

$$\leq ||M||.||\tilde{V}_2Q_2-V_2||+||E_m||$$

$$= ||M||.2\sin 0.5\theta(\tilde{V}_2,V_2)+||E_m|| \qquad \text{(Th. I-2, Co. I-2)}$$

$$\leq ||M||.2^{1/2}\sin\theta(\tilde{V}_2,V_2)+||E_m||$$

$$\leq 2^{1/2}||M||\left(\frac{||E_n||}{s_1(N)-||E_n||}\right)+||E_m|| \qquad \text{(Th. I-3)}$$

$$\leq 2^{1/2}(||\tilde{M}||+||E_m||)\ \left(\frac{||E_n||}{\tilde{s}_1(\tilde{N})-2||E_n||}\right) + ||E_m||$$

$$= ||E_m||\ \left(1 + \frac{2^{1/2}(\tilde{s}_1(\tilde{M})+||E_m||)}{\tilde{s}_1(\tilde{N})-2||E_n||}\right) \cdot \frac{||E_n||}{||E_m||}$$

$$= ||E_m||\ \left(1 + \frac{2^{1/2}((\tilde{s}_1(\tilde{M})/||E_m||)+1)}{(\tilde{s}_1(\tilde{N})/||E_n||)-2}\right)$$

<step 2.>

(i)

$$||e_{m21}|| = ||R^T\tilde{m}_{21}-m_{21}||$$

$$= ||(R^T\tilde{U}_2^T\tilde{m}_1 q_1-U_2^T m_1||$$

$$= ||R^T\tilde{U}_2^T(m_1+e_{m1})-U_2^T m_1||$$

$$= ||(\tilde{U}_2R-U_2)^T m_1+R^T\tilde{U}_2^T e_{m1}||$$

$$= ||m_1^T(\tilde{U}_2R-U_2)+e_{m1}^T\tilde{U}_2R||$$

$$= ||m_1^T(\tilde{U}_2\tilde{U}_2^TU_2-U_2)+e_{m1}^T\tilde{U}_2\tilde{U}_2^TU_2|| \qquad \text{(Th. I-1)}$$

$$\leq ||m_1||.||\tilde{U}_2\tilde{U}_2^TU_2-U_2||+||e_{m1}^T||$$

$$= ||m_1||\sin\theta(\tilde{U}_2,U_2)+||e_{m1}|| \qquad \text{(Co. I-1)}$$

$$\leq ||m_1||\ \left(\frac{||E_{m2}||}{s_r(M_2)-||E_{m2}||}\right) + ||e_{m1}|| \qquad \text{(Th. I-3)}$$

$$\leq (||\tilde{m}_1||+||e_{m1}||)\ \left(\frac{||E_{m2}||}{\tilde{s}_r(\tilde{M}_2)-2||E_{m2}||}\right) + ||e_{m1}||$$

$$= ||e_{m1}||\ \left(1 + \frac{(\tilde{s}_1(\tilde{m}_1)/||e_{m1}||)+1}{(\tilde{s}_r(\tilde{M}_2)/||E_{m2}||)-2}\right)$$

(ii)

$$
\begin{aligned}
||e_{n21}|| &= ||R^T\bar{n}_{21}-n_{21}|| \\
&= ||(R^T\bar{U}_2^T\bar{n}_1 q_1-U_2^T n_1|| \\
&= ||R^T\bar{U}_2^T(n_1+e_{n1})-U_2^T n_1|| \\
&= ||(\tilde{U}_2 R-U_2)^T n_1+R^T\bar{U}_2^T e_{n1}|| \\
&= ||n_1^T(\tilde{U}_2 R-U_2)+e_{n1}^T\tilde{U}_2 R|| \\
&= ||n_1^T(\tilde{U}_2\tilde{U}_2^T U_2-U_2)+e_{n1}^T\tilde{U}_2\tilde{U}_2^T U_2|| && \text{(Th. I-1)} \\
&\leq ||n_1||.||\tilde{U}_2\tilde{U}_2^T U_2-U_2||+||e_{n1}^T|| \\
&= ||n_1||\sin\theta(\tilde{U}_2,U_2)+||e_{n1}|| && \text{(Co. I-1)} \\
&\leq ||n_1||\left(\frac{||E_{m2}||}{s_r(M_2)-||E_{m2}||}\right)+||e_{n1}|| && \text{(Th. I-3)} \\
&\leq (||\tilde{n}_1||+||e_{n1}||)\left(\frac{||E_{m2}||}{\tilde{s}_r(\tilde{M}_2)-2||E_{m2}||}\right)+||e_{n1}|| \\
&= ||e_{n1}||\left(1+\frac{(\tilde{s}_1(\tilde{n}_1)/||e_{n1}||)+1}{(\tilde{s}_r(\tilde{M}_2)/||E_{m2}||)-2}\right)
\end{aligned}
$$

Phase II. Row compression.

(i)

$$
\begin{aligned}
||e_m|| &= ||T^T\bar{m}_1-m_1|| \\
&= ||(T^T\bar{U}_1^T R^T\bar{m}_{21}-U_1^T m_{21}|| \\
&= ||T^T\bar{U}_1^T(m_{21}+e_{m21})-U_1^T m_{21}|| \\
&= ||(\tilde{U}_1 T-U_1)^T m_{21}+T^T\bar{U}_1^T e_{m21}|| \\
&= ||m_2^T(\tilde{U}_1 T-U_1)+e_{m21}^T\tilde{U}_1 T|| \\
&= ||m_2^T(\tilde{U}_1\tilde{U}_1^T U_1-U_1)+e_{m21}^T\tilde{U}_1\tilde{U}_1^T U_1|| && \text{(Th. I-1)} \\
&\leq ||m_{21}||.||\tilde{U}_1\tilde{U}_1^T U_1-U_1||+||e_{m21}||
\end{aligned}
$$

$$= ||m_{21}||\sin\theta(\tilde{U}_1, U_1) + ||e_{m21}|| \qquad \text{(Co. I-1)}$$

$$\leq ||m_{21}|| \left(\frac{||e_{n21}||}{s_1(n_1) - ||e_{n21}||}\right) + ||e_{m21}|| \qquad \text{(Th. I-3)}$$

$$\leq (||\tilde{m}_{21}|| + ||e_{m21}||) \left(\frac{||e_{n21}||}{\tilde{s}_1(\tilde{n}_{21}) - 2||e_{n21}||}\right) + ||e_{m21}||$$

$$= ||e_{m21}|| \left(1 + \frac{(\tilde{s}_1(\tilde{m}_{21})/||e_{m21}||) + 1}{(\tilde{s}_1(\tilde{n}_{21})/||e_{n21}||) - 2}\right)$$

(ii)

$$||e_n|| = ||T^T\tilde{n}_1 - n_1||$$

$$= ||(T^T\tilde{U}_1^T R^T\tilde{n}_{21} - U_1^T n_{21}||$$

$$= ||T^T\tilde{U}_1^T(n_{21} + e_{n21}) - U_1^T n_{21}||$$

$$= ||(\tilde{U}_1 T - U_1)^T n_{21} + T^T\tilde{U}_1^T e_{n21}||$$

$$= ||n_2^T(\tilde{U}_1 T - U_1) + e_{n21}^T\tilde{U}_1 T||$$

$$= ||n_2^T(\tilde{U}_1\tilde{U}_1^T U_1 - U_1) + e_{n21}^T\tilde{U}_1\tilde{U}_1^T U_1|| \qquad \text{(Th. I-1)}$$

$$\leq ||n_{21}||.||\tilde{U}_1\tilde{U}_1^T U_1 - U_1|| + ||e_{n21}||$$

$$= ||n_{21}||\sin\theta(\tilde{U}_1, U_1) + ||e_{n21}|| \qquad \text{(Co. I-1)}$$

$$\leq ||n_{21}|| \left(\frac{||e_{n21}||}{s_1(n_1) - ||e_{n21}||}\right) + ||e_{n21}|| \qquad \text{(Th. I-3)}$$

$$\leq (||\tilde{n}_{21}|| + ||e_{n21}||) \left(\frac{||e_{n21}||}{\tilde{s}_1(\tilde{n}_{21}) - 2||e_{n21}||}\right) + ||e_{n21}||$$

$$= ||e_{n21}|| \left(1 + \frac{(\tilde{s}_1(\tilde{n}_{21})/||e_{n21}||) + 1}{(\tilde{s}_1(\tilde{n}_{21})/||e_{n21}||) - 2}\right)$$

[]

Theorem I-1.

Given two matrices $\tilde{V}_1$ and V_1, each with orthonormal columns, the matrix

$$Q=\tilde{V}_1^T V_1$$

minimizes the norm

$$||\tilde{V}_1 Q - V_1||_2$$

Proof:

We have,

$$||\tilde{V}_1 Q - V_1||_2 = \max_{x^T x=1} ||\tilde{V}_1 Q x - V_1 x||_2$$

Let $x^*(Q)$ be the maximizing x for some Q, then

$$||\tilde{V}_1 Q - V_1||_2 = ||\tilde{V}_1 Q x^* - V_1 x^*||_2$$

Since

$$V_1 x^* \subset \text{span}(V_1)$$

$$\tilde{V}_1 Q x^* \subset \text{span}(\tilde{V}_1),$$

and the minimum distance between the vector $V_1 x^*$ and the subspace $\text{span}(\tilde{V}_1)$ is given by the projection of $V_1 x^*$ onto the orthogonal complement of $\text{span}(\tilde{V}_1)$, the following inequality holds

$$||\tilde{V}_1 Q x^* - V_1 x^*||_2 \geq ||(I - \tilde{V}_1 \tilde{V}_1^T) V_1 x^*||_2$$

Clearly,

$$Q = \tilde{V}_1^T V_1$$

gives the equality, hence it minimizes the norm

$$||\tilde{V}_1 Q - V_1||_2$$

Q.E.D.

Definition I-2. Angle between subspaces [D1]

Given two subspaces X and Y, the angle $\measuredangle(X,Y)$ between them is defined by

$$\text{sine}(X,Y) = \|(I-P_X)P_Y\|_2$$

i.e.

$$\measuredangle(X,Y) = \sin^{-1}\text{e}\|(I-P_X)P_Y\|_2$$

where P_X and P_Y are the projection matrices of X and Y. □

Corollary I-1.

Given two matrices $\tilde{V}_1$ and V_1, each with orthonormal columns,

$$\min_Q \|\tilde{V}_1 Q - V_1\|_2 = \text{sine}(\text{span}(\tilde{V}_1), \text{span}(V_1))$$

Proof:

Let V_2 and $\tilde{V}_2$ be such that $[\tilde{V}_1, \tilde{V}_2]$ and $[V_1, V_2]$ are orthogonal matrices. From Theorem I-1,

$$\min_Q \|\tilde{V}_1 Q - V_1\|_2 = \|\tilde{V}_1\tilde{V}_1^T V_1 - V_1\|_2$$

$$= \|[\tilde{V}_1, \tilde{V}_2]\left[\begin{array}{c|c} \tilde{V}_1^T V_1 & 0 \\ \hline 0 & 0 \end{array}\right] - [V_1, V_2]\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]\|_2$$

$$= \|[\tilde{V}_1, \tilde{V}_2]\left[\begin{array}{c|c} \tilde{V}_1^T V_1 & 0 \\ \hline 0 & 0 \end{array}\right]\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} - [V_1, V_2]\left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]\begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}\|_2$$

$$= \|\tilde{V}_1\tilde{V}_1^T V_1 V_1^T - V_1 V_1^T\|_2$$

$$= \|(I-\tilde{V}_1\tilde{V}_1^T)V_1 V_1^T\|_2$$

$$= \text{sine}(\text{span}(\tilde{V}_1), \text{span}(V_1))$$

Q.E.D.

Theorem I-2.

Given two matrices $\tilde{V}_1$ and V_1, each with orthonormal columns, an orthogonal matrix Q that minimizes the norm

$$\|\tilde{V}_1 Q - V_1\|_2$$

is given by

$$Q = AB^T$$

where A and B are, respectively, the left and right singular vectors of the matrix

$$\tilde{V}_1^T V_1 \quad (=ASB^T)$$

Proof:

We have,

$$||\tilde{V}_1 Q - V_1||_2 = \max_{x^T x=1} ||\tilde{V}_1 Qx - V_1 x||_2$$

Let $x^*(Q)$ be the maximizing x for some Q, then

$$||\tilde{V}_1 Q - V_1||_2 = ||\tilde{V}_1 Qx^* - V_1 x^*||_2$$

Since

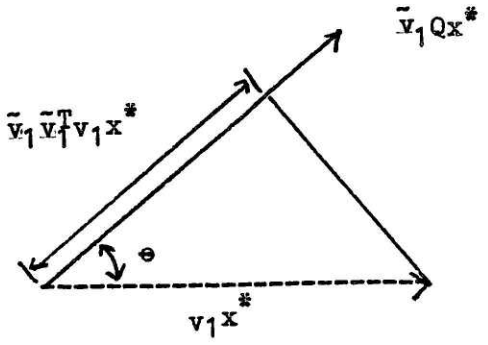
$$V_1 x^* \subset \text{span}(V_1)$$

$$\tilde{V}_1 Qx^* \subset \text{span}(\tilde{V}_1),$$

Q that minimizes the norm

$$||\tilde{V}_1 Q - V_1||_2$$

must be such that the vector $\tilde{V}_1 Qx^*$ is a unit vector along the projection of $V_1 x^*$ onto $\text{span}(\tilde{V}_1)$.



Figure

That is,

$$\tilde{V}_1 Q x^* = \frac{(\tilde{V}_1 \tilde{V}_1^T) V_1 x^*}{\cos(\theta)}$$

where

$$\theta = \sphericalangle\,(\tilde{V}_1 x^*, V_1 x^*)$$
$$= \cos^{-1}(\text{min. singular value of } \tilde{V}_1^T V_1) \quad \text{(from [G6])}$$

is the angle between the vectors $\tilde{V}_1 x^*$ and $V_1 x^*$. Hence,

$$Q = \frac{\tilde{V}_1^T V_1}{\cos(\theta)}$$

satisfies the requirement that the vector $\tilde{V}_1 Q x^*$ be a unit vector along the projection of $V_1 x^*$ onto span($\tilde{V}_1$), except that Q is not orthogonal in general. This can be seen by noting that since $\tilde{V}_1^T V_1$ has SVD

$$\tilde{V}_1^T V_1 = ASB^T$$
$$S = \text{diag}(\cos(\theta_1), \ldots, \cos(\theta_r))$$

and

$$\cos(\theta) = \text{min. singular value of } \tilde{V}_1^T V_1 = \cos(\theta_r)$$

we have,

$$Q = \frac{ASB^T}{\cos(\theta)}$$
$$= A \begin{bmatrix} \frac{\cos(\theta_1)}{\cos(\theta_r)} & & & 0 \\ & \ddots & & \\ & & \frac{\cos(\theta_{r-1})}{\cos(\theta_r)} & \\ 0 & & & 1 \end{bmatrix} B^T$$

which is <u>not</u> an orthogonal matrix in general. However, from

Theorem I-2a: $x^*=b_r$, the right singular vector corresponding to the smallest singular value of $\tilde{V}_1^T V_1$. Thus, any Q with the SVD

$$Q = A \begin{bmatrix} * & & & \\ & * & 0 & \\ & 0 & * & \\ & & & 1 \end{bmatrix} B^T$$

gives a vector $\tilde{V}_1 Q x^*$ of unit length along the projection of $V_1 x^*$ onto span($\tilde{V}_1$). Therefore, when Q is restricted to be orthogonal, the matrix

$$Q = AB^T$$

minimizes the norm

$$||\tilde{V}Q - V_1||_2$$

Q.E.D.

Theorem I-2a.

Given two matrices $\tilde{V}_1$ and V_1, each with orthonormal columns, an unit vector x that maximizes the norm

$$||(\tilde{V}_1 \tilde{V}_1^T V_1 - V_1)x||_2$$

is given by the right singular vector corresponding to the smallest singular value of

$$\tilde{V}_1^T V_1$$

Proof:

We have,

$$\max_{x^T x=1} ||(\tilde{V}_1 \tilde{V}_1^T V_1 - V_1)x||_2$$

$$= ||\tilde{V}_1 \tilde{V}_1^T V_1 - V_1||_2$$

$$= ||\tilde{V}_1 \tilde{V}_1^T [V_1, V_2] \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] - [V_1, V_2] \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right]||_2$$

$$= ||\tilde{V}_1\tilde{V}_1^T[V_1, V_2]\left[\begin{array}{c|c}I & 0\\ \hline 0 & 0\end{array}\right]\begin{bmatrix}V_1^T\\ V_2^T\end{bmatrix} - [V_1, V_2]\left[\begin{array}{c|c}I & 0\\ \hline 0 & 0\end{array}\right]\begin{bmatrix}V_1^T\\ V_2^T\end{bmatrix}||_2$$

$$= ||\tilde{V}_1\tilde{V}_1^T V_1 V_1^T - V_1 V_1^T||_2$$

$$(= (1-s_r^2)^{-1/2} \qquad [G6])$$

where s_r is the minimum singular value of $\tilde{V}_1^T V_1$.

Furthermore, given SVD of $\tilde{V}_1^T V_1$

$$\tilde{V}_1^T V_1 = ASB^T \qquad (SVD)$$

$$S = \text{diag}(s_1, \ldots s_r)$$

Let

$$x = b_r$$

the r-th right singular vector of $\tilde{V}_1^T V_1$, corresponding to the smallest singular value s_r. Then,

$$||(\tilde{V}_1\tilde{V}_1^T V_1 - V_1)x||_2^2$$

$$= ||(\tilde{V}_1\tilde{V}_1^T V_1 - V_1)b_r||_2^2$$

$$= (\tilde{V}_1 ASB^T b_r - V_1 b_r)^T(\tilde{V}_1 ASB^T b_r - V_1 b_r)$$

$$= (b_r^T BSA^T\tilde{V}_1^T - b_r^T V_1^T)(\tilde{V}_1 ASB^T b_r - V_1 b_r)$$

$$= (s_r a_r^T\tilde{V}_1^T - b_r^T V_1^T)(s_r\tilde{V}_1 a_r - V_1 b_r)$$

$$= s_r^2 a_r^T\tilde{V}_1^T\tilde{V}_1 a_r + b_r^T V_1^T V_1 b_r - s_r a_r^T\tilde{V}_1^T V_1 b_r - s_r b_r^T V_1^T\tilde{V}_1 a_r$$

$$= s_r^2 + 1 - 2s_r^2$$

$$= 1 - s_r^2$$

Thus,

$$||(\tilde{V}_1\tilde{V}_1^T V_1 - V_1)b_r||_2$$

$$= (1 - s_r^2)^{-1/2}$$

$$= \max_{x^T x=1} ||(\tilde{V}_1\tilde{V}_1^T V_1 - V_1)x||_2$$

Therefore, the right singular vector corresponding to the smallest singular value of $\tilde{V}_1^T V_1$ is the vector x that maximizes the norm

$$||(\tilde{V}_1 \tilde{V}_1^T V_1 - V_1)x||_2$$

Q. E. D.

Corollary I-2.

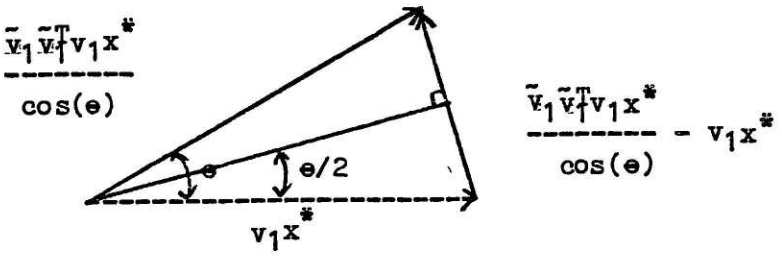
Given two matrices $\tilde{V}_1$ and V_1, each with orthonormal columns,

$$\min_{QQ^T=Q^TQ=I} ||\tilde{V}_1 Q - V_1||_2 = 2\sin(\theta/2)(\tilde{V}_1, V_1)$$

Proof:

From theorem *2.,

$$\min_{QQ^T=Q^TQ=I} ||\tilde{V}_1 Q - V_1||_2 = ||\underbrace{\tilde{V}_1 \tilde{V}_1^T V_1 x^*}_{\cos(\theta)} - V_1 x^*||_2$$



Figure

Since,

$$||\underbrace{\tilde{V}_1 \tilde{V}_1^T V_1 x^*}_{\cos(\theta)}||_2 = ||V_1 x^*||_2 = 1$$

and

$$\theta = \angle \left(\underbrace{\tilde{V}_1 \tilde{V}_1^T V_1 x^*}_{\cos(\theta)}, V_1 x^* \right)$$

we have,

$$\left\| \frac{\tilde{V}_1\tilde{V}_1^T V_1 x^* - V_1 x^*}{\cos(\theta)} \right\|_2 = 2\sin(\theta/2)(\tilde{V}_1, V_1) \qquad \text{(see Figure)}$$

i.e, the length of the base of an isoceles triangle with θ as the angle between the two congruent sides of unit lengths.

Q.E.D.

Theorem I-3. Sin θ theorem for SVD [W2,W3]

Let

$$\tilde{M} = M + E$$

and $M, \tilde{M}$ have SVD,

$$M = [U_1, U_2] \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$$

$$\tilde{M} = [\tilde{U}_1, \tilde{U}_2] \begin{bmatrix} \tilde{S}_1 & 0 \\ 0 & \tilde{S}_2 \end{bmatrix} \begin{bmatrix} \tilde{V}_1^T \\ \tilde{V}_2^T \end{bmatrix}$$

Define,

$$R_{11} = M\tilde{V}_1 - \tilde{U}_1\tilde{S}_1$$

$$R_{12} = M^T\tilde{U}_1 - \tilde{V}_1\tilde{S}_1$$

Assume there exist $a \geq 0$ and $d > 0$, such that

$$\tilde{s}_r \geq a+d \qquad \text{and} \qquad s_{r+1} \leq a$$

where r is the dimension of S_1 and $\tilde{S}_1$.

Let

$$e = \max(\|R_{11}\|, \|R_{12}\|)$$

Then for every unitary invariant norm,

$$\sin\theta(\tilde{V}_1, V_1) \leq e/d$$

$$\sin\theta(\tilde{U}_1, U_1) \leq e/d$$

where

$$\sin\theta(\tilde{V}_1, V_1) \hat{=} \|(I - \tilde{V}_1\tilde{V}_1^T)V_1V_1^T\| \; (= \|(I - \tilde{V}_2\tilde{V}_2^T)V_2V_2^T\|)$$

$$\sin\theta(\tilde{U}_1, U_1) \hat{=} \|(I - \tilde{U}_1\tilde{U}_1^T)U_1U_1^T\| \; (= \|(I - \tilde{U}_2\tilde{U}_2^T)U_2U_2^T\|)$$