



National Library  
of Canada

Acquisitions and  
Bibliographic Services Branch

395 Wellington Street  
Ottawa, Ontario  
K1A 0N4

Bibliothèque nationale  
du Canada

Direction des acquisitions et  
des services bibliographiques

395, rue Wellington  
Ottawa (Ontario)  
K1A 0N4

*Your file* *Votre référence*

*Our file* *Notre référence*

## NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

## AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

**ENHANCEMENTS TO MULTIDIMENSIONAL METHODS  
IN ANALYTICAL CHEMISTRY**

Stephen James Vanslyke

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

at

Dalhousie University

Halifax, Nova Scotia

May, 1994

© Copyright by Stephen James Vanslyke, 1994



National Library  
of Canada

Acquisitions and  
Bibliographic Services Branch

395 Wellington Street  
Ottawa, Ontario  
K1A 0N4

Bibliothèque nationale  
du Canada

Direction des acquisitions et  
des services bibliographiques

395, rue Wellington  
Ottawa (Ontario)  
K1A 0N4

*Your file* *Votre référence*

*Our file* *Notre référence*

THE AUTHOR HAS GRANTED AN IRREVOCABLE NON-EXCLUSIVE LICENCE ALLOWING THE NATIONAL LIBRARY OF CANADA TO REPRODUCE, LOAN, DISTRIBUTE OR SELL COPIES OF HIS/HER THESIS BY ANY MEANS AND IN ANY FORM OR FORMAT, MAKING THIS THESIS AVAILABLE TO INTERESTED PERSONS.

L'AUTEUR A ACCORDE UNE LICENCE IRREVOCABLE ET NON EXCLUSIVE PERMETTANT A LA BIBLIOTHEQUE NATIONALE DU CANADA DE REPRODUIRE, PRETER, DISTRIBUER OU VENDRE DES COPIES DE SA THESE DE QUELQUE MANIERE ET SOUS QUELQUE FORME QUE CE SOIT POUR METTRE DES EXEMPLAIRES DE CETTE THESE A LA DISPOSITION DES PERSONNE INTERESSEES.

THE AUTHOR RETAINS OWNERSHIP OF THE COPYRIGHT IN HIS/HER THESIS. NEITHER THE THESIS NOR SUBSTANTIAL EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT HIS/HER PERMISSION.

L'AUTEUR CONSERVE LA PROPRIETE DU DROIT D'AUTEUR QUI PROTEGE SA THESE. NI LA THESE NI DES EXTRAITS SUBSTANTIELS DE CELLE-CI NE DOIVENT ETRE IMPRIMES OU AUTREMENT REPRODUITS SANS SON AUTORISATION.

ISBN 0-612-05175-7

Canada

Name Stephen James Vanslyke

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Analytical Chemistry

SUBJECT TERM

0486

U·M·I

SUBJECT CODE

**Subject Categories**

**THE HUMANITIES AND SOCIAL SCIENCES**

**COMMUNICATIONS AND THE ARTS**

Architecture	0729
Art History	0377
Cinema	0900
Dance	0378
Fine Arts	0357
Information Science	0723
Journalism	0391
Library Science	0299
Mass Communications	0708
Music	0413
Speech Communication	0459
Theater	0465

Psychology	0525
Reading	0535
Religious Sciences	0527
Sciences	0714
Secondary	0533
Social Sciences	0534
Sociology of	0340
Special	0529
Teacher Training	0530
Technology	0710
Tests and Measurements	0288
Vocational	0747

**PHILOSOPHY, RELIGION AND THEOLOGY**

Philosophy	0422
Religion	
General	0318
Biblical Studies	0321
Clergy	0319
History of	0320
Philosophy of	0322
Theology	0469

Ancient	0579
Medieval	0581
Modern	0582
Black	0328
African	0331
Asia, Australia and Oceania	0332
Canadian	0334
European	0335
Latin American	0336
Middle Eastern	0333
United States	0337
History of Science	0585
Law	0398

**EDUCATION**

General	0515
Administration	0514
Adult and Continuing	0516
Agricultural	0517
Art	0273
Bilingual and Multicultural	0282
Business	0688
Community College	0275
Curriculum and Instruction	0727
Early Childhood	0518
Elementary	0524
Finance	0277
Guidance and Counseling	0519
Health	0680
Higher	0745
History of	0520
Home Economics	0278
Industrial	0521
Language and Literature	0279
Mathematics	0280
Music	0522
Philosophy of	0998
Physical	0523

**LANGUAGE, LITERATURE AND LINGUISTICS**

Language	
General	0679
Ancient	0289
Linguistics	0290
Modern	0291
Literature	
General	0401
Classical	0294
Comparative	0295
Medieval	0297
Modern	0298
African	0316
American	0591
Asian	0305
Canadian (English)	0352
Canadian (French)	0355
English	0593
Germanic	0311
Latin American	0312
Middle Eastern	0315
Romance	0313
Slavic and East European	0314

**SOCIAL SCIENCES**

American Studies	0323
Anthropology	
Archaeology	0324
Cultural	0326
Physical	0327
Business Administration	
General	0310
Accounting	0272
Banking	0770
Management	0454
Marketing	0338
Canadian Studies	0385
Economics	
General	0501
Agricultural	0503
Commerce-Business	0505
Finance	0508
History	0509
Labor	0510
Theory	0511
Folklore	0358
Geography	0366
Gerontology	0351
History	
General	0578

Political Science	
General	0615
International Law and Relations	0616
Public Administration	0617
Recreation	0814
Social Work	0452
Sociology	
General	0626
Criminology and Penology	0627
Demography	0938
Ethnic and Racial Studies	0631
Individual and Family Studies	0628
Industrial and Labor Relations	0629
Public and Social Welfare	0630
Social Structure and Development	0700
Theory and Methods	0344
Transportation	0709
Urban and Regional Planning	0999
Women's Studies	0453

**THE SCIENCES AND ENGINEERING**

**BIOLOGICAL SCIENCES**

Agriculture	
General	0473
Agronomy	0285
Animal Culture and Nutrition	0475
Animal Pathology	0476
Food Science and Technology	0359
Forestry and Wildlife	0478
Plant Culture	0479
Plant Pathology	0480
Plant Physiology	0817
Range Management	0777
Wood Technology	0746

Geodesy	0370
Geology	0372
Geophysics	0373
Hydrology	0388
Mineralogy	0411
Paleobotany	0345
Paleoecology	0426
Paleontology	0418
Paleozoology	0985
Polynology	0427
Physical Geography	0368
Physical Oceanography	0415

Speech Pathology	0460
Toxicology	0383
Home Economics	0386

**Engineering**

General	0537
Aerospace	0538
Agricultural	0539
Automotive	0540
Biomedical	0541
Chemical	0542
Civil	0543
Electronics and Electrical	0544
Heat and Thermodynamics	0348
Hydraulic	0545
Industrial	0546
Marine	0547
Materials Science	0794
Mechanical	0548
Metallurgy	0743
Mining	0551
Nuclear	0552
Packaging	0549
Petroleum	0765
Sanitary and Municipal System Science	0554
System Science	0790
Technotechnology	0428
Operations Research	0796
Plastics Technology	0795
Textile Technology	0994

**Biology**

General	0306
Anatomy	0287
Biostatistics	0308
Botany	0309
Cell	0379
Ecology	0329
Entomology	0353
Genetics	0369
Limnology	0793
Microbiology	0410
Molecular	0507
Neuroscience	0317
Oceanography	0416
Physiology	0433
Radiation	0821
Veterinary Science	0778
Zoology	0472

**Biophysics**

General	0786
Medical	0760

**HEALTH AND ENVIRONMENTAL SCIENCES**

Environmental Sciences	0768
Health Sciences	
General	0566
Audiology	0300
Chemotherapy	0992
Dentistry	0567
Education	0350
Hospital Management	0769
Human Development	0758
Immunology	0982
Medicine and Surgery	0564
Mental Health	0347
Nursing	0569
Nutrition	0570
Obstetrics and Gynecology	0380
Occupational Health and Therapy	0354
Ophthalmology	0381
Pathology	0571
Pharmacology	0419
Pharmacy	0572
Physical Therapy	0382
Public Health	0573
Radiology	0574
Recreation	0575

**PHYSICAL SCIENCES**

**Pure Sciences**

Chemistry	
General	0485
Agricultural	0749
Analytical	0486
Biochemistry	0487
Inorganic	0488
Nuclear	0738
Organic	0490
Pharmaceutical	0491
Physical	0494
Polymer	0495
Radiation	0754
Mathematics	0405
Physics	
General	0605
Acoustics	0986
Astronomy and Astrophysics	0606
Atmospheric Science	0608
Atomic	0748
Electronics and Electricity	0607
Elementary Particles and High Energy	0798
Fluid and Plasma	0759
Molecular	0609
Nuclear	0610
Optics	0752
Radiation	0756
Solid State	0611
Statistics	0463

**Applied Sciences**

Applied Mechanics	0346
Computer Science	0984

**EARTH SCIENCES**

Biogeochemistry	0425
Geochemistry	0996



DALHOUSIE UNIVERSITY

FACULTY OF GRADUATE STUDIES

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "Enhancements to Multidimensional Methods in Analytical Chemistry"

by Stephen Vanslyke

in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated June 28, 1994

External Examiner

Research Supervisor

Examining Committee

**DALHOUSIE UNIVERSITY**

July 1, 1994

Author: **Stephen James Vanslyke**

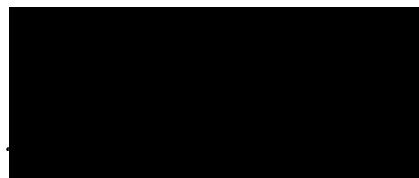
Title: **Enhancements to Multidimensional Methods  
in Analytical Chemistry**

Department: Chemistry

Degree: Ph.D.

Convocation: Fall, 1994

Permission is herewith granted to Dalhousie University to circulate and have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

A large black rectangular box redacting the author's signature.

Signature of Author

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

The author attests that permission has been obtained for the use of any copyrighted material appearing in this thesis (other than brief excerpts requiring only proper acknowledgment in scholarly writing) and that all such use is clearly acknowledged.

# CONTENTS

---

Table of Contents.....	iv
List of Figures.....	vii
List of Tables.....	xii
Abstract.....	xiii
Notation.....	xiv
Abbreviations and Symbols.....	xv
Acknowledgments.....	xviii
<b>1 Introduction</b>	
1.1 Multidimensional Methods.....	1
1.2 Intelligent Instruments.....	3
1.3 Chemometrics and Instrumentation.....	6
1.4 Signals, Noise and Filters.....	10
1.5 Summary.....	23
<b>2 The Kalman Filter</b>	
2.1 Introduction.....	24
2.2 Recursive Filters.....	27
2.3 Kalman Filters.....	29
2.3.1 The Discrete Kalman Filter Algorithm.....	29
2.3.2 Example: One Cycle of Kalman Filtering Reaction-Rate Data.....	36
2.3.3 Diagnostic Abilities of the Kalman Filter.....	39
2.4 Kalman Filters in Analytical Chemistry.....	43
2.4.1 Noise Removal.....	44
2.4.2 Determination of Kinetic Parameters.....	45
2.4.3 Resolution of Overlapped Responses.....	45
2.4.4 Calibration with Drift Correction.....	46
2.4.5 Model Identification and Improvement.....	47
2.5 Parallel Kalman Filter Networks for Kinetic Methods of Analysis.....	49
2.5.1 Theory.....	50

2.5.2	Experimental .....	55
2.5.3	Results and Discussion.....	56
2.6	Conclusions.....	66
<b>3</b>	<b>The Kalman Filter and Ordered Data Sets</b>	
3.1	Introduction.....	68
3.2	Beer's Law.....	69
3.3	Extracting Information from Bilinear Data.....	72
3.4	Deducing the Number of Factors.....	83
3.5	Ordered Data Sets.....	85
3.6	Parallel Kalman Filter Networks for Peak Purity Analysis .....	88
3.7.1	Theory.....	90
3.7.2	Experimental Section .....	103
3.7.3	Results and Discussion.....	105
3.7	Conclusions.....	118
<b>4</b>	<b>Limitations of EPCIA</b>	
4.1	Introduction.....	120
4.2	Experimental.....	121
4.3	Fundamental Limitations.....	123
4.3.1	Effect of Peak Shape .....	123
4.3.2	Effect of Spectral Correlation.....	127
4.3.3	Effect of Chromatographic Resolution .....	132
4.3.4	Effect of Concentration Ratio.....	134
4.3.5	Combined Effects.....	136
4.3.6	Experimental Results .....	140
4.3.7	Other Variables.....	140
4.4	Experimental Limitations of EPCIA.....	144
4.4.1	Systematic Deviations.....	146
4.4.2	Random Deviations.....	164
4.4.3	Effect of Systematic and Random Noise on EPCIA .....	170
4.5	Conclusions .....	186



<b>5</b>	<b>Self-Modeling Curve Resolution</b>	
5.1	Introduction.....	188
5.2	Traditional Methods for Mixture Analysis.....	190
5.2.1	Classical Least-Squares .....	190
5.2.2	Multicomponent Analysis .....	191
5.3	Factor Analysis Methods for Mixture Analysis.....	192
5.3.1	Target Factor Analysis.....	194
5.3.2	Self-Modeling Curve Resolution .....	196
5.3.3	Iterative Target Testing.....	198
5.4	Iterative Target Testing at EPCIA.....	203
5.4.1	Experimental .....	204
5.4.2	Effect of Chromatographic Resolution .....	205
5.4.3	Effect of Spectral Correlation.....	214
5.4.4	Effect of Concentration Ratio.....	215
5.4.5	Three-Component Systems .....	218
5.5	Conclusions.....	220
<b>6</b>	<b>Conclusions and Future Work</b>	
6.1	Conclusions.....	222
6.2	Future Work.....	224
	<b>Appendix A Program Listing for HP-PCA.BAS.....</b>	<b>226</b>
	<b>Appendix B Program Listing for Target.M.....</b>	<b>243</b>
	<b>References.....</b>	<b>250</b>

## LIST OF FIGURES

---

1.1	A two-component chromatogram with Gaussian peaks at $R_s=1.0$ .....	8
1.2	Hypothetical one-dimensional separation (top), and two-dimensional separation (bottom) of a ten component mixture.....	11
1.3	Schematic illustration of a symmetric digital filter. ....	15
1.4	Smoothing filters applied to a noisy signal.....	19
1.5	Derivative filtering of a noisy signal. ....	21
2.1	An overview of digital filters for state parameter estimation.....	25
2.2	The Kalman filter applied to the fourth point of a data sequence. The innovation is the difference between the measured value of this point and its predicted value .....	33
2.3	Summary of discrete Kalman filter equations.....	37
2.4	An example of least-squares fitting where the model is not consistent with all the data points.....	41
2.5	Strategy employed in implementing the parallel Kalman filter network .....	53
2.6	Effect of the data range on parameter estimates by the Kalman filter network .....	58
2.7	Effect of the number of data points on the precision of the parameter estimates.....	59
2.8	Effect of measurement noise on the precision of the parameter estimates.....	61
2.9	Typical reaction curves for the reduction of 12-molybdophosphoric acid to the heteropoly blue at three temperatures .....	62
2.10	Comparison of calibration curves obtained with (A) the Kalman filter network and (B) the fixed-time method at three different temperatures...	64

<b>3.1.</b>	<b>Pictorial representation of a bilinear data set. The data matrix <math>D</math> is the inner product of a matrix of pure concentration profiles, <math>C</math>, and pure spectral profiles <math>S</math>.</b> .....	<b>73</b>
<b>3.2.</b>	<b>One-component data are shown in two coordinate frames: (a) the measurement axes corresponding to absorbances at individual wavelengths <math>\lambda_1</math> and <math>\lambda_2</math>, and (b) the first two principal component axes <math>e_1</math> and <math>e_2</math>.</b> .....	<b>76</b>
<b>3.3.</b>	<b>Two-component data are shown in two coordinate frames: (a) the measurement axes, and (b) the first two principal component axes</b> .....	<b>79</b>
<b>3.4.</b>	<b>Result of principal components analysis on the data shown in Figure 3.1.</b> .....	<b>82</b>
<b>3.5.</b>	<b>Concentration profiles are shown for a data set with three absorbing components (top). The overall rank of this data is three, but the local rank varies from zero to three (bottom) depending on the elution time</b> ...	<b>86</b>
<b>3.6.</b>	<b>Strategy for implementation of parallel Kalman filter networks.</b> .....	<b>91</b>
<b>3.7.</b>	<b>Chromatographic elution profiles and their projections into <math>A^2</math> space</b> .....	<b>93</b>
<b>3.8.</b>	<b>Parallel Kalman filter network for recursive PCA application.</b> .....	<b>97</b>
<b>3.9.</b>	<b>Merging zones continuous flow apparatus for studies of dye coelution with spectra of dyes inset.</b> .....	<b>104</b>
<b>3.10.</b>	<b>Results of application of Kalman filter PCA algorithm to a single component elution profile (simulated).</b> .....	<b>107</b>
<b>3.11.</b>	<b>Results of application of Kalman filter algorithm to a simulated two-component elution profile.</b> .....	<b>108</b>
<b>3.12.</b>	<b>Evolution of the one-component model for the data in Figure 3.7c.</b> .....	<b>109</b>
<b>3.13.</b>	<b>Results of application of Kalman filter algorithm to a simulated two-component elution profile near limiting conditions.</b> .....	<b>111</b>
<b>3.14.</b>	<b>Results of application of Kalman filter algorithm to a simulated three-component elution profile.</b> .....	<b>113</b>
<b>3.15.</b>	<b>Absorbance matrix obtained from the coelution of organic dyes.</b> .....	<b>115</b>

<b>3.16.</b>	<b>Results of the application of the recursive PCA algorithm to the data in Figure 3.15.....</b>	<b>116</b>
<b>4.1.</b>	<b>Results of the application of the EPCIA algorithm to simulated two-component data. The top panels show the combined (solid line) and the individual (dashed lines) elution profiles for (a) Gaussian; (b) exponentially modified Gaussian; and (c) triangular profiles.....</b>	<b>124</b>
<b>4.2.</b>	<b>Results for the application of the EPCIA algorithm to simulated rectangular concentration profiles.....</b>	<b>126</b>
<b>4.3.</b>	<b>Effect of spectral correlation (Gaussian profiles) on the maximum rms innovation for two different chromatographic resolutions (<math>R_S</math>).....</b>	<b>129</b>
<b>4.4.</b>	<b>Effect of spectral correlation (PAH spectra) on the maximum rms innovation for spectral regions of 200-300 nm and 200-400 nm.....</b>	<b>131</b>
<b>4.5.</b>	<b>Effect of the chromatographic resolution on the maximum rms innovation obtained for two component mixtures. <math>\theta</math> is the angle between the two spectral vectors.....</b>	<b>133</b>
<b>4.6.</b>	<b>Dependence of the maximum rms innovation on the concentration of the minor component. ....</b>	<b>135</b>
<b>4.7</b>	<b>Results for Figure 4.6 plotted on a logarithmic scale for two different noise levels. ....</b>	<b>137</b>
<b>4.8.</b>	<b>Contour surface of the maximum rms innovation obtained for a minor component (10:1 ratio) as a function of chromatographic resolution and spectral angle. ....</b>	<b>139</b>
<b>4.9.</b>	<b>Spectrochromatogram of a mixture of phenanthrene and fluoranthene.</b>	<b>141</b>
<b>4.10.</b>	<b>Elution profile at 256 nm (top) and rms innovations sequence (bottom) for data in Figure 4.9. ....</b>	<b>142</b>
<b>4.11.</b>	<b>Nonlinear instrumental response is shown for a set of para-xylene calibration solutions. The measured absorbances at six wavelengths (see inset) are compared with the absorbances predicted by Beer's law.....</b>	<b>150</b>
<b>4.12.</b>	<b>Schematic diagram of a diode array spectrometer. ....</b>	<b>151</b>
<b>4.13.</b>	<b>The absorbances measure by the diode array for a calibrated photographic step tablet. ....</b>	<b>154</b>

4.14.	The calibration curve for para-xylene at 278 nm has large deviations from Beer's law (top). These errors are consistent with the effects of polychromatic and stray radiation modeled by Equation 4.8 (bottom)...	160
4.15.	Typical errors resulting from polychromatic radiation and stray light estimated from Equation 4.8. ....	161
4.16.	The difference between spectra acquired on the leading and trailing edges of the elution profile of phenanthrene .....	163
4.17.	The precision of absorbance reading versus sample absorbance for para-xylene solutions in a 1cm cuvette and a 30 $\mu$ L flow cell.....	166
4.18.	Noise power spectra for absorbance readings of 2 AU (top) and 0 AU (bottom) calculated from 2048 replicate 100 ms readings. ....	168
4.19.	Stopped flow apparatus used to evaluate the effect of systematic and random noise on EPCIA.....	172
4.20.	Absorbance spectra of methyl orange and praseodymium chloride in acidic solution. Symbols mark the wavelengths used for EPCIA .....	173
4.21.	Elution profile of methyl orange (top), and rms innovation sequences for one (solid line) and two (dashed line) component models.....	175
4.22.	Largest innovations from one- and two-component models plotted for each data matrix against the maximum absorbance. The expected magnitude of the random noise ( $3\sigma$ ) is given for comparison.....	176
4.23.	Results from PCA of the data shown in Figure 4.21. The scores of the first and second principal component are shown.....	177
4.24.	Innovations from the one-component model applied to $\text{PrCl}_3$ ; $\text{PrCl}_3$ between 436 and 452 nm; and methyl orange. ....	180
4.25.	Elution profile of methyl orange with the neutral density filter in place (top). The effect of heteroscedastic noise on the innovation sequences is shown (bottom) for one- and two-component models. ....	183
4.26.	Results from PCA of the data shown in Figure 4.25. The scores of the first (dashed line) and second (solid line) principal components are shown. ....	184
4.27.	Largest innovations from the one- and two-component models plotted for each data matrix for (a) 0.1 s and (b) 4.0 s integration.....	185

5.1.	A schematic illustration of the iterative target transform .....	199
5.2.	Starting profiles generated by EPCIA, varimax rotation and needle search are compared to the true concentration profiles of an overlapped ( $R_s = 0.35$ ) two-component peak .....	206
5.3.	The effect of chromatographic resolution on the starting profiles. Profiles generated by EPCIA, varimax rotation and needle search compared to the true concentration profiles: (a) sum of squared residuals, (b) systematic error in peak width, (c) systematic error in retention time .....	208
5.4.	Concentration profiles predicted by ITT (solid line) compared with the true profiles (dashed line) for a two-component peak with $R_s = 0.2$ . .....	211
5.5.	Results from ITT on a two-component peak with $R_s = 0.1$ : (a) the extracted spectrum of phenanthrene, (b) the true spectra of phenanthrene and triphenylene .....	212
5.6.	The concentration profile predicted by ITT for the major component compared to the true profiles .....	217
5.7.	The concentration profiles predicted by ITT compared to the true profiles. Results from EPCIA and needle search starting profiles are shown.....	219

## LIST OF TABLES

---

1.1	Summary of digital filters.....	23
3.1.	Comparison of eigenvectors produced by traditional principal components analysis and the Kalman filter method.....	117
4.1	Results from fitting Equation 4.4 to the para-xylene calibration data.....	153
4.2	Results from fitting Equation 4.7 to the para-xylene calibration data.....	158
4.3	Results from fitting Equation 4.8 to the para-xylene calibration data.....	159
4.4	Improvements in S/N from ensemble averaging a 2 AU signal.....	169
4.5	Spectral derivatives of the samples studied.....	181
5.1	The effect of chromatographic resolution on the concentration profiles predicted by ITT is given for each type of starting profile.....	209
5.2	The effect of chromatographic resolution on the pure spectra profiles predicted by ITT is given for each type of starting profile. The SSR is the sum for both components.....	213
5.3	The effect of spectral resolution on the concentration profiles predicted by ITT is given for EPCIA and varimax starting profiles.....	214
5.4	The effect of concentration ratio on elution profiles predicted by ITT of EPCIA starting profiles.....	216

## ABSTRACT

---

This work presents new methods for extracting information from multivariate data sets based on the Kalman filter, a digital filter for recursive estimation of parameters associated with a linear model. Parallel Kalman filter networks are used to take advantage of the diagnostic properties of the Kalman filter, namely its ability to detect the extent and nature of modeling errors in real-time.

The potential of the network is demonstrated for reaction-rate methods of analysis, using data from the molybdenum blue method for the determination of phosphate. These Kalman filter models, implemented in a quasi-continuous form, describe first-order reactions with a range of rate constants. The best model for a given set of data is selected by examining the innovation sequences. This algorithm successfully corrects for errors arising from variations in the pseudo-first-order rate constant yielding improved concentration estimates.

The ability of Kalman filter networks to perform recursive principal components analysis (PCA) is also demonstrated. Application to absorbance matrices such as those in chromatography with multisensor detection are considered. This network contains discrete models for describing one- and two-component bilinear responses. The model deviations can be used to elucidate the rank of the data set such that peak purity detection can be performed in real-time using an algorithm called evolving principal components innovations analysis (EPCIA). The fundamental and experimental limitations of this approach are examined for unresolved mixtures in liquid chromatography with UV-visible detection. The results can indicate the presence of minor impurities. Furthermore, the innovation sequences estimate the elution profiles of the individual components. These profiles were further refined with the iterative target transform.



## NOTATION

---

The discussions that follow will obey these conventions:

$x$  is a scalar value (i.e. a number) represented by a lower case letter, except for established symbols, like  $A$  for absorbance, listed on the following page.

$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$  is a vector, represented by a bold lower case letter.

$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \end{bmatrix}$  is a matrix, represented by a bold upper case letter.

The  $\hat{\phantom{x}}$ , called 'hat', above a quantity signifies an estimated quantity, e.g.,  $\hat{\mathbf{X}}$ .

The superscript  $^{-1}$  indicates the inverse of a matrix, e.g.,  $\mathbf{X}^{-1}$ .

The superscript  $^T$  indicates the transposed matrix or vector:

$$\mathbf{X}^T = [\mathbf{X}]^T = \begin{bmatrix} x_{1,1} & x_{2,1} & x_{3,1} \\ x_{1,2} & x_{2,2} & x_{3,2} \end{bmatrix}$$

## ABBREVIATIONS AND SYMBOLS USED

---

A	absorbance
$A_{meas}$	measured absorbance
$A_{pred}$	predicted (Beer's law) absorbance
b	sample pathlength
B	background absorbance
c	concentration
C	matrix of concentration profiles
CLS	Classical Least-Squares
$e'$	orthogonal innovations
E	error matrix
EFA	Evolving Factor Analysis
ECR	Effective Concentration Ratio
EMG	Exponentially Modified Gaussian
EPCIA	Evolving Principal Components Innovation Analysis
f	relative contribution of the stray light
FA	Factor Analysis
FWEFA	Fixed Window Evolving Factor Analysis
GC	Gas Chromatography
HPLC	High Performance Liquid Chromatography
ITT	Iterative Target Transform
LC	Liquid Chromatography
MCA	Multicomponent Analysis
MS	Mass Spectrometry
$n_c$	number of observable components
$n_p$	number of principal components
$n_s$	number of samples (e.g. spectra)
$n_w$	number of wavelengths
PAH	polyaromatic hydrocarbon
ppm	parts per million
IR	infrared
k	rate constant
P	radiant power

PCA	Principal Components Analysis
$R_s$	chromatographic resolution
rsd	relative standard deviation
rms	root mean squared
S	matrix of pure spectral profiles
SMCR	Self-Modeling Curve Resolution
SSR	Sum of Squared Residuals
S/N	signal to noise ratio
T	Transmittance
$\mathbf{T}$	transformation matrix
TFA	Target Factor Analysis
$t_R$	retention time of a chromatographic peak
$t$	time
UV	ultraviolet region (200 to 380 nm)
Vis	visible region (380 to 780 nm))
$\mathbf{X}$	scores matrix (abstract chromatograms)
$\mathbf{Y}$	loadings matrix (abstract spectra)
W	baseline width of a peak
$\Delta$	spectral bandpass of the spectrometer
$\epsilon$	molar absorptivity
$\Lambda$	Squared eigenvalue matrix
$\lambda$	wavelength
$\theta$	angle (degrees)
$\sigma$	standard deviation
$\tau$	1 / rate constant.

#### **Kalman filter symbols**

e	innovation
$\mathbf{H}$	observation matrix
$\mathbf{K}$	Kalman gain matrix
$\mathbf{P}$	error covariance matrix
$\mathbf{Q}$	model noise covariance matrix
$\mathbf{R}$	measurement covariance matrix
$\phi$	state transition matrix
$\mathbf{v}$	measurement noise vector

<b>w</b>	system noise vector
<b>x</b>	state vector
<b>z</b>	measurement vector

## ACKNOWLEDGMENTS

---

My research would have never been successful without the assistance of many supportive individuals, to whom I am enormously grateful. Foremost, I would like to thank Dr. Peter Wentzell for the opportunities he has provided me. His ideas and inspiration made this project possible and his guidance and dedication brought it to completion. Also, I would like to acknowledge past and present members of his research group who have made significant contributions to this project, especially Stephen Hughes, Nils Sundin, Kevin Bateman and Deborah Ford.

Members of the Trace Analysis Research Centre, in particular Dr. Robert Guy and Dr. Louis Ramaley, are thanked for endless equipment loans, and insights on matters academic and otherwise. In addition, I am indebted to the Analytical Chemistry Group at the Institute of Marine Bioscience for training and instrument time, particularly Dr. Robert Boyd, Pearl Blay and Dr. Steve Pleasance.

Financial support for this work was provided by the Natural Science and Engineering Research Council. I am deeply grateful to Dalhousie University, the Killam Trust, and the Sumner Foundation for graduate fellowships.

Finally, heartfelt thanks go out to my family (for the unconditional support), Dean and Elizabeth (for a comfortable floor), Dan (for being a lifelong friend), and my favorite dinner guests Karl, Donna, Jane and Rob ( 'sometimes you'll find the light' ). I couldn't have done it without you.

## INTRODUCTION

---

### 1.1 MULTIDIMENSIONAL METHODS

Analytical chemists are often called upon to provide quantitative and qualitative assessments of exceedingly complex mixtures, such as those found in environmental or biological samples. This assessment involves three major stages: (1) the design of an experiment, (2) the measurement of an analytical signal, and (3) the extraction of information from the data. Questions of a larger scope can be answered with a combination of this information and previous knowledge. For many complex instruments, the importance of the measurement process alone is typically overestimated, as it is the chemist's ability to select appropriate operating conditions and interpret the wealth of data that finally determines the success of the analysis. This work examines how chemical information is best extracted from the measurements obtained from modern analytical instruments.

The purpose of each of these stages is illustrated here for the analysis of a water sample. In this hypothetical example, the sample is analyzed to learn more about the system under study, namely the water supply of a town. In the experimental design stage, we should decide what properties of the sample will be measured, and consider what is already known about the system. Then, a clear goal is set for the experiment; for example, determining the concentration of mercury in a sample that contains a number of other ions. Next, an analytical method is designed. For this example, it may be the atomic absorption of mercury measured at a suitable wavelength. The actual measurement is an

intensity. At this stage, the analysis is still incomplete. In this experiment, like most in analytical chemistry, the analyte concentration is not measured directly. Instead, the property of interest (mercury concentration) must be estimated indirectly. Thus the final stage is one of information extraction, where the raw data generated by the experiment are used to answer questions about the sample. Two common tools in this final stage are mathematical transforms and models. One example of a transform is the logarithm that converts measured intensities into absorbances. Others include averaging, differentiation, and normalization. The next step in our experiment is predicting a concentration from the calculated absorbance. This involves a model that incorporates prior knowledge of the relationship between the measured signal and the property of interest. The model can be derived from theoretical considerations, as well as measurements of calibration samples. The accuracy of this estimate depends on the model's validity for this particular sample. Similarly, providing error bounds for the estimate requires an understanding of the measurement noise, and how it propagates through all the calculations. In summary, the "quality" of the estimate depends on the amount of information that the experiment provides, and our ability to reliably extract that information from the noisy measurements.

Unlike the atomic absorbance experiment, the analytical techniques used in this work produce more than one datum per experiment. In fact, experiments such as those that combine chromatography and spectroscopy produce a data matrix, which often contains thousands of points. Although many more points are taken, the end goal of this analysis is still to provide specific answers about the chemical system being measured. At this point, one might ask what motivates chemists to produce these seemingly unmanageable data sets? This thesis will demonstrate advantages of multidimensional techniques including:

1. They are efficient. Multiple concentrations can be estimated with the results of a single experiment. Furthermore, complex properties that depend on multiple chemical species can be estimated with such measurements. One example is the near-IR spectroscopic analysis of grain for protein, starch, and moisture content<sup>1</sup>.
2. The precision of the estimates is improved by using multiple points of the data<sup>2</sup>. For instance, concentrations can be estimated from entire analyte spectra, rather than the absorbance at a single wavelength.
3. The accuracy of multipoint methods is often better than single point methods, since procedures can be designed to detect, and correct for, interferences that might otherwise invalidate a calibration model<sup>3</sup>.

This work demonstrates how digital filters, particularly the Kalman filter, enhance these properties of multidimensional methods.

## 1.2 INTELLIGENT INSTRUMENTS

Computers and semiconductor electronics have profoundly influenced the practice of analytical chemistry, altering the techniques of making measurements and interpreting them. A perfect example is the comparison of a scanning spectrometer to the diode array spectrometer used in this work. Scanning spectrometers, the more traditional design, have a single detector that measures the intensity at one wavelength interval chosen by a monochromator. To record a spectrum, the monochromator is scanned to sequentially measure the intensities over a range of wavelengths. In contrast, a diode array spectrometer can be built with no moving parts. Instead, a multichannel detector is used, with



each channel fixed at one wavelength interval of the spectrum. Each of the several hundred detectors is a photodiode. Remarkably, this entire diode array and its associated electronics fit on a chip the size of a postage stamp.

This spectrometer is also an example of an "intelligent" instrument, as it has a central processor that coordinates its operations. The instrument can control many aspects of the experiment including sample loading, measurement timing, error checking, and signal processing. The resulting measurements are stored on a laboratory computer for viewing and further analysis. Such automated instruments have freed the chemist from the labor of taking repetitive measurements. Besides, the computer is better for such tasks, as its timing is precise, its operations are repeatable, and its results are free of transcription errors. Consequently, the chemist's time can be better spent on the experimental design and interpretation stages.

Computerized instruments, like the diode array, generate data at a tremendous rate. The problem arising in their use is the data alone are not information and it is our interpretative abilities that often limit us. "The added mass of data is, at best, underused and, at worst, tends to obscure the hidden information rather than clarify it."<sup>4</sup> Furthermore, the time saved by using these "efficient" instruments is easily negated if the results are presented in an indecipherable form. This work uses digital filters to transform and interpret chemical measurements. These filters run on the same laboratory computers that store the digital measurements.

At the end of an analysis we want to have estimated properties that summarize the data. Typically, these are properties that identify or quantify the sample. The connection between the data and this information is not always obvious. Earlier, it was suggested that data can be transformed and combined

to efficiently produce these estimates. But what transformation or combination of the data should be used? It would be impossible to try every possibility, and even if we did, how would we define the "best" method? Chemometrics<sup>5-8</sup> considers questions like these. Massart has defined chemometrics as: "the chemical discipline that uses mathematical, statistical, and other methods employing formal logic (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum relevant chemical information by analyzing chemical data."<sup>9</sup> Chemometrics adapts methods from many fields, such as the digital filters used in this work, which originated in engineering.

The greatest benefits are gained when chemometrics is employed at every stage of the analysis. While techniques such as factor analysis can find hidden relationships in the data and solve difficult calibration problems, they don't create information. Simply put, the amount of information that comes out of an experiment cannot exceed the amount available from the data. For this reason, experimental design and response surface methods<sup>10</sup> are important areas of study. Another early consideration is choosing calibration samples<sup>11</sup> that result in accurate and precise results for the unknown samples.

This work has two themes. First, that the chemist's ability to interpret chemical data can not be replaced by a computer: "It is often underestimated by chemometricians what a good spectroscopist can do without us. We should never under-estimate the pattern recognition capabilities of people."<sup>12</sup> Often the results just need to be compressed or recast into a more interpretable form. With this in mind, the results from Kalman filtering of chromatographic data will often be presented in terms of prediction errors versus time such that the filtering results have similar features to a chromatogram. As will be seen in Chapter 3, interfering peaks produce characteristic features on this plot that are readily

interpreted by the chemist. The second theme of this work is that the most powerful and reliable methods are those which incorporate prior knowledge of the sample, experimental system, and measurement noise. The difficulty is usually in transferring the chemist's intuitive knowledge of these into a more mathematical form. This transformation will be explored in Chapters 4 and 5.

In summary, modern instruments are potentially very powerful, but the raw data are rarely used directly. Chemometric methods attempt to realize the potential of these instruments. The digital filters used in this work extract reliable estimates of desired parameters from chemical measurements. Choosing the best chemometric technique is like choosing the best analytical instrument - without considering the analytical problem under study the question is meaningless. Each technique has advantages and limitations. As in any field, this choice can be difficult for a new user due to the cryptic terminology and exaggerated claims of the literature.

### 1.3 CHEMOMETRICS AND INSTRUMENTATION

The amount of information an instrument can provide depends, ultimately, on the nature of the analytical signal produced. To generalize these abilities, instruments can be categorized as zero-, first-, or second-order<sup>13,14</sup> by the form of the resulting data. Examples of zero-order instruments are pH meters, single wavelength spectrophotometers, and any other instrument that produces a single datum per sample. The signal produced by such an instrument is typically related to analyte concentration with a calibration curve, that is used to predict the concentration of one chemical component. However, a zero-order method is incapable of detecting or correcting for interferences. This is a severe limitation

when it is not possible to design an instrument that is specific to the analyte, or to completely remove the interfering species from the sample.

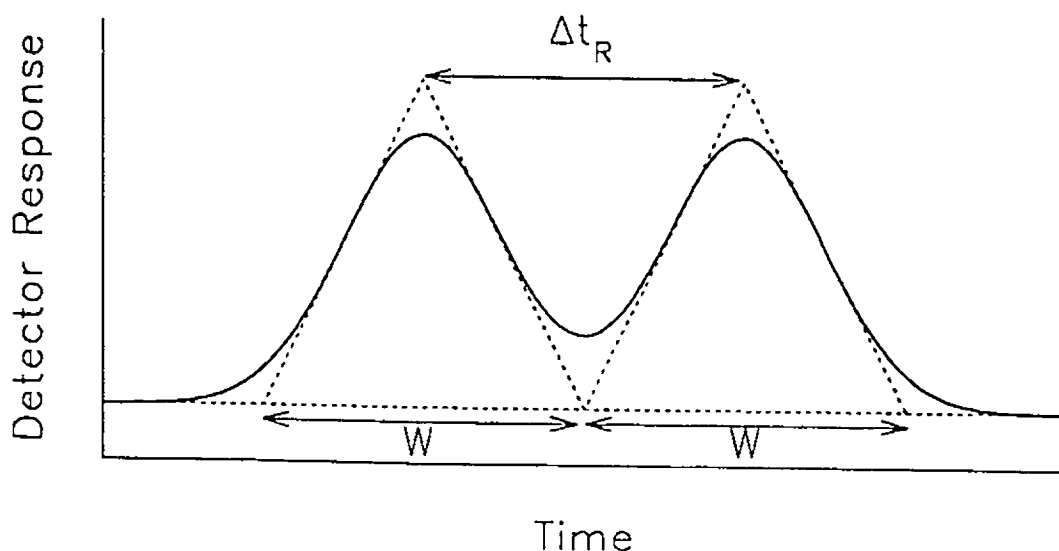
To deal with more complex samples, first-order techniques producing multiple analytical signals per sample may be used. These signals are reported as a function of a variety of variables such as time (chromatography and reaction-rate methods), energy (spectroscopy), and potential (electrochemistry). In chromatography, a physical separation of the analytes is necessary, while in spectroscopy the analytical signals are separated by a monochromator. As a result of this separation, these techniques can estimate multiple properties (concentrations) from the results of a single experiment.

First-order methods provide information by distinguishing among chemical species. Chromatography is a good case to consider, as the pure analyte signal is generally a single peak. In the following discussion, we will assume a constant peak shape and width along the chromatogram. The goal is to estimate the "peak capacity" of a column as the maximum number of resolvable components in a chromatographic run. Chromatographic resolution is defined as<sup>15</sup>

$$R_s = \frac{\Delta t_R}{W} \quad (1.1)$$

Where  $\Delta t_R$  is the difference in retention times of two peaks and  $W$  is the baseline width of the peaks. Figure 1.1 illustrates two Gaussian peak with  $R_s = 1.0$ . For a Gaussian peak with a standard deviation of  $\sigma$  Equation 1.1 is equivalent to:

$$R_s = \frac{\Delta t_R}{4\sigma} \quad (1.2)$$



**Figure 1.1** A two-component chromatogram with Gaussian peaks at  $R_s=1.0$ .

In this discussion, two peaks are considered as resolved at  $R_s = 1.0$ , that is when they are separated by at least one peak width.

The maximum number of components is separated when the peaks elute sequentially, at equally spaced intervals. With this approximation, Giddings<sup>16</sup> estimated typical peak capacities of 10 for gel methods, 50 for liquid chromatography and 200 for gas chromatography. This provides a rough estimation of the information content of these techniques. The information content of other methods, including the common spectroscopic techniques, have been estimated by Crozier and Reeve<sup>17</sup>. Thus, the resolving power of an instrument can be estimated when some assumptions are made about its typical signals.

Another consideration in first-order methods is efficiency: the amount of information produced divided by the time for the experiment. In chromatography, the length of the experiment is set by the time needed for a physical separation.

For any column, there is a compromise between obtaining rapid analyses and maintaining separation. By comparison, spectroscopic measurements can be very rapid, particularly for instruments that observe all the elements of a signal simultaneously. This category includes UV-visible spectrometers with diode array detectors and Fourier transform infrared spectrometers.

At this point, one might ask why a more complex instrument is necessary to solve analytical problems. To answer this, the problems arising in qualitative and quantitative assessments of mixtures will be reconsidered. As the set of potential compounds in a mixture grows, the odds that they can be differentiated with a first-order instrument decrease rapidly. Using chromatography as an example, the certainty in assigning a peak from a single retention time is limited. Similarly, it is difficult to quantify a peak consisting of two overlapped compounds, or even detect when this is occurring. These problems can be minimized by providing more information on each component of the sample, as well as a greater separation between components. To achieve this, second-order instruments rely on more than one method of separation<sup>18</sup>.

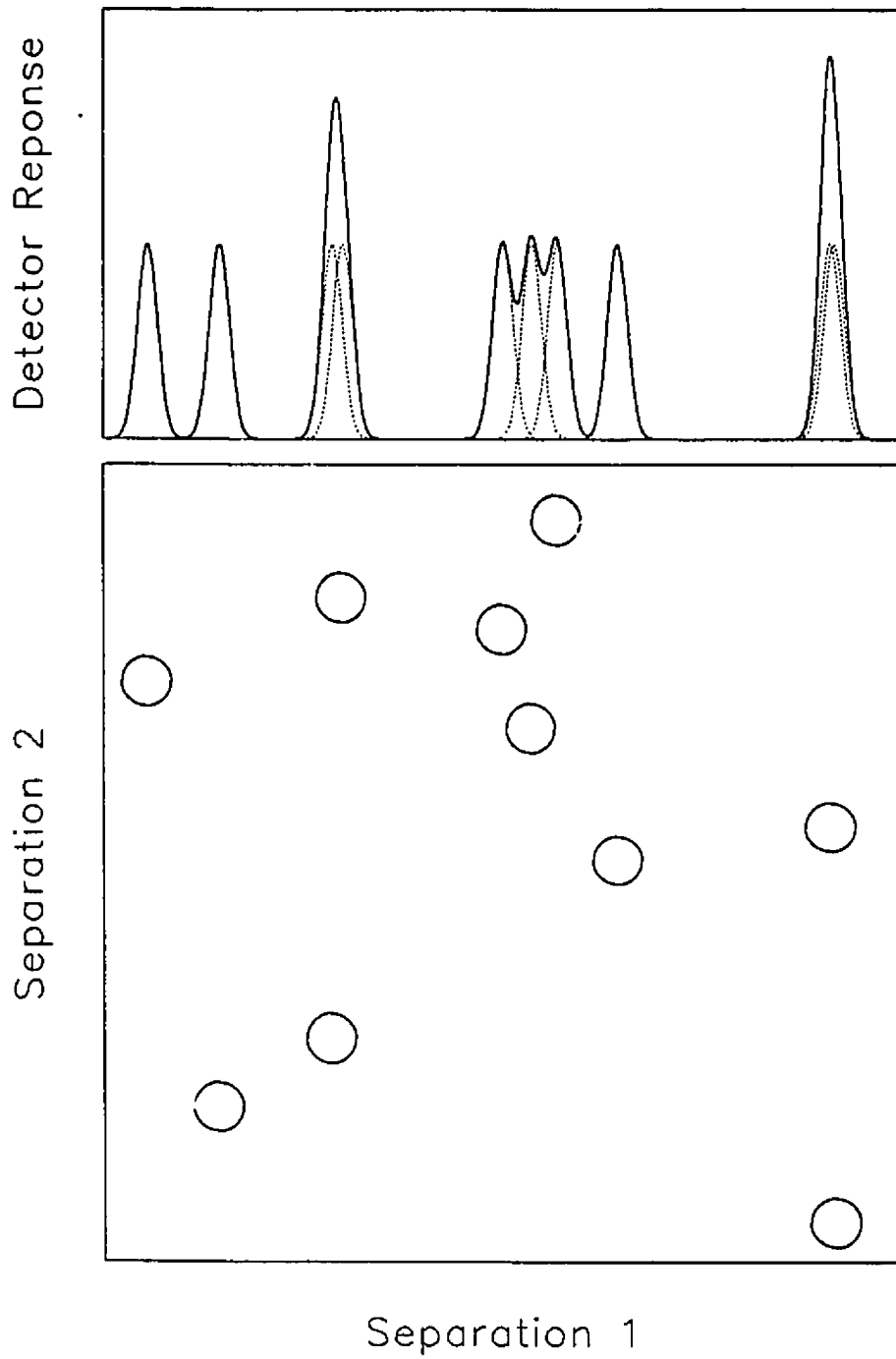
One of the best illustrations of a second-order instrument is gas chromatography combined with mass spectrometry (GC/MS), which first separates a mixture of components by retention time, and then separates by molecular mass. The resulting data forms a matrix. It is important to note that the two separation methods are different, and the separation generated in the first stage is maintained in the second. When the above conditions are met, the gain in peak capacity is multiplicative<sup>19</sup>. In contrast, when the two first-order techniques are used individually, the gains are additive. In practice, the gain in peak capacity falls somewhere in between these limits. The limiting factors in achieving a multidimensional separation are difficulties in obtaining truly

orthogonal separations, and in maintaining the first separation while achieving second separation.

Davis and Giddings<sup>20</sup> have developed a statistical theory of component overlap in multidimensional chromatograms. A pictorial representation of this is given in Figure 1.2. The top panel shows 10 peaks that are randomly distributed along the first axis of separation. Twenty single-component peaks could fit across this axis if they were spaced the right distance apart, that is, we are only working at half the peak capacity of the method. Despite this, the majority of the peaks are unresolved. Rosenthal has also shown that both in theory and in practice that: "the occurrence of overlapping components is far more prevalent than may have been previously realized"<sup>21</sup>. For comparison, the bottom panel show the same peaks after another random separation along a second axis. Here, the peaks are easily separated. A similar improvement could not be expected from simply doubling the number of theoretical plates for the first separation.

## 1.4 SIGNALS, NOISE AND FILTERS

The previous section emphasized that first- and second-order methods separate the signals from different chemical species. If an analyte signal is completely separated from other signals a univariate calibration can be used. Where the signal is a spectroscopic peak, the analyte concentration could be estimated from the absorbance at the peak maximum. The need for multivariate techniques, such as digital filters, becomes apparent when a more realistic view of a measurement is considered. Even in a well-designed experiment, measurements will contain noise as well as the pure analyte signal. Random



**Figure 1.2** Hypothetical one-dimensional separation (top), and two-dimensional separation (bottom) of a ten component mixture. See text for details.



noise causes differences among repetitive measurements. The random nature of this noise implies that its sign and magnitude vary over a sequence of measurements. This noise is reduced with averaging. In contrast, no amount of averaging can remove systematic (also known as cyclic) noise, since it repeats itself during each experimental cycle. This noise can result from instrumental artifacts or interfering chemical signals. One instrumental example of systematic noise is when the excitation source in a fluorescence experiment contaminates the analyte emission. This is a systematic interference, since the same noise would appear with each measurement of the sample. Chemical sources of systematic noise are also called interferences. They result when the instrument responds to chemical species other than the analyte, that is, when the instrument is not totally selective towards the analyte. The likelihood of this occurring increases for complex mixtures. Both random and systematic noise tend to obscure the analyte signal, making it difficult to identify and quantify.

Since random noise limits the precision of our estimates, it should be minimized through experimental design and signal processing. All attempts to extract information from a signal contained in noise are based on assumptions of how the two differ. A common model is white noise, defined to contain all frequencies in the same way that white light contains all the colors. Accordingly, some frequencies of this noise can be filtered out with no effect on the signal, just as a colored filter can remove portions of white light. While the sign and magnitude of each deviation is by definition unpredictable, a large collection of these can be used to characterize the noise. The magnitude of white noise has a Gaussian distribution<sup>22</sup>. This distribution is symmetric with a maximum at zero deviation, implying that the effect of random noise is as likely to be positive as negative. Another feature of this distribution is that its second central moment

can be estimated as the standard deviation. This allows us to define the signal to noise ratio (S/N) as the mean signal divided by the standard deviation.

White noise, and its associated Gaussian distribution, is only one possible form of noise. In practice, noise arises from many sources. These include the observation of a random event, like radioactive decay, (a Poisson distribution) and fluctuations in the instrumental response ( $1/f$  or flicker noise). Each instrument has its own noise characteristics. For example, the noise in a spectrometer results from fluctuations in the source intensity and the dark current of the detector, as well as noise from electronic components like amplifiers. Specific knowledge of these noise characteristics will help to design strategies for their removal.

If the measurement consists of a single datum, then there is no possibility of correcting for random noise, for although the overall character of the noise can be defined, there is no way of knowing its influence on a single value. At the other extreme, the average contribution of white noise is zero for an infinite number of measurements. The compromise is to repeat the experiment a number of times with the goal of reducing rather than eliminating the effects of noise. This process, called ensemble averaging<sup>23</sup>, involves summing repetitive measurements. Assuming they are constant for each measurement, the analyte signals will add coherently. In contrast, the repeated values of the random noise tend to cancel each other, so that its effect converges towards zero. In this case, ensemble averaging improves the S/N by up to  $n^{1/2}$ , where  $n$  is the number of replicate measurements. In summary, ensemble averaging a series of repetitive measurements will improve the precision of the estimated signal.

Ensemble averaging illustrates some desirable qualities of a noise-reduction method. Notably, it minimizes the influence of random variations

without degrading the underlying information of the signal. Unfortunately, it is only suitable for repeatable experiments. Generally, this limits its application to stable, nondestructive methods of analysis, and experiments where repetitive measurements can be obtained on a reasonable time scale. Consider that if the precision after 10 measurements is inadequate, then 30 or more measurements are needed to double the S/N.

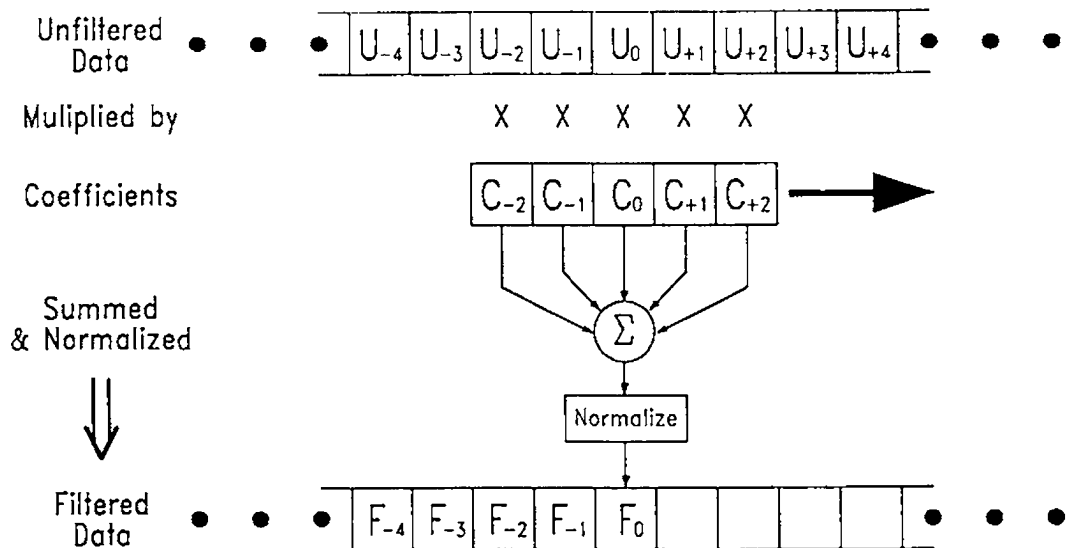
By definition, each higher-order measurement contains many data points. For example, one measurement of a sample can produce an entire spectrum. With some assumptions about the properties of the signal and noise, the S/N of this data can be improved by digital filtering. Unlike ensemble averaging, filtering can be applied to the signal measured in a single experiment.

The word 'filter' invokes many images to a chemist. One would certainly think of filters that remove large impurities from liquids. There are also optical filters that attenuate selected regions of light. To design a filter that passes analytical signal but suppresses noise, there must be some means of discriminating between the two. One way of discriminating is by frequency - assuming that the noise is at high frequencies, while the signal consists of mostly low-frequency or DC components. Filters that eliminate high frequencies are called smoothing or low-pass filters. For an electronic signal, this noise reduction is often accomplished with an electronic filter constructed from resistors and capacitors. As digital signals have become common in such diverse fields as analytical chemistry, music, and communications, digital filters have been developed for many operations traditionally done with electronic ones, including smoothing, differentiating, and integrating<sup>24</sup>.

A digital filter is an algorithm in which coefficients are convolved with a series of data points to obtain a new estimate of a particular data point

(Figure 1.3). The simplest example is a moving average filter, which can be used to smooth data collected at evenly spaced intervals. For example, each smoothed data point could result from the convolution of seven raw data points with the seven coefficients of the filter, divided by a normalization factor (although some texts define this operation as digital smoothing rather than digital filtering, no such distinct will be made in this work). The coefficients for this filter are all equal to unity, and the normalization factor is seven. After each smoothed data point is calculated from the data in the window, the filter moves forward, taking in one new point at the beginning and dropping the oldest point off the end before repeating the calculation.

The moving average filter described above works best when the signal being filtered does not change significantly within the seven-point window. When the signal is varying rapidly, peak shapes and heights will be distorted



**Figure 1.3** Schematic illustration of a symmetric digital filter.

causing information to be lost. For many data sets, a smooth curve better approximates the underlying function. Accordingly, polynomial functions are often used to smooth localized regions of the data set. For example, a subset of seven data points could be fitted to the quadratic function

$$y(x) = a_2x^2 + a_1x + a_0 \quad (1.3)$$

For a number of reasons, including random errors in measuring  $y$ , it is unlikely that such a curve will go through all the data points in the window. Still, the parameters  $(a_2, a_1, a_0)$  should be adjusted to give the greatest agreement between the predicted and actual points. For normally distributed errors, this "best fit" is obtained by a least squares solution<sup>25</sup>. A variety of computational methods are available to find this, the oldest treatment was developed by Gauss in 1795 for astronomical data<sup>26</sup>.

More recently, Savitzky and Golay<sup>27</sup> introduced digital filters based on such least squares solutions to analytical chemistry in 1964. They used the central point of this polynomial curve as the smoothed point, since it combines information from the original point and neighboring points based on a polynomial model. Furthermore, they showed that the value of this smoothed point can be calculated by a digital filter to give a least squares solution to this problem. For example, a seven-point quadratic filter would have the coefficients  $[-2, 3, 6, 7, 6, 3, -2]$ , with 21 as the normalizing factor. In general, this approach can be used for polynomial functions over any odd number of points. While such filters do not increase the amount of information in the data, filtering can improve the S/N of a peak.

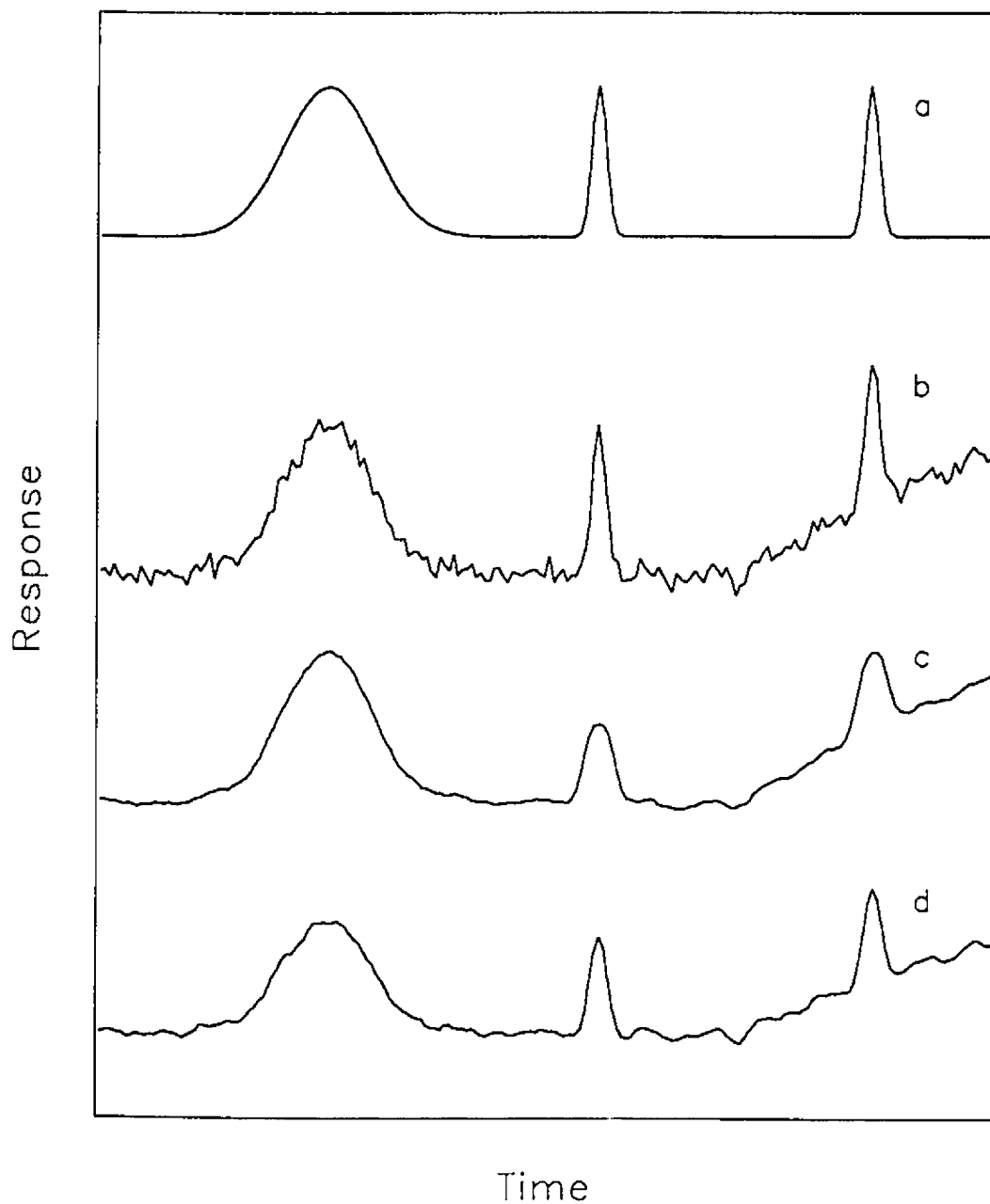
The choice of signal processing method depends on the prior knowledge available. At one extreme is ensemble averaging, which assumes nothing about

the signal except that it is repeatable. This method provides the best noise reduction under conditions of no signal distortion. Digital filters are often more useful since they can be applied to non-repetitive signals. These filters estimate the noise-free signal by using temporal correlation in the measurements. A convolution function defines how the data in the window are combined, thus determining the properties of the filter. Convolution functions were presented for low-pass filters, which suppress the high frequency components of the measurement. These filters incorporate more prior knowledge than ensemble averaging, as they assume the analytical signal is composed mostly of low frequency components. When this is true, these filters can successfully estimate the smooth underlying signal. If high-frequency components are also associated with the signal (sharp features) then excessive smoothing will distort the signal, resulting in a loss of information. Thus smoothing is a compromise between removing noise and changing the signal's shape.

The other extreme in prior knowledge is when the true form of the signal is already known. In this case, the problem is one of quantitation and detection rather than identification. Accordingly, the goal is to maximize the S/N, even at the expense of distorting the signal. This is achieved with a *matched filter*<sup>28</sup>. The convolution function of the matched filter is simply the noise-free signal. This is the optimal filter for a signal contaminated by white noise. This is reasonable, considering that (1) every point in the measurement is used, (2) points where the signal is strongest are given the highest weighting, and (3) points with only noise are given a weighting of zero. Thus, matched filters achieve the highest S/N ratio improvement, but require the most prior knowledge.

The methods discussed so far remove random noise from chemical measurements. This noise reduction increases the precision of properties derived from these measurements. All of these methods decrease the influence of random noise, but they are not designed to remove the systematic noise that often limits the accuracy of our estimates. Figure 1.4 illustrates how smoothing affects a noisy signal. This example contains both random noise and systematic noise (in the form of a sloping background on the last peak). Applying a moving average filter (seven point window) to this measurement reduces the random noise, but the systematic noise is essentially unchanged. Also note that this filter reduces the height of the narrow peaks and significantly distorts their shape. The wider peak is also distorted, but to a lesser degree since it contains mostly lower frequencies. The quadratic smoothing filter (seven point window) introduces less signal distortion, but is less effective in reducing the random noise. There are many types of digital filters for smoothing, and choosing among them is not a trivial problem<sup>29,30</sup>.

Unlike random noise, systematic noise repeats itself with every cycle of measurement. Thus, ensemble averaging will not reduce systematic noise or the biases it produces. The ability to detect and compensate for systematic noise depends on the nature of the measurement. Notably, zero-order methods are incapable of detecting or correcting for it. Their calibration models assume a totally selective response, but have no way of determining when this model fails. A major advantage of higher-order methods is their ability to detect systematic noise. An example is a first-order calibration based on a full spectrum of the analyte. Interferences could be detected by the unexpected peaks they produce in the measured spectra. Of course, this assumes that the analyte and interference don't have an identical response at every wavelength. In other



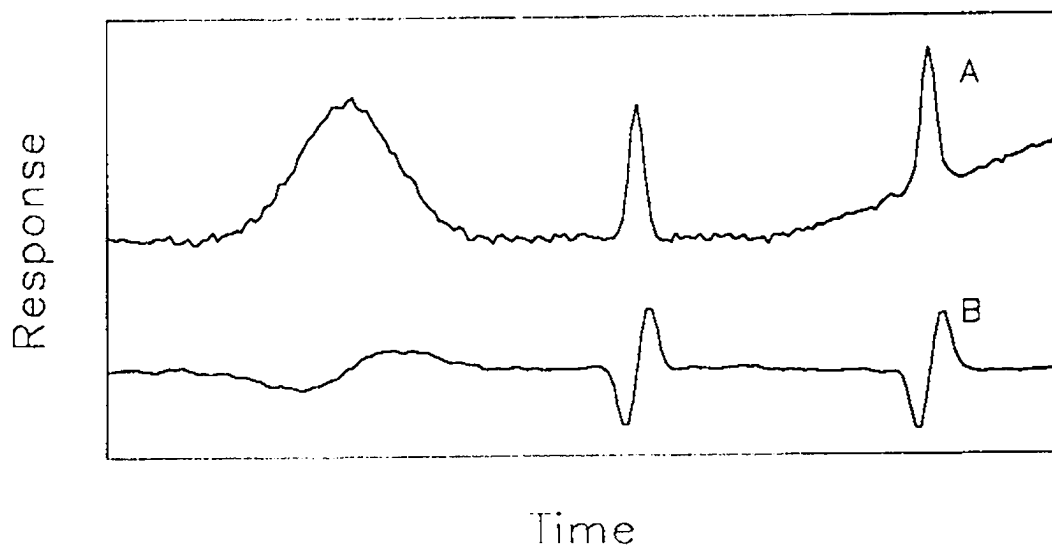
**Figure 1.4** Smoothing filters applied to a noisy signal: (a) noise-free signal; (b) signal plus random and systematic noise; (c) output from a moving average filter (seven point window); (d) output from a quadratic smoothing filter (seven point window).



words, the instrument has at least a partial selectivity to these species. Second-order instruments can detect systematic noise in a similar way, except using both dimensions of separation.

When systematic noise is a problem, there are two possible solutions: (1) an experimental approach that physically separates the analyte from interferences, and (2) a mathematical approach, such as digital filtering, that makes corrections at the data analysis stage. An example of the second approach is the use of derivative filters<sup>31</sup>. These are conceptually similar to smoothing filters in that they distinguish between the signal and noise by frequency. The difference is that derivative filters attenuate the low frequencies of the signal. For instance, a flat background has a derivative (slope) of zero, but a relatively sharp peak superimposed on this background has a significant derivative on its leading and trailing edges. Although the output of the derivative filter does not resemble the analyte signal, it can still be used for quantitative and qualitative purposes. When the signal has a linear relationship with concentration, then the derivative of the signals will also change linearly. Furthermore, the derivative signals are relatively unaffected by broad or sloping background. In practice, derivative filters are most suitable for dealing with the low-frequency systematic noise such as instrumental artifacts.

Figure 1.5 shows the effect of a derivative filter (seven point quadratic) on a signal contaminated with random and systematic noise. The second and third peaks give very similar signals in the filtered data, despite the sloping background. Also note that broad peak's intensity is attenuated, as it contains mostly lower frequencies. This is the opposite of what was observed in Figure 1.4 for the smoothing filter. The derivative filter is successful in certain applications because the systematic noise and the signal contain significantly



**Figure 1.5** Derivative filtering of a noisy signal: (A) signal plus random and systematic noise; (B) output from a derivative filter (seven point quadratic).

different frequencies. When the systematic noise is from a chemical interference this approach is less practical, particularly for an interference with properties very similar to the analyte. Still, there are many chemometric techniques for this situation. Again, the choice of technique depends on the form of the data, and the extent of previous knowledge. With first-order instruments, these corrections generally require some knowledge of the interfering species. One method for quantifying overlapped signal is linear regression, also known as curve fitting. This method assumes that the measurement is a linear combination of the individual signals, such as is the case in spectroscopy when Beer's law is obeyed. When the measured spectrum contains the absorbance of both analyte and interferences, the problem is one of estimating the contribution of the analyte to the overall measurement. This can be solved by linear regression when the spectra of the analyte and all of the interferences are known<sup>32</sup>. An equivalent solution can be obtained with the

Gram-Schmidt filter\*<sup>28</sup>. This filter is like the matched filter in that it optimally combines the data to maximize the signal to noise ratio, but it can deal with random as well as systematic noise. The derivation of this filter also requires a prior knowledge of the analyte and interfering signals. While the matched filter emphasizes the data with high S/N, this filter emphasizes data with partial selectivities between the analyte and the interference. In summary, there are methods of correcting for systematic interferences based on both linear regression and digital filtering. Prior knowledge of the analyte and interfering signal is usually required to apply these methods to first-order data, however.

The case of unknown interferences is often relevant in dealing with real samples, and more problematic to solve. Most of the solutions to this problem involve factor analysis, which will be outlined in Chapter 3. Briefly, these methods require a matrix of data. Such data can result from a single experiment with a second order technique, like liquid chromatography with a multiwavelength detector. Alternatively a matrix can be built with a first-order instrument using a set of calibration samples. Next, this matrix is decomposed into a number of factors that explain the systematic variance in the data. In this way, the data are assumed to be a linear combination of these factors, rather than a combination of the known spectra. The final stage involves building a model to predict the properties of interest in future samples. This is done by examining how the factors correlate with the value of this property. The advantages of this approach include: (1) it can compensate for unknown interferences which are included in the calibration set, (2) it can correct for

---

\*This filter is more commonly called the Kalman innovation filter. This name is not used here, to avoid confusion with the Kalman filter. Despite their similar names, these are very different approaches.

nonlinear responses, (3) it takes advantage of the entire data set, unlike approaches that use selected wavelengths. Thus, second-order instruments make it possible to detect and correct for unknown interferences.

## 1.5 SUMMARY

This chapter has introduced digital filters for extracting information from signal contaminated with both random and systematic noise. These are summarized in Table 1.1.

**Table 1.1** Summary of digital filters.

<i>Prior Knowledge</i>	<i>Random Noise Present</i>	<i>Random and Systematic Noise</i>
Signal can be measured repeatedly	Ensemble averaging	–
Frequency content of noise and signal differ	Smoothing filters	Derivative filters
Pure analyte response	Matched filter	Gram-Schmidt filter (requires knowledge of the interferences)

The following chapters discuss how another type of filter, the discrete Kalman filter, can be used to extract information from noisy signals of various types.

## THE KALMAN FILTER

---

### 2.1 INTRODUCTION

The fundamentals of digital filters and selected applications were presented in the introductory chapter. Figure 2.1 summarizes the different contributions to an experimentally measured signal and digital filters for treating each case. Smoothing and matched filters are appropriate when only random noise is present. By combining multiple values of the measurement, these filters can estimate the underlying signal with increased precision. The next class of filters, including the Gram-Schmidt and derivative filters, can also minimize the contribution of known systematic effects that would otherwise limit the accuracy of the filtered signal. A calibration or transformation step is still required to predict the properties of these systems. Usually, we are interested in properties like analyte concentration. This chapter introduces the Kalman filter<sup>33-36</sup> as an alternative method of estimating these properties of interest, or state parameters, directly from noisy measurements.

The Kalman filter combines many advantages of the digital filters introduced so far. One advantage of the digital smoothing and derivative filters is that they can be applied during the data collection, like their electronic counterparts. Recursive filters, like the Kalman filter, are also suited to real-time applications. Another attribute of polynomial filters is the flexibility that results from using polynomial equations with adjustable parameters, especially since these do not need to be known before the experiment. Similarly, the Kalman

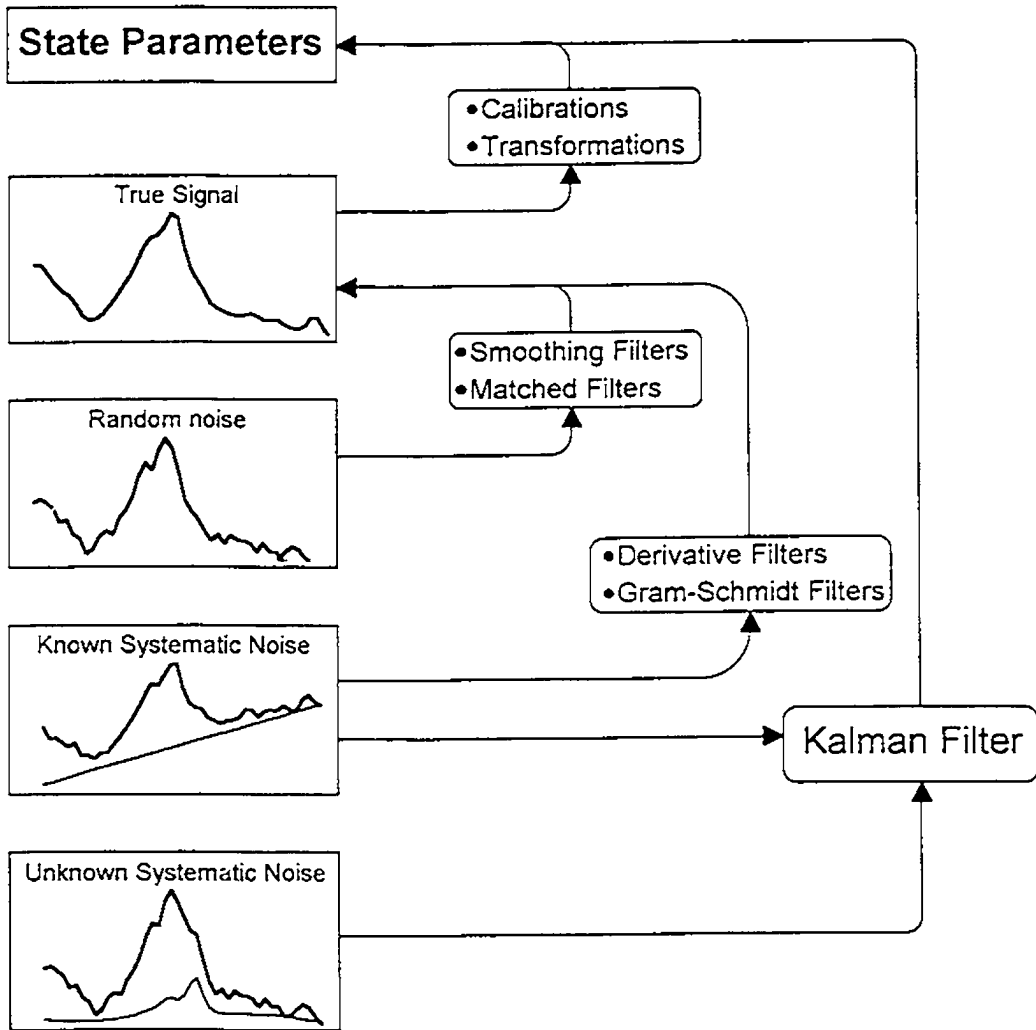


Figure 2.1. An overview of digital filters for state parameter estimation.

filter has adjustable parameters in its model of the system. In contrast, the Gram-Schmidt and matched smoothers have more exacting requirements for prior knowledge of the system. The strength of these approaches is that all the measurement data is incorporated into the estimates in a manner that results in optimal estimates of the signals. The Kalman filter also shares this ability to provide optimal estimates.

In the last case illustrated on Figure 2.1, an interference is present, but not included in the model. This interference will cause inaccurate results, that may not be identified with the output of a simple filter. The Kalman filter can often detect the presence of such unmodelled components, and discriminate against them. In summary, the advantages of using the Kalman filter include its flexibility, real-time operation, and ability to provide diagnostic information for detecting model errors.

The Kalman filter is fundamentally different from the filters introduced so far, in that it is a recursive filter<sup>37</sup>. Recursive filters include past values of their output as part of their current input. This could be regarded as 'feedback' in comparison to an electronic filter. A simple operation that can be carried with a recursive filter is integration: the output of the filter at point  $k$  is calculated by adding the signal's current value to the previous ( $k-1$ ) output of the filter. In contrast, the polynomial smoothing filter, discussed earlier, is nonrecursive since its only inputs are the measured values of the signal. Recursive smoothing filters are also used, but their design is more difficult and they are often less stable.

This chapter will demonstrate: (1) how **recursive filters** estimate signal properties in real-time; (2) what **state space** models are; (3) by what definition

these filters give **optimal estimates**; and (4) how Kalman filters aid in **model evaluation**.

## 2.2 RECURSIVE FILTERS

The concept of a recursive filter is demonstrated here with a simple example. The problem is one of estimating a fixed signal from a series of noisy measurements:

$$z_k = x + v_k \quad (2.1)$$

where  $z_k$  is the  $k^{\text{th}}$  value from the series of measurements  $z = \{ 16, 14, 17, \dots \}$ ;  $x$  is the true value of the signal, which is assumed to be constant; and  $v_k$  is the value of the noise. This is the same case as treated earlier by ensemble averaging, but here our goal will be to provide an estimate of the signal ( $\hat{x}_k$ ) after each measurement. One scheme for estimating the signal would be:

$$\hat{x}_1 = z_1 = 16$$

$$\hat{x}_2 = \frac{1}{2}z_1 + \frac{1}{2}z_2 = \frac{16+14}{2} = 15$$

$$\hat{x}_3 = \frac{1}{3}z_1 + \frac{1}{3}z_2 + \frac{1}{3}z_3 = \frac{16+14+17}{3} = 15.7$$

... ..

$$\hat{x}_{100} = \frac{1}{100}z_1 + \frac{1}{100}z_2 + \dots + \frac{1}{100}z_{100}$$

This algorithm is consistent with the previous digital filters in computing its output as the weighted sum of the inputs. For instance,  $\hat{x}_3$  is calculated by applying the convolution function  $[1/3, 1/3, 1/3]$  to the first three measurements. Although this method calculates the correct results, it is rather inefficient.



Like nonrecursive filters, recursive filters also refine the estimate with each new measurement. The difference is in the way that they calculate the estimate. The recursive solution computes the new estimate ( $\hat{x}_k$ ) using only the previous estimate ( $\hat{x}_{k-1}$ ) and the most recent measurement  $z_k$ . For example,

$$\hat{x}_1 = z_1 = 16$$

$$\hat{x}_2 = \frac{1}{2}\hat{x}_1 + \frac{1}{2}z_2 = \frac{16+14}{2} = 15$$

$$\hat{x}_3 = \frac{2}{3}\hat{x}_2 + \frac{1}{3}z_3 = \frac{2*15+17}{3} = 15.7$$

... ..

$$\hat{x}_{100} = \frac{99}{100}\hat{x}_{99} + \frac{1}{100}z_{100}$$

This recursive algorithm is practical for real-time operations as the number of calculations following each measurement is fixed. For example, measurements 1 to 99 are not explicitly involved in calculating the 100<sup>th</sup> estimate since any of the useful information they contain is already incorporated into the 99<sup>th</sup> estimate. The filter gain is the relative weight placed on the new information (provided by the measurement) versus the old information (contained in the prior estimate). For this example, the gain is simply  $1/k$ . This illustrates that recursive algorithms can estimate the true signal in real time from noisy measurements.

The above example shows the computational elegance of the recursive approach. It would be useful to extend this approach to more complex problems where the signals and noise statistics vary over the course of the experiment. A solution for estimating the current signal's value from a series of noisy measurements was found by Wiener<sup>38,39</sup> in 1949. He showed that a filter can

be designed to output an optimal estimate from a weighted sum of all the previous measurements. Unfortunately, such filters are not recursive, as they require all the weighting terms to be recalculated after each new measurement. Consequently the Wiener filter saw little practical use. The other limitation of the Wiener filter is the difficulty in extending it to include multiple inputs and outputs. Both of these limitations were overcome by R.E. Kalman<sup>40</sup> in 1960, who derived a recursive solution to the same problem.

The Kalman filter was originally presented in the engineering literature, so it uses the terminology of control theory. In this language, the Kalman filter is a state space method for modeling a system. In chemistry, the system could be a reaction, a mixture of chemical species, or an instrumental response. In modeling this system we seek a set of mathematical equations to account for its dynamic behavior, that is, to describe its state at any point in time. Although originally developed for time series, other independent variables such as wavelength can be used. The state vector contains the adjustable parameters of the model that summarize the properties of the system. The Kalman filter estimates the state vector (and hence the properties of interest) recursively from noisy measurements made on the system.

## 2.3 KALMAN FILTERS

### 2.3.1 The Discrete Kalman Filter Algorithm

In this section the equations used in each cycle of the Kalman filter will be described. The **measurement model** is defined as

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \quad (2.2)$$

where  $\mathbf{z}_k$  is an  $(m \times 1)$  vector which consists of  $m$  measurements made at interval  $k$ , and  $\mathbf{v}_k$  is the associated  $(m \times 1)$  measurement error vector assumed to be a white noise sequence. This vector approach allows the system to be monitored with an array of sensors, such as the diode array, as opposed to a single sensor. These measurements are used to estimate the  $(n \times 1)$  state vector  $\mathbf{x}_k$ . This vector contains the properties of interest, such as the contribution of each chemical species in a mixture. These state parameters do not need to be measured directly. For example, the initial concentration of a reaction mixture could be a state parameter, even when its value can only be inferred from absorbance measurements made during the experiment. The measurement vector and the state vector are related through the  $(m \times n)$  observation matrix  $\mathbf{H}_k$ .

Unlike the frequency-based approach, the state space method allows an explicit time dependence of the model; that is, the state parameters incorporate prior knowledge of the system. The Kalman filter then uses this knowledge to predict future states. Dynamic systems can be modeled at two stages. For the reaction-rate examples, that follow, the time-dependence will be included in the observation matrix<sup>41</sup>. The values of this matrix will change over time according to the system's model. For example, a reaction-rate model could incorporate first-order kinetics. In modeling other systems, such as a drifting instrumental response<sup>36</sup>, there is no deterministic model for the changes. The temporal changes in these systems can be modeled as an uncertainty in propagating the state vector. Thus the Kalman filter is capable of incorporating prior knowledge of the experimental system's random and systematic behavior.

An example of a "system" would be a zero-order reaction<sup>42</sup> that produces a spectroscopically observable product. In the following example, the system is "observed" by measuring the product's absorbance ( $A_p$ ) once per second. According to the reaction rate model, this absorbance should increase linearly with time:

$$A_p(t) = m t + b + v_t \quad (2.3)$$

such that a plot of  $A_p(t)$  versus time has a slope  $m$  (related to the rate constant), a background of  $b$  and a contribution from random noise  $v_t$ . Recast in the form of the measurement model:

$$\mathbf{x} = \begin{bmatrix} m \\ b \end{bmatrix} \quad (2.4)$$

$$\mathbf{H} = [t, 1] \quad (2.5)$$

where  $\mathbf{x}$  is the state vector and  $\mathbf{H}$  is the observation matrix. Note that unlike a smoothing filter, which estimates the noise-free values of  $A_p(t)$ , the Kalman filter estimates two parameters ( $m$ ,  $b$ ) that describe the state of the system at any time. These state parameters are assumed to be constant, because the observation matrix described by Equation 2.5 already accounts for the time dependence of the absorbance. Later, methods for testing the validity of this model will be explored.

With each new measurement the Kalman filter calculates a **state estimate update**:

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_{k-1} + \mathbf{K}_k (\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k-1}) \quad (2.6)$$

where  $\hat{\mathbf{x}}_k$  is the updated ( $n \times 1$ ) state vector

$\hat{\mathbf{x}}_{k-1}$  is best estimate of the state vector before measurement  $k$

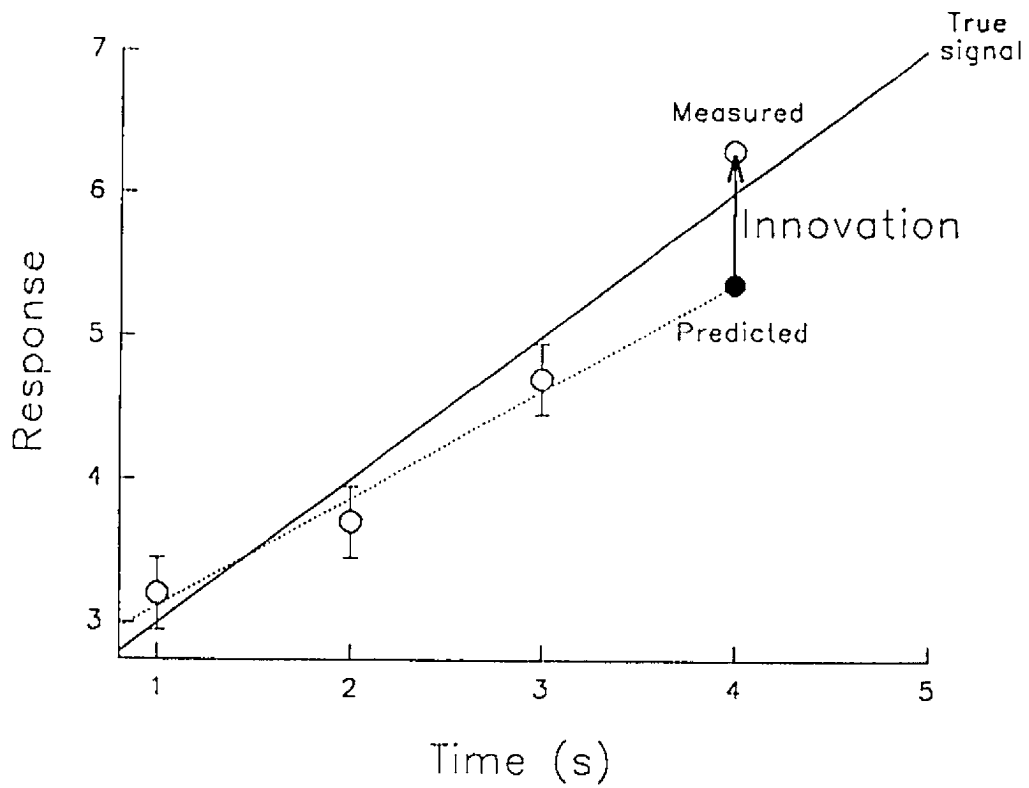
$\mathbf{z}_k$  is the ( $m \times 1$ ) measurement vector at point  $k$

$\mathbf{K}_k$  is the ( $n \times m$ ) Kalman gain matrix

$\mathbf{H}_k$  is the ( $m \times n$ ) observation matrix.

As before, only the current measurement is used explicitly in this equation. The extent to which this new measurement changes the state vector depends on the innovation (Figure 2.2), defined as the difference between the actual measurement ( $\mathbf{z}_k$ ) and the predicted measurement ( $\mathbf{H}_k \hat{\mathbf{x}}_{k-1}$ ). A correct prediction gives an innovation of zero, which causes no changes in the state vector. Otherwise, the state vector changes by a weighted portion of the innovation. The extent and direction of these changes are dependent on  $\mathbf{K}_k$ , the gain of the filter. As in previous filters, these weights define the properties of the filter. Unlike the filters discussed in Chapter 1, the Kalman filter gain usually changes over the course of an experiment. Next, we will consider what properties we want this filter to have, which leads us to the definition of the Kalman gain.

The goal of the Kalman filter is to provide "optimal" estimates of the state vector. If measurements made on the zero-order reaction were error free, then only two measurement cycles would be needed to calculate the true values of the slope and intercept. Experimentally, each measurement has an uncertainty, due to the noise sequence,  $v_k$ . We cannot know the individual values of this sequence, but instead we can only summarize their statistics. For the Kalman filter the measurement covariance matrix  $\mathbf{R}_k$  contains the noise statistics associated with the measurement of the system  $\mathbf{z}_k$ . In this example, the



**Figure 2.2.** The Kalman filter applied to the fourth point of a data sequence. The innovation is the difference between the measured value of this point and its predicted value.

absorbance is measured at one wavelength so  $\mathbf{R}_k$  simply contains the variance of this absorbance reading.

When the measurements are only approximations to the true values, then so are all calculations resting upon them. Thus, our goal is to develop the best approximation to the true state vector as possible. This is equivalent to minimizing the error in the estimate  $[\mathbf{x}_{true} - \hat{\mathbf{x}}_k]$ , but in practice this can not be calculated since  $\mathbf{x}_{true}$  is unknown. Still, this sequence of errors in the state vector can be described by its statistics in the same way as the measurement error, resulting in an error covariance matrix,  $\mathbf{P}$ . In summary, when there are random errors in the observation of the system, such as the absorbance measurements, then there will be random errors in any parameters that are estimated from them. The random errors in the measurements are contained in the measurement covariance matrix,  $\mathbf{R}$ . These errors can be evaluated experimentally. The error covariance matrix,  $\mathbf{P}$ , indicates the extent that these measurement errors are propagated to the state parameters.

Now consider the desirable properties for an estimate. First, the state vector is expected to converge on its true value, such that the average estimation error is zero. In other words, the Kalman filter should provide an accurate estimate. Second, these estimates should be precise, such that the errors  $[\mathbf{x}_{true} - \hat{\mathbf{x}}_k]$  are as small as possible. The values on the diagonal of  $\mathbf{P}$  summarize the expected size of these errors, expressed as variances. If we have no knowledge about the initial values of the state parameters,  $\mathbf{x}_0$ , before the experiment is conducted, they are set to zero. This absence of knowledge is expressed in the initial value of the error covariance matrix,  $\mathbf{P}_0$ , by setting its diagonal elements to infinity - approximated by a large value such as  $10^{30}$  in computations. The ideal experiment would yield the exact values of the state

vector, such that the elements of  $\mathbf{P}_{final}$  are zero. Thus the goal of estimating the state vector  $\mathbf{x}$  in an "optimal" way can be defined more exactly: the Kalman filter should use the measurements to minimize the uncertainty of the estimated state vector. To do this, it must employ the gain  $\mathbf{K}$  that minimizes  $\mathbf{P}$  along the diagonal. This definition results in a minimum mean square error estimate<sup>43</sup>.

Equation 2.6 updates the state vector for point  $k$ , using a gain of  $\mathbf{K}_k$ . In the Kalman filter, this gain is calculated to give an optimal estimate (as defined above) of the state vector. The ( $m \times n$ ) **Kalman gain matrix** is computed as

$$\mathbf{K} = \mathbf{P}_k \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_k \mathbf{H}_k^T + \mathbf{R}_k)^{-1} \quad (2.7)$$

This gain considers three things: (1) the uncertainty in the measurement made on the system, which is contained in  $\mathbf{R}$ , the measurement error matrix; (2) the uncertainty in the present state vector stored in the estimated covariance matrix,  $\mathbf{P}$ , which reflects the amount of information already gleaned from previous measurements; and (3) the observation matrix,  $\mathbf{H}$ , which indicates the magnitude by which the state parameters should be changed, and in what direction.

After computing a new state vector, the next step is to perform the **error covariance update**:

$$\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k) \mathbf{P}_k^- \quad (2.8)$$

In the notation used here, the "-" superscript of  $\mathbf{P}^-$  indicates that this is the estimated covariance matrix before assimilating the new measurement at interval  $k$ , and  $\mathbf{I}$  is the ( $n \times n$ ) identity matrix. This equation decreases the variance associated with each state parameter. Since the new estimates include information from the  $k^{\text{th}}$  measurement, the new value of  $\mathbf{P}$  is smaller to reflect



this improvement. In this work, a more computationally stable form of Equation 2.8 is used<sup>44</sup> which propagates the square root of the covariance.

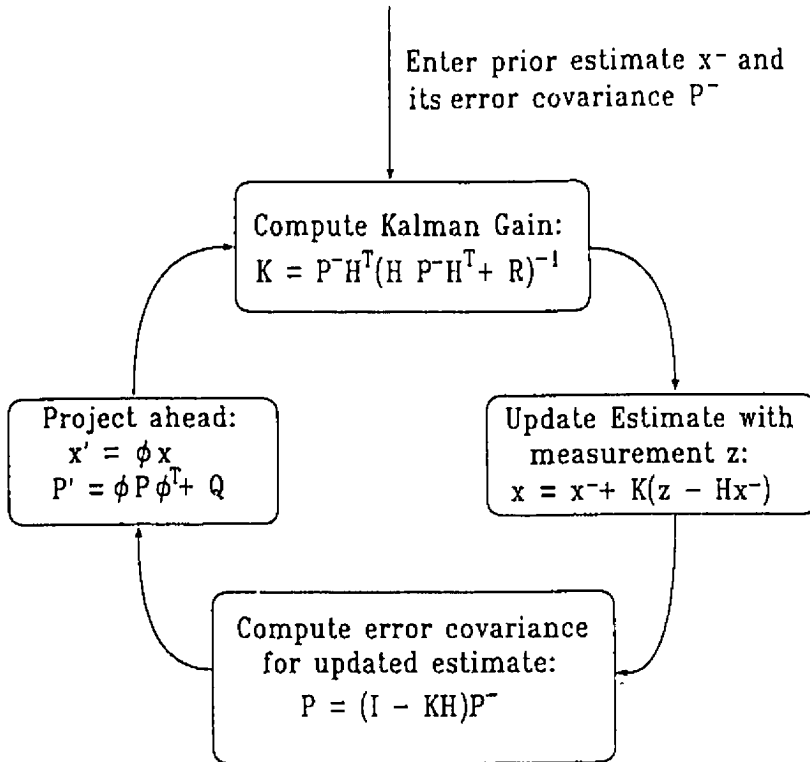
Figure 2.3 summarizes the recursive cycle of the Kalman. The "project ahead" stage was not used in our example since the state parameters were assumed to be constant. Otherwise, the model requires a state transition matrix ( $\phi$ ), to describe the manner in which the state changes between measurements. Any uncertainty in this relationship is expressed in the matrix  $\mathbf{Q}$ . In this case,  $\phi$  is equal to the identity matrix and  $\mathbf{Q}$  is equal to all zeros as the state vector is assumed to be constant. Note that random changes in the state, such as those introduced by instrument drift, deteriorate the model's ability to predict the future states of the system. Consequently this uncertainty increases  $\mathbf{P}$ , counteracting the benefits of previous measurements.

### 2.3.2 Example: One Cycle of Kalman Filtering Reaction-Rate Data.

This example will work through one cycle of the Kalman filter algorithm applied to the zero-order reaction data shown in Figure 2.2. Specifically, the following calculations update the state vector to include the fourth measurement. The test data were generated with Equation 2.3 using  $m = 1$ ;  $b = 2$ , and  $t = [1, 2, 3, 4, \dots]$ , that is, one measurement per second. Gaussian distributed noise was added to this data to simulate a measurement error,  $v$ , with a standard deviation of 0.2.

After three cycles of Kalman filtering

$$\hat{\mathbf{x}}_3 = \begin{bmatrix} 0.75 \\ 2.37 \end{bmatrix} \quad \mathbf{P}_3 = \begin{bmatrix} 0.020 & -0.040 \\ -0.040 & 0.093 \end{bmatrix}$$



$R$ ( $n \times n$ )	covariance of noise	$x$ ( $n \times 1$ )	state vector
$\phi$ ( $n \times n$ )	state transition matrix	$P$ ( $n \times n$ )	error covariance matrix
$Q$ ( $n \times n$ )	covariance of model noise	$H$ ( $m \times n$ )	observation matrix
$z$ ( $m \times 1$ )	experimental measurement	$K$ ( $n \times m$ )	Kalman gain matrix

**Figure 2.3.** Summary of discrete Kalman filter equations.

this state vector is based on the first three noisy measurements. The variances associated with its state parameters (the slope and background) are on the diagonal of the error covariance matrix  $\mathbf{P}$ . Thus the estimated state parameters are

$$\hat{\mathbf{x}}_3 = \begin{bmatrix} 0.75 \pm \sqrt{0.020} \\ 2.37 \pm \sqrt{0.093} \end{bmatrix}$$

where the error is expressed as one standard deviation. Equivalently,

$$m_{est} = 0.75 \pm 0.14 \quad (m_{true} = 1)$$

$$b_{est} = 2.37 \pm 0.30 \quad (b_{true} = 2)$$

To update this state vector to include the fourth measurement requires the observation matrix,

$$\mathbf{H}_4 = [t \ 1] = [4 \ 1]$$

and the Kalman gain matrix calculated from Equation 2.7 with  $\mathbf{R}_4 = (0.2)^2$  gives:

$$\mathbf{K}_4 = \begin{bmatrix} 0.3 \\ -0.5 \end{bmatrix}$$

The observation matrix multiplied by the state vector predicts the value of the fourth point. The innovation is the difference between the measured response and this predicted response:

$$\mathbf{z} - \mathbf{H}\mathbf{x} = 6.30 - [4 \ 1] \begin{bmatrix} 0.75 \\ 2.37 \end{bmatrix} = 6.30 - 5.37 = 0.93$$

The Kalman gain multiplied by this innovation is used to update the state parameters. Finally the error covariance matrix is updated with Equation 2.8 giving,

$$\hat{\mathbf{x}}_4 = \begin{bmatrix} 1.03 \\ 1.90 \end{bmatrix} \quad \mathbf{P}_4 = \begin{bmatrix} 0.008 & -0.020 \\ -0.020 & 0.060 \end{bmatrix}$$

or equivalently

$$m_{est} = 1.03 \pm 0.09$$

$$b_{est} = 1.90 \pm 0.2$$

Note that both estimated values have improved. The estimated slope was increased and the estimated background has been decreased. The signs of these changes, in relation to the innovation, were defined by the Kalman gain vector. Furthermore, the uncertainty in both estimated values is decreased by including another measurement.

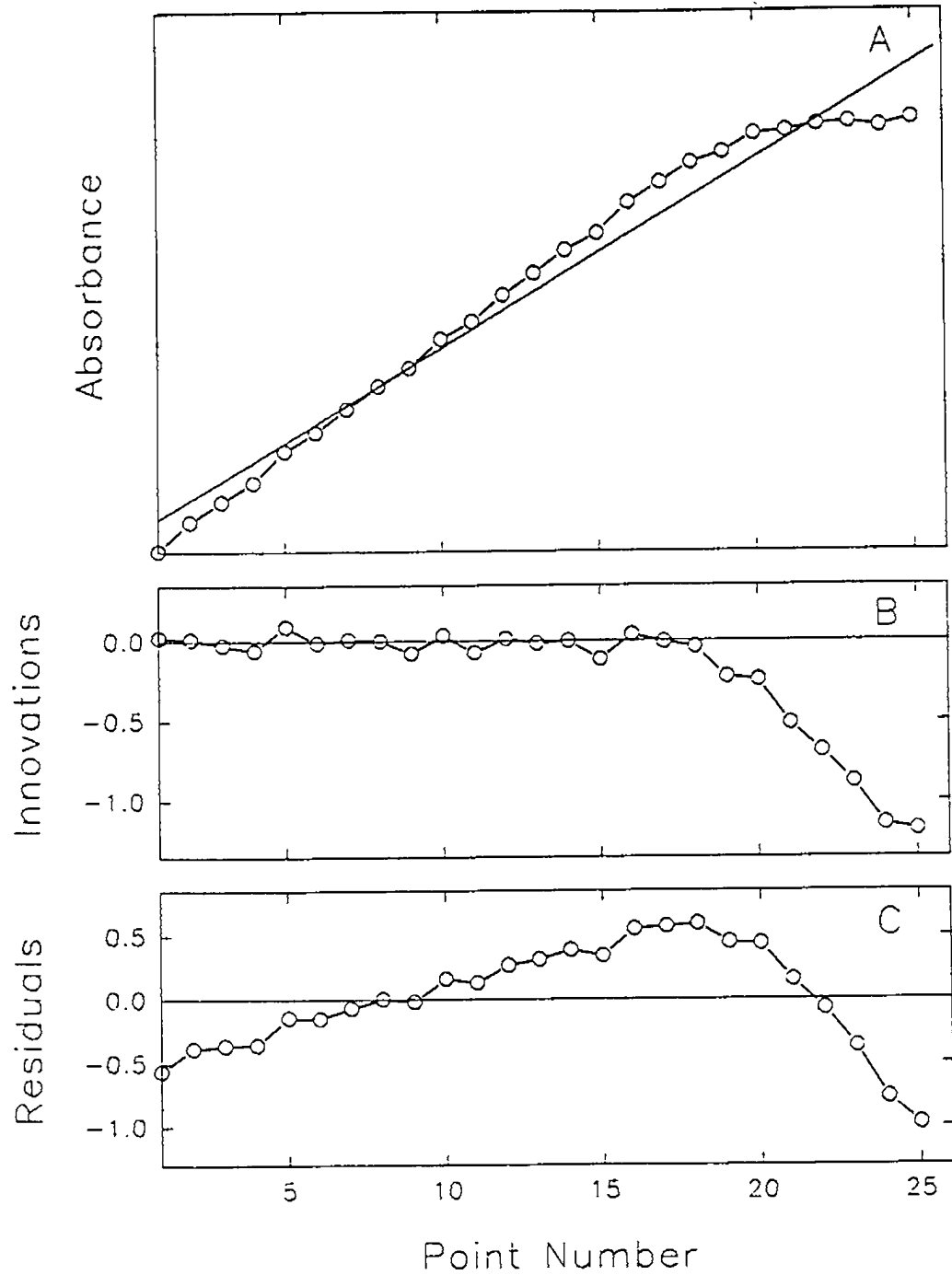
### 2.3.3 Diagnostic Abilities of the Kalman Filter

The Kalman filter used above had a static model with no prior estimates of the state parameters. In other words, the background and reaction rate were assumed to be constant, and their values were unknown before the experiment. In such cases, the final estimates of the Kalman filter are identical to those calculated by a traditional least-squares fit<sup>38</sup>. Thus, in this application, the Kalman filter can be regarded as a recursive means of computing least-squares estimates. Since the final estimate of the state vector is the most precise, it might be asked if there is any advantage to calculating its intermediate values

during the experiment. This thesis will emphasize that the Kalman filter provides important diagnostic information on the fit, thus indicating when the incoming data are not accurately described by the model.

One such indicator is the innovation, the difference between the measurement predicted by the current state parameter and the actual measurement. This is already calculated for each measurement in the state estimate update (Equation 2.6). If the noise in the measurements is a white noise sequence, then the innovation sequence should be too<sup>38</sup>. Therefore, the innovation sequence can be examined to verify it has the appropriate statistical characteristics. In this way, the assumptions of the model can be assessed in its use.

Figure 2.4 continues the example of modeling a zero-order reaction with the Kalman filter. In this example, the zero-order model fails to describe the data after about 15 points. The innovations (Figure 2.4B) appear random for the first 15 points, as expected, since the model is correctly predicting the measured data. Thus, we assume the state vector is converging on the true values  $[m, b]$  with increasing precision. After point 15, a change is observed: the magnitude of the innovations increases, and they are consistently negative. This indicates that the measured data are systematically lower than the predicted values. The state vector must also be changing in an attempt to minimize prediction errors of future measurements. This suggests that data collected after about point 15 are not described by the same state vector as the previous ones. Thus the ability of the final state vector to summarize the behavior of this system is questionable because the zero-order model is not consistent with all the data. The advantage of the Kalman filter is that it flags these erroneous results, and identifies the local region of model failure.



**Figure 2.4.** An example of least-squares fitting where the model is not consistent with all of the data points: (A) raw data and best-fit line; (B) Kalman filter innovations from recursive estimation; (C) residuals from a batch least-squares fit.

In contrast, the application of traditional least-squares yields only one set of estimated parameters. These can be substituted into the reaction rate model to calculate the fitted line (Figure 2.4A). The distance of the experimental points from the points predicted by this equation are known as the residuals. Unlike the innovations, all of these residuals are calculated from one set of the fitted parameters. For example, the residual for experimental point  $A_{15}$  would be calculated from parameters fitted to all 25 experimental points, even though some of these data are inconsistent with the proposed model. In contrast, the innovation for  $A_{15}$  is calculated from  $\hat{\mathbf{x}}_{14}$ , a state vector fit to the first 14 points. Although this state vector is based upon fewer points than the regression estimates, the points it uses are all in a region where the model is obeyed. Figure 2.4C illustrates that while the residuals can indicate a poor fit of the model, they don't suggest that the model was correct for the bulk of the measured points. The diagnostic properties of the Kalman filter become increasingly important with complex data and models.

The diagnostic features of the Kalman filter have been exploited with the *adaptive* Kalman filter which can be used to compensate for model errors. The adaptive Kalman filter was introduced into the analytical chemistry literature by Rutan and Brown<sup>45,46</sup> who showed that in certain cases errors resulting from spectral components not included in a model could be corrected.

In summary, the Kalman filter is based on the following principles: (1) **recursive estimation** - since the computation time and storage requirements of the algorithm are fixed, this is a practical approach for laboratory computers and real-time applications; (2) **state space estimation** - this allows the properties of interest to be estimated directly, and the time-dependent behavior to be included in the model; (3) **optimal estimation** - by definition, the Kalman

gain results in a minimum variance estimate; and (4) **model selection and evaluation** - the Kalman filter can evaluate prediction errors in real-time with its innovations. The diagnostic abilities of the Kalman filter will be explored further in this thesis.

## 2.4 KALMAN FILTERS IN ANALYTICAL CHEMISTRY

Since its introduction to engineering in 1960 the Kalman filter has been applied in many fields including navigation<sup>47</sup>, physics<sup>36</sup>, and biotechnology<sup>48</sup>. Early work in chemistry includes applications by Seelig and Blount<sup>49</sup> (1976) to voltammetry data, and Poulisse<sup>50</sup> (1979) to multicomponent determinations in UV spectroscopy. Poulisse noted the practical advantages of directly coupling the measurement device to a computer, and considered how theoretical results in Kalman filter theory apply to the design of analytical experiments. From 1980 on, applications of the Kalman filter to analytical chemistry have steadily increased, likely reflecting the increasing availability of laboratory computers and computer-controlled instruments.

For a comprehensive review of the Kalman filter in analytical chemistry the reader is directed to Brown<sup>33</sup> (1986) and subsequent biannual reviews of the field of chemometrics<sup>8</sup>. Brown grouped applications into four areas: noise removal; peak resolution; detection and compensation of instrumental drift; and model identification and improvement. Another useful summary by Rutan<sup>34</sup> illustrated Kalman filtering approaches to five problems: resolution of overlapped responses; removal of variable background responses; calibration with drift correction; determination of kinetic parameters; and estimation of electrochemical charge-transfer parameters. The following discussion will



summarize applications that are typical of each of these areas, as well as those which illustrate advantages of the Kalman filter over more traditional approaches.

### 2.4.1 Noise Removal

The nonrecursive digital filters that Savitzky and Golay introduced to chemistry are commonly used for smoothing and differentiation. The Kalman filter can be used for the same applications<sup>51</sup>, with the state parameter simply being the noise-free value, or the numerical derivative of the instrumental response. Lavagnini<sup>52</sup> *et al* used the Kalman filter to smooth voltammograms and estimate protonation constants from titration curves. In both cases the results compared favorably with those obtained with more traditional methods. They also noted that the Kalman filter was computationally efficient and, in the case of smoothing, it had fewer parameters to optimize than other methods. As was demonstrated for polynomial filters in Figure 1.4, smoothing is a compromise between noise reduction and signal distortion. When the signal's characteristics change over the course of the experiment, such as peak width in chromatography, so do the properties of the optimum smoothing filter. These changes can be incorporated into the Kalman filter model. Filters can also be designed which adapt to changes in the incoming data<sup>53,54</sup>. These are useful in cases where the system's behavior is not well characterized before the experiment. Similarly, Lilley<sup>55</sup> demonstrated a self correcting smoothing routine, based on the Kalman filter, that prevented over-smoothing of transient peaks.

### 2.4.2 Determination of Kinetic Parameters

An earlier example illustrated a Kalman filter for zero-order reactions. The Kalman filter has been applied to a wide range of reaction-rate studies<sup>33,34,41,56</sup> to determine rate constants, rate laws, and analyte concentrations. One example is simultaneous kinetic determinations<sup>41</sup>, which take advantage of differences in the rate constants for parallel reactions to determine more than one analyte simultaneously. In such determinations, the Kalman filter often outperforms graphical and initial rate methods, particularly when the first-order rate constants are very similar. Velasco *et al* used such a Kalman filter for the determination of three phenols<sup>57</sup>. Kalman filtering has also been used for the simultaneous determination of species following different kinetics<sup>58</sup>, specifically, simultaneous first- and second-order reactions with the same reagent. Furthermore, the Kalman filter model for reaction-rate methods can include the absorbances at multiple wavelengths<sup>59,60</sup>, thus allowing spectral as well as kinetic differences to be observed.

### 2.4.3 Resolution of Overlapped Responses

The introductory chapter noted that overlapped responses often occur in first-order methods like spectroscopy, chromatography and electrochemistry. If the instrumental responses of all the individual components are known, their concentrations can be estimated from the overlapped response<sup>61,62</sup>. Although this multicomponent analysis cannot resolve components with identical responses, it has been applied in UV spectroscopy when there are no specific wavelengths for the analytes<sup>50</sup>, and in square wave voltammetry with highly

overlapped responses<sup>63</sup>. Other examples include spectrophotometric determination of five metals in hair<sup>64</sup>; and of active constituents in analgesics<sup>65</sup>. Advantages of the Kalman filter include its ability to process data in real-time, to handle complex models, and to investigate interactions among the different components<sup>66</sup>.

In many of these determinations the analyte signal is superimposed on a background signal. If this background is reproducible then it can be explicitly included in the measurement model, but often the background is poorly characterized. Inductively coupled plasma-atomic emission spectrometry data is one case where variable and nonlinear backgrounds can limit the accuracy of determination. Kalman filtering of such data<sup>67</sup> has several advantages over conventional processing techniques, like three-point corrections, including lower detection limits and greater reliability. Powerful background correction routines have also been developed for fluorescence detection with thin layer chromatography<sup>68</sup>, and infrared monitoring of atmospheric pollutants<sup>69</sup>.

#### **2.4.4 Calibration with Drift Correction**

The process of calibration requires summarizing an instrument's response to a series of known samples, often with a linear calibration graph. These estimated parameters are subsequently used to predict the properties of unknown samples, assuming that the instrumental response is unchanged. In general, calibration parameters change slowly over time due to random fluctuations in experimental conditions and systematic changes like lamp and column aging. Thijssen *et al* studied methods of on-line drift compensation that can be incorporated into intelligent analyzers. They designed a Kalman filter<sup>70</sup> to

predict analyte concentrations from flow injection analysis data, evaluate the accuracy of the calibration model, and schedule recalibrations. This self-monitoring approach has also been applied to graphite furnace atomic absorption spectrometry<sup>71</sup> for detecting, correcting and forecasting drift. Adaptive Kalman filtering has also been used to discriminate against outliers in small calibration sets<sup>72</sup>, again, an important stage in automating an analytical process.

#### **2.4.5 Model Identification and Improvement**

In resolving overlapped responses, a set of single-component responses (e.g. spectra) are fit to the analytical signal of an unknown mixture. These single-component responses are usually measured experimentally for pure samples. The multicomponent model is only correct when all the components of the unknown are present in the model, and their responses are the same as those in the pure calibration samples. As noted earlier, the innovations produced by the Kalman filter are useful for detecting modeling errors. For example, peak width changes or shifts between single-component voltammograms and multicomponent voltammograms produce a correlated innovation sequence<sup>62</sup>. This innovation sequence has a sinusoidal appearance instead of being flat, indicating nonoptimal performance of the Kalman filter. Similarly, the sensitivity of fluorophores to the polarity of their environment can cause difficulties in fitting overlapped fluorescence responses<sup>73</sup>. In such cases, the innovations sequence can be an indicator of peak shifts in a sample relative to the calibration spectra. In addition, the innovations can also be used to estimate both the direction and the degree of this shift.

The goal of adaptive filtering<sup>45,46,74,75</sup> is to prevent deviant points or modeling errors from corrupting the parameter estimates. To achieve this, outliers or regions of model failure are identified by their large innovations, and then these points are effectively rejected from the model. For this approach to succeed, the model information must be accurate for some region of the data. Rutan<sup>46</sup> has demonstrated the use of adaptive filters in modeling gas-liquid partition coefficients. Another difficult problem that has been tackled is estimating concentrations when responses for all the contributing species may not be known<sup>76</sup>. In such cases the filter is often restarted several times with different initial guesses, a process that can be automated with simplex optimization<sup>77</sup>.

A major problem in Kalman filtering atomic emission spectra is instrumental shifts in wavelength positioning. Such shifts will cause modeling errors over the entire spectrum that result in a structured innovation sequence. In correcting these errors, van Veen<sup>67,78</sup> used the flatness of the innovation sequence as a criterion. Specifically, the summed innovation was found to go through a minimum for the best correction. This differs from adaptive Kalman filtering in that none of the data is being rejected, rather, the innovations are used in designing an accurate model for the entire data set. This model optimization was an iterative process of adjusting the wavelength correction, filtering the data, and evaluating the innovations. As a result, the Kalman filter's capacity for real-time filtering was relinquished. In the work that follows<sup>79</sup>, the innovation sequence will be used to select the best model to fit experimental observations. This will be achieved in real-time, by running a parallel network of filters. The use of parallel and block sequential filters for processing speech signals has been described in the engineering literature<sup>80</sup>, but this is the first

application of this type we have encountered. The parallel Kalman filter will be applied to reaction-rate data in which, like the atomic emission example given above, the model failure is usually global rather than localized.

## 2.5 PARALLEL KALMAN FILTER NETWORKS FOR KINETIC METHODS OF ANALYSIS

Analytical methods based on kinetic responses have been widely applied for many years<sup>81-83</sup>. The determination of a large variety of chemical species is possible through the direct or indirect measurement of kinetic parameters associated with appropriate reactions. To simplify the experimental and computational aspects of kinetic methods, most strategies use conditions under which zero- or pseudo-first-order kinetics apply. The former is limited to cases such as the determination of enzymes and catalysts, whereas the latter is more widely applicable and is the focus of this work.

One of the disadvantages of kinetic methods is that they are susceptible to experimental factors affecting the rate constant,  $k$ . This problem is aggravated by the fact that many of these effects are nonlinear so that even minor changes in conditions such as temperature, pH and ionic strength can dramatically affect the rate constant<sup>84</sup>. To combat these problems, several approaches have been developed to provide compensation at various stages of the experiment<sup>85-93</sup>. Some of these have been compared in the literature, both for their susceptibility to variations in the rate constant and their performance in the absence of those variations<sup>94</sup>. The focus in the development of these methods has been on correcting for sample-to-sample variations rather than variations within a single run because: (a) the former problem is more tractable,

and (b) between-sample variations are likely to be a bigger problem due to factors such as matrix variations and long-term temperature drift.

In this work, an approach to correct for between-sample variations in the pseudo-first-order rate constant is described. This method is based on the use of a parallel Kalman filter network. A modification of the linear Kalman filter, the extended Kalman filter, was used by Corcoran and Rutan to compensate for variations in rate constants<sup>92,95</sup>, but this suffered from some of the difficulties normally associated with nonlinear parameter estimation. The approach presented here utilizes a set of discrete models (the "network") with the linear Kalman filter in an attempt to compensate for changes in kinetic parameters. Kalman filter networks exhibit several important advantages, including speed, simplicity, stability, and a parallel algorithm. These features become particularly important in view of advances in digital signal processing chips and parallel computing architectures.

### 2.5.1 Theory

An analytical reaction which follows pseudo-first-order kinetics will exhibit an exponential change in some response parameter with time. For the purpose of simplifying the discussion, we will assume that the reaction of the analyte with a reagent produces a product which can be observed spectrophotometrically (although any response linear with concentration would be appropriate). The change in absorbance with time can then given by:

$$A_t = \Delta A (1 - \exp(-kt)) + B \quad (2.9)$$

where  $A_t$  is the absorbance at time  $t$ ,  $\Delta A$  is the absorbance change due to the product at  $t = \infty$ ,  $k$  is the pseudo-first-order rate constant, and  $B$  is the

background absorbance term. The presence of  $B$  in this model can account for either one of two effects: a constant absorbance by an interfering species in the matrix, or experimental variations in the measurement of  $t = 0$ . In the first case,  $\Delta A$  is the quantity sought, whereas  $\Delta A + B$  is of interest in the second case. If both the background absorbance and the time delay are unknown, accurate compensation is not possible for a first-order model.

Traditional kinetic methods, which are still widely used, seek to measure parameters such as  $A_t$  or  $dA_t/dt$ , but these methods can result in substantial errors if there are significant sample-to-sample variations in  $k$ <sup>84</sup>. Optimized methods<sup>85-88</sup> attempt to minimize this problem at the data collection level by selecting optimum measurement conditions, but these have limited effectiveness. Multipoint computational methods<sup>89-93</sup> attempt to compensate at the data analysis stage by applying Equation 2.9 to estimate  $\Delta A$ , which should be independent of  $k$ . These methods should, in theory, be able to provide total compensation for between-sample variations in the rate constant. The application of a parallel Kalman filter network falls into this last category.

Equation 2.9 provides a linear system model amenable to the Kalman filter only if  $k$  is accurately fixed, and a nonlinear model otherwise. One method of treating nonlinear models is with the *extended* Kalman filter<sup>92,95</sup> but this is quite sensitive to initial estimates, requires several passes through the data, and is not guaranteed to provide optimal estimates. The adaptive Kalman filter, which can correct for certain types of model errors, is generally ineffective for nonlinear models since the model is invalid at all points on the curve. Another course of action utilizing the innovations sequence is possible, however. If a number of different models are employed, each with a slightly different value of  $k$ , then a series of linear systems results, each of which can be applied to the



data through the Kalman filter. By examining the innovations sequence, specifically the running sum of squares of innovations for each model, the best model can be selected and the corresponding parameters extracted. This strategy is illustrated in Figure 2.5.

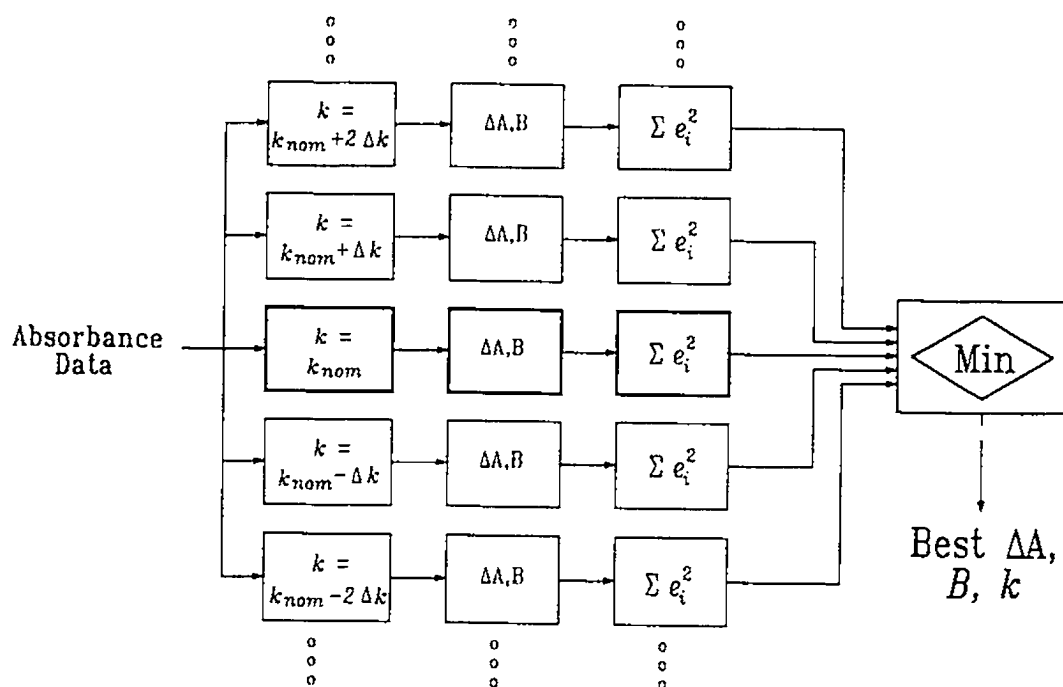
The measurement models used to implement the strategy shown in Figure 2.5 are all of the form:

$$\hat{A}_{ij} = \left[ (1 - \exp(-k_j t)) , 1 \right] \begin{bmatrix} \Delta \hat{A} \\ \hat{B} \end{bmatrix} + v \quad (2.10)$$

where  $A_{ij}$  is the predicted absorbance for point  $i$  (corresponding to time  $t$ ) and model  $j$ , and  $\Delta A$  and  $B$  are the model estimates of the absorbance change due to product and the background absorbance, respectively. The quantity  $v$  is the measurement error, which is characterized by the measurement noise variance,  $\mathbf{R}$ , in the Kalman filter model. The value of  $\mathbf{R}$  was assumed to be constant for all experiments and estimated from previous measurements. In principle,  $\mathbf{R}$  could be obtained more precisely as a function of absorbance, but only an approximate value is needed. The value of the pseudo-first-order rate constant for model  $j$  is given by:

$$k_j = k_{nom} + j \Delta k \quad (2.11)$$

where  $k_{nom}$  is the nominal rate constant for the reaction measured under expected experimental conditions, and  $\Delta k$  is a predetermined increment. In a symmetric network, the quantity  $j$  is an integer extending from  $-(m-1)/2$  to  $(m-1)/2$ , where  $m$  is the number of models. The values of  $m$  and  $\Delta k$  are determined by the expected deviation of  $k$  from its nominal value, the precision of the data, the desired accuracy of the model equations, and computational



**Figure 2.5.** Strategy employed in implementing the parallel Kalman filter network.

constraints. Typically, a range of  $k_{nom} \pm 40\%$  was employed in this work with  $m = 41$ , but these values can vary considerably with changing conditions. Alternative distributions of models (e.g. Gaussian distribution of  $k$ 's) may also be effective, but were not examined in this work.

Several other matrices need to be defined for the Kalman filter algorithm. The state transition matrix is the identity matrix in this case, and the noise vector associated with the state vector was assumed to be zero. The diagonal elements of the error covariance matrix were initially assigned large values ( $10^{30}$ ) and the off-diagonal elements were initialized to zero. The state vector was initialized to zero.

After initialization of the required matrices, absorbance measurements made during the course of the reaction were processed in sequence. Following each measurement  $A_j$ , the innovation for each model  $j$  was computed as the difference between the measurement and the predicted measurement based on the current state vector estimate for the model:

$$e_{ij} = A_i - \hat{A}_{ij} \quad (2.12)$$

The state vector for each model was then updated according to the Kalman filter algorithm. A decision parameter,  $D_{ij}$ , was updated for each model after each measurement according to:

$$D_{ij} = D_{(i-1)} + e_{ij}^2 \quad (2.13)$$

Values of  $D$  were initialized to 0 and Equation 2.13 was not used until  $i = 10$ . While this was not essential, it was intended to avoid the excessively large innovation values associated with the first few measurements as the Kalman filter converges to reasonable estimates of the state parameters. For optimal

estimation, the innovation sequence should be a white noise sequence, so the expectation value of  $D$  is  $nR$  ( $n$  = number of measurements processed,  $R$  = noise variance) if the model is perfectly correct, and should be significantly larger otherwise. After a fixed number of measurements, the optimum model was selected as that which exhibited the smallest value of  $D$ . Alternatively, the best model could be evaluated by examination of  $D$  at each step and data acquisition terminated when the corresponding covariance matrix indicated that error estimates in the parameters were satisfactory. The parameters ( $\Delta A$ ,  $B$ ,  $k$ ) from the best model were extracted and employed in calibration or measurement.

### 2.5.2 Experimental

**Reagents.** All solutions were prepared from reagent-grade chemicals (unless otherwise specified) in distilled water and stored in polyethylene bottles to prevent contamination with silicon from glass. A phosphate stock solution of 100 ppm phosphorus was prepared by dissolving 0.4393 g of primary standard  $\text{KH}_2\text{PO}_4$  and diluting to 1 L. Working standards of 1, 2, 3, 4 and 5 ppm were prepared from this stock. The molybdate reagent consisted of 0.30 M Mo(VI) prepared from 5.3 g of ammonium molybdate in 1 L of 1.0 M nitric acid. A 0.2% (w/v) L-ascorbic acid solution was prepared fresh daily.

**Apparatus.** A diode array spectrophotometer (Hewlett-Packard HP 8452A) with a 1 cm quartz cell and thermostated cell holder was used for measurement of the absorbance of the reaction product. Complete mixing of the reagents in the cell was ensured by the use of a magnetic stirring unit. The spectrophotometer was

interfaced to an IBM-compatible computer via an IEEE-488 parallel interface and data were acquired with the manufacturer's software.

**Procedure:** All solutions and the cell compartment were thermostated at the appropriate temperature. 1.00 mL of the molybdate solution and 1.00 mL of the phosphate standard were placed in the cell and allowed to mix for approximately one minute. Upon the addition of 0.500 mL of ascorbic acid with an Eppendorf<sup>®</sup> pipette, the data acquisition was initiated. The absorbance at 660 nm was measured every 0.2 s for 30 s.

**Computational Aspects.** All calculations were carried out on an 16 MHz IBM-PC/AT compatible personal computer with a math coprocessor using double precision arithmetic. The software was written in our laboratory using Microsoft QuickBASIC version 4.5. The Kalman filter program used the standard algorithm, with the modified covariance update equation for numerical stability.

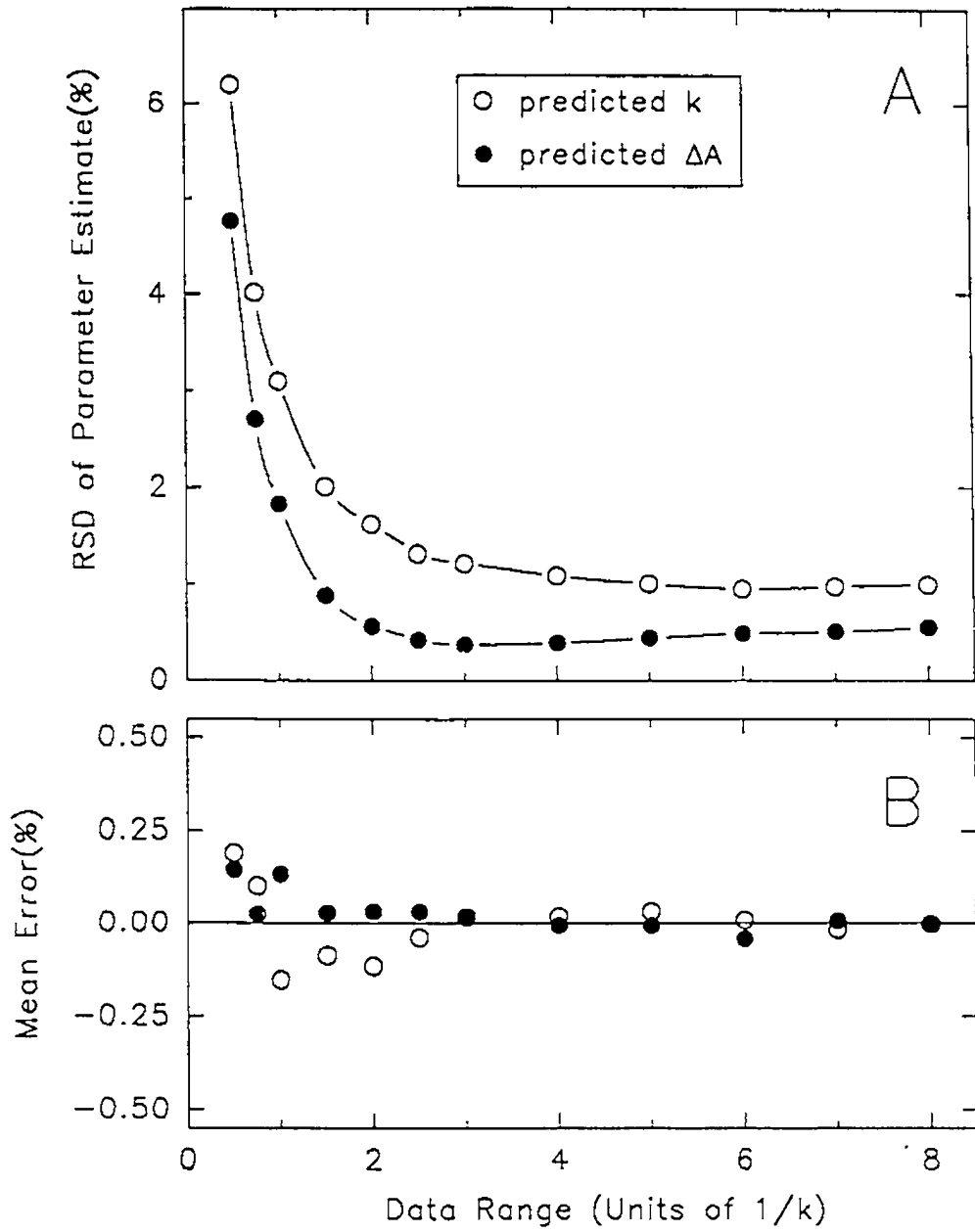
### 2.5.3 Results and Discussion

**Simulation Studies.** To examine the limitations of the parallel Kalman filter approach, a series of computer simulations were conducted to study the effect of three parameters: data range, the number of data points, and measurement noise. These were considered to be the primary factors affecting filter performance. A reference set of conditions was chosen to reflect realistic experimental data and each study considered the effect of varying one of the three factors from these reference conditions. The reference conditions consisted of 100 data points with  $\Delta A = 1$ ,  $B = 0.05$ , a noise level of 1% RSD (Gaussian), and a data range of  $2.5\tau$  ( $\tau = 1/k$ ). For each set of conditions evaluated in the simulations, 500 data sets were generated and the mean error

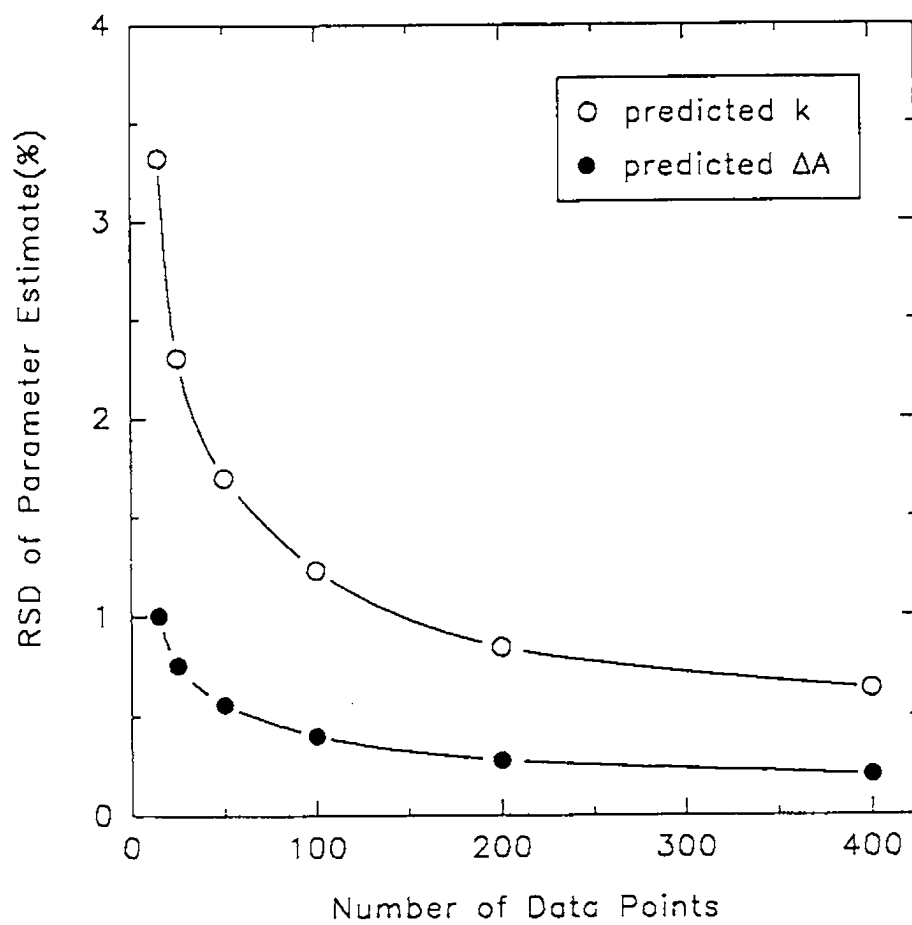
and RSD of the estimates of  $\Delta A$  and  $k$  were determined (errors in  $B$  correlated well with errors in  $\Delta A$ ). In all cases, 81 models were used with rate constants encompassing  $\pm 20\%$  of the true value. This number of models is probably excessive for real applications, but allowed sufficient resolution to avoid quantization error in the rate constant estimates.

The data range (*i.e.* the number of half-lives over which measurements are taken) is an important parameter in multipoint kinetic methods since a sufficient portion of the reaction curve needs to be used to extract estimates of  $\Delta A$  and  $k$  independently. To examine the effect of this parameter on the Kalman filter algorithm, the data range was varied from  $0.5\tau$  to  $8\tau$ . The results are shown in Figure 2.6. Figure 2.6B shows the deviation of the mean estimates of  $\Delta A$  and  $k$  from their true values. In this case, as with the other simulation studies, no significant bias was found. Figure 2.6A shows the RSD in the estimates as a function of data range. The most precise estimates for both parameters were obtained when the data range was greater than  $2\tau$ , but useful results can be obtained below this threshold. The reliability of the estimates diminishes rapidly at short times, however, as independent estimation becomes a problem. In this study, the number of points was kept constant as the data range was expanded, effectively increasing the sampling interval. Since this reduces the number of points in the region of maximum curvature, parameter estimates are not as good as they might be expected to be at long durations if the sampling interval had remained constant.

When the data range is fixed at  $2.5\tau$  and the number of points is varied, similar behavior results as shown in Figure 2.7. Only the RSD of the estimates has been plotted in this case since the mean estimates were again centered on the true values. Although the precision of the estimates becomes poorer as the



**Figure 2.6.** Effect of data range on parameter estimates by the Kalman filter network: (A) effect on mean error, (B) effect on precision.



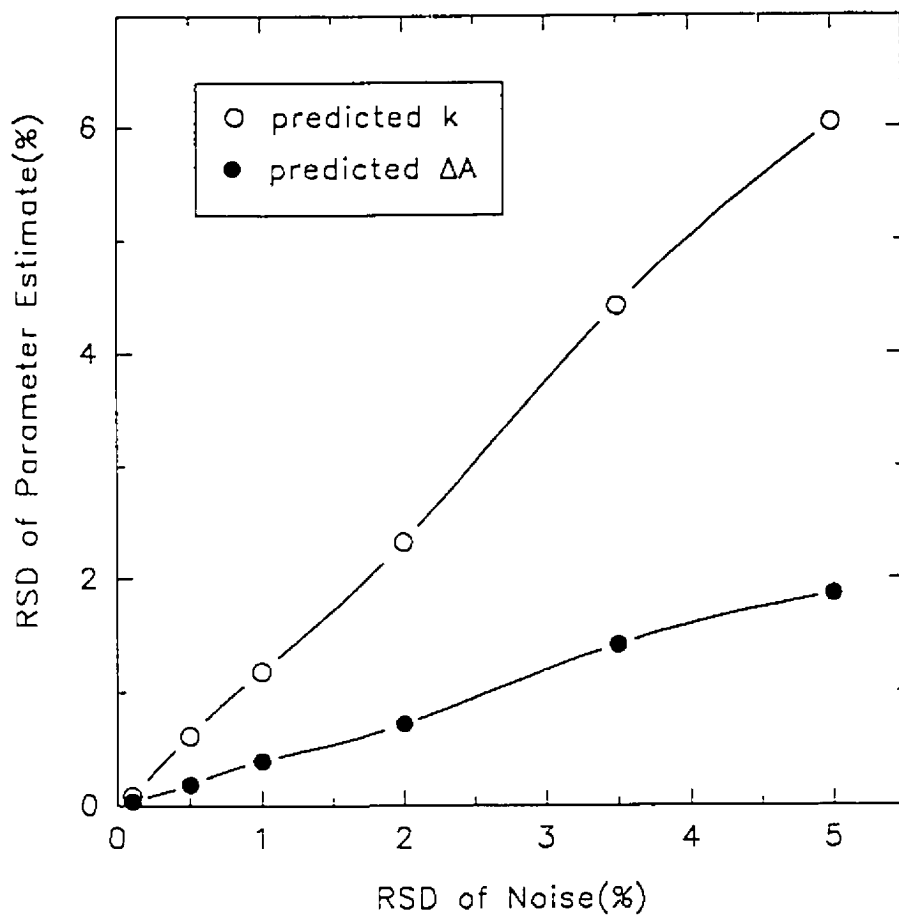
**Figure 2.7.** Effect of the number of data points on the precision of the parameter estimates.



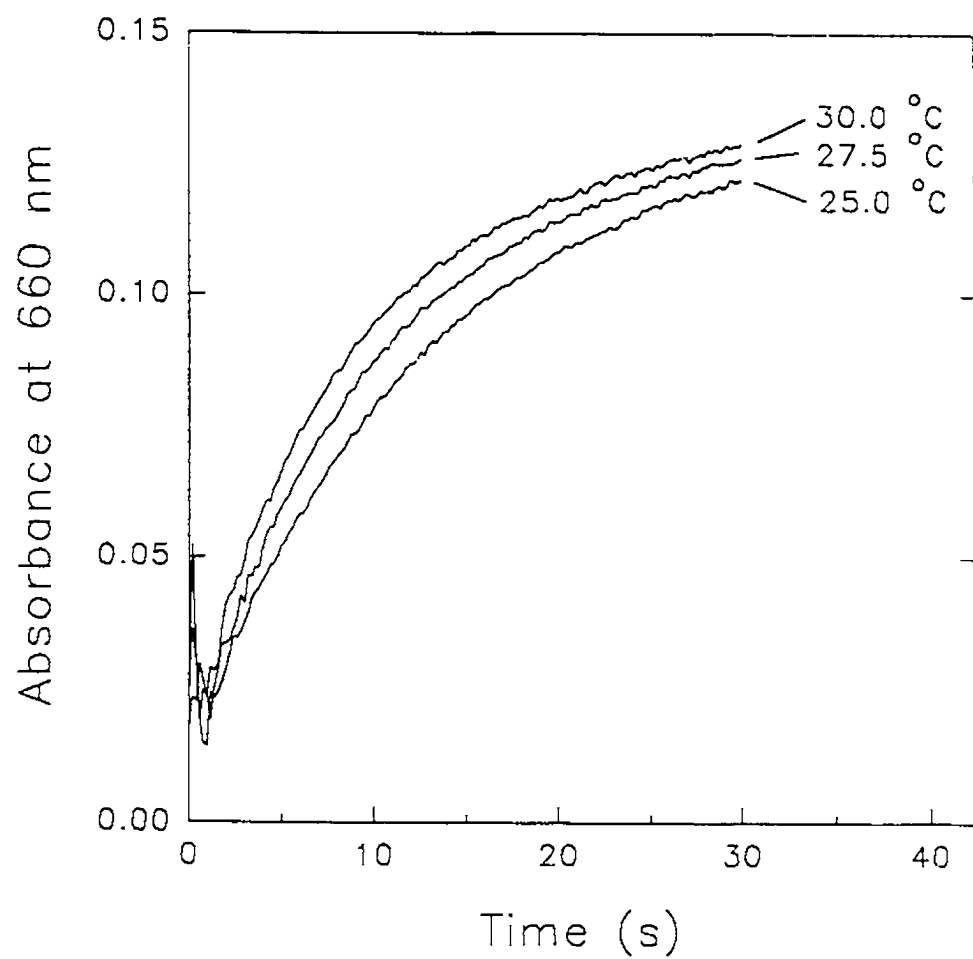
number of points decreases, remarkably few points are required for reliable estimation of  $\Delta A$ . Even with only 15 points, the precision is roughly the same as the measurement error.

The final simulation study examined the effect of measurement noise on the reliability of the estimates. As shown in Figure 2.8, the precision of the estimates exhibits a roughly linear dependence on the precision of the measurements. This linear system behavior is anticipated and a desirable feature for processing analytical data.

***Molybdenum Blue Reaction.*** In order to evaluate the performance of the parallel Kalman filter network on experimental data, the molybdenum-blue reaction<sup>96</sup> for the determination of phosphate was used. The pseudo-first-order rate constant for this reaction under the conditions employed ranged from 0.08 to 0.12 s<sup>-1</sup> over the temperature range of 25 to 30 °C. To determine the effectiveness of the Kalman filter approach for minimizing between-sample variations in the rate constant, calibration data were obtained at three different temperatures (25.0, 27.5, and 30.0 °C). Typical reaction curves are shown in Figure 2.9. Five concentrations of phosphate were used for the calibrations and each solution was run in duplicate at each temperature. The nominal rate constant for the reaction was determined by fitting a typical reaction curve at the middle temperature to Equation 2.9 by nonlinear least-squares. This value was used as  $k_{nom}$  for data at all three temperatures. Calculations with the Kalman filter used 41 models encompassing a range of  $\pm 40\%$  around the nominal rate constant. A data range of  $1.5\tau$  (80 points) was employed. The calibration parameter plotted was  $\Delta A + B$  rather than  $\Delta A$  to help compensate for uncertainty in the reaction starting time that resulted from the manual addition of reagents.



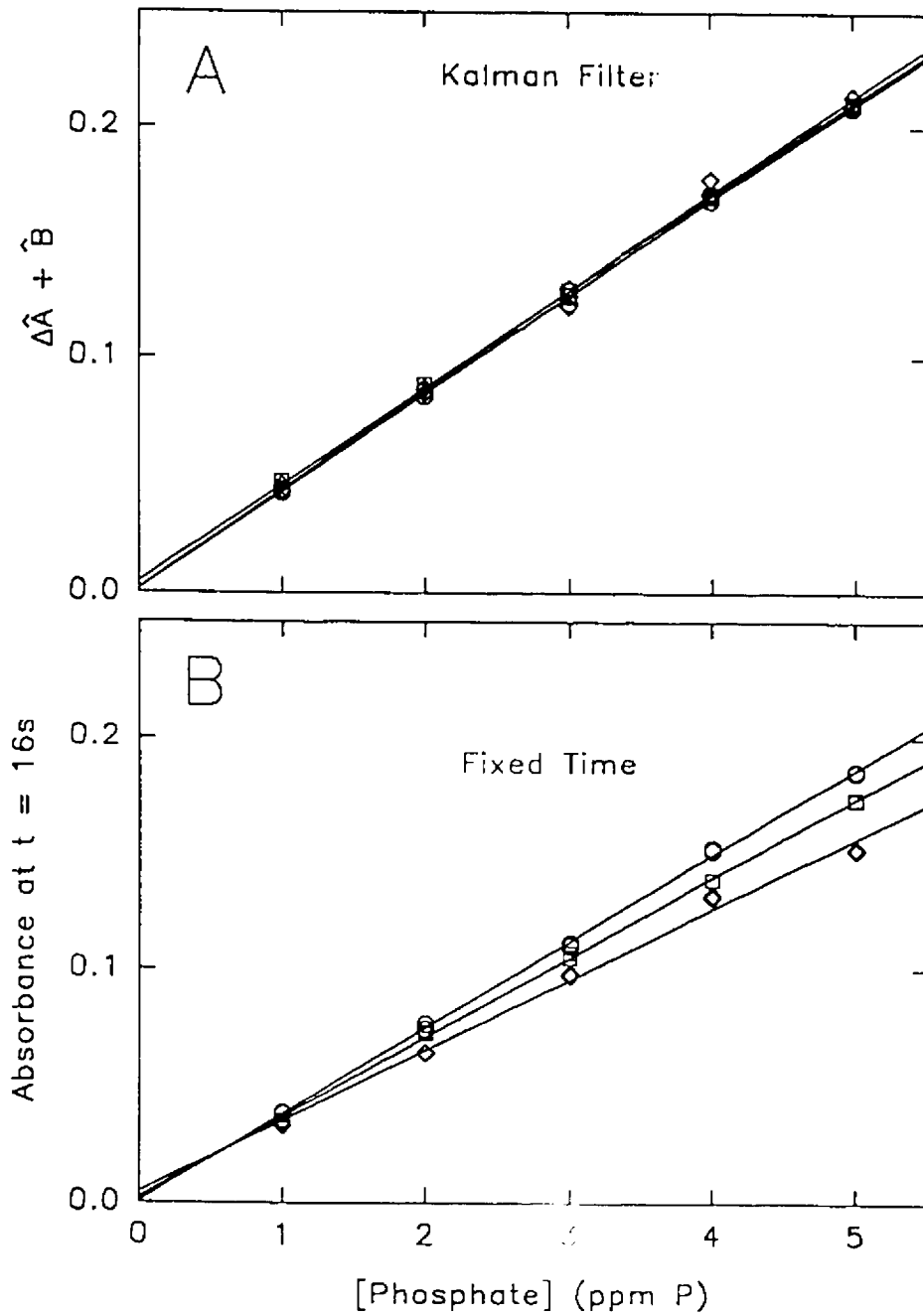
**Figure 2.8.** Effect of measurement noise on the precision of the parameter estimates.



**Figure 2.9.** Typical reaction curves for the reduction of 12-molybdophosphoric acid to the heteropoly blue at three temperatures.

Results of this study are shown in Figure 2.10. Figure 2.10A shows calibration curves obtained at three different temperatures with the Kalman filter method. For comparison, Figure 2.10B shows calibration results obtained using the fixed-time method over an equivalent time period. It is clear that the response parameters extracted with the Kalman filter are largely temperature independent while those for the fixed-time method show the expected systematic variations. Calibration plots obtained by nonlinear least-squares were essentially identical to those generated from the Kalman filter network. Reliable estimates of the pseudo-first-order rate constants were also obtained from the parallel filter. Using 80 points ( $1.5\tau$ ), the estimates were  $0.082 \pm 0.007$ ,  $0.10 \pm 0.01$ ,  $0.129 \pm 0.005 \text{ s}^{-1}$  at 25.0, 27.5, and 30.0 °C, respectively (precision estimates are one standard deviation based on the ten samples at a given temperature). The precision of these estimates improved when the data range was doubled, with values of  $0.077 \pm 0.005$ ,  $0.096 \pm 0.005$ , and  $0.115 \pm 0.003 \text{ s}^{-1}$ . For comparison, estimates obtained through nonlinear least-squares were  $0.081 \pm 0.007$ ,  $0.105 \pm 0.006$ , and  $0.128 \pm 0.005 \text{ s}^{-1}$  with 80 points, and  $0.076 \pm 0.004$ ,  $0.095 \pm 0.004$  and  $0.110 \pm 0.003 \text{ s}^{-1}$  with 160 points. The two methods also showed good correlation for changes in the rate constant estimates between individual data sets.

**Evaluation.** As demonstrated in the previous section, the parallel Kalman filter approach provides compensation for between-sample variations in the pseudo-first-order rate constant that can cause problems with traditional kinetic methods (fixed time, variable time, initial rate, derivative). It should also be superior in this respect to optimized methods (optimized fixed time<sup>85</sup>, optimized derivative<sup>86-88</sup>), although to a lesser degree. Other workers have compensated for variations in the rate constant by employing experimental measurements of



**Figure 2.10.** Comparison of calibration curves obtained with (A) the Kalman filter network and (B) the fixed-time method at three different temperatures:  $\circ = 30.0^\circ\text{C}$ ,  $\square = 27.5^\circ\text{C}$ ,  $\diamond = 25.0^\circ\text{C}$ . A data range of about  $1.6\tau$  was used in both cases.

temperature or pH<sup>92,97</sup>. These methods are quite effective, but in different category from the current study since they employ information assumed to be unavailable. Further comparisons will therefore be limited to other multipoint kinetic methods that aim for total compensation.

Comparisons among multipoint kinetic methods are difficult because of the many parameters that can influence the effectiveness of the algorithms. Nevertheless, the relationship of the Kalman filter network to some of the other methods can be considered. To date, actual implementation of these multipoint methods for routine use has been somewhat limited, probably because of the complexity of the algorithms relative to simple single parameter measurements. The emphasis in this work has been on the development of a robust method which can be employed as a digital filter in real-time. A disadvantage of many multipoint methods, including nonlinear regression, multilinear regression<sup>89</sup>, and the extended Kalman filter<sup>92,95</sup> is the need for multiple iterations and reliable initial estimates. With the Kalman filter network, only a single pass of the data is needed, so it may be employed by signal processing hardware as data are acquired. While the Kalman filter network cannot easily deal with more than one nonlinear parameter, other linear variations to the model, such as sloping backgrounds, are easily accommodated.

Nonlinear regression is clearly the most effective means of nonlinear parameter estimation in that multiple nonlinear parameters may be continuously varied to obtain the optimal estimates. The Kalman filter network is a more "brute force" approach to this process, limited to a discrete estimate of a single nonlinear parameter. However, the Kalman filter network is simpler and not prone to divergence or arithmetic failure. Convergence of the filter is virtually independent of the initial estimates of  $\Delta A$  and  $B$ , which is sometimes not true for

nonlinear regression methods. While a reasonably reliable estimate of  $k$  is required for the filter network, this is usually available for a reaction that is routinely used. Another feature of the Kalman filter network is that the best parameter estimates and their associated errors can be continuously provided during data acquisition, which is not true for nonlinear regression methods. Additionally, the computation time per cycle and storage requirements are independent of the number of data points to be processed with the filter network. The network implemented in this work required about 6 s for 80 points and 41 two-parameter models. For comparison, a nonlinear least-squares program based on a steepest-descent algorithm required 7 to 35 s with reasonable initial estimates. It is likely that the performance of nonlinear regression could be considerably improved through optimization of the algorithm, but the same is true for the filter network. Enhancements in the latter case could include: (1) utilization of more efficient variants of the Kalman filter algorithm, (2) precalculation of covariance matrices, (3) implementation using digital signal processors, and (4) implementation on a machine with a truly parallel architecture to take advantage of the naturally parallel structure of the algorithm. Although the last two circumstances are not widely exploited at the present time, they will no doubt become more important in the future.

## 2.6 CONCLUSIONS

In conclusion, the Kalman filter network has several advantages when applied to kinetic methods based on first- or pseudo-first-order kinetics. The algorithm is relatively simple and well-defined, requiring only a single pass of the data and providing continuous estimates of model parameters and their associated errors.

It is fast enough to be implemented in real-time and usage of computational resources is fixed. Finally, it is well-suited to trends in computing towards vector processing.



## THE KALMAN FILTER AND ORDERED DATA SETS

### 3.1 INTRODUCTION

This chapter introduces factor analysis<sup>4,5,98-102</sup> as a tool for examining multivariate data. Factor analysis encompasses a family of techniques suitable for modeling, interpreting and predicting chemical data. An overview of this field by Malinowski<sup>103</sup> includes applications in analytical, physical and medicinal chemistry. In the work presented here, factor analysis has been applied to spectroscopic data obtained from chemical mixtures, focusing on liquid chromatography with multiwavelength detection as an example. The problem of mathematically resolving overlapped chromatographic profiles will be addressed. This process consists of three steps: detection of peak overlap, identification of individual analytes, and quantitation of components. This chapter outlines and demonstrates an approach to the first problem, one of peak purity analysis, that combines the advantages of Kalman filtering with factor analysis.

The aim of chromatography is to separate the components of a mixture on the basis of their physical and chemical properties. Such results are used to identify and quantify the chemical components of a mixture. Chromatography using a detector measuring a single property, such as thermal conductivity or absorbance at one wavelength, produces first-order data<sup>11</sup>. From this, chromatographic peaks are identified by their retention times. To deal with more difficult problems, chemists often combine a chromatographic separation with first-order (multisensor) detection<sup>18</sup>, to produce second-order data. A wide

range of first-order instrumental techniques have been used for chromatographic detection<sup>104</sup>, including UV-visible absorbance<sup>105,106</sup> and fluorescence<sup>107</sup>, IR absorbance<sup>108</sup>, atomic emission<sup>109</sup>, and mass spectrometry<sup>110</sup>. The advantage of these detection schemes is that peak identity can be based on both retention time and the detector signal.

As the sample mixture becomes more complex, the likelihood of having overlapped chromatographic peaks increases. The presence of two (or more) chemical components that are not separated into distinct peaks complicates the interpretation of the data. With first-order data, the quantitation of the analyte will be inaccurate if the overlapped impurity also contributes to the detector response. This can also be a problem with second-order data when the selectivity of the detector is limited. Furthermore, the spectrum collected on the chromatographic peak is a composite of two chemical components, which would confuse most library searching routines. Thus, accurate identification and quantitation of a chromatographic peak relies on a knowledge of the number of chemical components it contains. To this end, an algorithm has been developed in this work that identifies both the number of components under a particular peak, and where the contribution of each becomes significant.

## 3.2 BEER'S LAW

In this section the nature of data from chromatography/spectroscopy experiments will be considered. For simplicity, the following discussion considers only absorption spectroscopy, which is commonly encountered in liquid chromatography with diode array detection. Quantitation in absorbance spectroscopy is usually based on Beer's law:

$$A = \epsilon b c \quad (3.1)$$

where the *absorbance*,  $A$ , from the analyte depends on its *molar absorptivity*,  $\epsilon$ , the *path length* of the radiation,  $b$ , and the analyte *concentration*,  $c$ . Equation 3.1 is appropriate for a single species absorbing monochromatic radiation. The situation is more complicated for a chromatographic run where the detector makes multiple readings at multiple wavelengths.

First, consider the case where there is more than one substance absorbing monochromatic radiation. A row vector,  $c$ , is defined to contain the concentration of each absorbing substance in the mixture:

$$c = [c_1, c_2, c_3] \quad (3.2)$$

where element  $c_i$  is the concentration for  $i^{\text{th}}$  component. Similarly, the vector  $s$  is defined as a column vector of detector responses:

$$s = \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} \quad (3.3)$$

where  $s_i$  is the detector response to the  $i^{\text{th}}$  component at the chosen wavelength. These are the molar absorptivities per path length, which incorporate  $\epsilon$  and  $b$  into a single constant. Provided there is no interaction among the various species the total absorbance for a multicomponent system is given by

$$\begin{aligned} A &= \epsilon_1 b c_1 + \epsilon_2 b c_2 + \epsilon_3 b c_3 \\ &= c s \end{aligned} \quad (3.4)$$

Note the absorbance is a scalar value, as it is the total absorbance at a single wavelength. Next, this equation is extended for the case of a first-order detector, such as the diode array spectrometer, that measures the absorbance at multiple wavelengths. This requires a matrix of sensitivities,  $S$ , that has one row for each wavelength such that  $s_{ij}$  is the detector response at the  $j^{\text{th}}$  wavelength for the  $i^{\text{th}}$  analyte. Thus the spectrum can be calculated as,

$$\begin{bmatrix} A_1 & A_2 & A_3 & A_4 \end{bmatrix} = \begin{bmatrix} C_1 & C_2 & C_3 \end{bmatrix} \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & S_{1,4} \\ S_{2,1} & S_{2,2} & S_{2,3} & S_{2,4} \\ S_{3,1} & S_{3,2} & S_{3,3} & S_{3,4} \end{bmatrix}$$

$$\underset{(1 \times n_w)}{\mathbf{a}} = \underset{(1 \times n_c)}{\mathbf{c}} \underset{(n_c \times n_w)}{\mathbf{S}} \tag{3.5}$$

where  $n_w$  is the number of wavelengths and  $n_c$  is the number of absorbing components in the mixture. Equation 3.5 calculates the absorbance spectrum for a multicomponent sample. Usually, multiple spectra will be collected during a chromatographic run, with each new spectrum adding another row to the data matrix. Accordingly, we define the concentration matrix  $C$ , with element  $c_{ij}$  containing the concentration of the  $j^{\text{th}}$  absorbing species at the time of the  $i^{\text{th}}$  scan of the diode array. Thus the general form of Beer's law for a chromatographic run with first-order detection is,

$$\underset{(n_s \times n_w)}{\mathbf{D}} = \underset{(n_s \times n_c)}{\mathbf{C}} \underset{(n_c \times n_w)}{\mathbf{S}} + \underset{(n_s \times n_w)}{\mathbf{E}} \tag{3.6}$$

where  $\mathbf{D}$  is the data matrix

$\mathbf{C}$  is the matrix of concentration profiles

$\mathbf{S}$  is the matrix of spectral profiles

$\mathbf{E}$  is the matrix of experimental errors

$n_s$  is the number of samples in time (i.e. spectra)

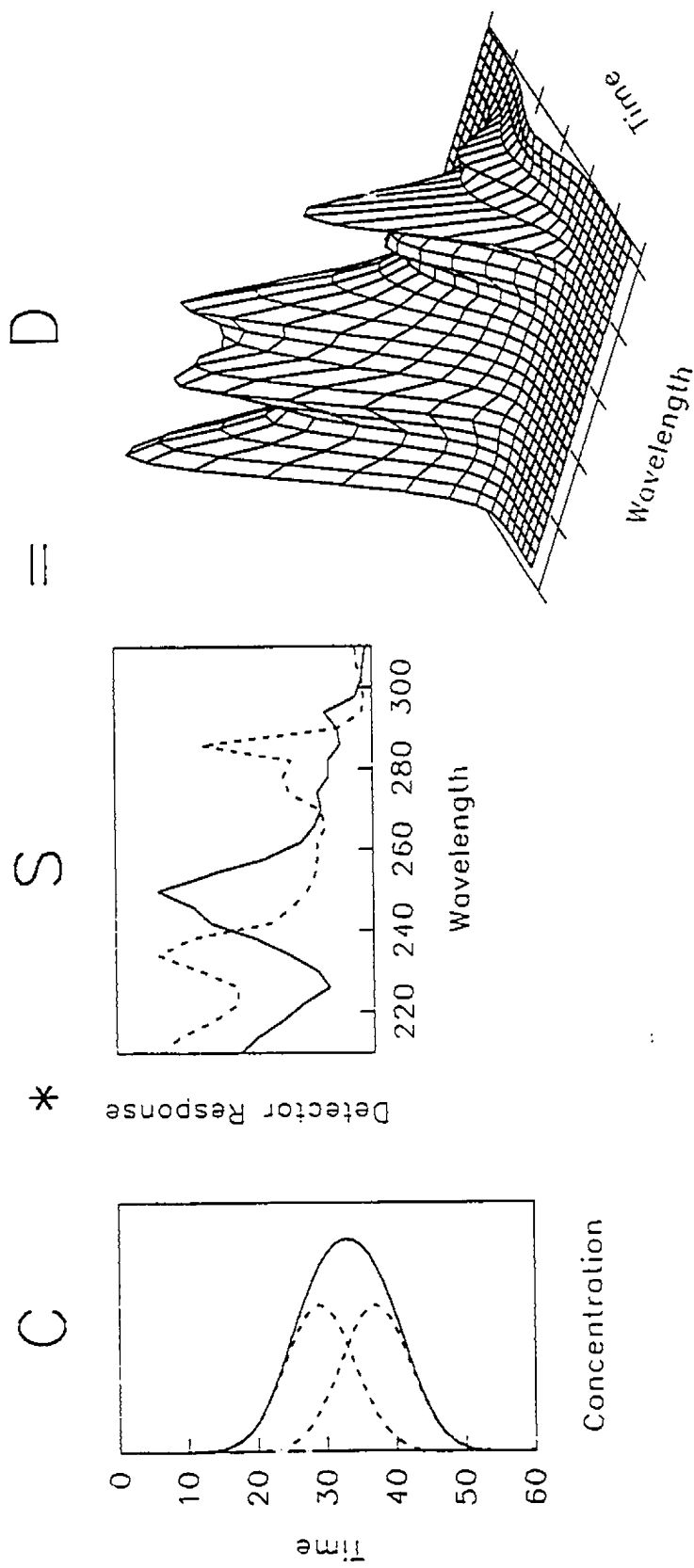
$n_w$  is the number of wavelengths  
 $n_c$  is the number of absorbing components

Figure 3.1 illustrates these matrices for a mixture of two chemical species. The data matrix  $\mathbf{D}$  can be considered to be a collection of chromatograms measured at different wavelengths, or equivalently, a collection of spectra acquired at different points during the elution profile.

Equation 3.6 gives bilinear data where the rows and columns of a mixture response are the sums of the responses of the individual components. For now, two assumptions will be made: that this model accurately describes the spectrochromatogram, and the noise in matrix  $\mathbf{E}$  is randomly distributed with uniform variance. The validity of these assumptions for the diode array spectrometer will be assessed in Chapter 4. By no means are these bilinear data structures unique to LC/UV. Data from other chromatographic systems (e.g. LC-fluorescence and GC-MS), or other techniques such as kinetics and spectrophotometric titrations can be modeled in a similar fashion. The only restriction is that the detector response follows a linearly additive model like Equation 3.6. Due to the wide range of analytical systems that produce these results, there is considerable interest in techniques that extract data from bilinear matrices.

### 3.3 EXTRACTING INFORMATION FROM BILINEAR DATA

Some common goals of these multidimensional analytical techniques are the identification and quantitation of some or all of the chemical species in the sample. This is most difficult when the analytes are not clearly resolved in either dimension. That is, the chromatographic peaks are overlapped and completely



**Figure 3.1.** Pictorial representation of a bilinear data set. The data matrix **D** is the inner product of a matrix of pure concentration profiles, **C**, and pure spectral profiles **S**.

selective detector channels are unavailable or unidentified. When this occurs, there are two possible solutions. The first is an instrumental solution, such as improving the separation and detection schemes. The alternative is a mathematical solution, like curve resolution<sup>111,112</sup>, which extracts the matrices **C** and **S** from a bilinear data matrix. Clearly, if the analyst succeeds with a curve resolution approach, which avoids repeating or modifying the instrumental method, a significant savings in time and cost can be realized. Even if the results of curve resolution are not entirely satisfactory, they can often serve as a guide in modifying the experimental approach.

As before, the appropriate method for extracting information from experimental results depends on the goal of the experiment and the extent of prior knowledge. When pure spectra are known for all of the chemical species present, curve fitting can estimate their chromatographic profiles. This process involves using least-squares methods to fit a linear combination of the pure spectra to the data matrix<sup>9</sup>, thus estimating the contribution of each of the chemical species. A similar strategy can be used to estimate the spectral matrix, **S**, from a set of individual concentration profiles for the components. In either case, only one matrix (**C** or **S**) is extracted from the measurement data through the complete knowledge of the complementary matrix. When this "known" matrix is incomplete, such as the when the number of absorbing species is underestimated, the accuracy of the results will suffer<sup>32</sup>. Therefore, the requirements for prior knowledge generally preclude applying this approach to unknown mixtures.

Soft-modeling is an alternative means of modeling the data when there is insufficient knowledge for curve fitting; that is, when there are unidentified components. The advantage of this method is that it requires no assumptions to

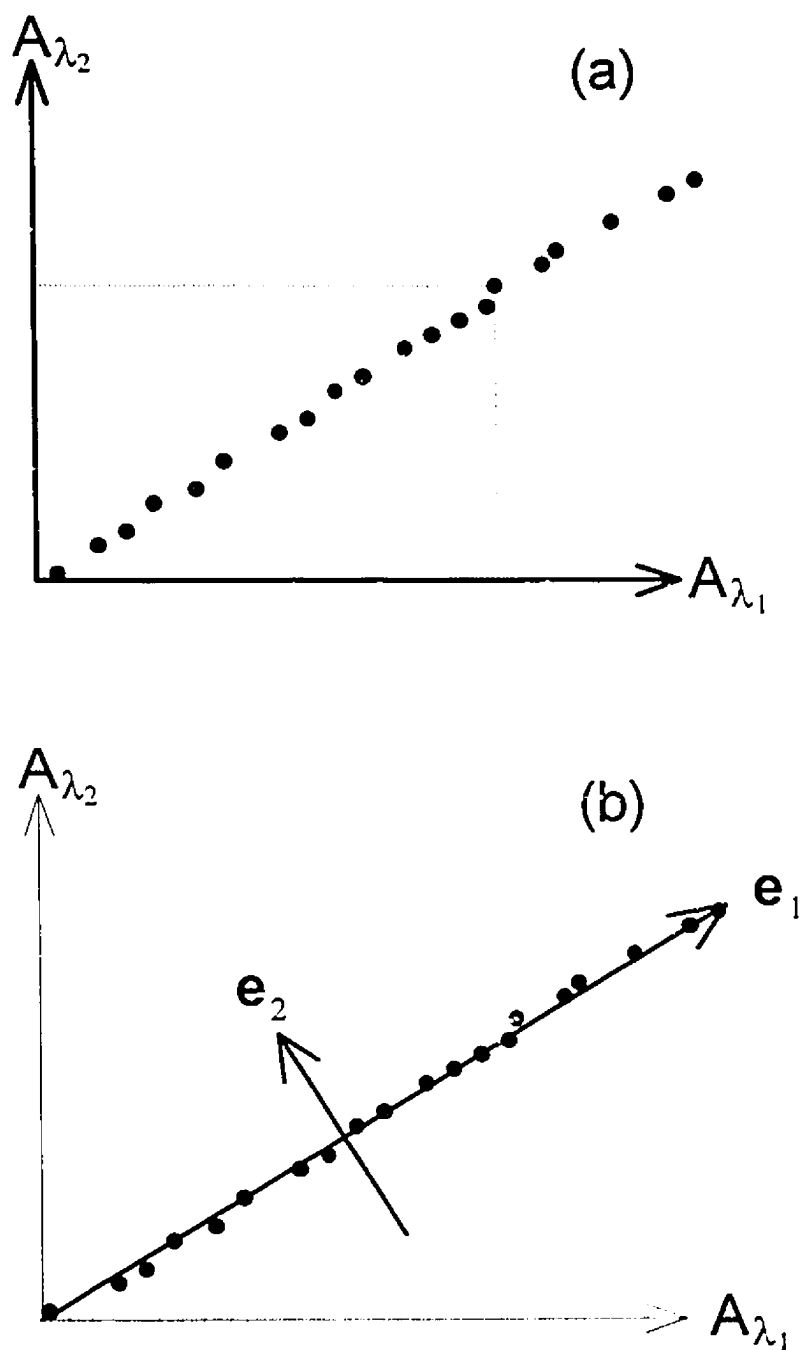
be made regarding the number of components, or the shapes of component spectra or elution profiles. With soft-modeling, principal components analysis (sometimes referred to as factor analysis) is used to decompose the data matrix, usually of large dimensions, into a smaller number of "significant" factors which can generally be associated with components of the mixture. In this way, the data matrix,  $\mathbf{D}$ , is represented as the inner product of two other matrices: the scores matrix,  $\mathbf{X}$ , and the loadings matrix  $\mathbf{Y}$ , such that

$$\mathbf{D} = \mathbf{X} \mathbf{Y} \quad (3.7)$$

Note that like Equation 3.6 this is a bilinear model. The matrix  $\mathbf{S}$  in Equation 3.6 has one row for each absorbing species, by comparison, the matrix  $\mathbf{Y}$  has one column for each "factor". Two fundamental problems in factor analysis are: (1) defining the mathematical properties of the factors and (2) relating these mathematical (or abstract) factors to the true (or chemical) factors. Principal components analysis<sup>100,102</sup> (PCA) is commonly used for the first problem.

PCA is a method for describing the space of the original data with a set of new axes, also known as a basis set. This will be illustrated here with graphical examples. More thorough mathematical treatments are given in the references<sup>103</sup>. The first example considers the absorbance measurements of twenty samples ( $n_s = 20$ ) acquired at twenty wavelengths ( $n_w = 20$ ), resulting in a 20 by 20 data matrix  $\mathbf{D}$ . These samples contain varying amounts of the same analyte, with no other absorbing species. Figure 3.2a shows a plot of each sample in a two-dimensional subspace of the original 20-dimensional space. Here the coordinate axes are the instrumental responses for the sample at the two wavelengths. The basis set contains all 20 wavelengths. The data points are highly correlated in this space, as they are in the full space of twenty





**Figure 3.2.** One-component data are shown in two coordinate frames: (a) the measurement axes corresponding to absorbances at individual wavelengths  $\lambda_1$  and  $\lambda_2$ , and (b) the first two principal component axes  $e_1$  and  $e_2$ .

wavelengths. The first step of PCA is to calculate the **eigenvectors**, which are the new set of axes. The direction of each eigenvector is chosen to successively account for the maximum amount of variance in the data while retaining perpendicularity with all previously selected eigenvectors. The first eigenvector,  $e_1$  in Figure 3.2b, is a linear combination of all of the measurement vectors,  $A_{\lambda_i}$ . The position of the samples along this new axis describe the variance in the data much better than the individual absorbance values would alone. Still, not all the points fall on the line described by  $e_1$  so a second eigenvector  $e_2$ , perpendicular to  $e_1$ , is included to account for the residual variance. For the full data set, this process continues until twenty eigenvectors have been calculated.

These eigenvectors form a new basis set in which each sample can be described by its coordinates. For this one-component data, there is a wide range of values for the samples projected onto this first eigenvector. In contrast, the distribution of the sample points about the second eigenvector is much smaller. **Eigenvalues** quantify the relative importance of each eigenvector in describing the data. For this one-component example, the first eigenvalue would be much larger than the second, and then the following eigenvalues would continue to decrease slowly.

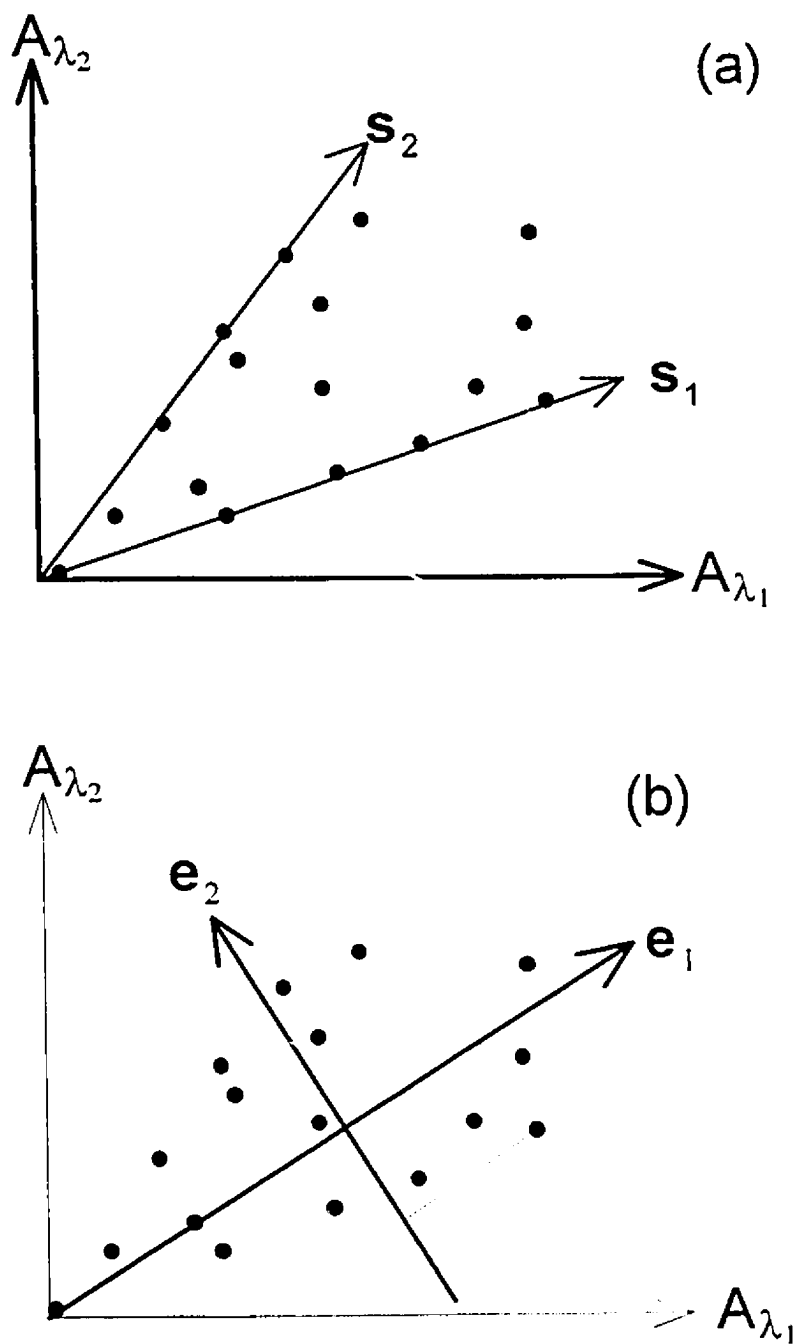
Since this first data set contains experimental noise, all twenty of the eigenvalues are nonzero. This indicates that every eigenvector is required to reproduce the data exactly. In practice it is generally not necessary to model the original data exactly; instead it is normally sufficient to model the data within experimental noise. In this case, only the first eigenvector contains useful information. The remaining factors simply reproduce the experimental noise, so they are discarded. In this way the model is compressed to

$$\hat{\mathbf{D}}_{(n_s \times n_w)} = \mathbf{X}_{(n_s \times n_p)} \mathbf{Y}_{(n_p \times n_w)} \quad (3.8)$$

Where  $n_s$  is the number of samples,  $n_w$  is the number of wavelengths, and  $n_p$  is the number of principal components. In this example  $n_p = 1$ , since only one principal component was required to describe the data. A central problem in PCA is determining the number of principal components required to adequately describe the data from a mixtures.

The important idea presented here is that soft modeling provides a new basis set for describing the data. This basis set, found through PCA, defines a vector space. Each sample is then represented as a point in this space, where its position reflects its chemical properties. There are many advantages to using this space over the original measurement space.

Figure 3.3a shows the case where the samples contain mixtures of two absorbing species. The instrumental response to the two pure components is indicated by the spectral vectors  $s_1$  and  $s_2$ , which are assumed to be unknown before the experiment. Note that the instrument is only partially selective to each chemical species, as they both absorb at wavelengths  $A_1$  and  $A_2$ . Figure 3.3b shows how the first two eigenvectors are required to describe these data. Accordingly, the original set of twenty eigenvectors can be divided into a set of two principal components that summarize the chemical information, and a set containing 18 eigenvectors associated with experimental noise. The first two eigenvectors define a plane. Since the illustration is only in two dimensions, it appears coplanar with the two original axes, but this is not usually true when all twenty wavelengths are included. As before, the position of the points in this



**Figure 3.3.** Two-component data are shown in two coordinate frames: (a) the measurement axes, and (b) the first two principal component axes.  $s_1$  and  $s_2$  are the spectral responses for pure samples of the two chemical components.

twenty-dimensional space can be summarized by their coordinates in the new basis set:

$$\hat{D} = X Y$$

$$\begin{bmatrix} d_{1,1} & \cdots & \cdots & d_{1,20} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ d_{20,1} & \cdots & \cdots & d_{20,20} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ \vdots & \vdots \\ x_{20,1} & x_{20,2} \end{bmatrix} \begin{bmatrix} y_{1,1} & \cdots & \cdots & y_{1,20} \\ y_{2,1} & \cdots & \cdots & y_{2,20} \end{bmatrix}$$

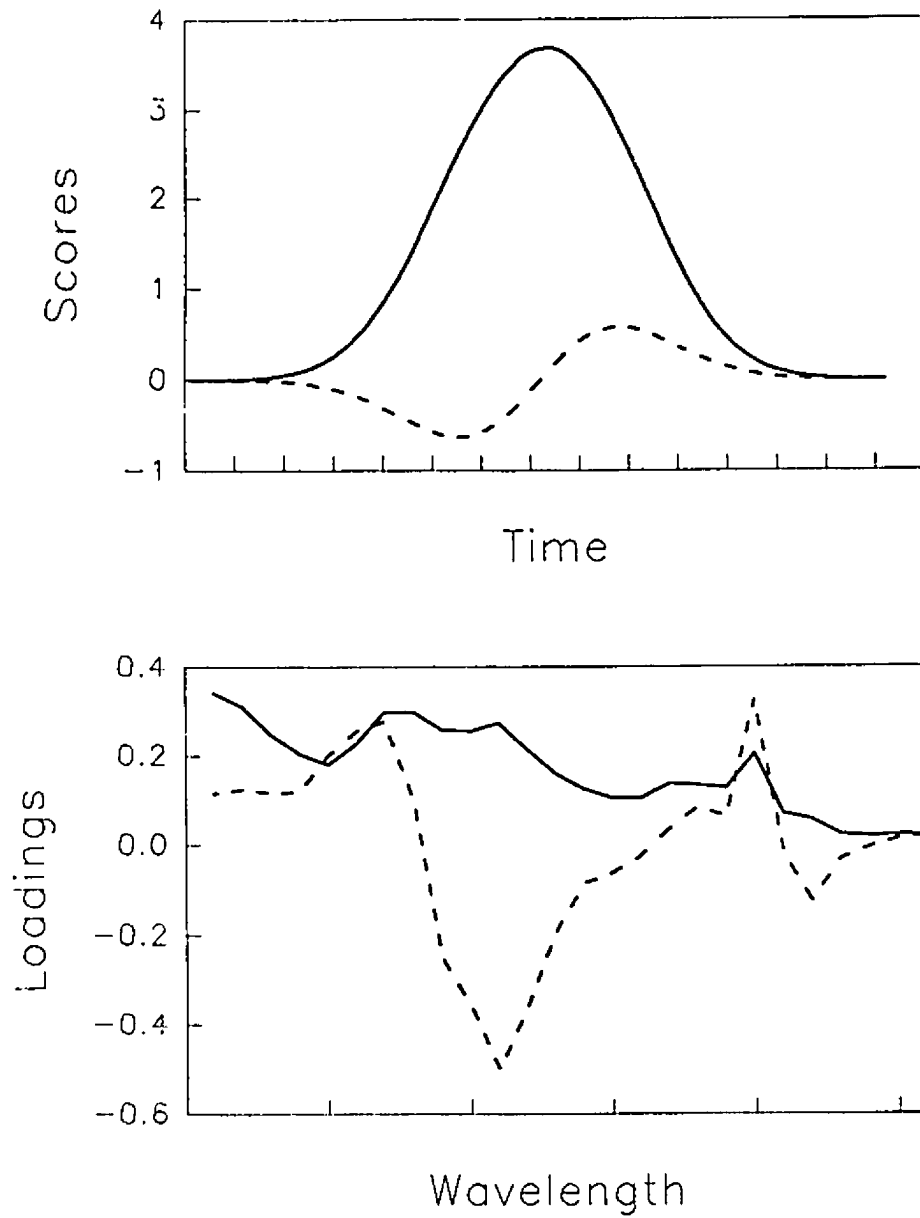
There are several advantages to this data compression. First, it is a more efficient way of representing the data. For this example, the data are reproduced by multiplying a matrix containing the coordinates of each point, with a matrix describing the direction of the eigenvectors in terms of the original measurement vectors. Thus 80 values (two 40 element matrices) contain the information extracted from a 400 element data matrix. The second advantage is the noise reduction achieved by discarding the vectors that represent primarily experimental error<sup>113,114</sup>. Calculations using this new basis set will propagate less noise.

The rank of a matrix<sup>115</sup> is defined as the minimum number of independent vectors it contains. For this two-component chemical system, each row of the data matrix  $D$  is some combination of the two spectral vectors, and thus the matrix has a rank of two. Accordingly it can be described by two principal components. For bilinear data, the number of observable components will be given by the true rank of the data matrix, but the apparent rank is often much larger due to experimental noise<sup>116</sup>. Thus an important stage in PCA is deciding which factors contain useful information rather than noise. An understanding of the chemical system and the measurement device is critical at this stage. In

practice this is a difficult problem, but it is a crucial step in applying factor analysis to chemical data. In this work, the number of principal components will be estimated with Kalman filter networks on the basis of little or no information other than the multivariate data set itself.

It is important to realize that while the set of factors obtained by PCA describe the data matrix as well as the real factors, they are not identical. For instance, the direction of the second eigenvector in Figure 3.3b is negative compared to some of the original axes, such as  $A_{\lambda,1}$  in this example. Figure 3.4 shows the scores and loadings matrices for the data shown in Figure 3.1. Since the columns of matrix  $\mathbf{X}$  are associated with the concentration vectors, they are often called abstract chromatograms. Likewise, the rows of  $\mathbf{Y}$  are called abstract spectra. In summary, PCA calculates a row matrix  $\mathbf{X}$  and a column matrix  $\mathbf{Y}$  that describe the data within experimental error using a bilinear model. This reduction is guided by mathematical goals and as such it may not be chemically relevant. Accordingly, these vectors are collectively called abstract factors. There are many advantages to using this new basis set. Three common applications using factor analysis are:

1. **Pattern recognition:** The principal components summarize useful information contained in the data<sup>117-119</sup>. This is convenient for visualizing the data, since the similarities or differences between samples can be evaluated with principal component plots. Since the eigenvectors are by definition orthogonal, they will contain unique information. In contrast, for the raw data set there are 190 different plots of the pairs of wavelengths that often contain highly correlated information.



**Figure 3.4.** Result of principal components analysis on the data shown in Figure 3.1. The first (solid line) and second (dashed line) principal components are shown.

2. **Prediction:** The properties of the samples can be calibrated and predicted as a function of their coordinates in this new space<sup>3,32,120,121</sup>. This procedure, called principal component regression, has been applied to many difficult calibration problems.
3. **Mixture Analysis:** In most cases the true factors (the pure component spectra) are not the same as the abstract factors, but they share the same space. In the two-component example above (Figure 3.3) the pure component solutions have been restricted to a two-dimensional plane instead of the original space of twenty noisy variables. The goal in mixture analysis<sup>111,112,122</sup> is to find a transformation that converts the abstract factors into real factors.

For all of these applications, choosing the correct size of the factor space is very important. In the analysis of mixtures, the number of factors should be the same as the number of observable chemical components in the mixture, but the determination of this value is complicated by random noise and other experimental artifacts.

### 3.4 DEDUCING THE NUMBER OF FACTORS

This section considers some different approaches for determining the number of chemical components that contribute to the spectra of chemical mixtures<sup>123-127</sup>. As was mentioned previously, this number would be the rank of the data matrix in the absence of experimental error. In practice, measurement noise inflates the eigenvalues of unnecessary eigenvectors to small, but nonzero, values<sup>128</sup>. So, while important factors typically have larger eigenvalues, it is difficult to



answer the question: "how large is a 'significant' eigenvalue?". Naturally, the magnitude of eigenvalue for noise vectors depends on the noise characteristics of the measurement system. When the magnitude of the noise is unknown before factor analysis empirical methods<sup>123,125</sup> must be used. These methods are based on the presence of trends in the progression of eigenvalues, often it is assumed that since meaningful factors explain a substantial amount of the variance in the data, their inclusion in the model is associated with a large drop in the next eigenvalue. In other words, their application will significantly reduce the amount of unexplained variance in the data. In contrast, the addition of an eigenvector that only explains random noise will not reduce the eigenvalue by as great a proportion since the noise is basically uncorrelated with the vector. These methods should be used with care since their assumptions may not hold true for all experimental measurements<sup>129</sup>.

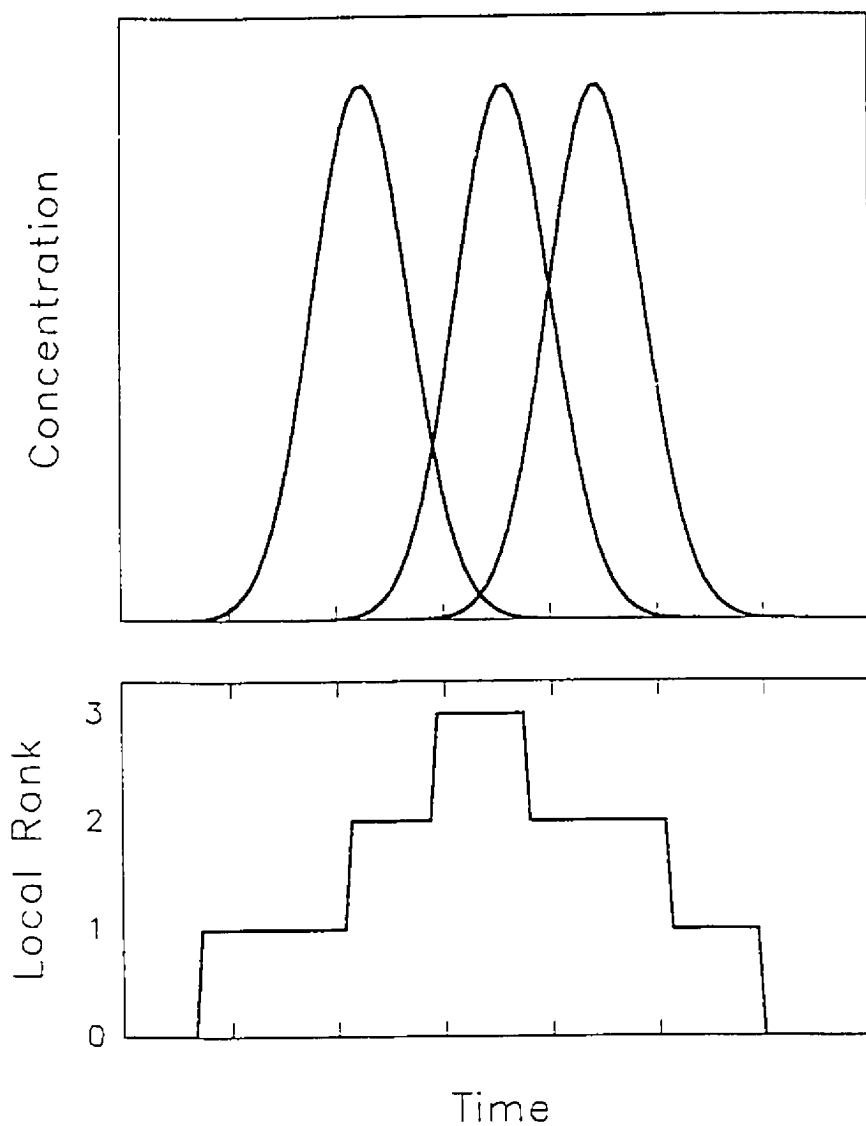
Another approach is based on examining the properties of the eigenvectors themselves. Here the assumptions are similar to those used in smoothing filters, namely that eigenvectors associated with chemical information will contain lower frequencies than those associated with noise<sup>130,131</sup>.

If the magnitude of the experimental noise is known or can be estimated from the data set, then it can be used to estimate the number of principal components<sup>126,132,133</sup>. Eigenvectors can be added to the model until its fitting errors are reduced to same size as the expected experimental error. The Kalman filter approach presented in this chapter falls into this category since it uses the innovations to indicate when systematic errors are occurring.

### 3.5 ORDERED DATA SETS

One weakness of the factor analysis methods mentioned so far is they do not take advantage of one important characteristic of chromatographic data - it is ordered in time. An ordered data set is one in which the contributions of underlying factors evolve continuously with time or some other variable in a manner that is consistent with the physical phenomenon being observed. For example, the chromatographic elution of components in a mixture will occur in a particular order. A spectrophotometric titration is another example of a process leading to an ordered data set since it involves the disappearance of one species and the appearance of others as the titrant is added, resulting in compositional changes that occur in a predictable fashion. In contrast, a series of samples for calibration measurements will not generally result in an ordered data set since there is no restriction on the relationship between two consecutive samples measured (this is especially true for multivariate calibration, where many components are unknown).

An example of noise-free ordered data from chromatography is shown in Figure 3.5. The overall rank of this data is three, since there are three absorbing species present, but local rank as well as global rank should be considered<sup>134</sup>. The local rank varies from zero, in the baseline regions, to the maximum of three. The eigenvalues that result from "batch" factor analysis do not reflect this information. These eigenvalues remain unchanged when rows or columns of the data matrix are interchanged, thus they cannot reflect the ordered nature of the data. When ordered data sets also involve multivariate measurements, such as UV-visible spectra, the temporal structure of the data



**Figure 3.5.** Concentration profiles are shown for a data set with three absorbing components (top). The overall rank of this data is three, but the local rank varies from zero to three (bottom) depending on the elution time.

can be exploited advantageously in the data analysis. One way to do this is through evolving factor analysis (EFA) <sup>135-140</sup> and related techniques<sup>141-144</sup>.

In principal components analysis (PCA), a data set is analyzed to determine its intrinsic dimensionality, or rank, and then expressed in terms of the product of a scores matrix and a loadings (or eigenvector) matrix. With EFA and related methods, subsets of the original data matrix are examined by PCA. The rank of these smaller matrices is often less than the overall rank and can provide information about the structure underlying the data. In EFA, the data matrix grows as a function of time (or other ordinal variable) as a new row of measurements is added to the matrix at each measurement interval. By plotting a suitable parameter, such as the logarithm of the eigenvalue for each eigenvector, changes in the rank (*i.e.* the appearance of new components) can be detected as a function of the ordinal variable (time, pH, etc.). There are two principal benefits to EFA: (1) it indicates where individual components begin to appear in the data sequence, permitting regions of spectral purity to be assessed and aiding in the extraction of pure component profiles, and (2) it improves the reliability of rank estimation by allowing relative comparisons to be made as opposed to the evaluation of a single figure of merit.

A variation of EFA is fixed-window evolving factor analysis<sup>142</sup> (FWEFA). In EFA, the size of the data matrix analyzed continues to grow as each new set of measurements is acquired, but for FWEFA the size of the data matrix remains constant. As the most recent measurements are acquired, the oldest ones are dropped, resulting in a data matrix that is a "window" of the entire matrix. The size of this window remains fixed as it "slides" along the data, and the resulting rank can increase and decrease depending on how many components are present in the window at a particular time. The main advantage of FWEFA is

that it can more clearly show the appearance and disappearance of components and provide an indication of where the peak analyte concentrations occur. The matrices employed are also smaller, which decreases calculation time but also reduces the noise rejection capabilities somewhat over EFA. It is also necessary to select the optimal window size for FWEFA.

An alternative to these approaches, which will be referred to as evolving principal component innovation analysis (EPCIA), will be described in the following section. This method exploits the best features of EFA and FWEFA, and is capable of real-time operation through its implementation via the Kalman filter.

### 3.6 PARALLEL KALMAN FILTER NETWORKS FOR PEAK PURITY ANALYSIS

The problem of mathematically resolving overlapped chromatographic profiles is an old and difficult one in chemical analysis. It consists of essentially three steps: detection of peak overlap, identification of individual analytes, and quantitation of components. A variety of methods have been proposed for the first of these problems, the simplest involving the monitoring of response ratios at two detector settings<sup>145</sup> (e.g., absorbances at two wavelengths). This approach, while straightforward, suffers from a number of disadvantages, including a susceptibility to a sloping background and the requirement of a pre-existing knowledge of wavelengths to be used. Its biggest drawback, however, is that it doesn't identify the number and nature of coeluting analytes or provide quantitative results.

In recent years, a number of algorithms based on PCA have been described for curve resolution in HPLC with UV-visible diode array detection<sup>111,112</sup>. One of the drawbacks of chromatographic curve resolution based on PCA is that calculations are generally performed after all of the data have been acquired. Chromatographic regions of interest must first be selected manually or automatically and subjected to PCA to determine the number of components. Thorough analysis requires interrogation of all chromatographic peaks. A simple peak purity test can be performed to screen particular areas of interest, but this suffers from the problems previously noted. A useful alternative would be the ability to carry out PCA recursively, i.e. while the data are being acquired. This would allow the determination of the number coeluting components in real-time and also act as a preprocessing step for self-modeling curve resolution. The development of a real-time PCA algorithm was the objective of this work.

The possibility of conducting recursive principal components analysis is made difficult by the fact that the usual PCA procedure is already iterative. To be capable of real-time implementation, a recursive procedure needs to maintain a static cycle time for each new data point obtained. One solution to the problem is to use the principles of adaptive Kalman filtering. The adaptive Kalman filter can be used as a recursive linear least-squares estimation procedure that has some built-in features to compensate for model errors. The strength of the adaptive Kalman filter is that it provides diagnostic information on model validity. If a parallel network of filters is employed, each with a different model, this information can be used to select the best model to fit experimental observations. This is the basis of the work described here<sup>146</sup>, in which initial results for a recursive PCA method are presented. The recursive PCA

procedure developed here is limited to two-component models, but extensions to higher dimensionality are possible. Computer simulated chromatographic profiles and experimental data from coeluting dyes are used to demonstrate the capabilities of the algorithm.

### 3.7.1 Theory

The use of parallel Kalman filter networks for kinetic methods of analysis was described in Chapter 2. The general scheme is illustrated in Figure 3.6. The incoming data sequence is applied to the inputs of a number of Kalman filters, each with a different model. These models may be used to handle data from nonlinear systems by introducing small variations in nonlinear parameters between adjacent models (quasi-continuous case) as was the case in Chapter 2, or they may represent distinct alternatives (discrete case). The recursive PCA application described here employs the latter form. In either case, the application of each filter provides new estimates of the state parameters for the corresponding model. These parameters are used to evaluate the performance of each model in terms of its consistency with actual observations. A useful measure of model performance is the innovations sequence which has been employed for adaptive Kalman filter algorithms<sup>45,46</sup>. The innovation is defined simply as the difference between the actual and predicted measurement (for multiple measurements in a single cycle, the innovation is a vector). The innovation differs from the residual normally used in modeling problems in that it is calculated after each measurement on the basis of current model parameters, whereas residuals are calculated in a batch procedure. The innovations

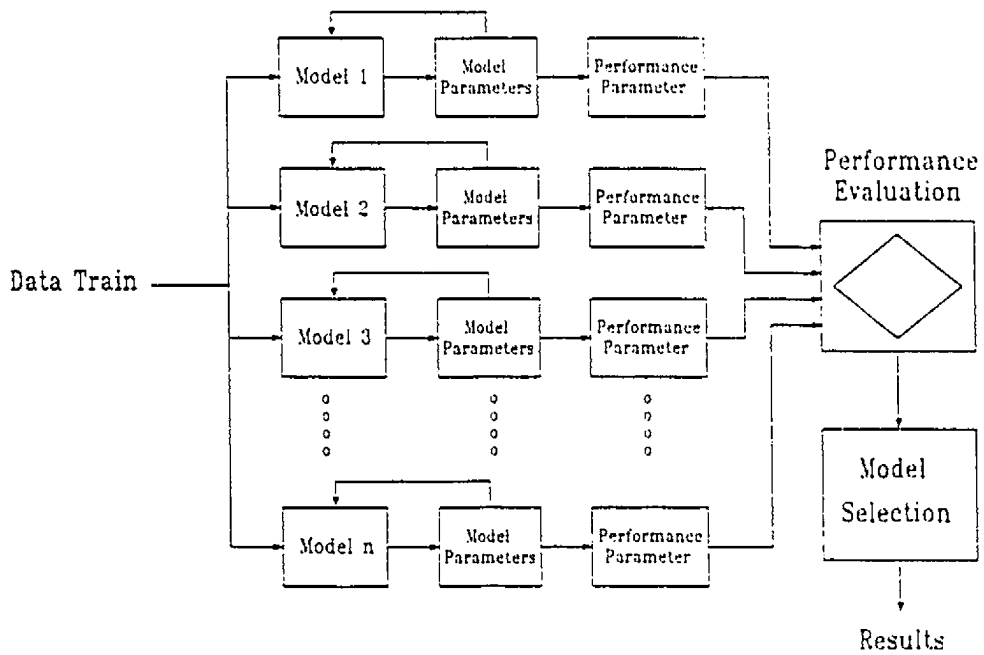
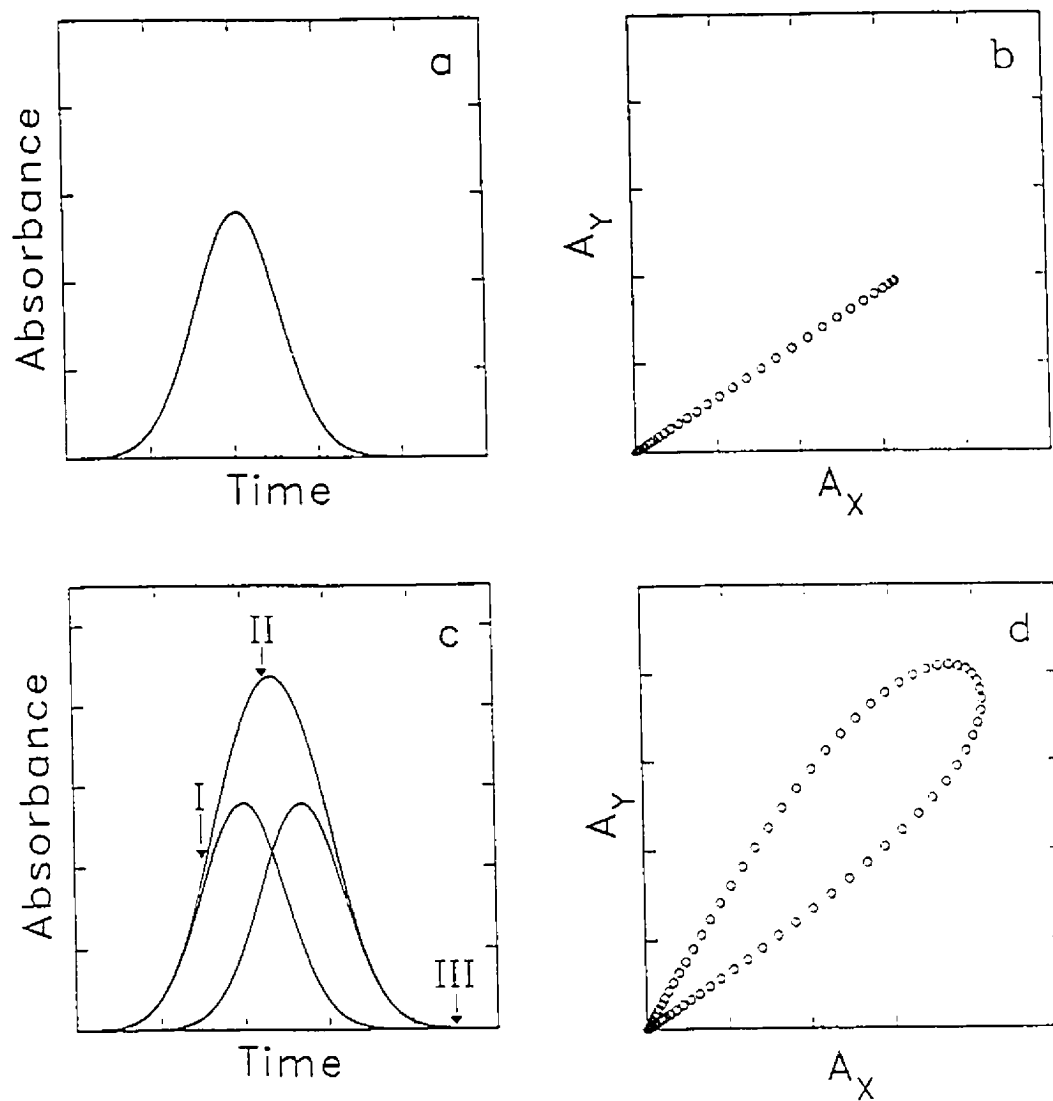


Figure 3.6. Strategy for implementation of parallel Kalman filter networks.



sequence is particularly useful where model deviations are localized, since it indicates regions of model validity.

The application of the Kalman filter to chromatographic peak resolution has been previously described by Brown and coworkers<sup>147,148</sup> and by Hayashi *et al*<sup>149</sup>, but these methods require a knowledge of individual component spectra or elution profiles and differ from the present work which seeks to function in the absence of this information. The principles of the recursive PCA approach are best illustrated with an example. Figure 3.7 shows synthetic chromatograms obtained for the elution of one- and two-component mixtures. For simplicity, Gaussian, noise-free peaks have been assumed, but this is not a requirement of the algorithm. It is also assumed that the two components are not completely overlapped and have sufficiently different spectral profiles. These are requirements of most curve resolution methods. Shown adjacent to the two representative chromatograms in Figure 3.7 are plots of absorbance measurements at two wavelengths for each sampled point. The wavelengths selected are the absorbance maxima of the two hypothetical components. Plots such as these (which will be referred to as  $A^2$  plots) illustrate the principles of both the absorbance ratio and PCA methods for peak purity assessment. It is clear that the one-component mixture gives a constant  $A_2/A_1$ , while the two-component case does not. Methods based on PCA extend this principle further by recognizing that in an  $n$ -dimensional absorbance space ( $A^n$ ), the one-component case will always fit a linear model within experimental error. Likewise, a two-component chromatogram will fit a planar model in  $A^n$  space. Therefore, the minimum number of components in an overlapped region can be determined from the intrinsic dimensionality (*i.e.* rank) of multivariate absorbance



**Figure 3.7.** Chromatographic elution profiles and their projections into  $A^2$  space: (a) & (b), one-component profile and its projection; (c) & (d), two-component profile and its projection.

data. Obviously, the PCA method is more general than simple approaches such as those based on absorbance ratios.

Two notable drawbacks of usual methods based on PCA are the need for batch processing of the data and difficulty of accurately determining the dimensionality of the data. The former problem has already been mentioned. The latter results from the difficulty of distinguishing residual eigenvectors arising from experimental error from those that represent true components. This problem is further complicated by experimental realities such as a sloping background that may appear as additional chromatographic components. Such features, while of interest in quantitation, can be misleading in the detection of peak overlap. Part of the difficulty in the determination of the true number of chromatographic components is that rank is normally assessed on the basis of one or more scalar quantities<sup>103,150,151</sup> that ignore information available in the temporal structure of the data. Clearly, deviations from an  $n$ -dimensional model due to the presence of additional components or a sloping background should exhibit characteristic patterns if the evolution of the model is examined. This behavior can be detected if PCA is performed recursively.

The principle of operation of recursive PCA or evolving principal components analysis (EPCIA) is that an  $n$ -dimensional data set projected onto an  $n+1$  dimensional space (or higher) should always give a fit to a multilinear function which is satisfactory within experimental error. Thus, the one-component data in Figure 3.7 will give a satisfactory fit to a straight line at any two wavelengths, but the fit for the two-component data should be unsatisfactory for at least certain pairs of wavelengths. Both data sets should give a good fit to a planar model in any  $A^3$  space, but a three-component data set should not. This strategy can be extended to higher dimensions, although visualization

becomes more difficult. On this basis, the intrinsic dimensionality of the data set can be deduced by selecting a number of wavelengths and fitting data in lower subspaces comprised of various wavelength combinations. A one-component model should exhibit comparable residuals for both linear and planar models, but a two-component data set should exhibit excessive residuals for the linear model at certain wavelength combinations. Furthermore, the resulting models will serve as a means to construct good approximations to the principal components vectors. For example, the linear models should be projections of the first eigenvector, and the planar models should all be contained in the planes defined by the first two eigenvectors. The correspondence between the true eigenvectors and the reconstructed vectors may not be exact, since the multilinear least-squares models are generally constructed assuming no errors in the  $x$ -direction<sup>152</sup>, but the correspondence should be close under the right conditions.

The use of multiple models of lower dimensionality offers no particular advantages over traditional PCA except when used in conjunction with the Kalman filter. The Kalman filter can be used to generate fits to linear and planar models recursively. This increases the potential for real-time implementation of the algorithm and observation of model evolution. Before the algorithm is initiated,  $n$  wavelengths at conveniently spaced intervals are selected. For the one-component model, the absorbance at one wavelength (designated  $A_x$ ) is selected as the independent variable and a series of  $n-1$  models of the form,

$$\hat{A}_{ij} = \alpha_i A_{xj} + \beta_i \quad (3.9)$$

are used for the Kalman filter. In this equation,  $A_{ij}$  represents the predicted absorbance at wavelength  $i$  for measurement  $j$ ,  $A_{xj}$  is the measured absorbance

at the wavelength chosen for the independent variable, and  $\alpha_i$  and  $\beta_i$  are parameters associated with the one-dimensional model. The parameter estimates evolve as each measurement is processed by the Kalman filter. The model may be limited to the trivial case of single parameter ( $\alpha_i$ ) if a zero intercept is assumed, but this will not always be the case. These models are used as  $n-1$  elements of the parallel filter network. Likewise,  $n-2$  two-dimensional (planar) models of the form,

$$\hat{A}_{ij} = \alpha_i A_{xj} + \beta_i A_{yi} + \gamma_i \quad (3.10)$$

are also used. In this case, absorbances at two wavelengths (arbitrarily designated as  $A_x$  and  $A_y$ ) are needed as independent variables and three parameters are estimated. Models of higher dimensionality are also possible, but were not employed in this initial work. The  $2n-3$  models described are sufficient to indicate one, two, or more than two coeluting components.

The strategy for implementation of the models in the parallel Kalman filter network is shown in Figure 3.8. As each set of absorbance values for a single chromatographic point is received, a set of innovations corresponding to the  $2n-3$  models is calculated using predicted measurements according to,

$$e_{ij} = \hat{A}_{ij} - A_{ij} \quad (3.11)$$

where  $e_{ij}$  is the innovation for measurement  $j$  at wavelength  $i$ , and  $\hat{A}_{ij}$  and  $A_{ij}$  are the predicted and measured absorbances, respectively. In each cycle, there are  $n-1$  innovations calculated for the one-dimensional models and  $n-2$  for the two-dimensional models. The magnitude of the innovations should be close to the measurement noise level if the correct model is used, but significant deviations should be observed otherwise. Thus, the absolute value of the innovations

## Parallel Filter Network for Coelution Profile Analysis

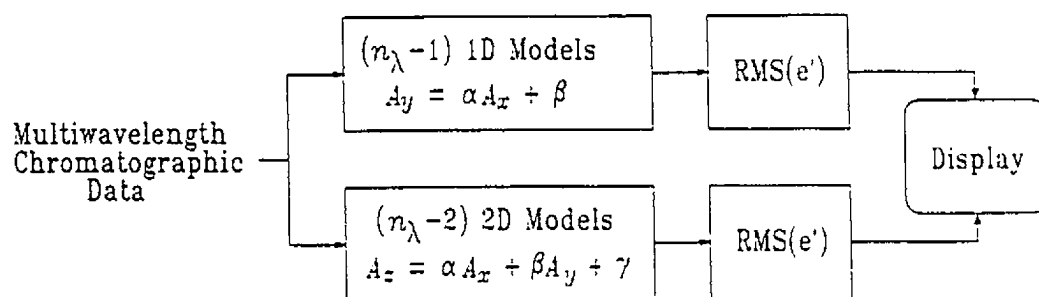


Figure 3.8. Parallel Kalman filter network for recursive PCA application.

sequence can be used to indicate when the dimensionality of the data does not match the dimensionality of the model. Not all of the innovations will provide this information, however, since the wavelengths used in a particular model may not be appropriate for observing model variations (e.g. if there is no absorbance for the dependent variable). One way around this problem is to examine the maximum absolute innovation for each set of models, but this will be sensitive to outliers in the data. An alternative approach is to calculate the root mean square (rms) value of innovations for each set of models:

$$\text{rms}(e_k) = \sqrt{\frac{\sum e_l^2}{m_k}} \quad (3.12)$$

where the summation is over the number of models of order  $k$ ,  $m_k$ . The rms values can be plotted in real-time along with the chromatogram and the sequence should remain fairly flat as long as the dimensionality of the data is a subset of the model space.

A number of practical problems need to be addressed in implementing the above strategy. The first is the selection of wavelengths which will act as independent variables in the models. The selection is not entirely arbitrary since the independent variable will exhibit a certain amount of uncertainty along its axis. In the case of a one-dimensional model, this means that if the wavelength selected for the x-variable shows no absorbance for any component, the least-squares fit will result in a nearly vertical line for those models whose dependent variable is non-zero. Although this may result in a valid least-squares fit, the innovations (measured in the y-direction) will be excessively large. One way to minimize this problem is to ensure that the wavelengths chosen as the independent variables in an  $n$ -dimensional model exhibit significant absorbance

for some portion of the data. In practice, the wavelengths exhibiting the highest absorbance when the peak is first detected are used. This does not eliminate the problem entirely since the magnitude of the innovations will still increase with the slope of the line. In the two-dimensional case, the problem is compounded by the likelihood of high correlation between the independent variables. A more robust solution uses innovations measured orthogonally from the model rather than vertically. For the one-dimensional model, it can be shown that the orthogonal innovation is given by,

$$e'_{ij} = \frac{-\alpha_i A_{xj} + A_{yj} - \beta_i}{\sqrt{\alpha_i^2 + 1}} \quad (3.13)$$

and likewise for a two-dimensional model,

$$e'_{ij} = \frac{-\alpha_i A_{xj} - \beta_i A_{yj} + A_{ij} - \gamma_i}{\sqrt{\alpha_i^2 + \beta_i^2 + 1}} \quad (3.14)$$

Extension to higher dimensions is straightforward. These modified innovations were used in place of the usual values for model evaluation (Equation 3.12) since they should more accurately reflect the true model errors. The modified values were not used in the Kalman filter algorithm, however, since they would lead to erroneous results.

Another practical aspect of the implementation of the EPCIA algorithm concerns its activation. In practice, the Kalman filter network is not activated until a peak is detected, although in principle the baseline region could be used if appropriate independent variables could be selected in advance. Once



activated, the rms innovations for each model (Equation 3.12) may exhibit high values for the first two or three points as the model converges on reasonable parameter estimates. Display of these points could be suppressed, but this was not done for the results presented here and is generally not necessary.

***Relationship to Principal Components Analysis.*** In view of the importance of the connection drawn between PCA and the Kalman filter algorithm developed here, a more detailed discussion of this relationship is warranted. There are obvious computational differences between the traditional batch processing method for performing PCA and the multilinear approach presented here. This means that the resultant vectors are not necessarily identical, but the differences should be small enough to be inconsequential. It is known that the Kalman filter will provide model equations that are virtually identical to the traditional least-squares method as long as the diagonal elements of the covariance matrix are initially set to large values relative to estimated measurement error<sup>38,50</sup>. In this work, a ratio of  $>10^{10}$  was normally used for the diagonal elements in order to achieve the least-squares solution (off-diagonal elements were initially set to zero).

The relationships between the PCA eigenvectors and the Kalman filter results are as follows. For the set of one-dimensional models given by Equation 3.9, the vector resulting from the combination of all models into  $A^n$  space corresponds to the first eigenvector obtained by traditional PCA methods if the absorbance data were mean-centered at each wavelength. As this eliminates residual eigenvectors resulting from an offset at particular wavelengths, it is preferred for peak purity analysis. If mean-centering of the absorbance vectors is not carried out prior to batch PCA, the first eigenvector will correspond to the

vector generated by the models in Equation 3.9 if the  $\beta_i$ 's are forced to zero. In either case, the first eigenvector from the Kalman filter ( $\mathbf{e}_1$ ) will be given by:

$$\mathbf{e}_1 = [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{n-1} \quad 1]^T / \sqrt{1 + \sum_{i=1}^{n-1} \alpha_i^2} \quad (3.15)$$

The first  $n-1$  components of the vector are the projections of the  $n-1$  dependent variables of the models, while the last represents the wavelength selected as the independent variable. The denominator simply serves to normalize the vector. To find the second eigenvector, both the one- and two-component models are required. This is because the two-component model only defines the plane containing the first two eigenvectors and not the vectors themselves. Generally, if it is known that two-components are present, a knowledge of the plane of the first two eigenvectors is sufficient to perform self-modeling curve resolution. Nevertheless one method of obtaining the actual eigenvectors is presented here. As in the case of the one component models, omission of the offset parameter (in this case  $\gamma_i$ ) will lead to the PCA result for data which are not mean-centered for absorbance. The plane containing the first two eigenvectors will also contain the vectors  $\mathbf{v}_1$  and  $\mathbf{v}_2$ .

$$\mathbf{v}_1 = [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{n-1} \quad 1 \quad 0]^T$$

$$\mathbf{v}_2 = [\beta_1 \quad \beta_2 \quad \dots \quad \beta_{n-1} \quad 0 \quad 1]^T$$

The  $(n-1)^{th}$  component of the vectors corresponds to the first independent variable and  $n^{th}$  component to the second. These vectors correspond to the intersection of the planar model with the  $(n-1)$  dimensional subspaces along the axes of  $A_x$  and  $A_y$ . After normalization of  $\mathbf{v}_1$  to  $\mathbf{n}_1$ , a vector  $\mathbf{n}_2$  which is

orthonormal to  $\mathbf{n}_1$  is determined by projection of  $\mathbf{v}_2$  onto  $\mathbf{n}_1$ , subtraction, and normalization.

$$\mathbf{n}_1 = \mathbf{v}_1 / |\mathbf{v}_1| \quad (3.16)$$

$$\mathbf{p} = \mathbf{v}_2 - (\mathbf{v}_1^T \mathbf{v}_2) \mathbf{v}_1 \quad (3.17)$$

$$\mathbf{n}_2 = \mathbf{p} / |\mathbf{p}| \quad (3.18)$$

The vectors  $\mathbf{n}_1$  and  $\mathbf{n}_2$  are just one set of orthonormal vectors which can be used to define the plane. Ideally, the first eigenvector obtained from the one-dimensional models ( $\mathbf{e}_1$ ) will lie in this plane, but in practice there may be a slight elevation due to minor computational differences for the two models. To ensure the integrity of the two component plane, the first eigenvector is recalculated as its projection onto the plane defined by the orthonormal vectors.

$$\mathbf{e}'_1 = (\mathbf{e}_1^T \mathbf{n}_1) \mathbf{n}_1 + (\mathbf{e}_1^T \mathbf{n}_2) \mathbf{n}_2 \quad (3.19)$$

In all cases that we have checked, the difference between  $\mathbf{e}_1$  and  $\mathbf{e}'_1$  has been insignificant, but calculation of the projection is more robust. Calculation of the second eigenvector,  $\mathbf{e}_2$ , in the two-dimensional subspace is now trivial.

$$\mathbf{e}_2 = -(\mathbf{e}_1^T \mathbf{n}_2) \mathbf{n}_1 + (\mathbf{e}_1^T \mathbf{n}_1) \mathbf{n}_2 \quad (3.20)$$

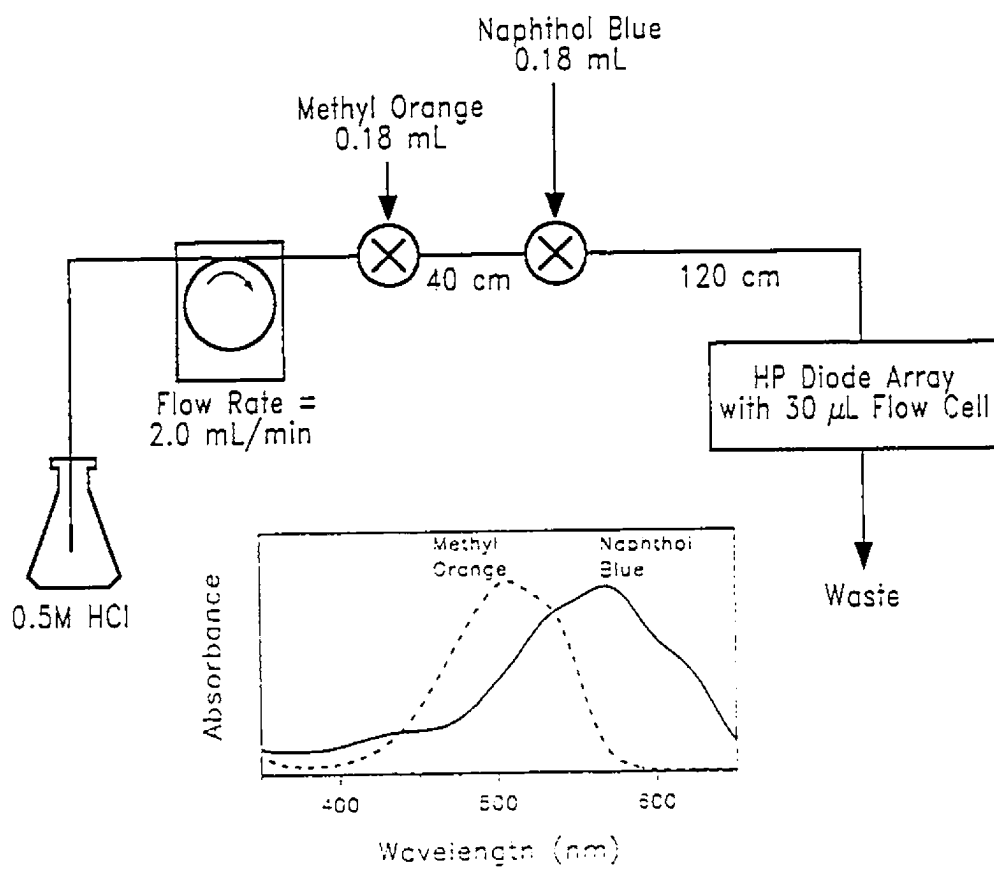
Extension of these principles to systems of higher dimensionality should be straightforward, but will not be considered here.

### 3.7.2 Experimental Section

**Simulation Studies.** Generation of absorbance vs. wavelength vs. time data to test the Kalman filter algorithm was carried out with a program that allowed a variety of conditions to be simulated. For simplicity, Gaussian shapes were assumed for spectral and concentration profiles. Normally distributed random values were added to simulate measurement noise. Further details of conditions used accompany results presented in the Results and Discussion section.

**Dye Coelution.** In the absence of an HPLC/photodiode array system, experimental results for coeluting components were obtained using a continuous flow system in a merging zones configuration as shown in Figure 3.9. Dye solutions employed for the results presented here were 0.312  $\mu\text{M}$  methyl orange (acid orange 52, color index 13025; Fisher, Fair Lawn, NJ) and 1.17  $\mu\text{M}$  naphthol blue (Meldola's blue, basic blue 6, color index 51175; Pfaltz & Bauer, Waterbury, CT), both prepared in 0.5 M HCl. These concentrations were found to give a noise level suitable for testing the algorithm. The samples were injected simultaneously into the stream using two six-port two-way valves (Rheodyne model 5020, Cotati, CA) with 180  $\mu\text{L}$  sample loops. The injected samples were transported to the detector through 0.8 mm i.d. PTFE tubing by an 8-roller Ismatec SA peristaltic pump (Cole-Parmer, Chicago, IL). A Hewlett-Packard model 8452A photodiode array spectrometer (Hewlett-Packard, Palo Alto, CA) with a 30  $\mu\text{L}$  flow cell (Hellma Cells, Jamaica, NY) was used to acquire spectra at 1 s intervals for about 100 s after injection.

**Computational Aspects.** All calculations were carried out on a 16-MHz IBM PC/AT compatible computer with a math coprocessor. Programs were written in



**Figure 3.9.** Merging zones continuous flow apparatus for studies of dye coelution with spectra of dyes inset.

Microsoft QuickBASIC (Microsoft Corp., Redmond, WA). Implementation of the Kalman filter employed the standard algorithm<sup>41</sup> with double precision arithmetic. Principal components analysis was carried out using subroutines written in our laboratory and based on procedures outlined by Malinowski and Howery<sup>103</sup>. Three-dimensional displays of experimental data were generated with the program SURFER<sup>R</sup> (Golden Software, Golden, CO).

### 3.7.3 Results and Discussion

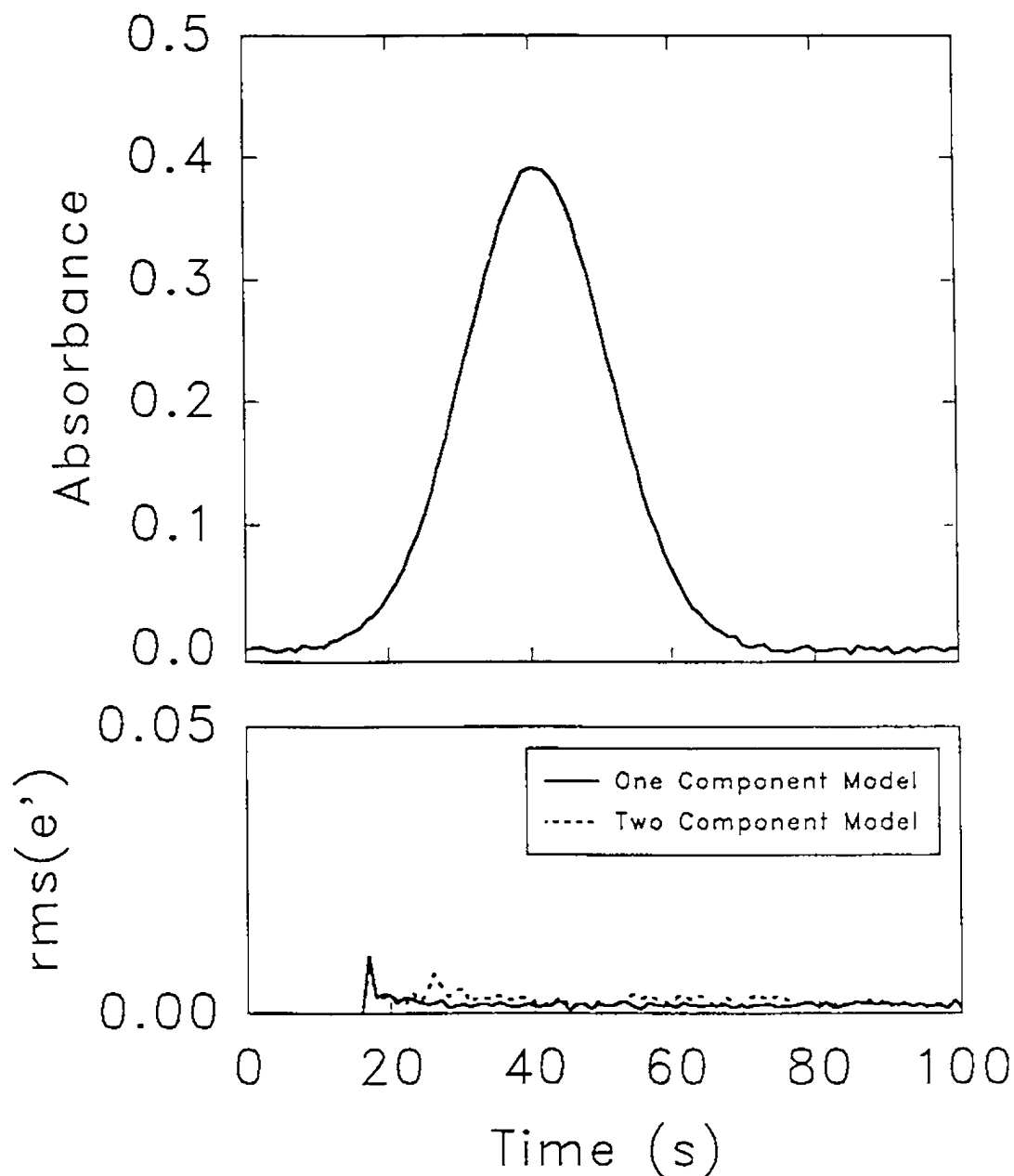
**Simulation Studies.** To provide an initial evaluation of the EPCIA algorithm, simulated chromatographic data were used. Because a large number of parameters will affect the performance of the algorithm (chromatographic peak shapes, spectral profiles, spectral and chromatographic resolution, number of components, component ratios, noise level, background absorbance, number of wavelengths used, etc.), only a limited subset of results is presented here to demonstrate the principles of the method. More complete studies to investigate the limitations of the algorithm are presented in Chapter 4.

The simulated experimental data presented here utilized Gaussian profiles in both the chromatographic and spectral domains for simplicity. Component 1 was assigned a wavelength maximum of 400 nm and component 2 a maximum at 500 nm. The width of both spectral peaks was  $\sigma = 100$  nm and equivalent molar absorptivities were assumed. The concentration ratio and chromatographic resolution of components were varied between studies. Chromatographic peak widths of  $\sigma = 10$  s were used with a simulated spectral sampling rate of 1 s. The noise level (typically 0.5% RSD) was computed as a

percentage of the maximum of the entire absorbance matrix. For all of the results presented here, 10 wavelengths at equally spaced intervals were used.

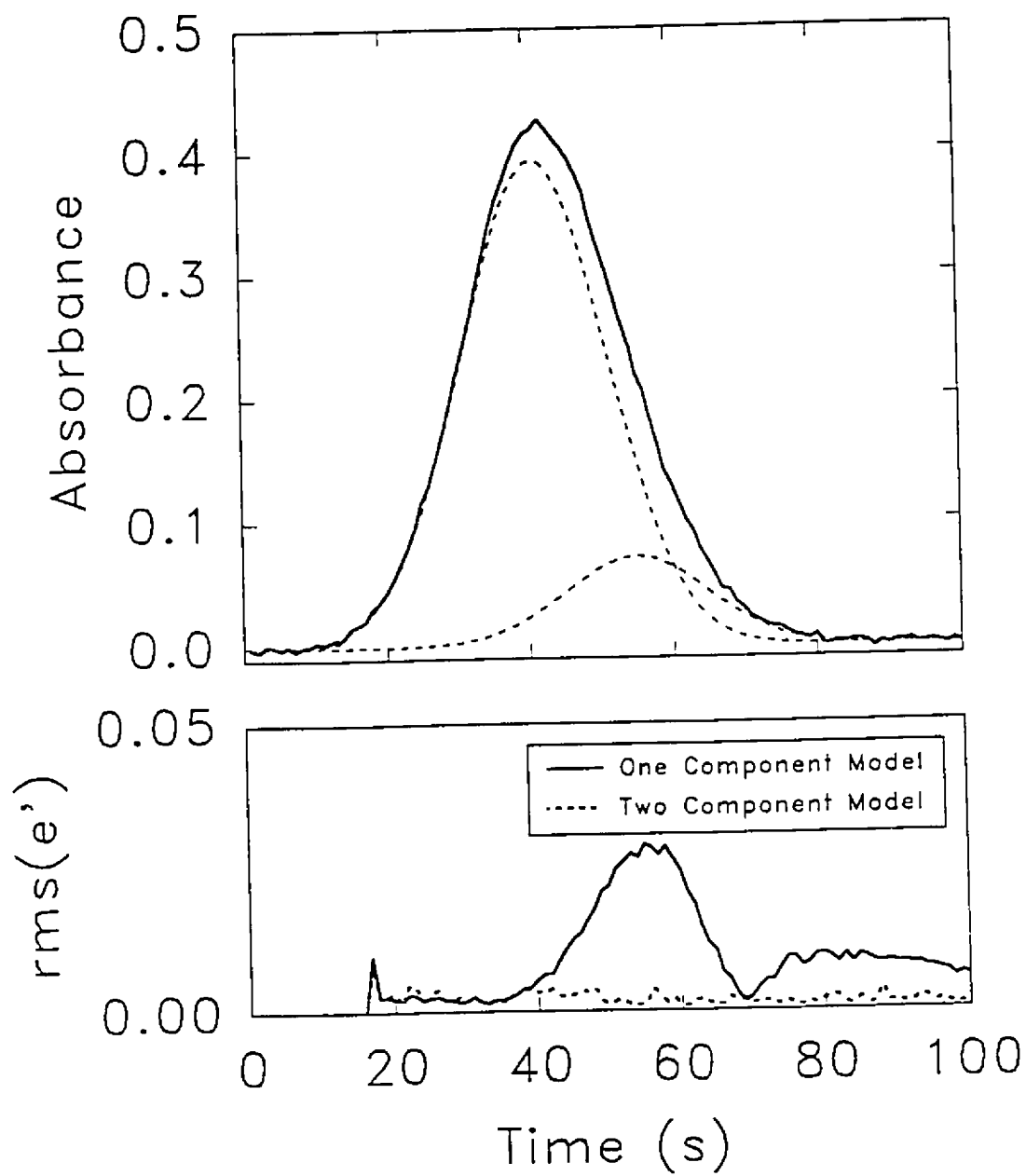
Figure 3.10 shows typical results obtained with the parallel Kalman filter network for the elution of a single component. The top part of the figure shows the chromatographic trace at the absorbance maximum while the bottom portion shows the rms orthogonal innovations for the one- and two-component models. Note that both models indicate acceptable performance, verifying that there is only one component present. In contrast, Figure 3.11 shows results for two eluting components (3:1 ratio, 0.35 resolution, 0.1% noise). Under these conditions, the two-component model gives a fairly flat innovations trace, while the trace for the one-component model indicates significant model deviations. Furthermore, the point at which the innovations sequence begins to diverge for the one-component model reveals where the second component begins to appear. This information is not available from batch PCA and is significant because it allows a key set of factors to be identified for target transformation<sup>103,111,153</sup>. This could expedite the generation of component elution profiles considerably.

In order to illustrate how the EPCIA algorithm functions, the evolution of a single one-dimensional model for a two-component data set is shown in Figure 3.12. One  $A^2$  data space for the data in Figure 3.7c is shown. Lines in Figure 3.12 correspond to the one-dimensional model at various points throughout the elution of the peak and are labeled to correspond to points indicated in Figure 3.7c. Initially, when only one component is present, the linear model fits the observed data relatively well and the innovations are small. As the second component introduces curvature into the data, the least-squares fit must accommodate this nonlinearity and the model begins to track the

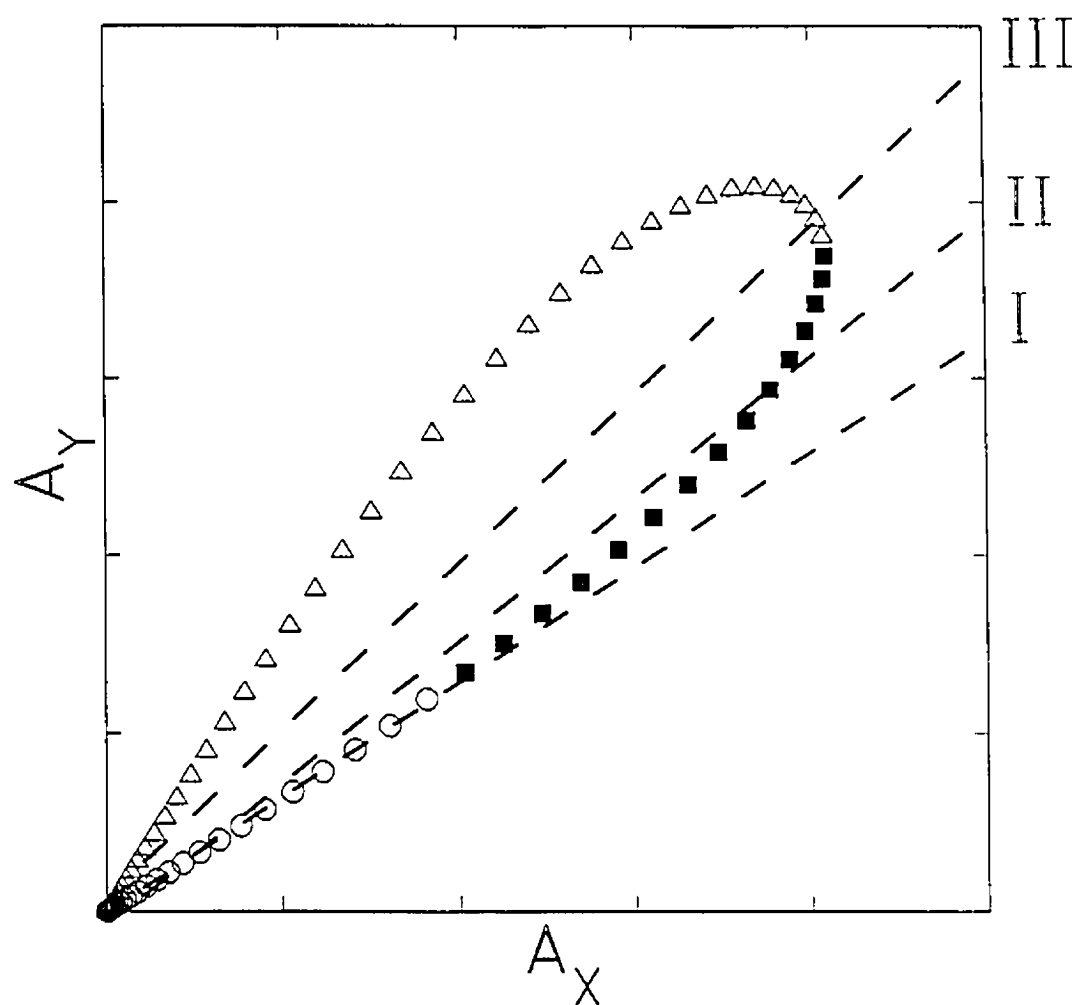


**Figure 3.10.** Results of application of Kalman filter PCA algorithm to a single component elution profile (simulated). The top trace shows the chromatographic signal at the wavelength of maximum absorbance. The bottom trace shows the sequence of rms orthogonal innovations for each model type.





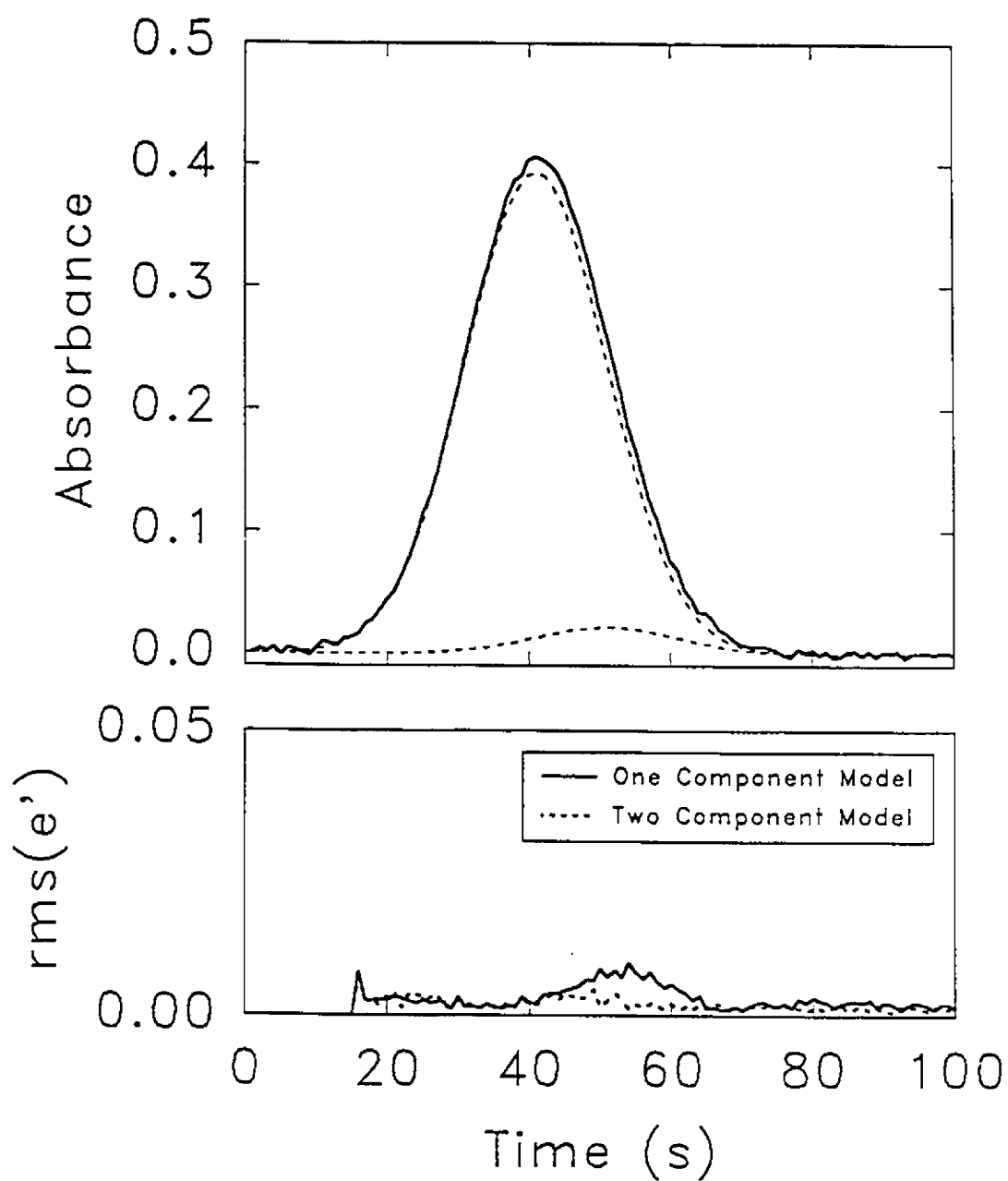
**Figure 3.11.** Results of application of Kalman filter algorithm to a simulated two-component elution profile.



**Figure 3.12.** Evolution of the one-component model for the data in Figure 3.7c. The model equations (dashed lines) are shown at three stages: initial (after the open circles, I), intermediate (after the filled squares, II), and final (III). Corresponding points on the elution profile are indicated in Figure 3.7c.

measurements more poorly, leading to larger innovations. Although the model is no better when the signal returns to the baseline, measurements near the origin do not exhibit large deviations and so the innovations return to their original level.

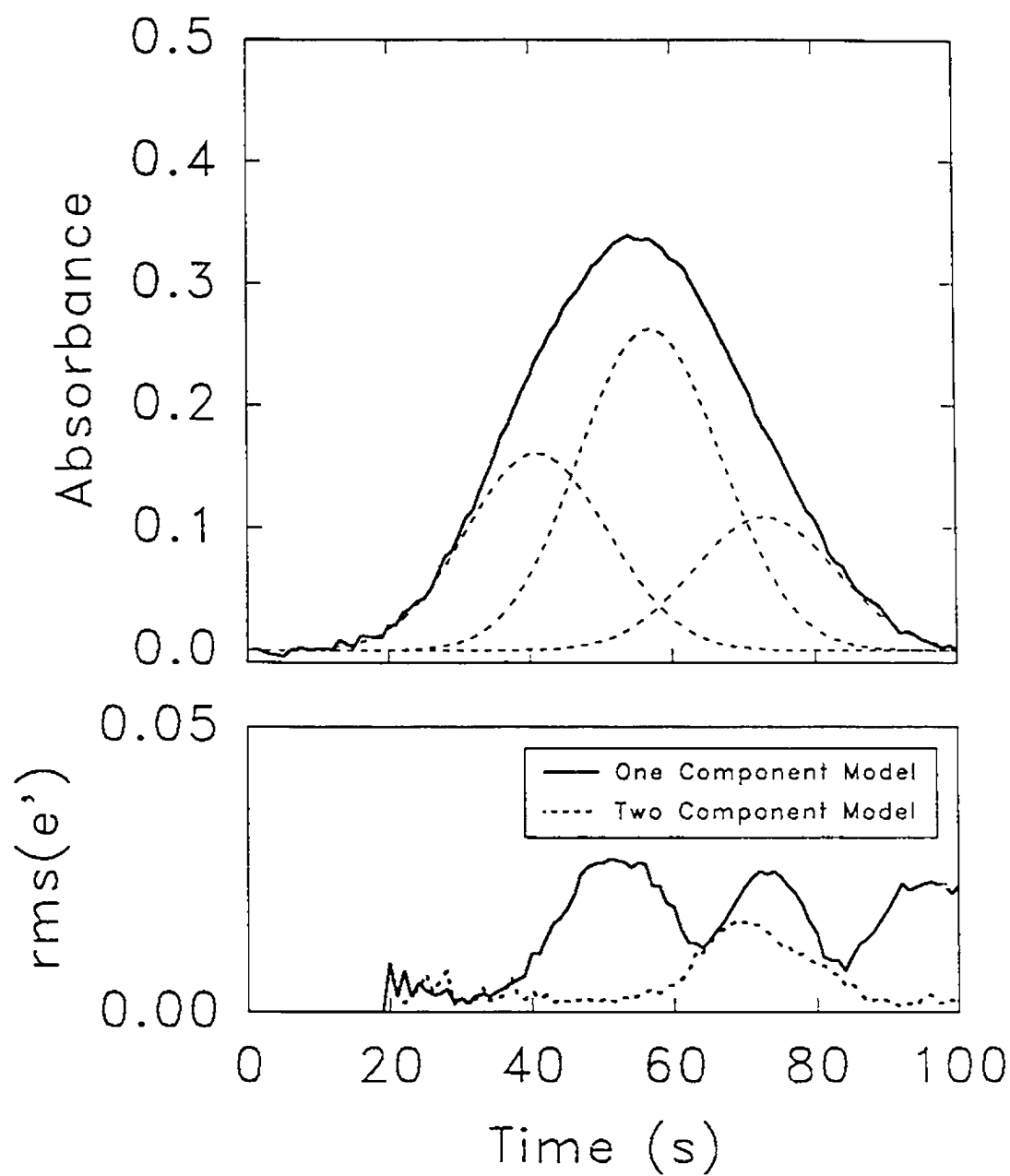
As an indication of the limitations of the EPCIA algorithm, Figure 3.13 shows results obtained with a 10:1 component ratio and a resolution of only 0.25. The second component can still be detected in this case. The ability of the algorithm to detect minor components is very dependent on the noise level, as expected. Generally it was found that when the recursive algorithm failed to distinguish a second component, visual inspection of the data in the plane resulting from the first two eigenvectors also suggested only one component. As anticipated, performance of the method also improves with chromatographic and spectral resolution, and with the number of wavelengths used. The latter effect arises from the increased likelihood of selecting wavelengths with maximum discriminating ability, and smoother traces for the rms innovations. Improvements in results achieved by increasing the number of wavelengths are quickly limited by the spectral correlation of the two components, however. The order of component elution (*i.e.* minor component first or second) affects the shape of the innovations trace but does not significantly diminish the ability of the algorithm to determine the dimensionality of the data set in most cases. In some extreme cases, the rms innovations of the two-component model exhibit a small disturbance when the second component is detected (*i.e.* the reverse of the usual case) but this due to the fact that the planar model "floats" around its primary axis until the necessary points are obtained to more rigidly define the second eigenvector.



**Figure 3.13.** Results of application of Kalman filter algorithm to a simulated two-component elution profile near limiting conditions.

An example of a simulated three-component mixture is shown in Figure 3.14. In this case, the third component was assigned a wavelength maximum of 300 nm with  $\sigma = 100$  nm. The concentration ratio ( $c_1:c_2:c_3$ ) is 1:3:1 and components elute in order with a resolution of 0.4 between adjacent peaks. Other conditions are as previously given. Note that the trace of rms innovations indicates the successive failure of the one- and two-component models. It should be pointed out, however, that failure of the two-component model was not always observed for three-component mixtures, depending on the relationships among spectra and elution profiles. It is believed that this problem has its roots in the correlation between wavelengths selected for independent variables. Solutions include a more careful selection of wavelengths or imposition of a complete set of models with all wavelength combinations. Other options also exist, but may not be necessary as the pattern of the innovations for the one-dimensional model will indicate the presence of a third component in most cases.

The computational performance of the parallel Kalman filter network is currently limited by its implementation in serial fashion but is still quite acceptable. Cycle times of about 0.1 s are not difficult to achieve with one- and two-component models at ten wavelengths. This is in a range suitable for most chromatographic applications. The efficiency of the serial implementation will diminish as the models of higher dimensionality are added and the number of wavelengths is increased. The highly parallel nature of the algorithm can exploit trends in computing towards vector processing, however, and this should dramatically reduce computation time.



**Figure 3.14.** Results of application of Kalman filter algorithm to a simulated three-component elution profile.

**Experimental Results.** Since simulated experimental data often imposes deterministic and stochastic characteristics which are not observed in practice (e.g. Gaussian profiles, uncorrelated noise), the EPCIA algorithm was also applied to experimental data from the coelution of organic dyes. One of the data sets used in this study, obtained with the apparatus in Figure 3.9, is shown in Figure 3.15. Dye concentrations were reduced to a level which gave a relatively noisy signal (approximately 3% baseline noise relative to the absorbance maximum). It also appears from the figure that the noise exhibits some correlation, possibly due to pump pulsations. The ratio of peak heights (methyl orange to naphthol blue) was 2:1 and the resolution (determined by individual injection) was about 0.4. Ten wavelengths at equally spaced intervals were used. Results of the application of the Kalman filter are shown in Figure 3.16. The presence of two components in the elution profile is clearly indicated by the rms innovations sequence even though the noise level is quite high.

**Comparison with PCA.** A comparison between the eigenvectors computed by the usual batch PCA procedure and those determined by the parallel Kalman filter network is given in Table 3.1. The basis of the comparison is the angle between the eigenvectors calculated through the batch PCA procedure and the Kalman filter network. Results under various simulation conditions for a ten-dimensional space (ten wavelengths) are shown. The first eigenvector used from the Kalman filter network was that obtained from the combination of one- and two-component models rather than from the one-component model alone, but differences were insignificant.

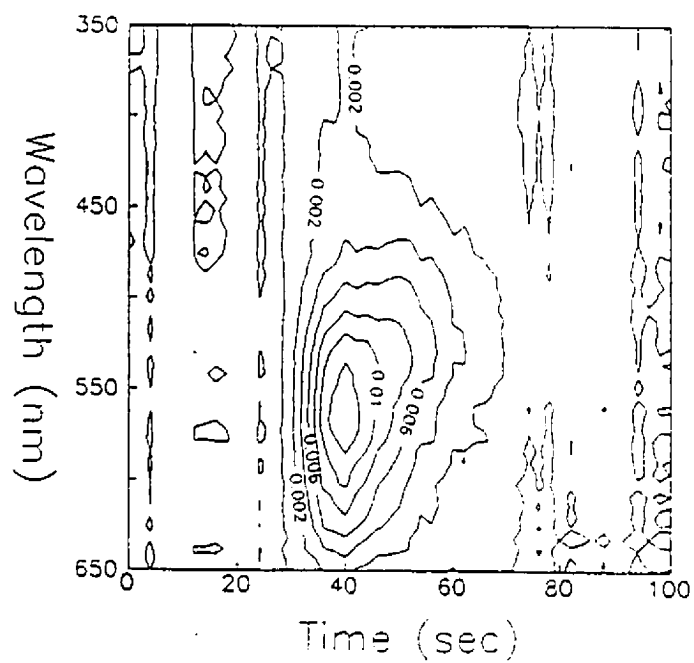
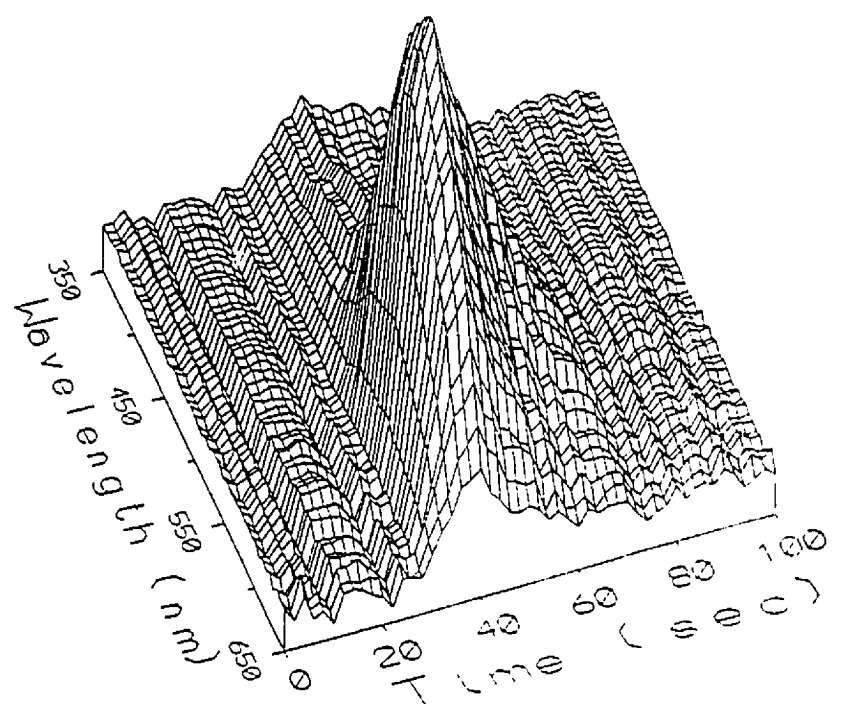
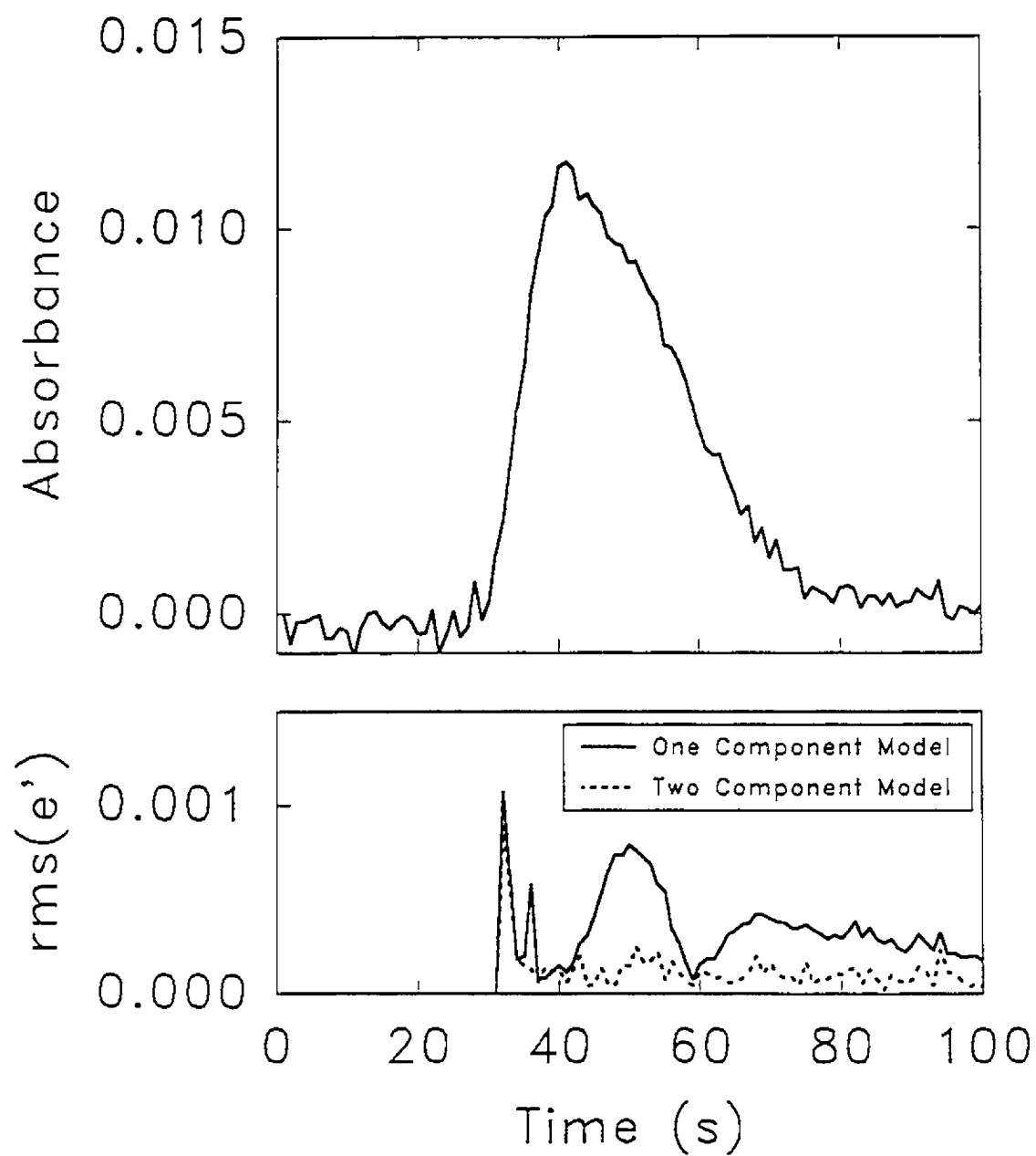


Figure 3.15. Absorbance matrix obtained from the coelution of organic dyes.





**Figure 3.16.** Results of the application of the recursive PCA algorithm to the data in Figure 3.15.

**Table 3.1.** Comparison of eigenvectors produced by traditional principal components analysis and the Kalman filter method.

<i>no. of comp.</i>	<i>concn ratio</i>	<i>resoln</i>	<i>noise level, %</i>	<i>angle between eigenvectors, deg</i>	
				<i>first eigen-vector</i>	<i>second eigen-vector</i>
1	-	-	0.5	0.01	53.0
2	3:1	0.35	0.5	0.06	1.37
2	3:1	0.35	2.0	0.49	7.99
2	10:1	0.25	0.5	0.02	6.39
2	5:1	0.2	0.5	0.08	3.12
2	1:5	0.2	0.5	0.06	3.02

Agreement between the two PCA methods, while not perfect, is very good in most cases. One exception is where an attempt is made to determine the second eigenvector for the one-component data set. In this case, the second eigenvector is defined purely by noise and is of no real consequence, however. Certainly the agreement between batch and recursive procedures should be good enough to permit further calculations, such as self-modeling curve resolution, to be carried out.

The recursive PCA method does not directly provide eigenvalues or the row and column matrices associated with batch PCA, but these can be easily determined (if necessary) once the eigenvectors are assigned. Eigenvalues are not as essential for determination of rank with the recursive algorithm since this information is provided by the rms innovations sequence.

### 3.7 CONCLUSIONS

The initial studies presented here have demonstrated the viability of performing principal components analysis recursively through the use of a parallel Kalman filter network. Application to the problem of chromatographic peak purity analysis has shown how the rank of a data matrix can be deduced while the data are being acquired. Although more extensive studies are required to fully explore the potential and limitations of the EPCIA approach, several important advantages are apparent. First, because the algorithm is recursive and parallel with a fixed cycle time, it should be significantly faster than traditional PCA methods, especially when implemented on parallel computing architectures. The speed advantage does not result from a more computationally efficient algorithm, but rather because data analysis is performed while data are being acquired. A second advantage of the recursive approach is that it provides information on the temporal evolution of models. This is particularly useful in cases such as chromatography and titrimetry where certain types of behavior can be anticipated. To obtain equivalent information by batch PCA, numerous subsets of the data would have to be processed independently. The information provided by EPCIA should be particularly useful in resolving ternary component mixtures by window factor analysis<sup>144</sup> since it identifies regions in which certain models are valid. Furthermore, it can help diagnose model deviations arising from factors such as a sloping background. Absolute information on model deviations is readily provided by the rms innovations sequence, which should approximate measurement noise when the model is valid. Finally, the flexibility of the Kalman filter models allows for a variety of processing options to be

exercised, simultaneously if desired. Inclusion of the offset term in the models, for example, will have the same effect as mean-centering the absorbance data prior to batch PCA. Unlike some approaches, however, the absorbance data are not normalized, so the measurement noise information is retained at its original magnitude.

In spite of these advantages, the utility of the Kalman filter network in cases where real-time data processing is not required remains to be explored. This will be addressed in Chapter 4. The algorithm is also likely to become less useful as the number of factors to be extracted becomes large, since the number and complexity of models become more difficult to handle. Nevertheless, it may allow techniques such as self-modeling curve resolution to be more readily implemented in real-time.

#### 4.1 INTRODUCTION

In this chapter, some aspects of the application of EPCIA to unresolved two-component mixtures in liquid chromatography with multiwavelength UV-visible diode array detection are discussed. The treatment is limited to two-component mixtures because of the difficulties involved in studying interactions of many variables in mixtures of more components, and because this case is especially important for peak purity assessment in quality control<sup>142</sup>. Simulated and experimental data are used to demonstrate how the performance of the technique is affected by spectral correlation, chromatographic resolution and peak shape, concentration ratios, and other factors.

The principles of the EPCIA algorithm were described in the previous chapter and will only be treated briefly here. The approach used is to consider each spectrum in the ordered data set as a point in the  $n$ -dimensional absorbance space, where  $n$  is the number of wavelengths employed (typically 20 to 50). With factor analysis based methods, this is the space decomposed by PCA. With EPCIA, a modified treatment produces results that closely approximate those from PCA, but allows the use of the Kalman filter for real-time implementation. The philosophy of the EPCIA approach is that the chromatographic elution of a one-component mixture will give rise to a data set that is intrinsically one-dimensional in an  $n$ -dimensional space and therefore will project onto straight lines in the  $(n-1)$  two-dimensional subspaces. The rms of

the orthogonal innovations should therefore approximate the standard deviation of the absorbance measurements at all points across the elution profile; *i.e.* it should be a flat line when the measurement errors have a constant variance. The appearance of a spectrally different second component incompletely resolved from the first will lead to a two-dimensional data set in the absorbance space that will not project onto straight lines. Prediction errors will then become larger than anticipated, leading to an increase in the rms innovations where the second component appears. In this way, the second component is detected. Although the procedure described refers to the implementation of a one-component model, it can be extended with relative ease for the extraction of higher principal component vectors. In this work, since the emphasis is on detecting impurities under chromatographic peaks, higher order models are not extensively used.

In the results presented here, the rms sum of the orthogonal innovations obtained from the EPCIA algorithm, designated as  $\text{rms}(e')$ , is evaluated as a tool for detecting incomplete resolution of chromatographic peaks. In some cases, this parameter is plotted below the chromatogram to illustrate how it deviates from the baseline in the presence of a second component. In other cases, the maximum value of  $\text{rms}(e')$  is plotted to indicate its sensitivity to various experimental factors.

## 4.2 EXPERIMENTAL

The EPCIA algorithm was characterized primarily through simulation of spectral and chromatographic responses, since this allowed maximum flexibility in the conditions and permitted the effect of variables to be independently assessed.

Gaussian profiles were generally used in both domains, although in some cases spectra of polycyclic aromatic hydrocarbons (PAH's) were employed. The PAH's (anthracene, fluoranthene, phenanthrene, and triphenylene) were obtained from Aldrich Chemical Co. (Milwaukee, WI) with a minimum 98% purity. In all simulations, Gaussian noise was added at varying levels. The EPCIA algorithm was written in Microsoft QuickBASIC (v. 4.5), and run on an 80486 based computer under DOS 5.0. Analysis of a complete data set generally took only a few seconds.

Experimental results presented were obtained using an HP 8452A diode array spectrometer (Hewlett-Packard, Palo Alto, CA) as the detector. For chromatographic studies, this was equipped with a 30  $\mu$ L flow cell (Hellma Cells, Jamaica, NY) at the end of a C<sub>18</sub> reversed-phase column (10 x 0.46 cm) with a 20  $\mu$ L injection valve (Rheodyne, Cotati, CA). Para-xylene calibration solutions were prepared in ethanol to cover a range of  $1 \times 10^{-4}$  to 1 % (v/v). For some experiments, a photographic step tablet (#3, Eastman Kodak, Rochester, NY) was used to reduce the source intensity. This strip was graduated with 21 neutral density filters ranging from approximately 0.05 to 3.05 AU. For the stopped-flow studies, the solutions were transported through the flow cell using two peristaltic pumps (Ismatech MS Reglo, Cole Parmer, Chicago, IL). Methyl orange solution from 2 to 15 ppm and praseodymium chloride solutions from 1 to 8 % (w/v) were both prepared in 0.1 M HCl.

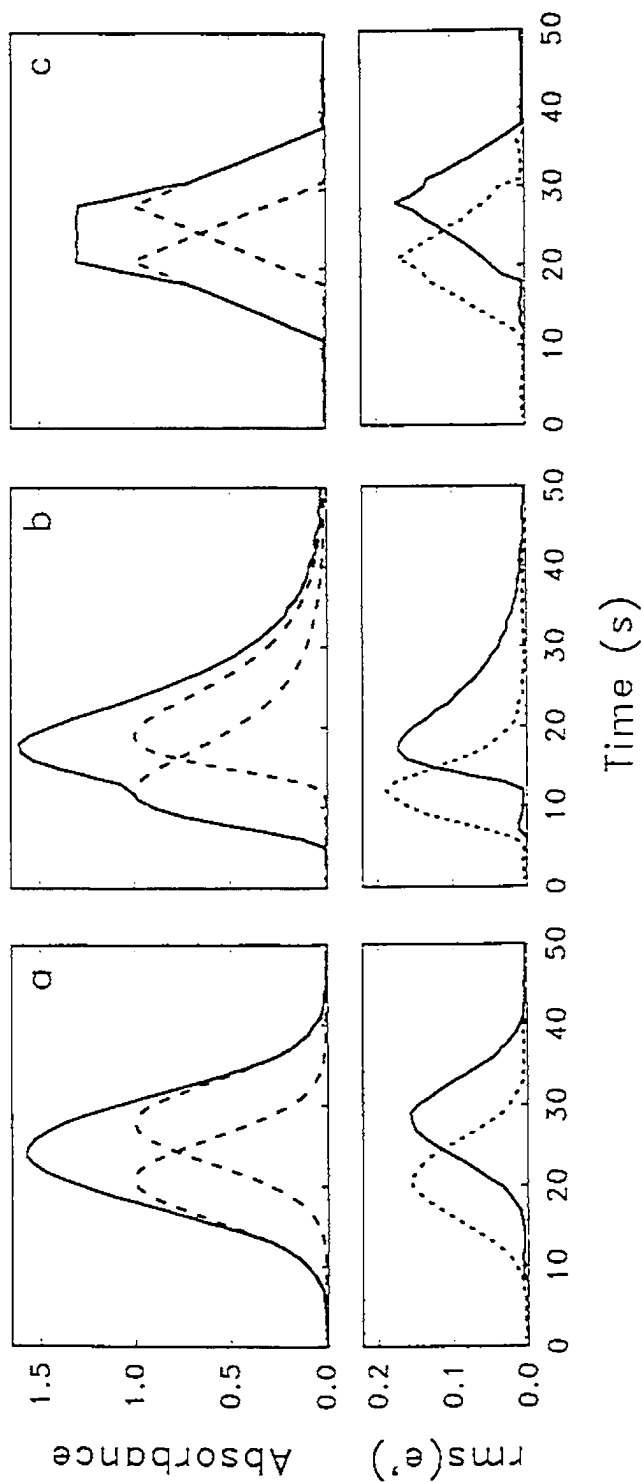
## 4.3 FUNDAMENTAL LIMITATIONS

### 4.3.1 Effect of Peak Shape

The ultimate objective of curve resolution methods in chromatography is to obtain the pure component elution profiles and spectra. In the application of EPCIA, it was noted that the shape of the rms innovations sequence closely resembled the pure component elution profiles. To investigate this further, a series of simulation results were generated using different elution profiles. Profiles used were Gaussian, exponentially modified Gaussian (EMG)<sup>154</sup>, and triangular. The individual component profiles, as well as the overall chromatographic peak (with 0.5% Gaussian noise added) are shown in the top part of Figure 4.1. The spectra used for the two components were Gaussian with  $\lambda_{\max 1} = 450$  nm and  $\lambda_{\max 2} = 550$  nm, a width of  $\sigma = 100$  nm, and equal heights. For the filter network, 21 wavelengths were used at equally spaced intervals between 300 and 700 nm. The bottom part of the Figure 4.1 shows the rms innovations sequences obtained when the filter was used in the forward and reverse directions. It is necessary to filter the data in both directions to obtain information on both components. In each of the three cases, it is apparent that the shape of the innovations sequence closely approximates the true peak shapes. Small differences are evident, however, particularly in the case of the EMG peaks, and these have been found to depend on the shape of the peak and concentration ratio of the two components.

The similarity between the innovation sequences and the peak shapes can be understood by considering the evolution of the first principal component vector generated by the Kalman filter network. As the first component begins to

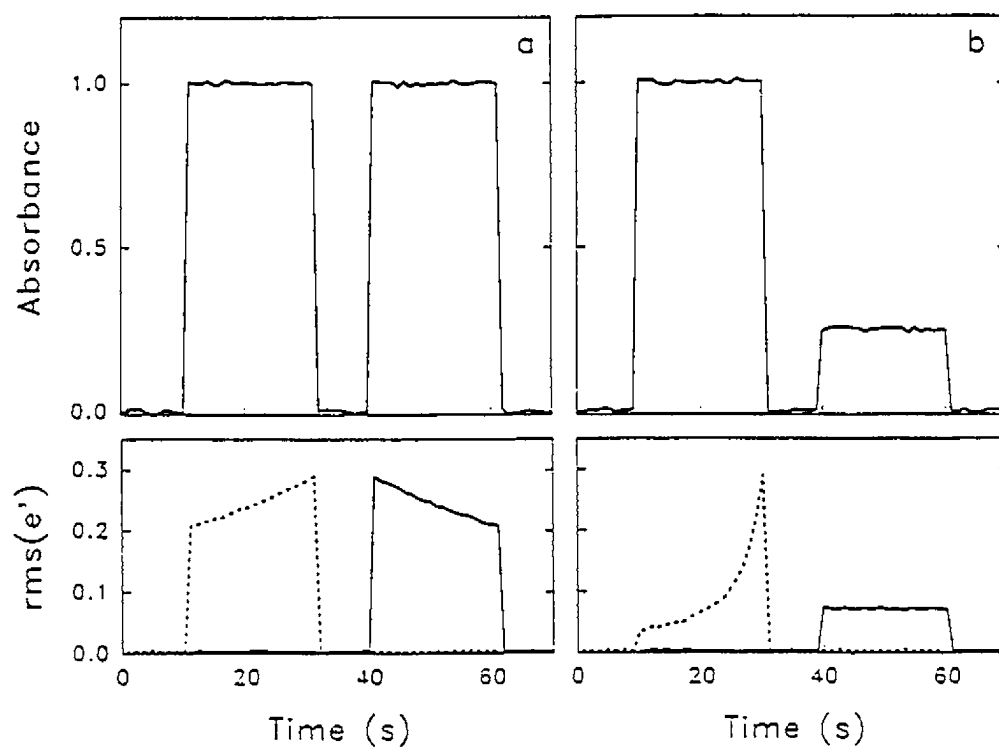




**Figure 4.1.** Results of the application of the EPCIA algorithm to simulated two-component data. The top panels show the combined (solid line) and the individual (dashed lines) elution profiles for (a) Gaussian; (b) exponentially modified Gaussian; and (c) triangular profiles. The bottom panels show the rms innovations obtained by filtering in the forward (solid line) and reverse (dashed line) directions.

elute, the first eigenvector is aligned with its spectrum in the  $n$ -dimensional absorbance space. If this vector were "locked in" before the appearance of the second component (as is done in adaptive Kalman filtering), then the innovations would approximate the difference between the spectrum of component 1 and the total spectrum for future points. This difference depends on the spectrum of the second component and the concentration ratio of the two components. Since the spectrum of the second component is fixed for a given mixture, the difference reflects the shape of the concentration profile for the second component. In reality, EPCIA does not fix the position of the first eigenvector, but allows it to continue to adjust once the second component appears. Nevertheless, the eigenvector lags the mixture spectrum in absorbance space and the effect is similar to results obtained if it had been fixed. Thus an approximation to the elution profile of the second component is obtained. This approximation is most valid when the second component is the minor component and the changes in its concentration are gradual. Of course, these arguments extend to the component eluting first when the filter is passed through the data in the reverse direction.

The effects of peak shape and component concentration on the accuracy of the approximation of the innovations sequence are clearly illustrated in Figure 4.2. The simulated concentration profiles for this case were rectangular functions with a concentration ratio of 1:1 for Figure 4.2a and 4:1 for Figure 4.2b. As before, the figure compares the rms innovations sequences for forward and reverse filtering with the true elution profiles. In Figure 4.2a, both concentration profiles are poorly estimated because of the abrupt changes arising from the rectangular peak shapes. In Figure 4.2b, however, the profile of the second component is more accurately modeled in spite of its unusual shape because



**Figure 4.2.** Results for the application of the EPCIA algorithm to simulated rectangular concentration profiles. The top panels show profiles with (a) 1:1; (b) 4:1 concentration ratios. The bottom panels show the resulting rms innovations for the forward (solid line) and reverse (dashed line) filters.

the large influence of the major component effectively fixes the first eigenvector. This observation is significant because it is usually the elution profile of the minor component of a mixture that is more difficult to estimate. Such profiles may be particularly useful as starting points for methods such as iterative target transformation factor analysis<sup>151</sup>. Furthermore, better elution profiles should be possible by coupling the results of EPCIA with principles of adaptive Kalman filtering.

It should be pointed out that in order to obtain good peak shape approximations, the EPCIA models should not contain the intercept term (constant) as a model parameter. This assumes that the pure spectral vectors pass through the origin, but the inclusion of the additional term permits two ways for the models to adapt to the appearance of an additional component and does not allow for the calculation of a reliable difference spectrum.

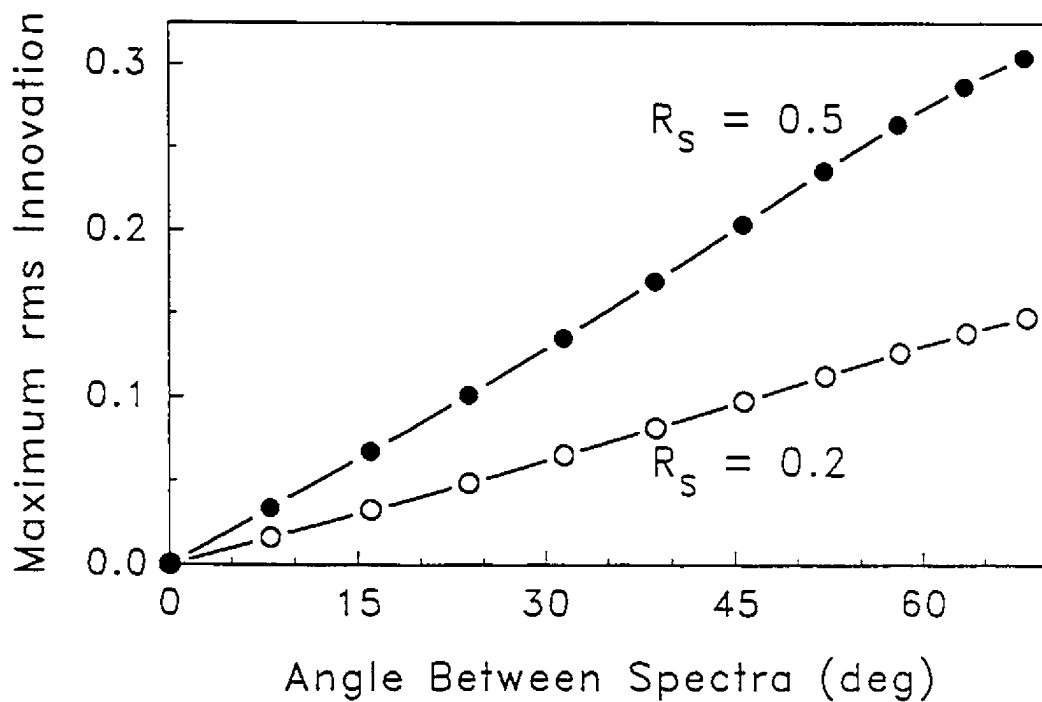
#### **4.3.2 Effect of Spectral Correlation**

The ability of the EPCIA algorithm to detect the presence of an unresolved impurity depends primarily on the similarity of the two spectra, the chromatographic resolution, the concentration ratio, and the noise level. With this method, the baseline level of the rms innovations sequence in the absence of a second component should be approximately equal to the noise level in the absorbance measurements. This provides a reference point when assessing limitations imposed by the first three factors; that is, one can examine whether or not the maximum in the rms innovations sequence exceeds the baseline level. Unfortunately, there is an interaction among these factors that can make this difficult to do in practice. Nevertheless, the three factors were examined

independently to see if an indication of their general influence could be obtained.

To examine the effect of spectral correlation on the EPCIA algorithm, a series of simulations were run in which the similarity of the two spectra was varied. The simulated spectra were Gaussian with  $\sigma = 50$  nm and  $\lambda_{\max 1} = 400$  nm. The position of  $\lambda_{\max 2}$  was varied to adjust the spectral correlation, and wavelengths were sampled at 10 nm intervals beginning at  $2\sigma$  before the first peak (*i.e.* 300 nm) and ending at  $2\sigma$  after the second peak (*i.e.*  $\lambda_{\max 2} + 100$  nm). Although the number of wavelengths sampled varies with spectral correlation in this approach, it is preferable to adjusting the sampling interval each time. As discussed later, the number of wavelengths used does not generally affect the magnitude of the maximum rms innovation, but the sampling interval can be important when sampling becomes sparse. Gaussian peaks were also used for the chromatographic profiles ( $\sigma = 5$  s with a sampling interval of 1 s) and results are reported for chromatographic resolutions of 0.5 and 0.2 ( $R_S = \Delta t / 4\sigma$ ). Conditions were set so that the maximum absorbance of the pure component profiles was unity and the noise level was 0.001%. The noise level was low in these studies to allow the observation of the systematic variations in innovations sequence.

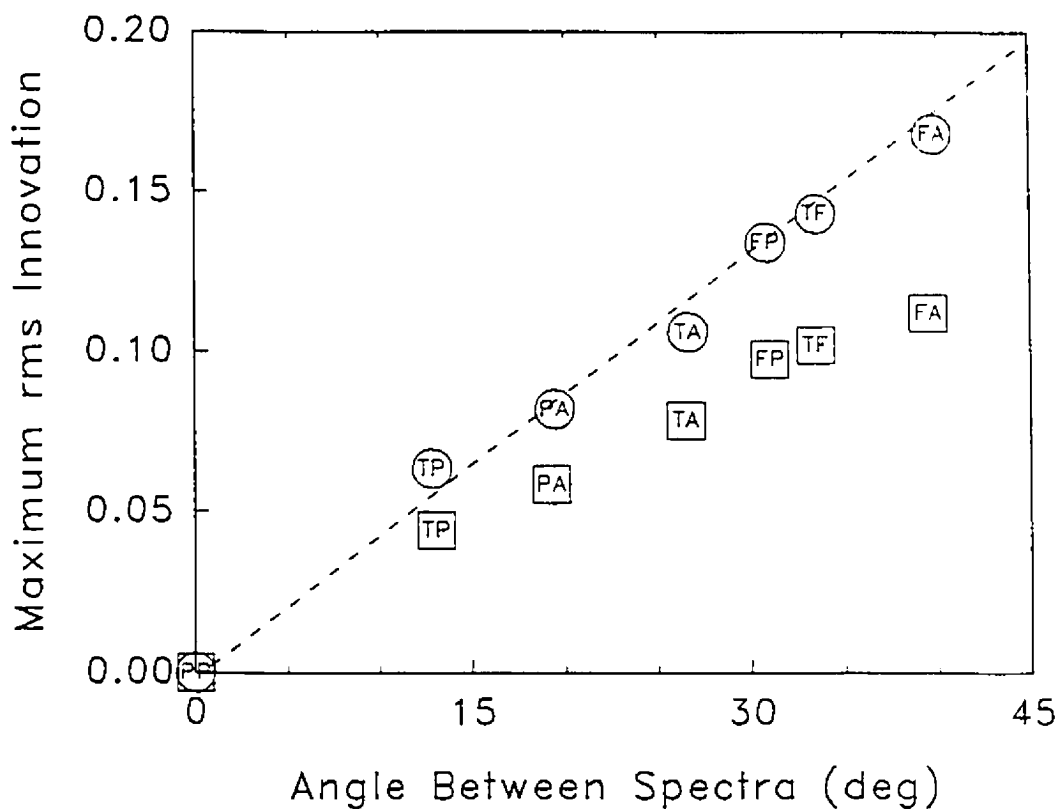
Figure 4.3 shows the results of the simulation studies. The maximum of the rms innovations sequence (forward filtering) for the two different chromatographic conditions is plotted as a function of spectral similarity. The spectral correlation on the abscissa is represented as the angle between the pure component spectra in absorbance space,  $\theta$ , rather than with the correlation coefficient since the latter proved to be less linear. As expected, the maximum rms innovation (*i.e.* the peak of the rms innovations sequence), which is directly



**Figure 4.3.** Effect of spectral correlation (Gaussian profiles) on the maximum rms innovation for two different chromatographic resolutions ( $R_S$ ).

related to the detectability of the second component, decreases as the spectral similarity increases ( $\theta$  decreases). The relationship is approximately linear with  $\theta$  over most of the range, with a slope that is dependent on chromatographic resolution as shown. At spectral angles above those shown in the figure, the line shows some curvature as the result of the fact that the spectra are almost completely resolved and there is a region of little or no absorbance between them.

When real spectra are considered, the study of the effects of spectral correlation are complicated by the fact that the spectra do not normally have the same shape. In simulating the effects with real spectra, it is desirable to normalize the two spectra so that the effects of spectral similarity are isolated from changes in the effective concentration ratio (see section 4.3.4). Normalization can be done on the basis of the area or maximum of the spectra, with somewhat different results. A second set of simulations was carried out with the same conditions as in Figure 4.3 (with  $R_S = 0.35$ ), but using the measured spectra of polycyclic aromatic hydrocarbons (PAH's) rather than pure Gaussians. These were normalized to the same maximum absorbance and the maximum rms innovations were determined for various pairs of spectra. Two different wavelength ranges, 200-300 nm and 200-400 nm, were used with a sampling interval of 4 nm. The results are shown in Figure 4.4 along with the curve predicted on the basis of pure Gaussian spectral peaks employed under similar conditions for the smaller wavelength range. All of the pairs follow the predicted behavior closely, although there are some deviations due to differences in shape. When the spectral range is increased, however, there is a decrease in the maximum rms innovation. This is because the PAH spectra show little or no absorbance between 300 and 400 nm and there is a



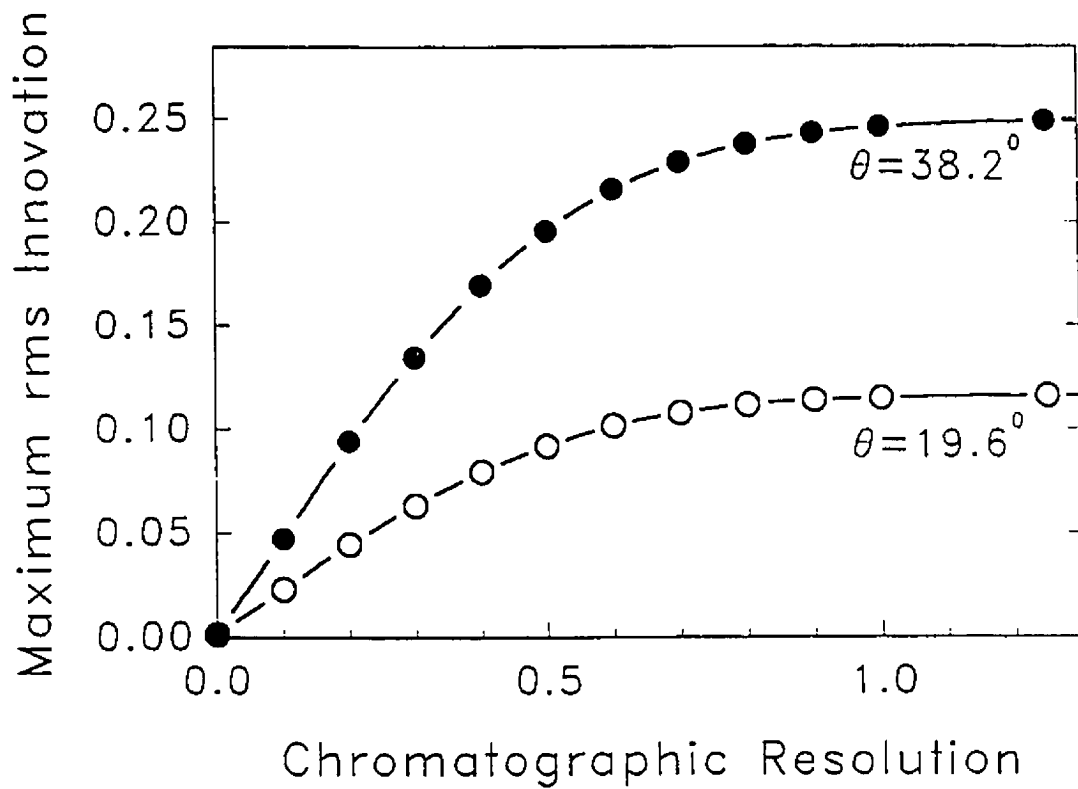
**Figure 4.4.** Effect of spectral correlation (PAH spectra) on the maximum rms innovation for spectral regions of 200-300 nm (O) and 200-400 nm (□). The dashed line represents predictions based on a Gaussian peak shape. A = anthracene; F = fluoranthene; P = phenanthrene; T = triphenylene.



corresponding reduction in the individual innovations for that region and therefore the rms sum. For this reason, selection of a wavelength range for the algorithm should avoid empty regions of the spectrum.

### 4.3.3 Effect of Chromatographic Resolution

In some ways, understanding the effect of chromatographic resolution on the EPCIA algorithm is more important than for spectral correlation since the latter is fixed but the former can be changed by varying the separation environment. As before, a series of simulations were carried out to evaluate this effect. The spectra used for these simulations were Gaussian with  $\sigma = 100$  nm and  $\lambda_{\max 1} = 450$  nm. Two values of  $\lambda_{\max 2}$  (500 and 550 nm) were used to confirm the consistency of the behavior under different conditions. The resolution of the Gaussian chromatographic peaks ( $\sigma = 5$  s, sampling interval = 1 s), was varied by changing the peak separation. As before, the pure component profiles were normalized to a height of unity and the noise level was 0.001%. The results of the study are presented in Figure 4.5. It will be noted that the maximum rms innovation varies in an exponential manner with chromatographic resolution. The curve is approximately linear up to a resolution of about 0.5 and bends over as the peaks approach complete separation. The maximum rms innovation obtained will be determined by the spectral correlation. Although the peak shapes obtained under real chromatographic conditions are often not pure Gaussians, it is expected that the behavior will be similar for peaks that do not deviate radically from this model.



**Figure 4.5.** Effect of the chromatographic resolution on the maximum rms innovation obtained for two component mixtures.  $\theta$  is the angle between the two spectral vectors.

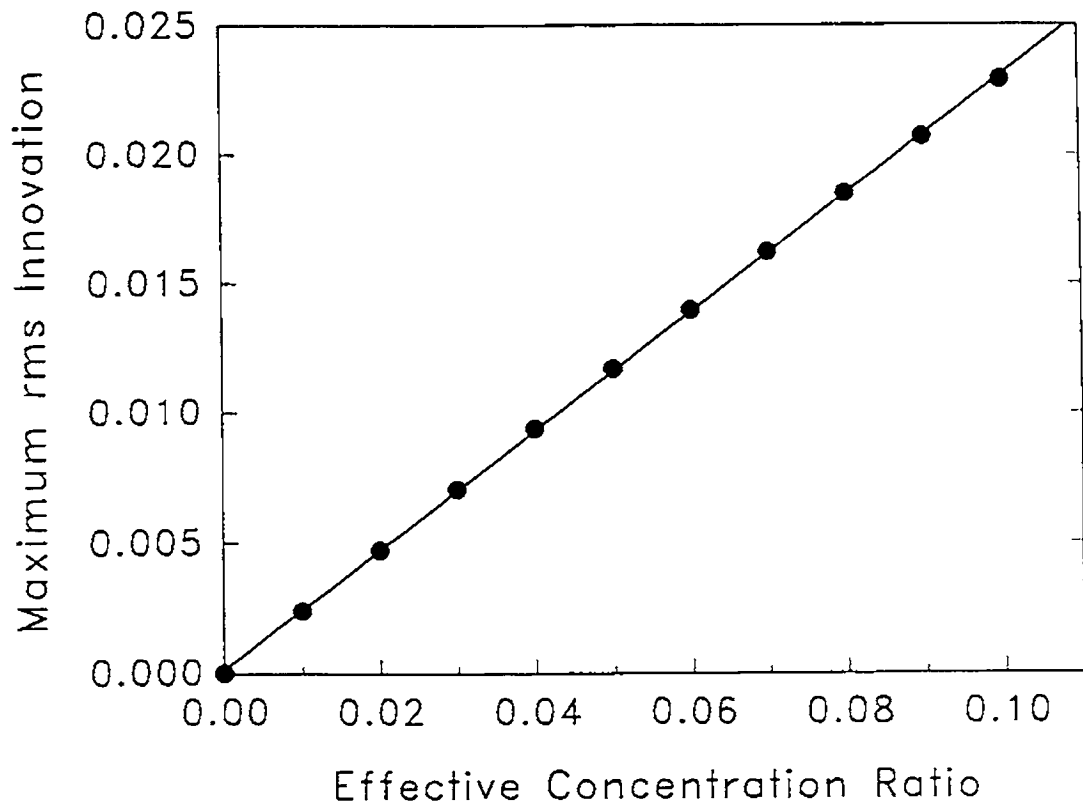
#### 4.3.4 Effect of Concentration Ratio

Up until this point, most of the simulations have considered mixtures of equal concentrations. This may seem an unusual choice, since the normal objective of applying methods such as EPCIA to two component mixtures is to detect the presence of the minor component. A 1:1 mixture provides a good reference point, however, since by definition it represents the maximum attainable concentration of the minor component. The effect of concentration ratio can then be evaluated with this as the limit. In discussing concentration ratio, the concept of an *effective* concentration ratio (ECR) will be implied. This is necessary because, in all curve resolution methods, it is not the absolute concentration of a component that is important, but rather the product of the concentration and the absorbance spectrum. Therefore, the ECR in this work is defined as,

$$\text{ECR} = \frac{\varepsilon_{\max 2} C_2}{\varepsilon_{\max 1} C_1} \quad (4.1)$$

As before, Gaussian profiles were used to simulate chromatographic and spectral conditions. Spectra were centered at 450 and 550 nm with a width of  $\sigma = 100$  nm. The chromatographic resolution was 0.35 ( $\sigma = 5$  s, sampling interval = 1 s). To study the effect of concentration ratio, spectra of both components and the elution profile of the first component were set to a maximum of unity. The maximum of the elution profile of the second component was then varied. The absolute noise level was fixed at values of  $10^{-4}$  and  $10^{-5}$ .

The results of the simulations are presented in Figure 4.6 for an ECR between 0 and 0.1. The relationship between maximum rms innovation and



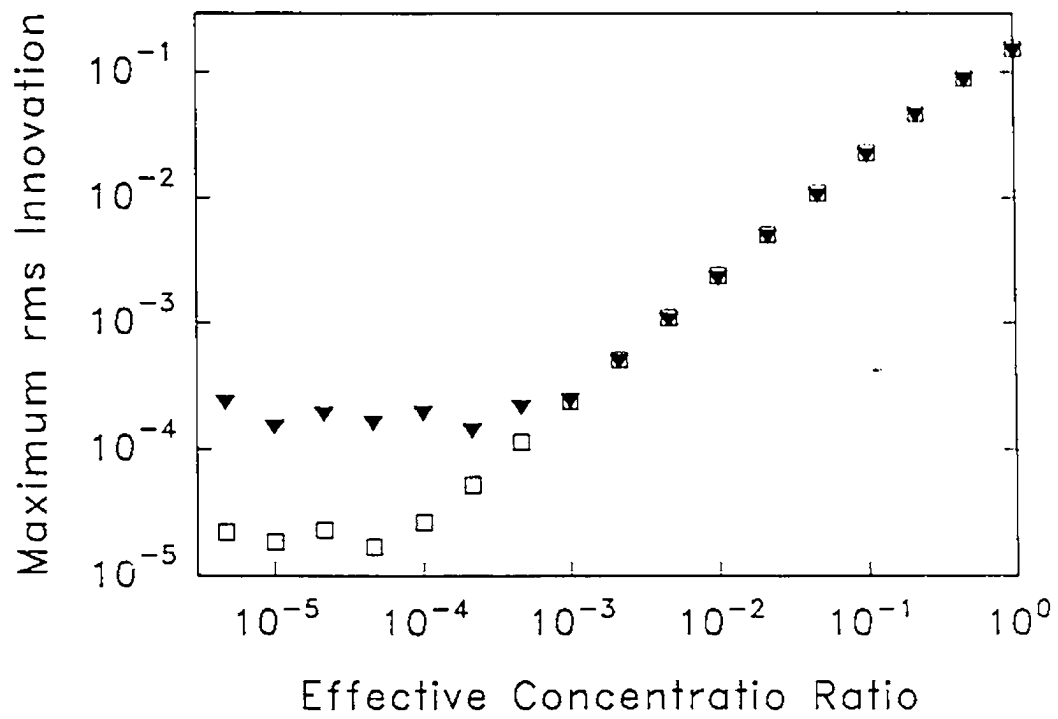
**Figure 4.6.** Dependence of the maximum rms innovation on the concentration of the minor component.

concentration of the minor component is linear over the range, although a slight deviation is evident at the high end. This curvature becomes more obvious at concentration ratios above 0.1, but the relationship continues to be approximately linear even up to a ratio of unity. The linearity is potentially very useful, since it means that the relative contribution of a minor impurity in a series of mixtures can be determined without actually identifying it. It also implies that if the maximum rms innovation is known for a given concentration ratio, it may be predicted for other ratios.

Another useful representation of the concentration ratio dependence is given in Figure 4.7, which displays the same results on a logarithmic scale. This figure shows that the linear relationship is valid over a wide range and also clearly shows that the EPCIA algorithm will fail to detect the presence of the minor component when the maximum rms innovation reaches the level of the noise. As soon as this happens, the curve becomes flat. This is illustrated in the figure for two different noise levels. Other workers in this area have found that a baseline noise level of about  $10^{-4}$  AU. is reasonable for most instruments<sup>138,139,155,156</sup> and we have observed similar limitations in this work. Changes in the chromatographic resolution and spectral correlation should affect the slope of the line in Figure 4.6 and the intercept (but not the slope) of the linear portion in Figure 4.7. This will naturally have an effect on the minimum detectable concentration of the minor component.

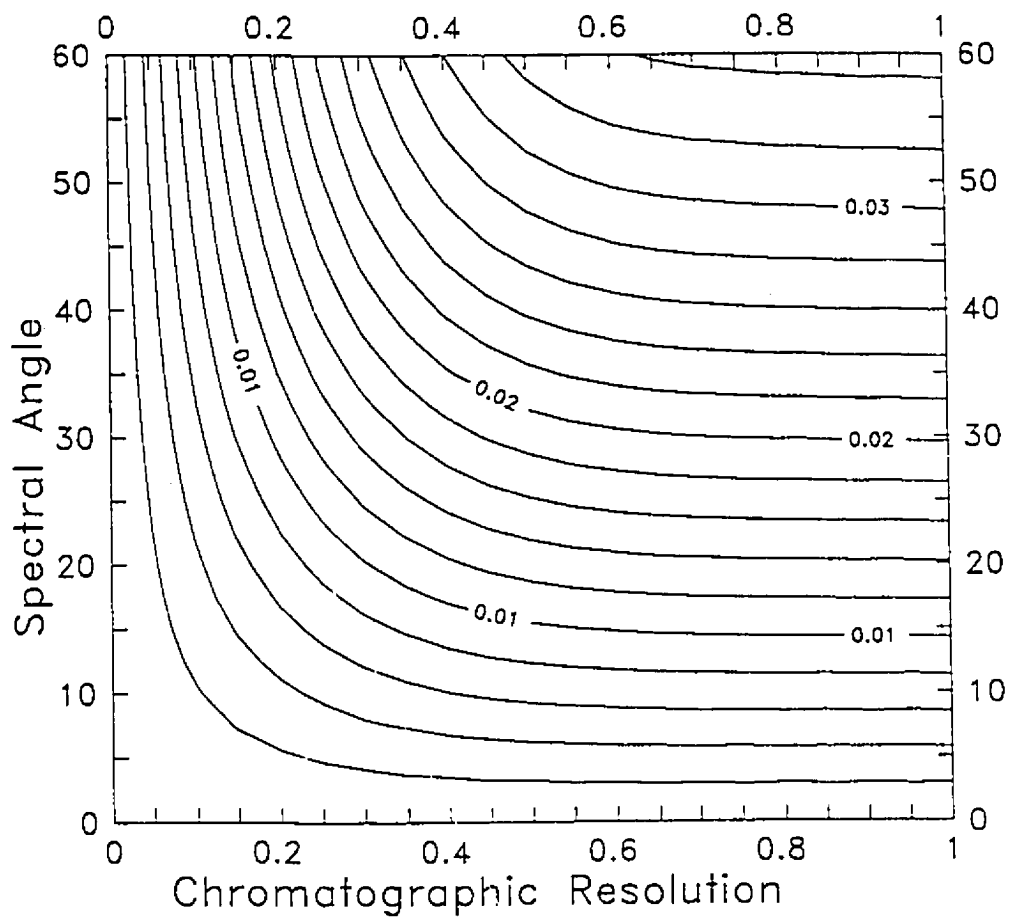
#### **4.3.5 Combined Effects**

The key to assessing the limitations of the EPCIA algorithm is being able to predict the maximum rms innovation and determining if this is larger than the



**Figure 4.7** Results for Figure 4.6 plotted on a logarithmic scale for two different noise levels. Noise: ( $\blacktriangledown$ )  $10^{-4}$ ; ( $\square$ )  $10^{-5}$  AU.

baseline noise. This is a difficult task because the performance of the algorithm is influenced by interactions among the variables discussed in the preceding sections. Nevertheless, an attempt can be made to estimate the maximum rms innovation for a given situation if certain approximations are made. Figure 4.8 is a contour plot showing how the maximum rms innovation for the second component varies as a function of chromatographic resolution and spectral angle when the ECR is 0.1. To generate this plot, Gaussian chromatographic and spectral profiles were assumed and the maximum absorbance of the first component was taken to be unity. The use of this surface can be illustrated with the following example. Suppose the two components in a mixture have a spectral angle of  $\theta = 40^\circ$ , a chromatographic resolution of 0.3, an ECR of 0.05, and a maximum absorbance of 0.1. On the contour surface, the first two conditions give a maximum innovation of about 0.018, but this must be decreased by a factor of 10 since the maximum absorbance is not 1 but 0.1 (this is actually the combined maximum, but the true value for the first component will be very close to this). The prediction must be further reduced by a factor of 2, since the ECR is 0.05 and not 0.1, so that the final estimate for the maximum rms innovation is 0.009. Of course, this is only accurate to the degree that the assumptions made in generating the surface are valid, but the method is a good starting point for assessing the utility of the EPCIA algorithm for a given application. Obviously, in many applications, the characteristics of a potential impurity are not known, but these results provide an indication of the limitations of the algorithm.



**Figure 4.8.** Contour surface of the maximum rms innovation obtained for a minor component (10:1 ratio) as a function of chromatographic resolution and spectral angle. See text for details.

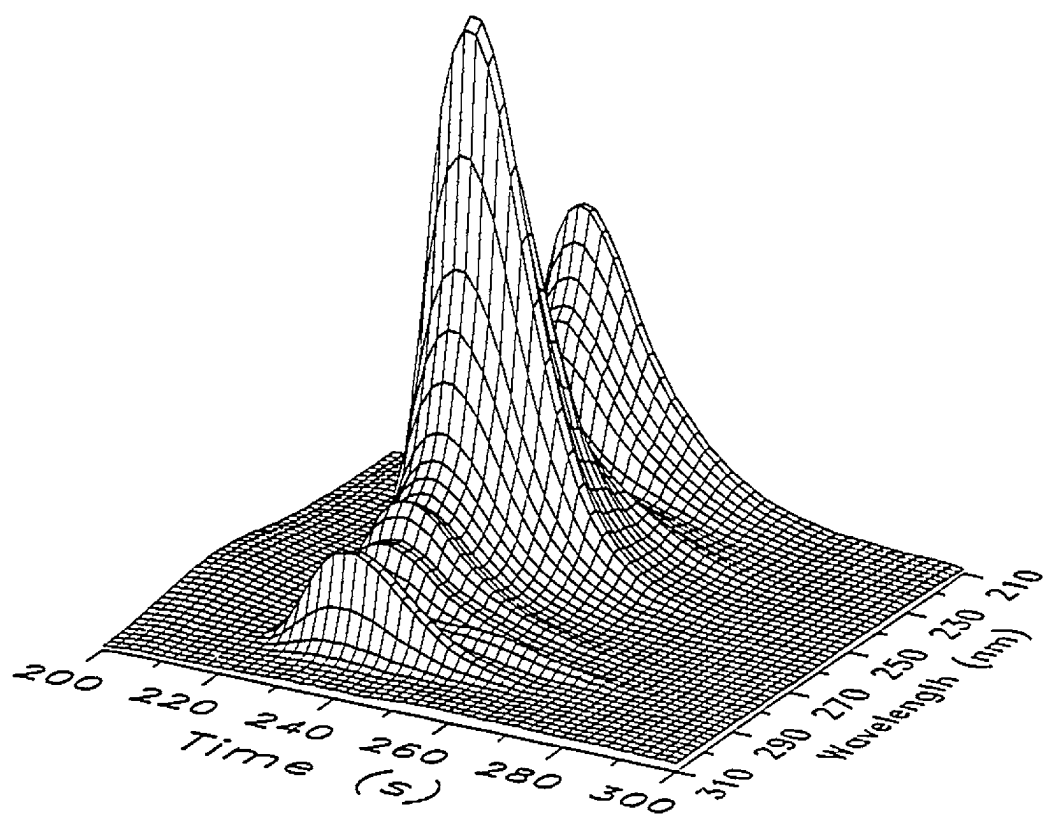


#### 4.3.6 Experimental Results

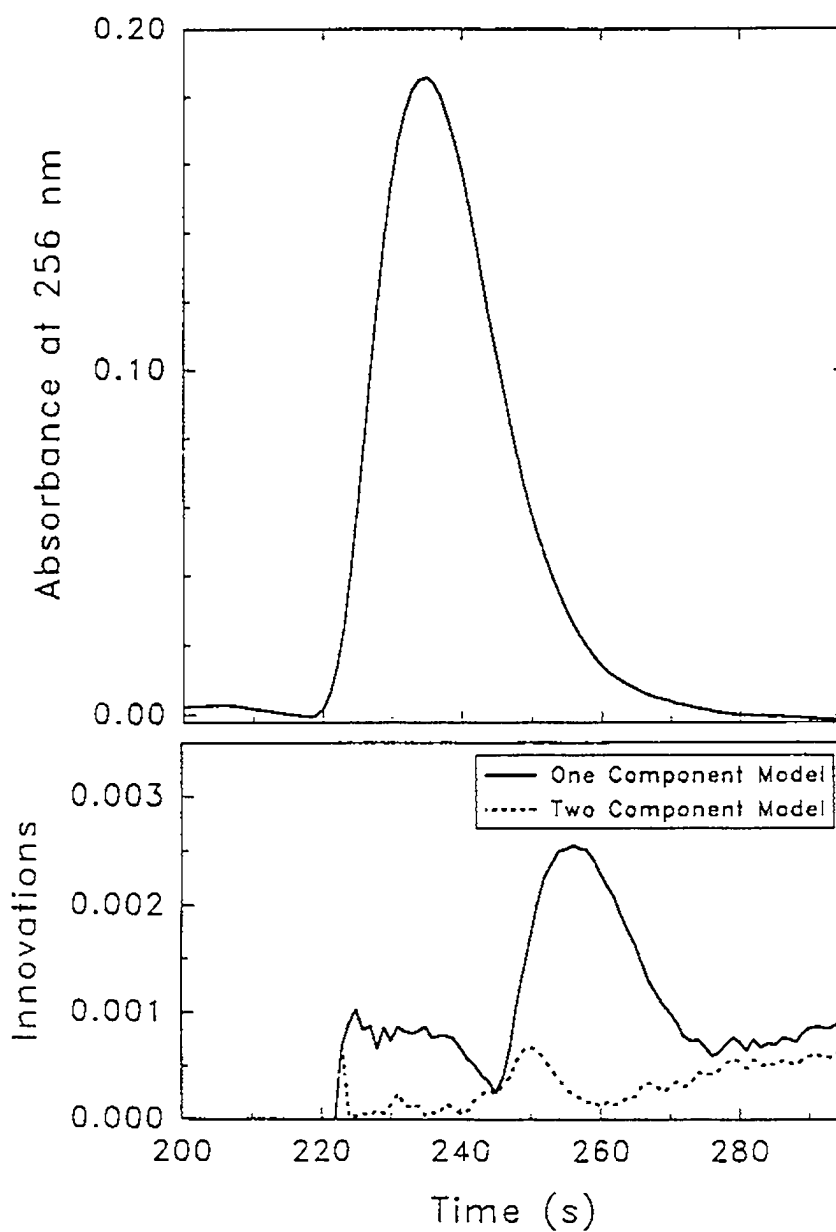
To illustrate the utility of the contour plot presented in Figure 4.8, a real chromatographic mixture consisting of phenanthrene (3.00  $\mu\text{g/mL}$ ) and fluoranthene (0.25  $\mu\text{g/mL}$ ) was examined. The sample volume injected was 20  $\mu\text{L}$  and the eluent was 100% acetonitrile at a flow rate of 0.5 mL/min. Figure 4.9 shows the spectrochromatogram obtained, together with the rms of the orthogonal innovations at each point. The plot of  $\text{rms}(e')$  shows a peak on the trailing edge of the elution profile, indicative of the presence of a second component (the initial perturbation is believed to be due to the baseline irregularity). Under these conditions, the effective concentration ratio was about 1:11, the chromatographic resolution was approximately 0.6, the spectral angle was  $37^\circ$  and the maximum absorbance was 0.18. Applying these conditions to Figure 4.10, the predicted maximum rms innovation is about 0.004, whereas the maximum value observed in Figure 4.10 is about 0.0025. Although this is slightly low, the agreement is very good considering that neither the chromatographic nor spectral peaks are Gaussian, as was used for the simulation studies. This demonstrates that Figure 4.8 can act as a guide for assessing the limitations of the EPCIA algorithm.

#### 4.3.7 Other Variables

Throughout this discussion, three of the variables which have been largely ignored are the chromatographic sampling interval, the number of wavelengths employed in the filter, and the wavelength selected as the independent variable. The reason for this is that the values are not critical to the performance of the



**Figure 4.9.** Spectrochromatogram of a mixture of phenanthrene and fluoranthene.



**Figure 4.10.** Elution profile at 256 nm (top) and rms innovations sequence (bottom) for data in Figure 4.9. Innovations sequences for one (solid line) and two (dashed line) component models are shown.

algorithm as long as certain minimum conditions are met. In terms of chromatographic sampling, it is important to maintain an acquisition rate high enough to avoid distorting the peak through undersampling. A minimum of 20 points across the peak (or 5 points/ $\sigma$  for Gaussian peaks) has been used in this work. As long as this minimum is maintained and sampling occurs at equally spaced intervals, the maximum rms innovation is not significantly altered by the sampling rate.

A minimum requirement for the number of wavelengths used must also be met, but spectra have a greater variability than elution profiles and the lower limit will depend on the spectral features. The range of wavelengths used should encompass the regions of greatest spectral difference between the two components, but exclude regions where neither component absorbs. Intervals should also be frequent enough to capture any sharp features in the spectrum. As an alternative to selecting wavelengths at fixed intervals, particular wavelengths representative of the two spectra could be selected, but this requires prior knowledge of the components. In any case, it should be noted that the spectral angle ( $\theta$ ) and ECR described earlier refer to the wavelengths selected and not to the entire spectrum. There is no upper limit on the number of wavelengths, although it will increase the computational burden. The use of more than a minimum number of wavelengths should not change the maximum rms innovation obtained for the same range, but it should decrease the baseline noise in the rms innovations sequence by a factor of  $\sqrt{n}$  if spectral noise is uncorrelated. This may have some effect in determining the lower concentration limit detectable. All pairs of spectra will have an optimum set of wavelengths based on their correlation, but in the absence of prior knowledge, 20 to 50

wavelengths at equally spaced intervals over the useful range of the UV-visible spectra of the components should be sufficient.

The choice of the wavelength used for the independent variable in the EPCIA algorithm will affect the maximum rms innovation obtained. In the results presented here, the wavelength of maximum absorbance of the first component encountered was employed. In the absence of prior knowledge of the component spectra, using the wavelength of highest absorbance should produce the largest rms innovation for the minor component. This follows because, these points will exhibit a high leverage when plotted in absorbance space. This inhibits the model from adjusting rapidly when the second component appears and therefore maximizes the deviations of the new points. The selection of this wavelength is easily made after the data have been collected. When using the filter for real-time data analysis, the selection is not as reliable because it is based on incomplete information, but this should not seriously affect the performance of the filter.

#### 4.4 EXPERIMENTAL LIMITATIONS OF EPCIA

In the derivation of the EPCIA algorithm and the simulation studies, the spectroscopic data were assumed to follow a bilinear model. When Beer's law is obeyed this holds true, and the number of observable chemical species equals the number of principal components predicted by EPCIA. Unfortunately, there are experimental conditions where a plot of analyte concentration versus measured absorbance does not produce the straight line and zero intercept that Beer's law predicts. This will occur when the assumptions, from which the law was derived<sup>157</sup>, are violated. These systematic errors often reflect

concentration-dependent changes in the physical and chemical environment of the analyte. Systematic deviations also result from instrumental errors<sup>158</sup>, in which the absorbance reported by the spectrometer differs from the true absorbance of the sample. It is important to understand these effects, because they increase the rank of the data in a manner that could be misattributed to chemical components. In summary, nonideal chemical or instrumental behavior can cause EPCIA to overestimate the number of chemical components in a mixture.

Another experimental consideration is random noise, since it sets a baseline value for the innovation sequence. As was shown earlier, a peak in the innovation sequence for a one-component model will signal when the spectra have contributions from more than one chemical species. This signal, and the associated chemical impurity, is only observable when it exceeds the baseline noise. Thus, high levels of experimental noise can cause EPCIA to underestimate the true number of chemical species. In the computer simulations, random deviations from Beer's law were generated by adding white noise of constant variance to the data. The noise characteristics of experimental data are typically more complex, as will be shown for absorbance measurements collected with the diode array spectrometer in which the noise increases with absorbance. In some cases these changes in the noise can be mistaken for additional chemical species. Therefore, the properties of the noise are important because they can cause the number of chemical species in a mixture to be over- or underestimated.

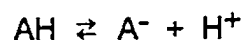
#### 4.4.1 Systematic Deviations

**Chemical and Physical Deviations.** The derivation of Beer's law requires some assumptions about the physical and chemical state of the analyte and the method of measurement. It is based on a monochromatic beam of radiant energy passing through the absorbing sample. Within this sample are absorbers (atoms, molecules, ions, etc.) that absorb photons and reduce the radiant energy of the beam. According to Beer's law the absorbance of a sample varies as

$$A = abc \quad (4.2)$$

since the number of absorbers that a photon passes as it travels through the sample increases linearly with the path length  $b$  and the concentration of the absorbers  $c$ . The absorptivity  $a$ , is a proportionality constant that depends on the wavelength of the incident radiation and the nature of the absorber. This constant reflects the probability of a photon interacting with an absorber, or equivalently, the effective cross section of the absorber. When  $c$  is expressed in units of moles per liter and  $b$  in centimeters then the proportionality constant is called the molar absorptivity,  $\epsilon$ .

Apparent deviations from Beer's law can occur when the concentration of the absorbers is only a portion of the total analyte concentration. This will occur when the analyte is involved in equilibria such as



where more than one analyte species exists in solution. The effect of this equilibrium on Beer's law depends on the two analyte spectra. If only one of the species ( $AH$  or  $A^-$ ) absorbs radiation, then a plot of its absorbance versus the

net analyte concentration will be nonlinear, with either negative or positive deviations. This is an apparent deviation from Beer's law as the errors result from inaccurate values of the absorber concentration rather than a failure of the law itself. Although the plots of  $A_\lambda$  against  $C$  are nonlinear, in this case the  $A^2$  plots used in EPCIA would still be linear. This is because the relative absorptivities ( $A_i/A_j$ ) for the absorbing form of the analyte are unaffected by the equilibrium. In other words, the spectra still lie on a single spectral vector.

The other possibility is that both forms of the analyte ( $AH$  and  $A^-$ ) absorb radiation, but have different spectra. Again, this could cause positive or negative deviations from Beer's law. In this case, EPCIA would indicate that two components contribute to the spectra. While this is true, it might be confusing since we usually assume that each analyte has a unique absorption spectrum. Since the EPCIA algorithm has no way of knowing that the two absorbing species originate from the same analyte, the chemist's knowledge is required to overcome this ambiguity.

Along with its chemical state, the properties of an absorber also depend on its physical environment. This includes the effects of solvent, temperature, and electrolytes. All of these influence the molecular environment of the analyte and thus effect its absorption spectrum. In general, these properties do not vary greatly over a chromatographic peak, so they are not a major problem in EPCIA. What does change within ordered data sets is the concentration of the analyte. In Beer's law the absorptivity of a substance is assumed constant with respect to changes in concentration, but this requires that the absorbers act independently of each other. At high analyte concentrations or electrolyte concentrations this assumption may fail, due to solute-solute or solute-solvent interactions. The



resulting systematic deviations can increase the rank of the spectroscopic data and cause EPCIA to overestimate the number of chemical components present.

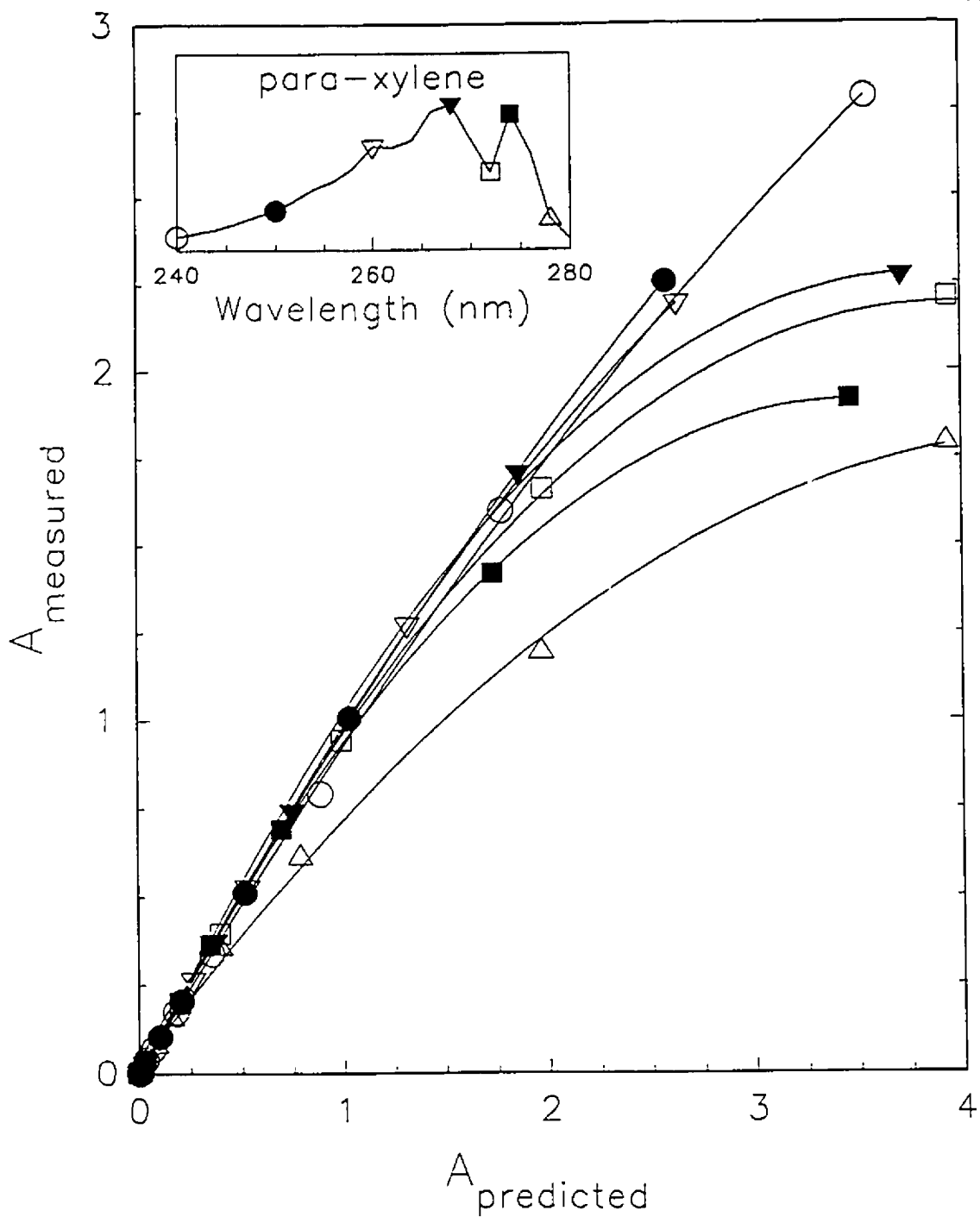
In summary, the presence of equilibria involving the analyte can cause both positive and negative deviations from Beer's law. These arise from nonlinear relationships between the net analyte concentration and the individual absorber concentrations. Despite these nonlinearities, EPCIA can often correctly identify the number of absorbing species. What EPCIA cannot do is identify the origin of each absorbing species since this requires chemical knowledge that it lacks. Deviations from Beer's law also occur when the molecular environment of the analyte changes with concentration. The resulting spectral changes are more problematic for EPCIA since their effects may be mistaken for a chemical impurity.

***Instrumental Errors.*** Unlike the chemical and physical effect discussed above, instrumental errors don't alter the true absorbance of the sample. Instead, they degrade the accuracy and precision of the measured absorbance values. This section considers errors associated with the spectrometer which limit the accuracy of the measured values. For EPCIA we are most concerned with instrumental errors whose magnitude is dependent on the absorbance of the sample; that is, errors that produce a nonlinear response in absorbance rather than a constant bias. There are many possible sources for these systematic errors, including nonidealities in optical design, detector response, and electronic components like amplifiers. For the diode array spectrometer, stray light and polychromatic radiation can cause significant deviations from Beer's law. The magnitude of each of these errors depends on the total absorbance and spectra of the samples.

Figure 4.11 compares measured absorbances of a series of para-xylene solutions to absorbances predicted by Beer's law. These predictions are based on the molar absorptivities of dilute solutions (0.5 AU or less). No systematic deviations were observed at these values, but at higher absorbances significant deviations from Beer's law occur. All of these deviations were negative (measured absorbance is less than predicted), which is typical of instrumental errors<sup>158</sup>. Furthermore, these deviations were generally smaller at wavelengths where the spectrum is relatively flat. This behavior will be examined in the following sections.

To understand the origins of systematic and random errors in absorbance measurements the design of the diode array spectrometer<sup>159-162</sup> will briefly be considered. The conventional design for a UV-visible spectrometer focuses light onto a monochromator that selectively transmits a narrow band of light. After this light passes through the sample cell its intensity is measured by a single detector. In contrast, a diode-array spectrometer (Figure 4.12) has the positions of the dispersive device (grating) and the sample reversed. This is called a reverse-optics design. Thus, polychromatic radiation passes through the sample first, and then it is dispersed onto a linear array of detectors. In this way, each element of the photodiode detector measures the intensity of a different region of the spectrum.

The spectrometer used in this work has an array of 316 diodes. The first diode is positioned to record the absorbance at 190 nm, the next at 192, 194, 196, ... ,820 nm. Thus the spectrum is digitized with 2 nm increments. Although each diode is centered on a particular wavelength, it actually measures the spectrum for a bandpass around this value. The width of this bandpass is related to the performance of the polychromator and the physical size of the



**Figure 4.11.** Nonlinear instrumental response is shown for a set of para-xylene calibration solutions. The measured absorbances at six wavelengths (see inset) are compared with the absorbances predicted by Beer's law. See text for details.

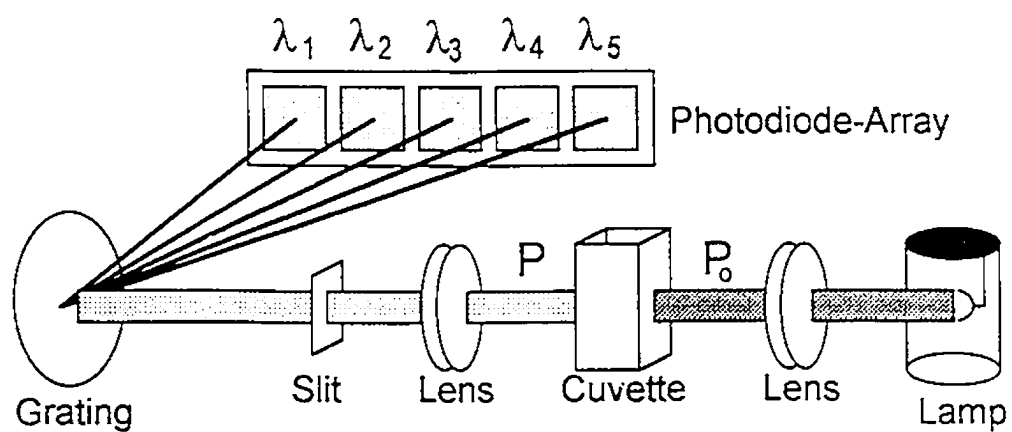


Figure 4.12. Schematic diagram of a diode array spectrometer.

photodiode. The spectral bandwidth of the polychromator usually depends on the width of its entrance slit. A narrow slit decreases the bandwidth, but also decreases the amount of light reaching the detector. The detector bandwidth indicates the region of the spectrum that gives a response for each diode. While the lower limit should be 2 nm for this detector, its actual bandwidth is somewhat wider because the sensing areas overlap for adjacent diodes<sup>160</sup>. The overall bandpass of the spectrometer depends on both the spectral and detector bandwidths, but it is limited by the larger of the two.

**Detector Linearity.** Spectrometers do not measure the absorbance of a sample directly, but instead measure the attenuation of the beam of power  $P_0$  to  $P$  by the absorbing solution. Then the absorbance is calculated as:

$$A = \log \frac{P_0}{P} = -\log T \quad (4.3)$$

where  $T$  is transmittance. To measure the power of the incident radiation, the detector converts the radiant energy of photons into an electrical signal, such as current, which is easier to quantify. An ideal detector would have a linear response with respect to the incident radiant power. To evaluate this relationship experimentally, the detector's response was measured for a range of photon fluxes. These were achieved with a photographic step tablet, a strip of photographic film with calibrated steps of silver density. Each step increases absorbance by a fixed amount. Thus a plot of measured absorbance against step number should be linear.

A mask was cut to expose individual steps of the tablet. Then, this mask was fixed in the sample compartment with its opening centered in the beam.

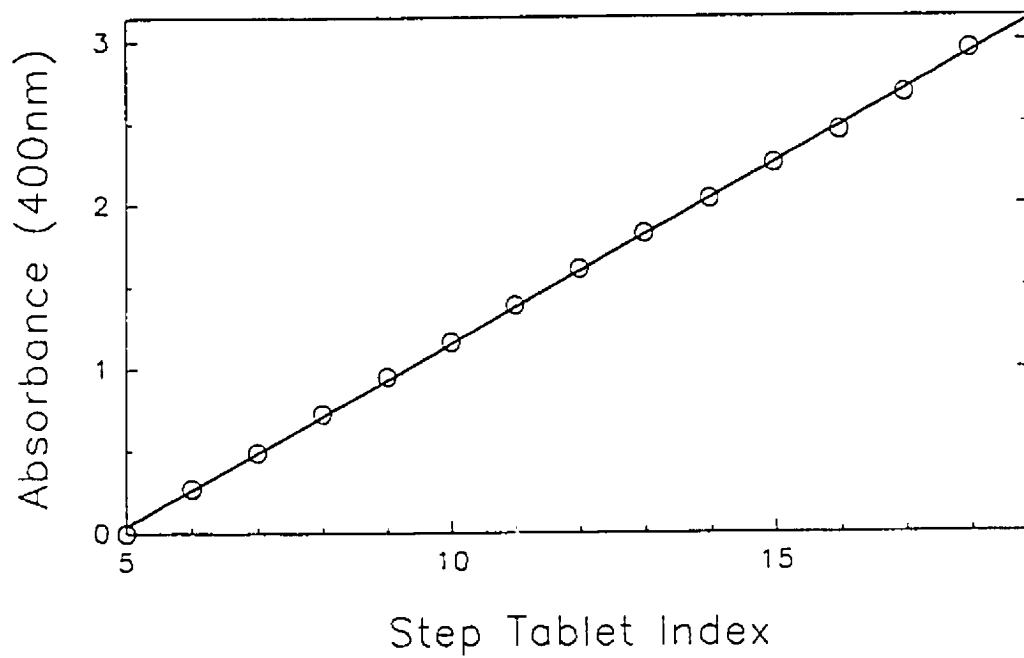
Measurements made over a range of 15 steps are shown in Figure 4.13. The linearity of the detector was evaluated with the equation

$$A = k n^r \quad (4.4)$$

where  $A$  is the measured absorbance,  $k$  is a system-dependent constant,  $n$  is the step tablet index and  $r$  is the response index of the detector. The response index should not be confused with a correlation coefficient which shares the same symbol. An ideal detector would have  $r$  equal to unity. The response index for the step tablet at 400, 600, and 800 nm was found to be  $0.96 \pm 0.01$  over a range of about 3 AU. Over smaller ranges of absorbance the response index approached unity. The linearity of the detector response could not be measured in the UV region with this experiment due to the film base absorbance. The response indices for the para-xylene solutions (Table 4.1) were much lower than those calculated for the step tablet, even though they are calculated for a smaller range of absorbance values.

**Table 4.1** Results from fitting Equation 4.4 to the para-xylene calibration data.

<i>Sample</i>	<i>Wavelength</i> (nm)	<i>Maximum</i> <i>absorbance</i>	<i>Response Index</i>
step tablet	600	2.95	0.96
para-xylene	240	2.76	0.88
para-xylene	250	2.26	0.91
para-xylene	260	2.19	0.85
para-xylene	268	2.27	0.65
para-xylene	272	2.10	0.53
para-xylene	274	2.19	0.62
para-xylene	278	1.79	0.67



**Figure 4.13.** The absorbances measure by the diode array for a calibrated photographic step tablet.

If the detector behavior is similar in the UV region to that observed from 400 to 800 nm, then nonlinear detector response is insufficient to explain the deviations observed for the chemical samples. Some attributes of para-xylene must cause Beer's law to fail at lower absorbances. One important difference between the step tablet and the chemical samples is the latter have nonuniform absorbance across their spectra, making them more susceptible to the effects of stray light and polychromatic radiation.

**Stray Light.** Stray light is defined<sup>163</sup> as detected light of any wavelength that is outside the bandwidth of the selected wavelength. Ideally, all the light of a given wavelength would be able to reach the appropriate diode, but a portion of this light is misdirected due to imperfections in the gratings and optics. Scattered light can also include light that has not passed through the sample and light dispersed by higher orders of the grating. These photons can then contribute to the signal at another region of the spectrum, perhaps hundreds of nanometers away, causing inaccurate results. Spectrometers with reverse-optics designs are susceptible to stray light as polychromatic light is projected onto the grating. The effect of the stray light on the measured absorbance is,

$$A_{\text{meas}} = -\log\left(\frac{P + P_s}{P_o + P_s}\right) = -\log\left(\frac{T_{\text{true}} + f}{1 + f}\right) \quad (4.5)$$

where  $A_{\text{meas}}$  = measured absorbance  
 $P_s$  = power of the stray radiation  
 $T_{\text{true}}$  = true transmittance of the sample  
 $f$  = relative contribution of the stray light ( $P_s / P_o$ )



Stray light causes negative deviations that increase nonlinearly with absorbance. These errors usually limit the maximum absorbance that the spectrometer can record. The level of stray light is dependent on instrumental design and the quality of the grating used, but it is also necessary to consider the properties of the sample. In general, it will affect a sample with a few narrow absorbance bands more than one with a broad flat spectrum, like the neutral density filters.

**Polychromatic Radiation.** Beer's law assumes that the incident radiation is monochromatic, but for reasons explained earlier, spectrometers always have a finite bandpass. For example, a diode centered at 500 nm might respond to photons in the range of 499 to 501 nm. All of these photons contribute to the measured response since they are indistinguishable to the detector. This polychromatic radiation causes both bias and nonlinearity in the measured absorbance.

Dose and Guiochon<sup>164</sup> have modeled the dependence of the spectrometer response on bandpass and the shape of the absorbing sample's spectrum. To do this, the detector signal is calculated by integrating the radiant energy over the bandpass. The model assumes that the incident power and diode response are constant over the bandpass of the instrument. Then the measured absorbance  $A_{\text{meas}}$  can be calculated as:

$$A_{\text{meas}} = -\log\left(\int_{\lambda_0 - \Delta/2}^{\lambda_0 + \Delta/2} 10^{-A(\lambda)} d\lambda\right) \quad (4.6)$$

where  $\Delta$  = the spectral bandpass (nm)  
 $\lambda_0$  = the detection wavelength  
 $A(\lambda)$  = the true absorbance at wavelength  $\lambda$

If the sample absorbance does not vary over the bandpass, such as would be the case for monochromatic radiation, then Equation 4.6 reduces to Beer's law. Deviations from Beer's law are appreciable when there are large changes in absorbance within the bandpass, such as on the sides of sharp spectral features. One rule of thumb is that the bandpass should be 1/10 the width of the peak to give an error of less than 0.5%. This rule assumes that the wavelength is set to the peak maximum. To provide a more general evaluation of the effects of stray light, Dose and Guiochon modeled the absorbance of the sample with a Taylor series in  $\lambda$  about the band's central wavelength  $\lambda_0$ . In the case of a first-order expansion the integral can be written in a closed form:

$$A_1 = A_0 - \log \left( \frac{\sinh(Ka_1\Delta/2)}{Ka_1\Delta/2} \right) \quad (4.7)$$

where  $A_1$  = absorbance predicted by a first-order expansion  
 $A_0$  = Beer's law absorbance  
 $K$  =  $\ln(10)$   
 $a_1$  = absorbance derivative ( $dA/d\lambda$ )

Thus the absorbance is partitioned into two terms: the Beer's law absorbance ( $A_0$ ) and a correction term that is nonlinear with concentration. Also note that the correction term for polychromatic radiation is always negative.

The values of  $a_1$  and  $A_0$  were estimated from para-xylene solutions with absorbances less than 0.5 AU. Equation 4.7 was then fit to a wider range of absorbance values to estimate the value of the spectral bandpass. This fit was evaluated for wavelengths where there was little curvature in the spectra, that is,

regions where the second derivative is small. These regions were chosen for two reasons: (1) they are on the sides of peaks where the absorbance changes rapidly with  $\lambda$ , (2) the first-order approximation is reasonable. In contrast, an accurate description of the peak maximum would require a second-order Taylor series expansion. Some typical results are given in Table 4.2

**Table 4.2** Results from fitting Equation 4.7 to the para-xylene calibration data.

<i>Wavelength</i> (nm)	<i>Maximum</i> <i>absorbance</i>	$\Delta_{fit}$	<i>Notes</i>
250	1.65	4.79	
250	2.75	7.20	poor fit
270	1.68	4.38	
270	2.10	7.04	poor fit

At absorbances below 1.7 AU Equation 4.7 adequately described the deviations from Beer's law. At higher absorbance the model failed to describe the measured absorbances and the estimated spectral bandpass increased. This is physically unreasonable since the bandpass should be independent of absorbance. Assuming the additional contribution at high absorbances was due to stray light, a composite model was designed which included the effects of stray and polychromatic radiation

$$A_{1, \text{stray}} = A_1 - \log \left( \frac{10^{-A_1} + f}{1 + f} \right) \quad (4.8)$$

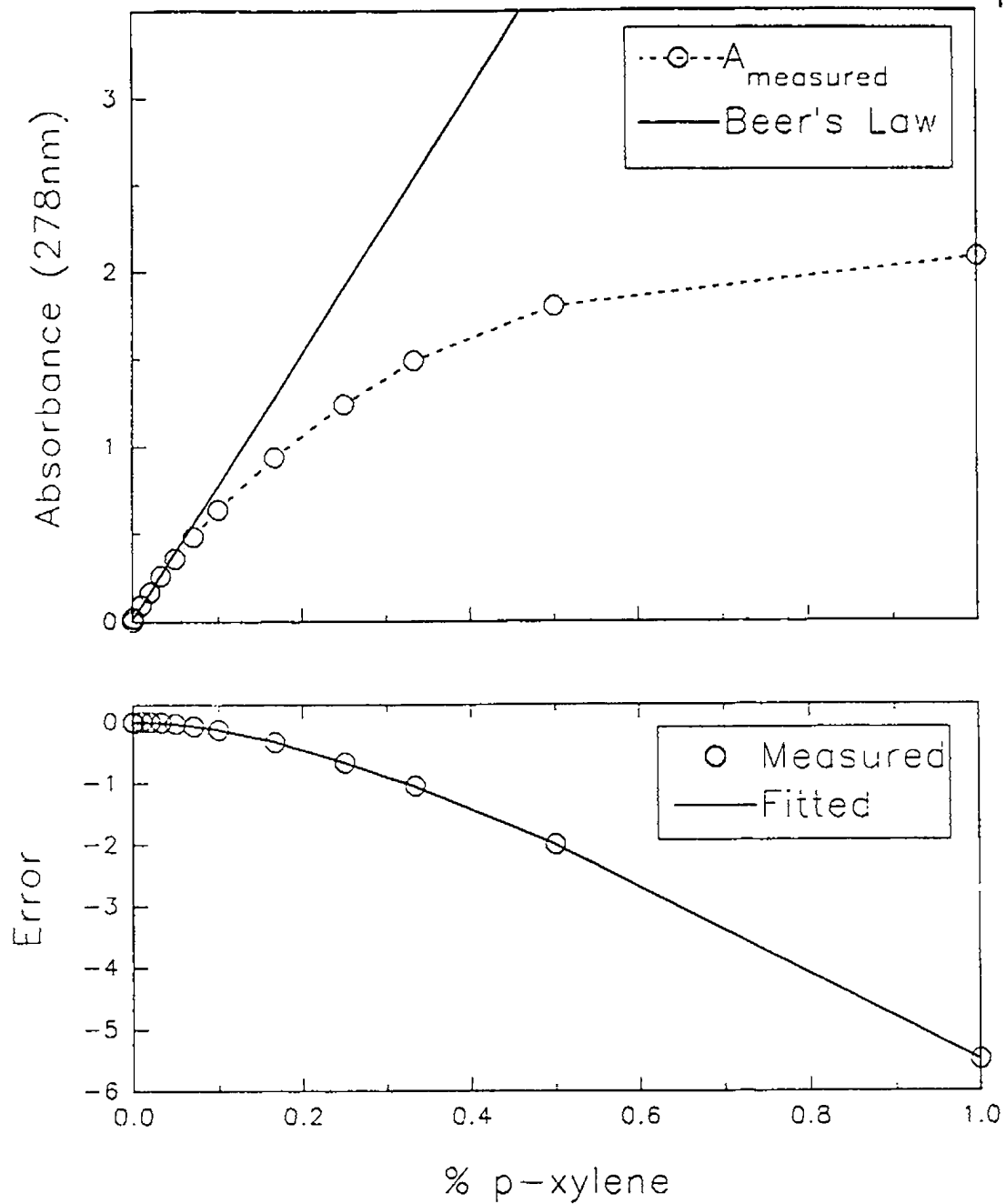
The result for this model are given in Table 4.3. Including the stray light term improved the model's performance at high absorbances. The results in Figure

4.14 are from fitting the entire calibration curve at 278 nm. The deviations from Beer's law were negative, and they increased with absorbance. Equation 4.8 accurately described these deviations over the full range of the experiment. In turn, the fitted values of the stray light and the polychromatic terms can be used to predict the relative contributions of the two effects. As an illustration, Figure 4.15 gives the errors predicted at a wavelength with a molar absorptivity of 10,000 and a first derivative of 1000 ( $a_1/A_0 = 0.1$ ). These are typical values for a peak with a width of 10 nm at half height. While both effects increase with absorbance, the stray light term is less important at low absorbances. This explains why Equation 4.7 alone was successful in modeling deviations at lower absorbance. At higher absorbances the stray light dominates.

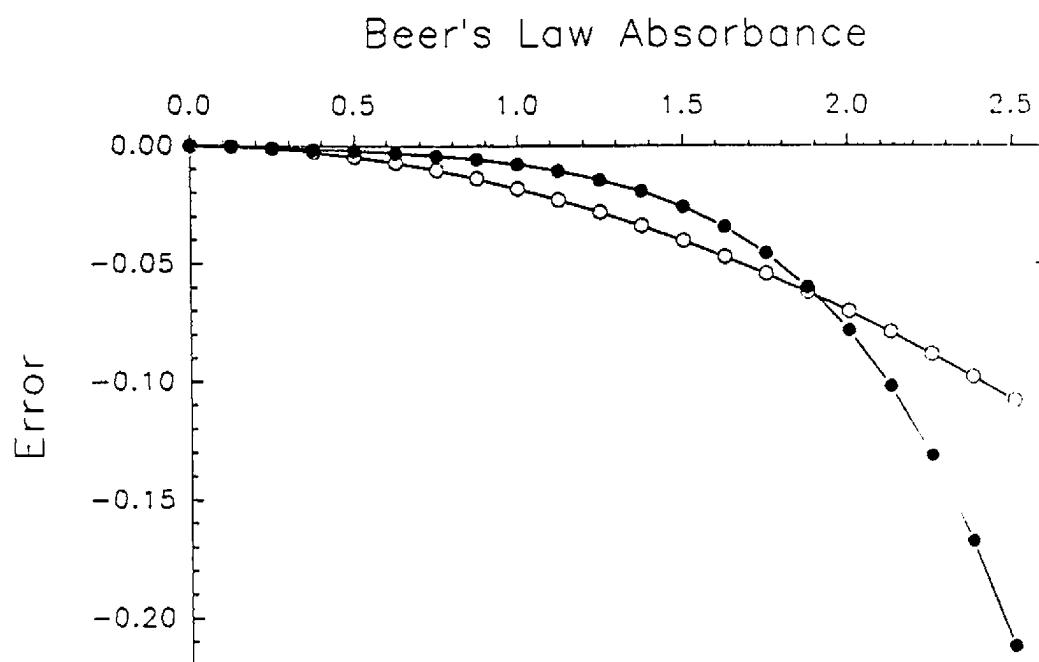
**Table 4.3** Results from fitting Equation 4.8 to the para-xylene calibration data.

<i>Wavelength</i> (nm)	<i>Maximum</i> <i>absorbance</i>	$\Delta_{fit}$ (nm)	$f_{fit}$
242	3.03	4.3	$0.93 \times 10^{-3}$
250	2.98	3.7	$1.1 \times 10^{-3}$
264	2.62	6.7	$2.8 \times 10^{-3}$
270	2.17	4.3	$5.9 \times 10^{-3}$
276	2.01	4.5	$8.8 \times 10^{-3}$
278	1.79	4.3	$7.9 \times 10^{-3}$
Average	2.54	$4.4 \pm 0.2$	$2.7 \pm 1 \times 10^{-3}$

**Scan Time.** The diode array spectrometer continuously exposes all the diodes to the spectrum during an absorbance reading. These diodes are reverse-



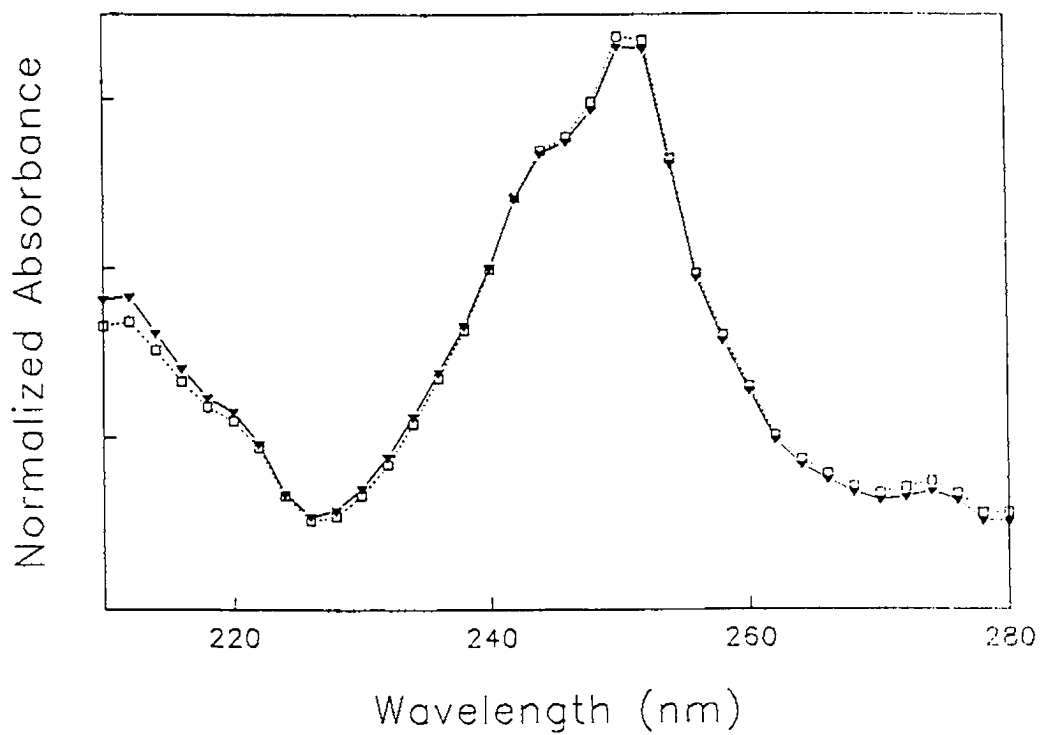
**Figure 4.14.** The calibration curve for para-xylene at 278 nm has large deviations from Beer's law (top). These errors are consistent with the effects of polychromatic and stray radiation modeled by Equation 4.8 (bottom).



**Figure 4.15.** Typical errors resulting from polychromatic radiation (O) and stray light (●) estimated from Equation 4.8. See text for details.

biased such that they don't conduct electricity until the incident photons produce charge carriers. Thus the charge produced by each diode circuit should be proportional to the incident power. Each diode has a capacitor to store this electrical charge. The process of reading these capacitors and converting their charge to a digital signal is carried out sequentially. Thus the electronic reading of the detectors, not a mechanical scanning of a monochromator, determines scan time of this instrument. This is analogous to a row of buckets collecting rain, where the water level in each bucket is recorded and then emptied in a continuous cycle. The instrument used in this work has a scan time of 100 ms for the full spectrum. Thus, to record a spectrum for 1 s, the diode array sums the results of ten 100 ms scans.

Instrumental artifacts arise when the sample absorbance changes rapidly during the scan time<sup>156</sup>, just as large changes of absorbance (as a function of wavelength) were a problem with polychromatic radiation. To understand this problem, consider a simpler system where there are ten diodes over the spectrum with a one second scan time. We start to record a spectrum at 1.0 s. The signal recorded at the first diode was integrated between 0 and 1.0 s. The signal at the next diode is also for a 1.0 s integration, but it is for the signal between 0.1 to 1.1 s. This continues to the last diode, whose signal is read at 2.0 s. There are no problems for static systems, since the signal is integrated for the same duration at each wavelength, but with dynamic systems the lag time between diode readings can alter the appearance of the spectra. On the rising edge of a chromatographic peak, the last diode read will be observing a higher concentration of the analyte than the first diode. On the trailing edge, the effect is reversed. The effect of scan time is illustrated in Figure 4.16 where two spectra obtained at different points on the elution profile of phenanthrene are



**Figure 4.16.** The difference between spectra acquired on the leading (dashed line) and trailing (solid line) of the elution profile of phenanthrene.



compared. The magnitude of these errors increases for narrow chromatographic peaks, but would decrease for an instrument with a faster scan time, such as one specifically designed for HPLC detection. Alternatively, mathematical corrections<sup>156,165</sup> can be applied.

#### 4.4.2 Random Deviations

Random noise limits the precision with which absorbance measurements can be made. There are a variety of potential noise sources for a spectroscopic measurement<sup>166,167</sup>. These include instrumental sources such as electronic components, and fundamental sources like the particle nature of light. In this discussion we will consider two general classes of noise:

1. Shot noise refers to random fluctuations associated with the counting of a random event. Poisson statistics predict a standard deviation of  $n^{1/2}$  for counting  $n$  events. In spectroscopy, this fundamental noise results from random arrival times of the photons at the detector. Another source of shot noise is the dark current from thermally generated charge in the diodes.
2. Flicker noise is noise due to fluctuations in the source intensity. This is a nonfundamental effect since it depends on the stability of the lamp. These fluctuations can be minimized by using double beam geometry, optical feedback, or source modulation. The standard deviation of flicker noise is proportional to the intensity of the beam. Fluctuations in the detector response or background, such as those occurring with temperature changes, also produce flicker noise.

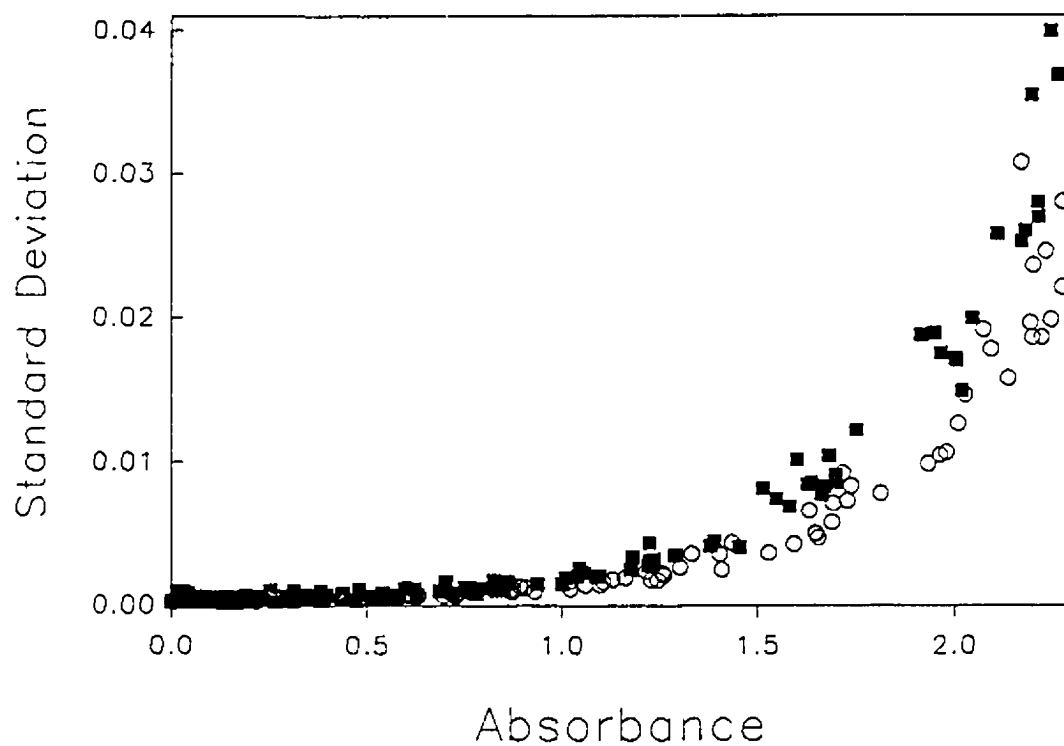
There are other possible noise sources for diode array spectrometers, including amplifier noise and limited resolution in the analog to digital conversion, but in a good instrumental design these are not the limiting noise sources.

Multiple values of a signal are needed to calculate a standard deviation but these are easy to acquire with the diode array since it measures the sample every 100 ms. Repetitive scans allow an average signal and noise estimate to be calculated at each point in the spectrum. These calculations are performed by the spectrometer's processing unit in real-time.

Instrumental noise can be characterized by its magnitude and frequency characteristics<sup>166</sup>. In the simulation studies, the variance of the noise was assumed to be constant. The precision of the diode array spectrometer was evaluated with the para-xylene solutions between 210 and 310 nm. The standard deviation was estimated from 60 scans of the diode array. Figure 4.17 shows that the noise is heteroscedastic, that is, its magnitude is dependent on the absorbance of the sample. At absorbances of 0.5 AU or less the noise is almost independent of sample absorbance. By 1.0 AU the noise is roughly doubled, and by 2.0 AU it has increased by more than an order of magnitude.

Highly absorbing samples reduce the number of photons reaching the detector, which degrades the precision of the absorbance measurement. The flow cell has a smaller aperture than the normal cuvette that also reduces the beam intensity. As a result, the precision of a sample measured in the flow cell is lower than for the standard cuvette.

***Frequency Characteristics.*** The magnitude and frequency of the noise can be examined in more detail by calculating its power spectrum from replicate



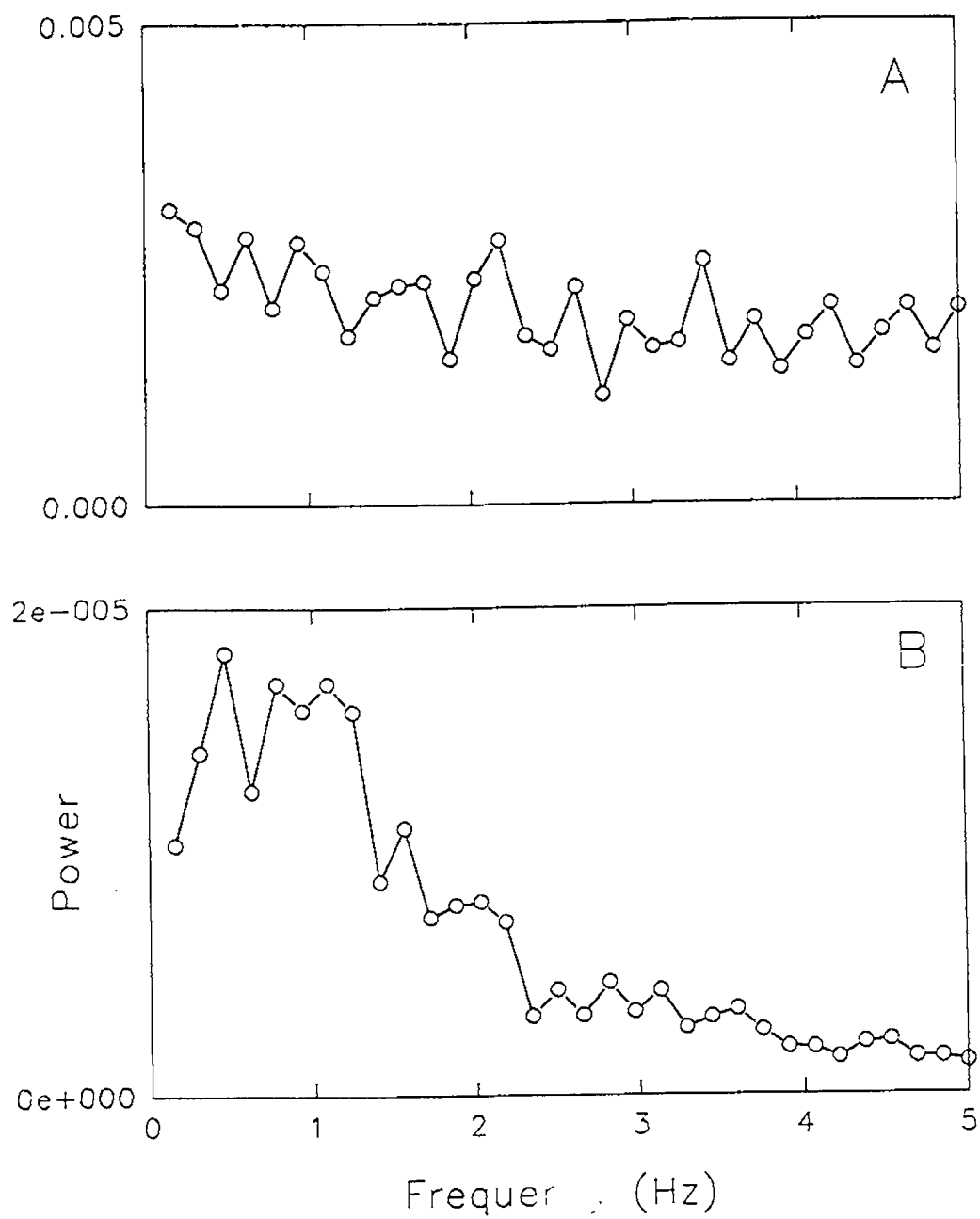
**Figure 4.17.** The precision of absorbance reading versus sample absorbance for para-xylene solutions in a 1 cm cuvette (O) and a 30  $\mu$ L flow cell (■).

measurements. Flicker noise is also known as  $1/f$  noise since the magnitude of its power spectrum varies inversely with frequency. That is, the fluctuations are predominantly low frequency. For shot noise, the magnitude of the noise is independent of frequency, giving a flat power spectrum. Another possibility is interference noise at specific frequencies, such as 60 Hz line noise.

The noise power spectrum of the diode array spectrometer was found to vary with sample absorbance. Figure 4.18B shows that the power spectrum of the blank reading contains predominantly low frequency noise, indicating that flicker noise is present. At higher absorbances, the magnitude of the noise increases and its distribution becomes less dependent on frequency (Figure 4.18A). Although flicker noise is still present in this reading, the shot noise appears to dominate. Both of these noise sources vary with incident power: flicker noise increases linearly with  $P$ , while shot noise increases as  $P^{1/2}$  where  $P$  is the number of photons measured. Thus the relative contribution of flicker noise increases with increasing photon counts. Hence flicker noise should be most apparent in the blank where the incident power is at its maximum.

**Ensemble Averaging.** When integration times longer than 100 ms are used, the absorbance signal reported by the diode array spectrometer is an ensemble average. For example, a 2.0 s absorbance reading is the average of 20 scans of the detector. Ensemble averaging of  $n$  scans should improve the S/N by a factor of  $n^{1/2}$  for a signal contaminated by white noise<sup>28</sup>. Table 4.4 compares this predicted improvement to the actual improvement for a sample with an absorbance of 2.0.

For short integration times, ensemble averaging gives the signal enhancements expected for white noise. As the integration times increase, the



**Figure 4.18.** Noise power: spectra for absorbance readings of 2 AU (top) and 0 AU (bottom) calculated from 2048 replicate 100 ms readings.

actual improvements are smaller than the predicted. This suggests that the shot noise is reduced by ensemble averaging, but the remaining flicker noise is harder to remove because it contains predominantly low frequencies. The efficiency of ensemble averaging also decreases for samples with low absorbance, such as the blank readings. For example, averaging 16 scans of the blank reading decreased the standard deviation from 3.7 to  $2.1 \times 10^{-4}$ ; this is a S/N enhancement of only 1.76.

**Table 4.4** Improvements in S/N from ensemble averaging a 2 AU signal.

<i>scans averaged</i>	$\sigma_{\text{absorbance}}$ ( $\times 10^{-3}$ AU)	<i>predicted</i> <i>improvement</i>	<i>actual</i> <i>improvement</i>
1	5.80	-	-
2	3.91	1.41	1.5
4	3.20	2	1.8
16	1.93	4	3.1
64	1.31	8	4.4
128	0.96	11.3	6.0

In summary, the spectrometer's noise characteristics depend on the incident power reaching the detector. When this power is at its maximum (low absorbances), flicker noise dominates and the noise power spectrum contains predominately low frequencies. As the power decreases, the relative error in the detector signal and the absorbance measurements increases. This noise results predominantly from random shot noise, and is therefore easier to remove with ensemble averaging.

#### 4.4.3 Effect of Systematic and Random Noise on EPCIA

The previous sections showed how experimental artifacts can cause deviations from Beer's law. In practice, peak purity detection is usually limited by these experimental and instrumental non-idealities<sup>156,165,168-171</sup> rather than fundamental mathematical ones. The magnitude of these deviations increase with absorbance, but the nature of each depends on the experimental conditions in a different way:

1. **Stray light** causes negative deviations. The size of these deviations depends primarily on the relative contribution of stray light ( $f$ ). Thus, there is no simple way to change the contribution of stray light to a given sample absorbance.
2. **Polychromatic radiation** also causes negative deviations from Beer's law, so its effects are difficult to distinguish from those of stray light. The size of these deviations depends on ( $\Delta$ ), the bandpass of the spectrometer, which is fixed for the diode array, but also on the shape of the spectrum. The effects of polychromatic radiation are most pronounced at wavelengths with large spectral derivatives.
3. **Random noise** results from instrumental sources like the lamp and the detector. It cause both positive and negative deviations. While the noise characteristics of the spectrometer components are fixed, their effect on the absorbance readings is not. For instance, the precision of the absorbance readings can be improved with ensemble averaging. Also, the precision decreases under experimental conditions that reduce the power of the blank reading ( $P_0$ ).

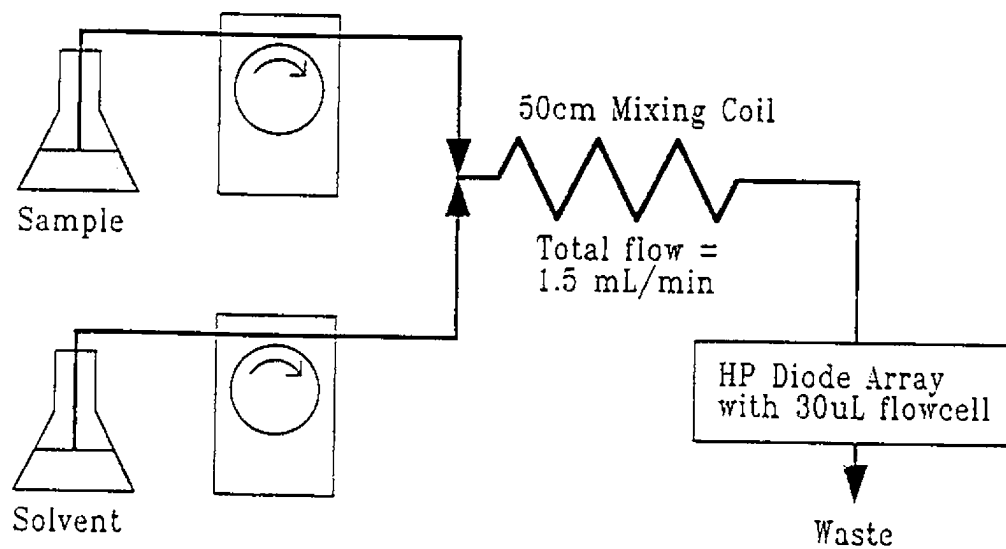
4. **Scan-time** causes problems for dynamic systems. Its magnitude and direction depend on the absorbance changes over time ( $dA/dt$ ). Thus it has no effect on measurements of a static system.

Experiments were designed to emphasize each of these noise sources so that their effects on EPCIA can be observed. To do this, a flowing system was constructed with two peristaltic pumps, one for controlling the flow of a solvent and the other for an analyte solution (Figure 4.19). The flow rates of the pumps were computer controlled. The analyte and solvent streams were combined in a mixing coil. The analyte/solvent flow ratio was varied over the experimental run to produce a Gaussian concentration profile, with a constant total flow rate. The two flowing streams were then combined in a mixing coil before reaching the flow cell.

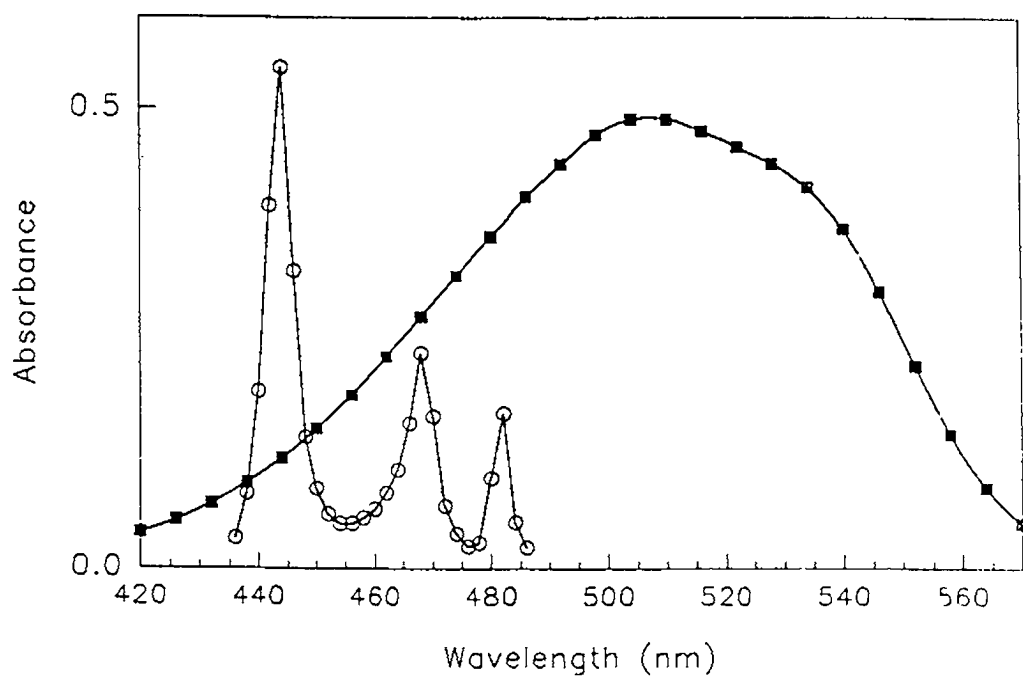
There are several advantages to using this setup. First, the system can be used in a stopped-flow mode where the pumps are turned off before the spectra are acquired. With the flow stopped, the analyte concentration in the cell should remain constant, thus eliminating scan-time effects. The stopped-flow also allows long integration times that increase the precision of the absorbance measurements and provides good variance estimates. A third advantage of this system is that it gives reproducible peak profiles for the analytes regardless of their concentration or the solvent used. This allows results from different analytes to be compared fairly.

In the initial studies, a stopped-flow approach was used with a 4 s integration. These conditions should emphasize the effects of stray and polychromatic radiation. Both methyl orange and praseodymium chloride were studied. Methyl orange has a relatively broad spectrum (Figure 4.20) in acidic





**Figure 4.19.** Stopped flow apparatus used to evaluate the effect of systematic and random noise on EPCIA.

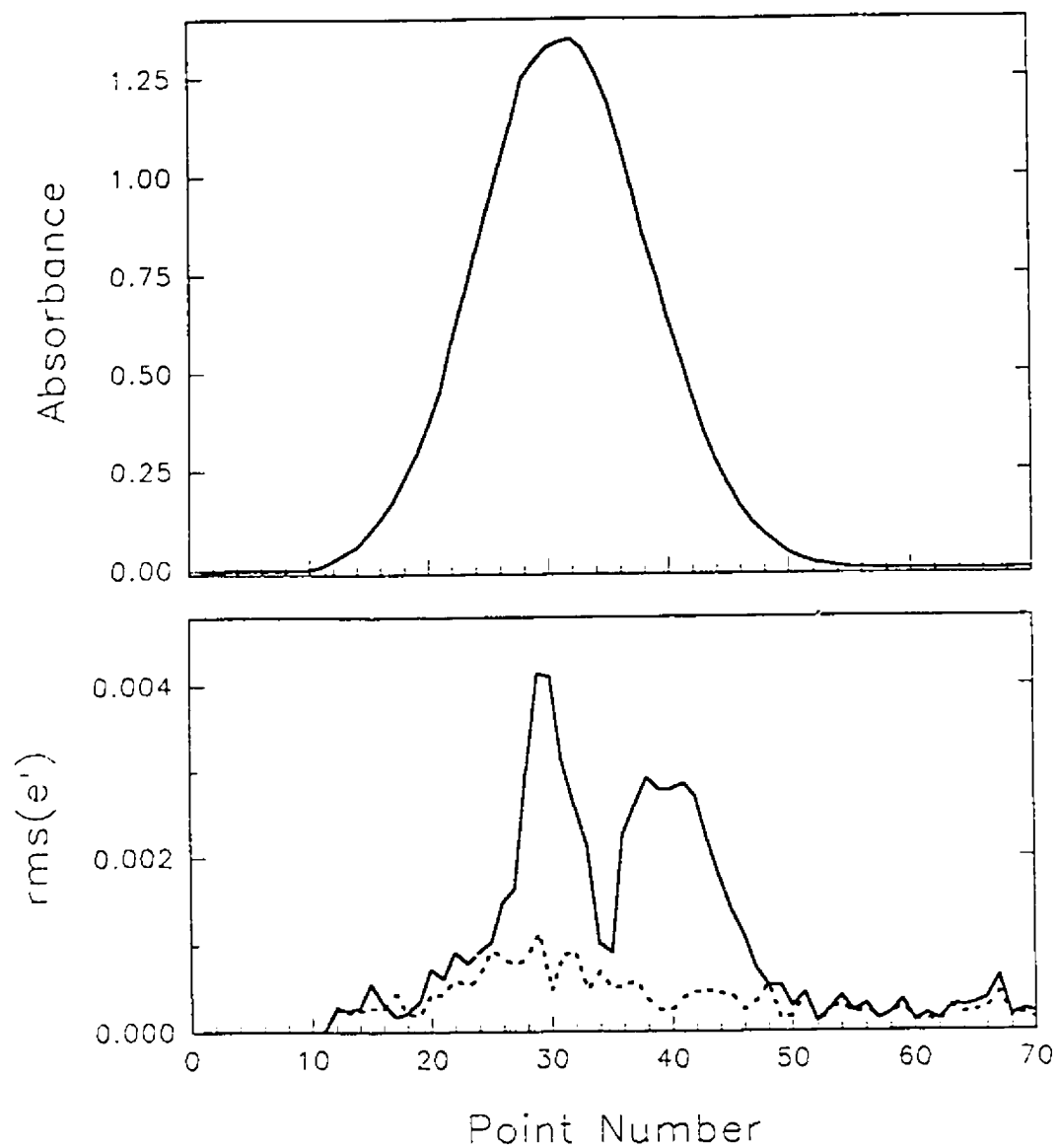


**Figure 4.20.** Absorbance spectra of methyl orange (■) and praseodymium chloride (○) in acidic solution. Symbols mark the wavelengths used for EPCIA.

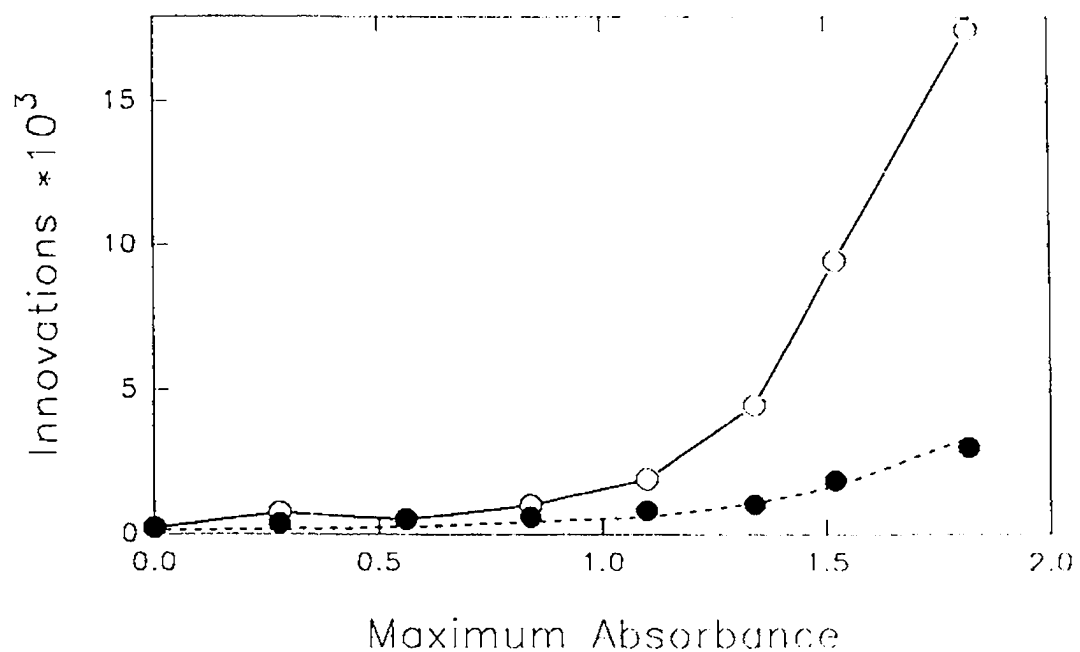
solution, which was recorded at 26 wavelengths (6 nm intervals). The praseodymium spectrum has sharper spectral features. It was also recorded at 26 wavelengths (2 nm intervals).

**Methyl orange.** EPCIA was initially evaluated for nine methyl orange samples with maximum absorbances from 0 to 2 AU. The resulting matrices each contained 70 spectra at 26 wavelengths. The concentration profile had a baseline width of roughly 40 points as seen on the top panel of Figure 4.21. For samples with maximum absorbances of 1 AU or less, the EPCIA results were comparable to those obtained for simulated one-component systems. The one- and two-component models had flat innovation sequences, consistent with the presence of one chemical component. For more concentrated samples, the one-component innovations increased beyond those of the two-component model. This behavior, shown on the bottom of Figure 4.21, was not seen for simulated one-component systems that contained perfectly bilinear data. The failure of the one-component model suggests that two components are present, but a double hump is not expected for two chemical components. This behavior continued for more concentrated solutions giving increasingly large one-component innovations (Figure 4.22). While the two-component innovations also increased, they did not exceed the level of the random noise.

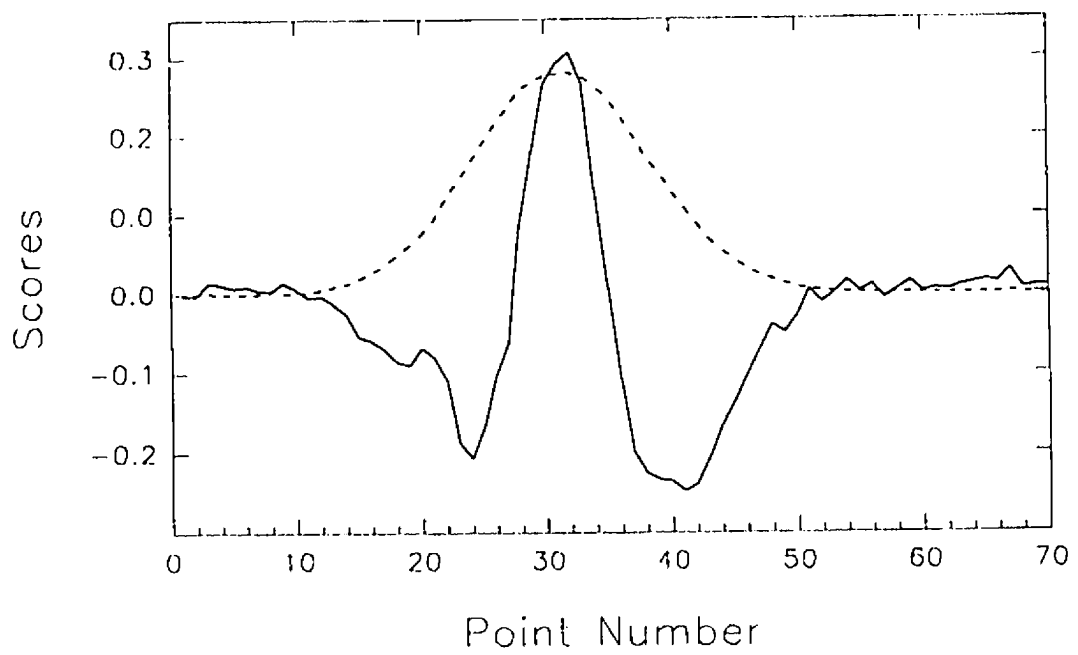
The second component that EPCIA detects is likely an instrumental artifact rather than a chemical component. This was investigated by performing principal components analysis on the data matrix. The scores plot (also called an abstract chromatogram) shown in Figure 4.23 illustrates the contribution of each principal component as a function of sample number. The first principal component (PC1) explains the majority of the variance in the data, and thus has



**Figure 4.21.** Elution profile of methyl orange (top), and rms innovation sequences for one (solid line) and two (dashed line) component models.



**Figure 4.22.** Largest innovations from one (O) and two (●) component models plotted for each data matrix against the maximum absorbance. For example, the result of Figure 4.21 are the two values at 1.4 AU. The expected magnitude of the random noise ( $3\sigma$ ) is given for comparison (dashed line).



**Figure 4.23.** Results from PCA of the data shown in Figure 4.21. The scores of the first (dashed line) and second (solid line) principal component are shown.

the shape of the concentration profile. This component models the average spectrum of methyl orange. If PC1 explained all the systematic variance, then PC2 would only contain random variations. For these data, the PC2 scores have a systematic trend, with the largest scores occurring at the top of the concentration profile. The most negative values of PC2 occur for samples where there is significant absorbance, but Beer's law is still obeyed. Thus PC2 explains the systematic differences that occur for highly absorbing samples.

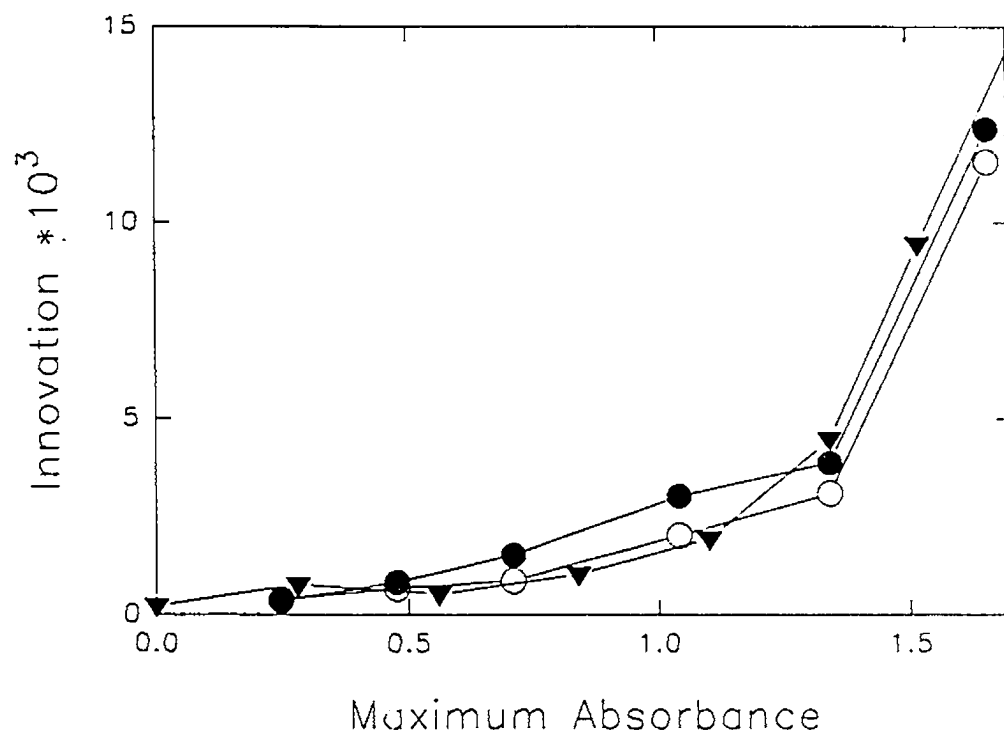
The double hump in the innovations also results from the nonlinearities, which affect Kalman filter differently since it is an evolving model. At the beginning of the concentration profile the filter estimates state parameters to predict the methyl orange spectra. Up to about point 25, these predictions are accurate, so the innovations reflect only the level of random noise. Then the spectra start to change systematically as the analyte absorbance increases beyond about 1 AU. These changes result in larger prediction errors for the model. The state parameters are adjusted in response to spectral changes resulting from deviation from Beer's law, but, because the model is incorrect, the prediction errors increase faster than the parameter adjustments. After the chromatographic peak reaches its maximum, the spectral vector begins to shift back to its original value, so the innovations begin to decrease as it comes closer to the EPCIA model. At some point on the return, the spectrum passes the EPCIA model momentarily, causing the minimum in the double hump. Then it overtakes the EPCIA model in the other direction, causing the innovations to increase again. Finally, the innovations decrease as the signal returns to the baseline. Thus the two components that EPCIA detects are not associated with different chemical species, but rather they are associated with different experimental conditions. In this sense, the two species detected by EPCIA are

methyl orange at low absorbance values and methyl orange at high absorbance values, which are spectroscopically different due to deviations from Beer's law. The results do not permit attribution of the deviations to chemical effects, stray light, or polychromatic radiation.

**Praseodymium chloride.** For comparison, praseodymium samples were run using the same concentration profile and integration time. EPCIA of these data gave comparable results for both the appearance and magnitude of the innovations (Figure 4.24). The praseodymium chloride was expected to show greater deviations due to polychromatic radiation, because its spectrum contains sharper features, but it did not. To investigate this, EPCIA was applied to only the largest spectral peak (436 to 452 nm) in the praseodymium spectrum. In this analysis, the one-component model failed at lower absorbance values (0.7 AU) than any of the previous runs, again showing the characteristic double hump. At higher absorbances the results were similar to those for methyl orange and the analyses using the full spectrum for praseodymium. This suggests that the effects of polychromatic radiation are only observable for wavelength ranges with predominantly large spectral derivatives, otherwise stray light dominates.

In evaluating the effect of the derivative, its sign is not important since the hyperbolic sine in Equation 4.7 is symmetric. Thus, the average magnitude of the derivative can be used for comparing spectra. Table 4.5 contains the spectral derivatives for samples with a maximum absorbance of 1 AU. Recall that the relative contributions of stray light and polychromatic radiation were shown in Figure 4.15 for a wavelength with a spectral derivative of 0.1, which is close to the average derivative for the largest praseodymium peak. The model predicted that the contribution of polychromatic radiation at 1 AU was  $-1.8 \times 10^{-2}$ , roughly twice the stray light. The dependence of the polychromatic





**Figure 4.24.** Innovations from the one-component model applied to PrCl<sub>3</sub> (O); PrCl<sub>3</sub> between 436 and 452 nm (●); and methyl orange (▼).

radiation on the spectral derivative is nonlinear, though, such that decreasing the derivative by half causes the contribution of polychromatic radiation to drop to  $-4.5 \times 10^{-3}$ , making it smaller than the contribution of stray light. Thus the effect of polychromatic radiation on the EPCIA is only significant when the average derivative of the spectrum is large. This was observed for selected regions of the para-xylene and praseodymium spectra. Each of these spectra had an average derivative that was roughly 10% of their maximum absorbance. Interestingly, these values are similar to the average derivative of a Gaussian peak with a standard deviation equal to the bandpass of the instrument.

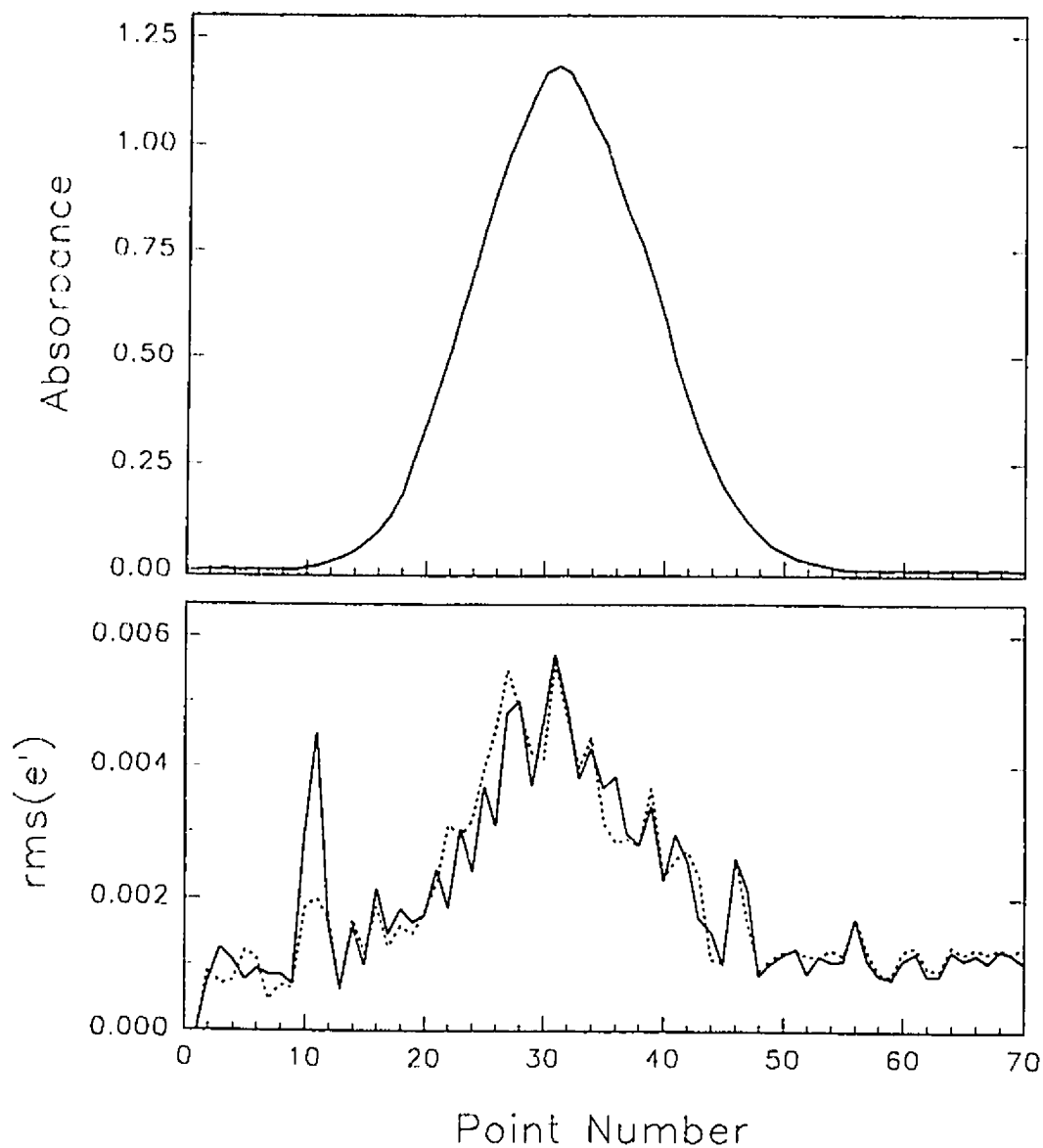
**Table 4.5** Spectral derivatives of the samples studied

<i>Sample</i>	<i>Wavelength Range</i> (nm)	<i>Average Derivative</i> (nm <sup>-1</sup> )
methyl orange	420 - 570	0.012
PrCl <sub>3</sub>	436 - 486	0.050
PrCl <sub>3</sub>	436 - 452	0.096
para-xylene	266 - 280	0.086
Gaussian ( $\sigma = 4.4$ nm)	$4\sigma$	0.098

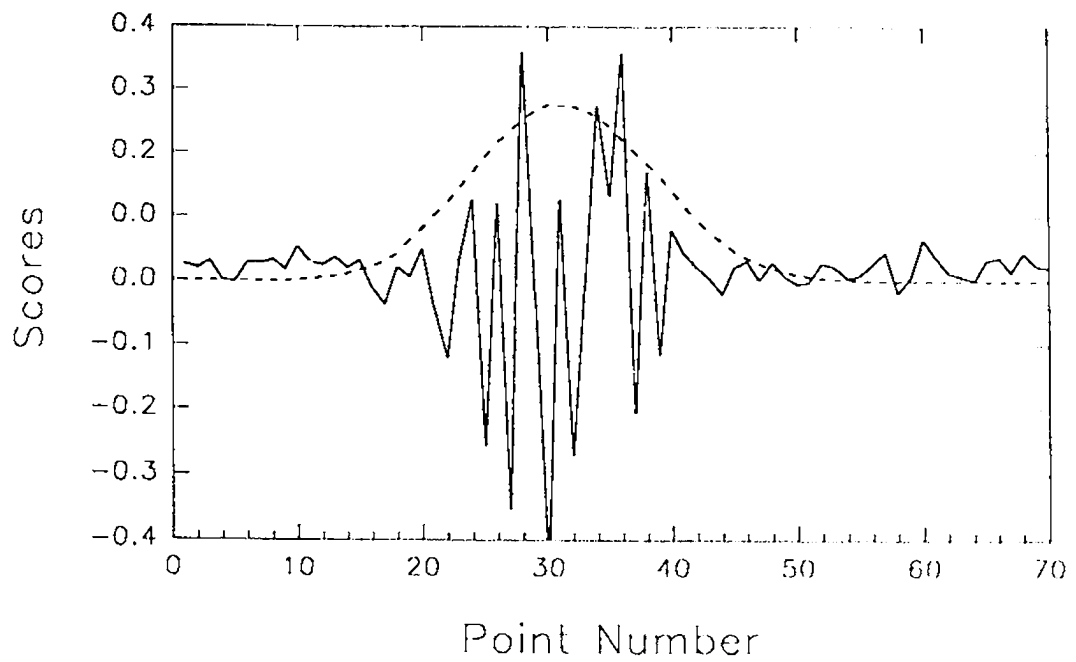
In summary, at low absorbance (<0.7 AU) the innovations from both the one- and two-component models were flat, with a magnitude reflecting the random noise. This indicates that the one-component model predicts the spectrum within the precision of the instrument and therefore the system contains one chemical component. At high absorbance (>1.5 AU), systematic deviations from Beer's law cause the one-component model to fail. These deviations were attributed to stray light since their contribution was independent of the spectral shape of the analyte. The effects of polychromatic radiation could

be observed for samples between 0.5 and 1.5 AU, but only when the average magnitude of the spectral derivative was large. Stray and polychromatic radiation increase the rank of the data and cause EPCIA to overestimate the number of chemical components present.

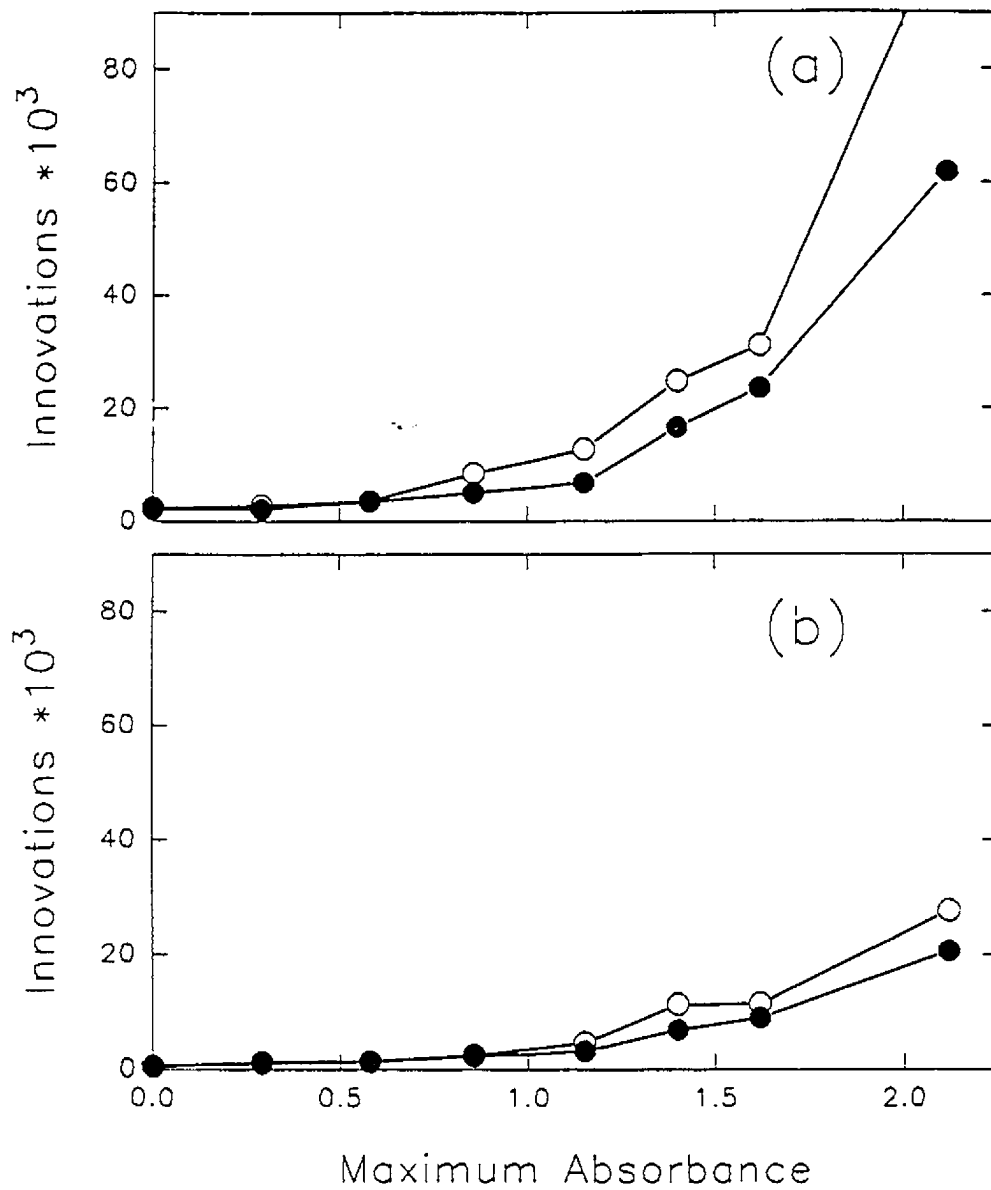
**Random noise.** In characterizing the spectrometer, the variance of the random noise was found to increase with absorbance. The effects of this heteroscedastic noise on EPCIA were not obvious in the previous stopped-fl- experiments since the systematic noise dominated. While the systematic sources of noise cannot be eliminated, the contribution of the random noise can be increased. Accordingly, the methyl orange experiments were repeated with two changes: (1) the integration time was reduced from 4 s to 0.1 s, and (2) a neutral density filter (1.3 AU) was used to reduce the intensity of the spectrometer source,  $P_0$ . This moves the blank reading and subsequent sample readings ( $P$ ) into a noisier range for the detector. The systematic effects are assumed to be unchanged since the model for polychromatic radiation is independent of power, and the neutral density filter would attenuate stray light to the same extent as the spectrometer source. Figure 4.25 shows how the noise affects EPCIA. The innovations for the one- and two-component models increased at high absorbances where the variance of the random noise increases. The scores plot (Figure 4.26) indicates that PC1 is similar to earlier results (Figure 4.23), but PC2 models only random noise in this case. Note that the values of PC2 scores, like the variance of the noise, increase with absorbance. The maximum innovations are shown in Figure 4.27a. Note that the scale is larger than the one used for the original methyl orange data (Figure 4.22) by a factor of five. At the highest absorbances, the one-component model



**Figure 4.25.** Elution profile of methyl orange with the neutral density filter in place (top). The effect of heteroscedastic noise on the innovation sequences is shown (bottom) for one (solid line) and two (dashed line) component models.



**Figure 4.26.** Results from PCA of the data shown in Figure 4.25. The scores of the first (dashed line) and second (solid line) principal components are shown.



**Figure 4.27.** Largest innovations from one (O) and two (●) component models plotted for each data matrix for (a) 0.1 s and (b) 4.0 s integration.

became slightly larger than the two-component model, due to effects of the nonlinearities.

If these large innovations are mostly due to random noise, they should be reduced by ensemble averaging. Figure 4.27b shows that ensemble averaging the spectra for a 4 s integration reduces the innovations significantly. In contrast, the innovations for methyl orange with 0.1 s integration but no neutral density filter were within 10% of the results for the 4.0 s integration. This suggests that, under normal operating conditions, the deviations of the one-component innovations are mostly due to systematic effects, such as stray light, that cannot be removed by averaging.

## 4.5 CONCLUSIONS

The results presented here have demonstrated that the EPCIA algorithm is a useful approach to the problem of peak purity assessment in chromatography when multiwavelength UV-visible absorbance detection is available. It has been shown that the shape of the rms orthogonal innovations sequence will approximate the shape of the elution profile of the minor component, and that the maximum of this sequence is proportional to the concentration of a given minor component over a wide range. This maximum also decreases with increasing spectral correlation and decreasing chromatographic resolution. The minor component will not be detected when the maximum rms innovation approaches the noise level of the absorbance measurements.

The effectiveness of the algorithm may be limited in practice by heteroscedastic noise, spectral scan time, baseline perturbations, and nonlinear response. For one-component systems with sample absorbance below about

0.7 AU on the HP8452A spectrometer, the innovations were attributable to random noise of essentially constant magnitude. With higher sample absorbances, nonlinear instrumental response became important. This caused a double hump in the innovations of the one-component model, but had little effect on the two-component model. This is different behavior than observed for a two-component chemical system, which gave a single peak in the one-component innovations. Heteroscedastic noise caused both the one- and two-component innovations to increase, but was only evident when neutral density filters were used to attenuate the spectrometer beam. The ability of EPCIA algorithm to detect impure chromatographic peaks could probably be improved by correcting for these nonidealities by experimental or computational means.



## SELF-MODELING CURVE RESOLUTION

---

### 5.1 INTRODUCTION

Second-order analytical techniques are designed to produce a matrix of data for each sample run. The theoretical advantages of these multidimensional methods are well documented, particularly for the case where many chemical components may be present in the run. In practice, these advantages will only be realized when suitable techniques for processing the resulting data have been established. This chapter considers techniques for extracting the pure concentration profiles from overlapped mixtures, which result from techniques such as chromatography, kinetics and titrations. This chapter uses liquid chromatography with multiwavelength detection as an example, but the techniques discussed are applicable to data from many other experiments. The goal of the techniques discussed in this chapter is to estimate the matrix of pure concentration profiles,  $C$ , from bilinear data of the form:

$$\underset{(n_s \times n_w)}{\mathbf{D}} = \underset{(n_s \times n_c)}{\mathbf{C}} \underset{(n_c \times n_w)}{\mathbf{S}} + \underset{(n_s \times n_w)}{\mathbf{E}} \quad (5.1)$$

where  $\mathbf{D}$  is the matrix of experimental data  
 $\mathbf{C}$  is the matrix of concentration profiles  
 $\mathbf{S}$  is the matrix of spectral profiles  
 $\mathbf{E}$  is the matrix of experimental errors  
 $n_s$  is the number of samples in time (i.e. spectra)  
 $n_w$  is the number of wavelengths  
 $n_c$  is the number of absorbing components.

There are many methods for extracting information from bilinear data, and the 'best' method is the one that takes advantage of all the prior information

available. The most commonly used information is a knowledge of the pure component spectra, that is, knowledge of the spectral matrix  $S$ . Three cases are considered:

1. Complete knowledge of the spectral profiles exists. The number of chemical components is known and a pure spectrum is available for each. This would be the case for a well-characterized mixture where the concentrations of the components are unknown.
2. Partial knowledge of the spectral profiles exists. In this case the components of the overlapped peak are assumed to be members of a larger set. Specifically, the spectrum of each pure component is contained in a library of known spectra, but components of the peak are otherwise unidentified. Accordingly, the problem is one of identification and quantitation.
3. No knowledge of the spectral profiles is available. This situation often occurs when a peak purity algorithm, such as EPCIA, detects a chromatographic peak containing more than one component. This chapter will focus on mathematical methods of extracting the pure concentration profiles from such data that don't require prior knowledge of the spectral profiles.

There are two general approaches for estimating concentration profiles. **Traditional methods** estimate the mixture spectra as a linear combination of known spectra, thus estimating the contribution of each component. Since they require prior knowledge of the pure spectra, these methods can only be applied to the first and second cases listed above. The other approach, based on **principal components analysis**, is to estimate the mixture as a linear

combination of the principal components. For example, the mixture spectra can be estimated as linear combinations of abstract spectra, or similarly, the individual concentration profiles can be estimated as linear combinations of the abstract chromatograms. Since the principal components are calculated directly from the experimental data, these factor analysis methods are suitable for the second and third cases listed above. A goal of this work is to use information obtained from EPCIA for mixture analysis, when there is no prior knowledge of the spectral profiles.

## 5.2 TRADITIONAL METHODS FOR MIXTURE ANALYSIS

### 5.2.1 Classical Least-Squares

Classical least-squares<sup>32,120</sup> (CLS) is the traditional method for treating the first case mentioned in the introduction, where the spectrum of each component in the mixture is known. Thus CLS is used for quantitative, rather than qualitative, spectral analysis. As was shown in Chapter 3, each element  $d_{ij}$  of the data matrix,  $\mathbf{D}$ , contains the total absorbance of the  $i^{\text{th}}$  sample at the  $j^{\text{th}}$  wavelength. Thus, a system of simultaneous equations can describe the contribution of each component to the mixture spectra. For a simple example with two components and two wavelengths, the following equations can be written for a single sample:

		component 1		component 2	
wavelength 1	$d_{i1} =$	$c_{i1} s_{11}$	+	$c_{i2} s_{21}$	(5.2)
wavelength 2	$d_{i2} =$	$c_{i1} s_{12}$	+	$c_{i2} s_{22}$	(5.3)

If pure component spectra are known, then these two simultaneous equations can be solved for the two unknown concentrations ( $c_{i1}$  and  $c_{i2}$ ) in this sample as:

$$\hat{\mathbf{C}} = \mathbf{D}\mathbf{S}^{-1} \quad (5.4)$$

In Equation 5.4 the concentration profiles are estimated with the inverse of the spectral matrix, which is only defined for square matrices where the number of equations equals the number of unknowns. Often a full spectrum is used in mixture analysis, such that there are more wavelengths than chemical components. This is called an overdetermined system, as the number of equations is greater than the number of unknowns. The least-squares solution for an overdetermined system is:

$$\hat{\mathbf{C}} = \mathbf{D}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1} \quad (5.5)$$

This gives the best fit of the pure-component spectra to the mixture spectra. The drawback to this method is it requires a knowledge of all the pure signals contributing to the measurement. If an absorbing component is not included in the spectral matrix  $\mathbf{S}$ , the model is incomplete. Furthermore, the accuracy of all the concentration estimates suffers with an incomplete CLS model<sup>32,172</sup>.

### 5.2.2 Multicomponent Analysis

The goal of multicomponent analysis<sup>173-175</sup> (MCA) is to extend the classical least-squares approach to a library of reference spectra. In this case the number of candidate reference spectra in the library exceeds the number of components in the overlapped peak. The library is made from a set of pure component spectra. In contrast, the experimental spectra often contain mixtures

of components in unknown proportions. Thus, the problem is to identify the components of the mixture and estimate the contribution of each. To achieve this, Equation 5.5 is solved with the matrix  $S$  containing the complete library of spectra. This calculates the least-squares fit of  $C$ , the estimated concentration profiles of each member of the library. Ideally, members of the library contained in the sample have reasonable concentration profiles and species that are absent have zeros for their columns of  $C$ . One difficulty in practice is the need for the library spectra to match the experimental spectra well<sup>175</sup>. In LC-UV this becomes a problem when the library spectrum for an analyte has been recorded in a different solvent than the one it elutes with during the experiment. Small differences between these spectra can lead to errors in quantitation, or worse, a misidentification of the number and nature of the components. Finally, in CLS, the MCA model is only accurate when all the components of the mixture are included in the model.

### 5.3 FACTOR ANALYSIS METHODS FOR MIXTURE ANALYSIS

In principal components analysis a data matrix, usually of large dimensions, is decomposed into two smaller matrices:

$$\hat{\mathbf{D}}_{(n_s \times n_w)} = \mathbf{X}_{(n_s \times n_p)} \mathbf{Y}_{(n_p \times n_w)} \quad (5.6)$$

where  $\hat{\mathbf{D}}$  is the factor reproduced matrix

$\mathbf{X}$  is the scores matrix

$\mathbf{Y}$  is the loadings matrix

$n_s$  is the number of samples in time (i.e. spectra)

$n_w$  is the number of wavelengths

$n_p$  is the number of principal components

As this reduction is guided by mathematical principles, namely the explanation of variance, the factors obtained may not be chemically relevant. As was shown in Chapter 3, the number of principal components is equals the number of spectroscopically observable components for perfectly bilinear data. In practice, the effects of systematic and random noise, as discussed in Chapter 4, can increase the number of principal components required to accurate model the data. In such cases the concentration matrix,  $C$ , has the same dimensions as the scores matrix  $X$ , and accordingly its columns are often considered abstract chromatograms. A similar connection exists between the spectral matrix,  $S$ , and the loadings matrix  $Y$ , such that its rows are considered abstract spectra. Thus a fundamental problem in factor analysis is deriving a transformation matrix,  $T$ , that converts the abstract factors to the true factors. This transformation matrix estimates the pure component spectra and elution profiles as

$$\hat{C} = XT \quad (5.7)$$

$$\hat{S} = T^{-1} Y \quad (5.8)$$

where  $T$  is the ( $n_c \times n_c$ ) transformation matrix. Methods of estimating  $T$  can be subdivided into two categories: (1) modeling methods and (2) self-modeling methods. Modeling methods<sup>176</sup> use a "hard" model of the true factors, such as Beer's law with pure component spectra. Modeling methods can also use a functional form of the concentration profile such as a peak shape for chromatography or equilibrium equations for titrations. These hard models require many assumptions about the chemical system understudy, that are often unavailable for an unknown sample. In contrast, self-modeling or "soft" methods use information extracted from the data set and mathematical constraints. One

example of a constraint would be that the estimated concentration profiles must be nonnegative. The advantage of self-modeling methods is that they can be applied with very little prior knowledge, the disadvantage is that the results can be ambiguous.

### 5.3.1 Target Factor Analysis

Target factor analysis<sup>150,175</sup> (TFA) is a tool for mixture analysis that can be used with only a partial knowledge of the spectral profiles. Thus, it is a factor analysis approach to solving the problem treated previously (Section 5.2.2) with multicomponent analysis. It is a hard-modeling method since it requires pure component spectra. The goal of TFA is to select which members of a spectral library contribute to the experimental spectra. This is equivalent to asking which of these candidate spectra, or targets, are true factors of the data set. The underlying assumption of this approach is that true factors can be accurately described by a linear combination of abstract factors.

There are three major steps in TFA. The first is to calculate the combination of abstract spectra that best describes the target:

$$\mathbf{t}_i = \mathbf{Y} \mathbf{s}_{target,i}^T \quad (5.9)$$

$(n_c \times 1)$        $(n_c \times n_w)$        $(n_w \times 1)$

$\mathbf{s}_{target,i}$  is the  $i^{th}$  target vector (i.e. the test spectrum), and  $\mathbf{t}_i$  is the  $i^{th}$  column of the transformation matrix. Then the predicted vector is calculated with the transformation:

$$\hat{\mathbf{s}}_i = \mathbf{t}_i^T \mathbf{Y} \quad (5.10)$$

$(1 \times n_w)$        $(1 \times n_c)$   $(n_c \times n_w)$

Where  $\hat{S}_i$  is the projection of the  $i^{th}$  target vector into the factor space. This is the best description of the candidate spectrum that can be calculated as a combination of the abstract spectra. If this candidate spectrum is a member of the mixture, the original target vector,  $s_{target,i}$ , and this predicted vector,  $\hat{S}_i$ , should be very similar. Thus the final stage of target factor analysis is to evaluate if the transformed abstract factors describe the target factor within experimental noise:

$$\hat{S}_i \stackrel{?}{=} s_{test,i} \quad (5.11)$$

Malinowski<sup>125</sup> has suggested statistical tests for equation 5.11 that consider errors in the data matrix and the target itself. As in any statistical test, there is a finite chance obtaining a false-positive result. For example, one of the library spectra may coincide with the factor space, even though it is not in the mixture.

If this process can be repeated to find all the component spectra, the entire transformation matrix can be deduced. Finally the concentration profiles can be predicted with Equation 5.7. The advantage of TFA is that it allows the targets to be tested individually. Therefore, results from TFA may still be valid even when the spectral library is incomplete. In some cases, the concentration profiles of the identified components can be estimated<sup>144,177</sup> despite the incomplete model for the mixture. This was not possible with the traditional methods. Another advantage of TFA is that it is more robust than MCA when differences exist between the library spectra and experimental spectra<sup>175</sup>. The disadvantage of TFA, like all the factor analysis techniques considered in this chapter, is that it requires some chromatographic separation.



### 5.3.2 Self-Modeling Curve Resolution

The term self-modeling curve resolution (SMCR) has been used by Delaney<sup>178</sup> to encompass techniques that: 'determine the number of components in an overlapped chromatographic peak as well as the spectrum and concentration profiles of each compound, without assumptions regarding peak shape, location, or identity.' SMCR attempts to find the transformation matrix that converts the abstract factors into true factors. The crux of the curve resolution problem is to find the 'best' transformation. This cannot be calculated in the least-squares sense since there is no 'hard' model to compare with the transformed results. Instead, the general approach is to constrain the solutions to those that are physically and chemically reasonable. Some common constraints are: (1) that all points in the component concentration profiles are nonnegative, (2) that all points in the component spectra are nonnegative, and (3) that the component concentration profiles are unimodal, that is, each has a single maximum.

The accuracy of results from SMCR depends upon choosing constraints that are reasonable for the experimental system. For example, the constraint of nonnegative spectra is not valid for measurements like optical rotation. Similarly, the constraint of unimodal concentration profiles is only applicable to certain types of ordered data, such as spectrochromatograms. Another practical consideration is how experimental noise, both random and systematic, can invalidate these assumptions. In making too rigid an assumption, the accuracy of SMCR may well suffer. Conversely, if the constraints are too general, the 'precision' of SMCR is limited, in that a wide range of solutions may be mathematically satisfactory. With these considerations in mind,

Gemperline<sup>111</sup> has cautioned that, "without a theoretical model to base the results on, an infinite number of reasonable solutions or 'best fits' exist; while constraints can narrow the set of feasible solutions somewhat, physically correct solutions are never guaranteed."

Many different approaches have been taken to attempt to solve the SMCR problem. Some of which are briefly summarized here. For a more extensive discussion of the area, the reader is directed to a review by Hamilton and Gemperline<sup>111</sup>, and tutorials by Vandeginste<sup>112</sup>, and Windig<sup>101</sup>. The first application of SMCR was reported by Lawton and Sylvestre<sup>179</sup> in 1971. They used SMCR to estimate the pure UV-vis spectra of a two-component mixture by assuming the resolved spectra had nonnegative absorbance and unique spectral regions. From these estimates in the spectral domain, the concentration profiles were calculated with the traditional least-squares approach. This technique has also been applied to other experimental systems, including GC/MS<sup>180</sup> and IR spectra of equilibrium mixtures<sup>181</sup>. Also, various attempts have been made to extend this approach to three components<sup>182,183</sup>.

The complementary approach is to solve for the elution profiles first, and then estimate the component spectra. This will be referred to as the concentration domain approach. Much of the work in the concentration domain has been based on the iterative target transform<sup>151,175,184-187</sup> (ITT) approach, which is outlined in the following section. Vandeginste and coworkers have noted several advantages<sup>188,189</sup> of working in the concentration domain as compared to the spectral domain. These include more accurate quantitative estimates and a lower sensitivity to noise. Also, concentration-domain methods have been extended to systems of greater complexity<sup>140</sup> including four-component chromatographic peaks and multicomponent equilibria.

The concentration and spectral domain approaches are not necessarily exclusive. SMCR techniques based on alternating regression<sup>190</sup> apply constraints in both domains concurrently. Another area of active research involves designing interactive or graphical methods that combine factor analysis techniques with the user's intuitive knowledge of the experimental system. Examples include the SIMPLISMA method developed by Windig and coworkers<sup>191</sup>, and the HELP method of Kvalheim *et al*<sup>192</sup>.

### 5.3.3 Iterative Target Testing

All the curve resolution results that follow will be calculated with the iterative target transform (ITT) applied in the concentration domain. This technique was independently developed by Gemperline<sup>151</sup> and Vandeginste<sup>185</sup> to model the concentration profiles of overlapped chromatographic peaks. As the name implies, ITT is related to target factor analysis, but there are two important differences between the methods. First, ITT is a self-modeling method that uses information from the data matrix, rather than an external reference, for the target. Second, ITT is an iterative method that generates progressively refined estimates of the transformation matrix. Figure 5.1 shows the four major steps to the ITT algorithm:

1. A starting profile is generated from the data matrix, often with a factor analysis-based method. This starting profile is a first guess at one of the underlying concentration profiles.
2. The starting target is projected into the factor space with the estimated transformation matrix:

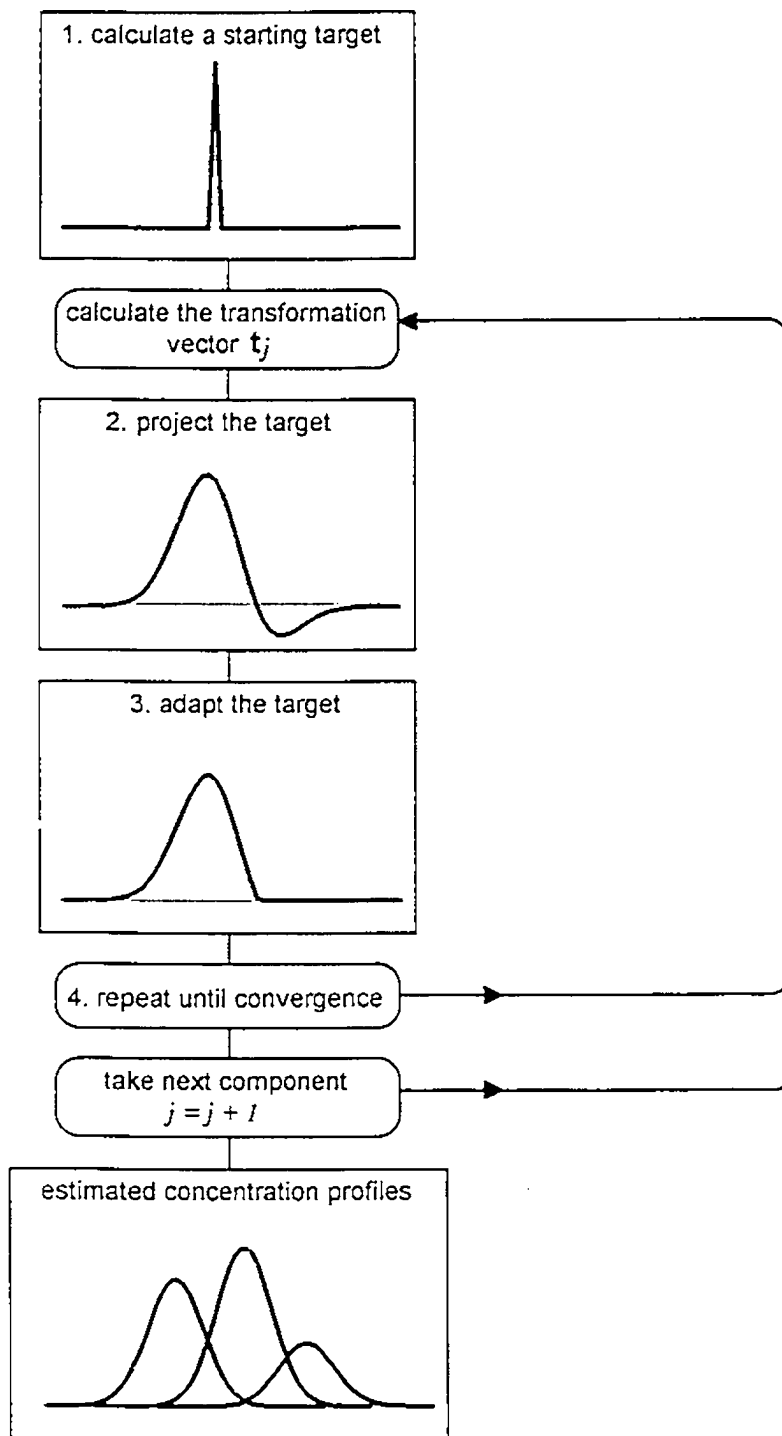


Figure 5.1. A schematic illustration of the iterative target transform.

$$\mathbf{t}_j = \Lambda^{-1} \mathbf{X}^T \mathbf{c}_{target,j} \quad (5.12)$$

$$\hat{\mathbf{c}}_j = \mathbf{X} \mathbf{t}_j \quad (5.13)$$

where  $\mathbf{t}_j$  is one column of the transformation matrix  
 $\Lambda$  is a diagonal matrix of the squared eigenvalues  
 $\mathbf{X}$  is the scores matrix  
 $\mathbf{c}_{target}$  is the target vector  
 $\hat{\mathbf{c}}_j$  is the predicted vector

These are the target testing equations for a concentration profile. The eigenvalue matrix serves to normalize the scores vectors to unit length. This was unnecessary for spectral targets (Equations 5.9 and 5.10) because the loadings matrix is, by definition, already normalized. If the starting profile is a true factor of the data matrix it will be unchanged by this projection. On the first few iterations of ITT, the predicted concentration profile usually changes significantly. Often the predicted vector is a better approximation to the true concentration profile than the starting profile, but it may still have some undesirable characteristics, such as regions of negative concentration.

3. The next step is to remove the undesirable characteristics of the projected target. This is called adapting or refining the target, and it is at this stage that constraints are applied. For example, regions of negative concentration may be set to zero.
4. The adapted target becomes the starting target for another iteration of the procedure. Ideally, this new starting point gives a better estimate of the transformation matrix and the concentration profile. Obviously there must

be some criteria to terminate the iterative process. Examples of termination criteria include: (i) the predicted target is not significantly different from the starting target, (ii) the current refinements on the predicted profile are larger than the previous one - suggesting that the profile is getting worse with continued transformation, and (iii) a maximum number of iterations has been reached.

A major advantage of ITT is that it estimates the concentration profile of each component individually. Thus, in principle, there are no restrictions on the number of components it can resolve from an overlapped chromatographic peak.

The main difference in the ITT approaches reported by Gemperline and Vandeginste lies in the method for calculating starting profiles from the experimental data. Gemperline used a **needle search** that starts with a set of vectors containing a unit vector for each retention time:

$$\mathbf{t}_1 = [1 \ 0 \ 0 \ \dots \ 0 \ 0]'$$

$$\mathbf{t}_2 = [0 \ 1 \ 0 \ \dots \ 0 \ 0]'$$

$$\vdots \quad \quad \quad \vdots$$

$$\mathbf{t}_{n_s} = [0 \ 0 \ 0 \ \dots \ 0 \ 1]'$$

These needle vectors could be considered to be very narrow Gaussian peaks. The next step was to select one of these starting vectors for each unresolved component. Ideally, needle vectors that matched the retention times of the true concentration profiles would be selected. Gemperline chose starting profiles from this set with target factor analysis.

Vandeginste used a method called **varimax rotation**, which originated from the application of factor analysis in the social sciences. This method assumes that the abstract factors are related to the true factors with an orthogonal rotation matrix. For a two-component case, this matrix is of the form:

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \quad (5.14)$$

This transformation matrix is used in Equation 5.7 to predict the concentration matrix. The angle  $\theta$  is chosen to maximize the varimax function:

$$v = \sum_{i=1}^{n_c} \left( \frac{1}{n_c} \sum_{j=1}^{n_s} c_{ij}^4 - \frac{1}{n_c} \left( \sum_{j=1}^{n_s} c_{ij}^2 \right)^2 \right) \quad (5.15)$$

for the normalized predicted vectors  $c_i$ . The varimax function maximizes the squared variance of this vector. The theory is that this function favors a rotated vector which combines large and small elements over one with all intermediate values. This rotation generally gives a smooth maximum in each column of the starting profiles.

Starting profiles can also be generated with **evolving factor analysis**<sup>135-139</sup> (EFA). This method uses a series of eigenvalues calculated from subsets of the experimental data matrix (see Section 3.5 for details), thus it takes advantage of the ordered nature of the data. Results from evolving factor analysis have been used in the SMCR of spectrochromatograms<sup>137</sup>, spectrophotometric titrations<sup>193</sup>, and pyrolysis mass spectra.

An alternative method for generating starting profiles, which has not been explored, is **EPCIA**. As was shown in Chapters 3 and 4, the rms orthogonal innovations from EPCIA can be used to estimate the concentration profiles from

a spectrochromatogram. Therefore, it would seem natural to use these as starting profiles for the ITT.

#### 5.4 ITERATIVE TARGET TESTING AND EPCIA

This section compares results of ITT starting profiles from the time-domain generated by EPCIA, varimax rotation and needle search. EFA was also considered in a preliminary study, but it showed a dependence on experimental parameters, such as the sampling rate and the length of the baseline, that was not seen for the other three methods. This behavior made it difficult to fairly assess EFA, and it was not included in subsequent work. Exclusion of EFA from consideration here does not imply that EFA is a poor method for generating starting profiles, but instead, it suggests that a more comprehensive study is required to understand its performance. The suitability of the starting profiles for ITT is also addressed. The goal of this study was to determine what effect the quality of the starting profile has on the final concentration profile generated by ITT. As all the results are for simulated data, this can only be considered a fundamental study of the different approaches. The effects of experimental nonidealities like heteroscedastic noise and nonlinear instrumental response are not considered here, but they are important considerations in the application of any SMCR method. In practice, the ability of a method to generate realistic starting profiles in the presence of such experimental difficulties may prove to be more important than the limitations considered here.



### 5.4.1 Experimental

The simulated data used in this study were generated in the same manner as those used to evaluate the fundamental limitations of EPCIA (see Section 4.2 for details). The concentration profiles were Gaussian profiles with a standard deviation of 15 s. These were sampled once per second to give 60 spectra across the chromatographic profile. While this sampling rate was higher than required for these techniques, it allowed subtle changes in the predicted retention times to be observed. The spectrochromatograms were generated by multiplying concentration profiles, of varying degrees of chromatographic overlap, with experimentally measured PAH spectra. The spectra of anthracene, fluoranthene, phenanthrene and triphenylene were recorded between 210 and 310 nm at 4 nm intervals. The chromatographic peak was recorded from  $4\sigma$  before the retention time of the first peak to  $4\sigma$  after the retention time of the last peak. Random noise was added with a standard deviation of 0.01 % of the maximum absorbance.

To examine the limitations of the different approaches, simulation studies were carried out on the effects of three parameters: chromatographic resolution, spectral correlation and effective concentration ratio. A reference set of conditions was chosen; each study considered the effect of varying one of the three parameters. The reference conditions used here were a two-component peak of phenanthrene and fluoranthene with a chromatographic resolution of 0.35 and an effective concentration ratio of unity.

The ITT algorithm was written in MatLab. The profiles were refined by setting the negative concentrations to zero on each iteration. The transform was

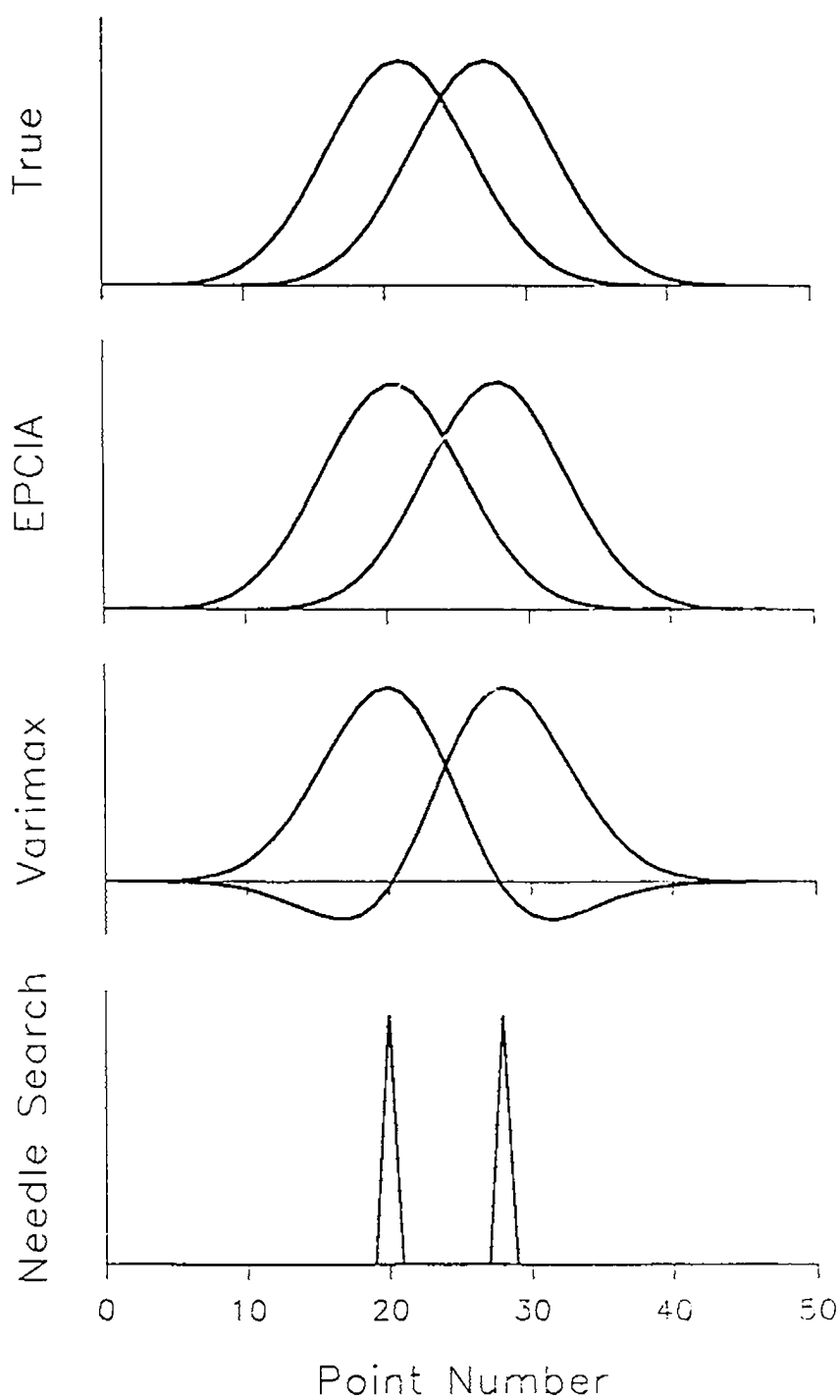
repeated ten times for each starting profile. Greater numbers of iterations were tested, but no significant improvements were obtained for these data sets.

#### 5.4.2 Effect of Chromatographic Resolution

**Starting Profiles.** Figure 5.2 shows starting profiles generated by EPCIA, varimax rotation, and needle search of an overlapped chromatographic peak. Each method has a characteristic shape. For EPCIA, the starting profiles are quite similar to the true concentration profiles. These profiles contain only positive values since they are the rms of the innovations. The varimax profiles are also similar to a Gaussian profile, except they contain regions of negative concentration. The needle search profiles provide reasonable estimates of the retention times, and by definition contain only positive values. Other than this, they do not resemble the concentration profiles. The quality of the starting profiles was evaluated with three metrics: (1) Euclidean distance, (2) retention time, and (3) peak width. Euclidean distance refers to the distance between the true concentration profile ( $c_{\text{true}}$ ) and the starting profile ( $c_{\text{start}}$ ). This distance was evaluated as the sum of the squared residuals (SSR):

$$\text{SSR} = \sum_{j=1}^{n_s} (c_{\text{true},j} - c_{\text{start},j})^2 \quad (5.16)$$

Both concentration vectors were normalized to unit area before this calculation. The retention time was taken to be the maximum of the concentration profile. The peak width was measured at 10% of the peak maximum. This measure was used instead of the standard deviation because the shape of the starting profiles often deviated significantly from a Gaussian peak.

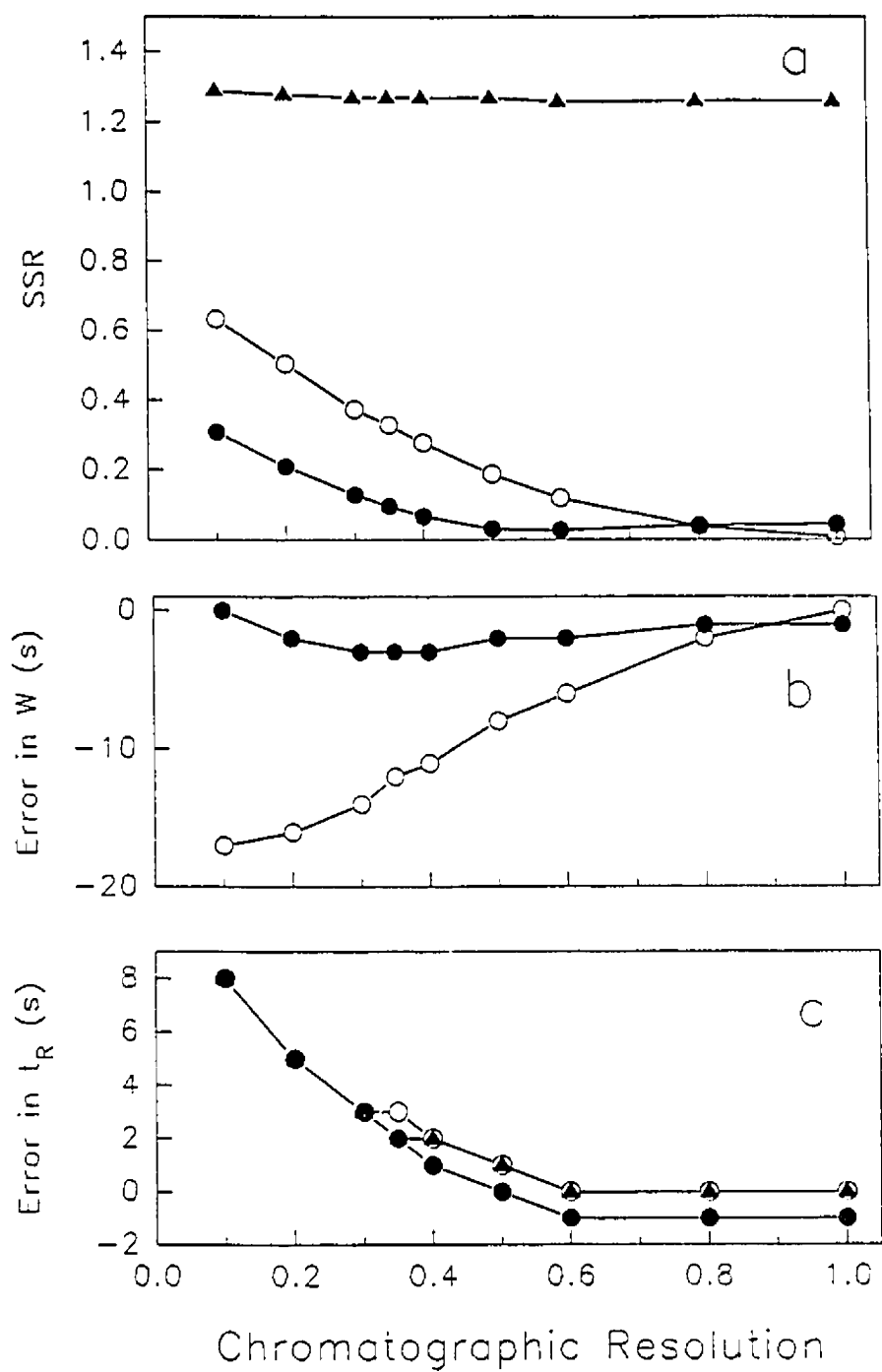


**Figure 5.2.** Starting profiles generated by EPCIA, varimax rotation and needle search are compared to the true concentration profiles of an overlapped ( $R_s = 0.35$ ) two-component peak. For comparison all peaks are scaled to unit height.

The quality of the starting profiles was compared for two-component peaks with chromatographic resolutions of 0.1 to 1. The results given in Figure 5.3 are for the fluoranthene concentration profile; the results for phenanthrene are not shown, but they were almost identical. The overall performance of all three methods for generating starting profiles decreased steadily for chromatographic resolutions of 0.5 or less. In this region the EPCIA gave the lowest SSR, followed by varimax rotation. The needle search vectors had a large SSR for all the cases studied, reflecting the significant differences between the shape of the starting profiles and the true profiles. With chromatographic resolutions above 0.7, the starting profiles from varimax rotation were closer to the true profiles, but EPCIA also gave good results.

These methods were found to give very different estimates of the peak widths. The needle search method does not really estimate peak width, as it gives only an estimated retention time. The varimax rotation gave accurate estimates of peak width when the two peaks were completely separated, but these estimates got steadily worse for overlapped peaks (Figure 5.3b). Two changes were observed in the varimax profiles as the chromatographic resolution was decreased: the profiles showed greater regions of negative concentration, and the peak widths were underestimated. In contrast, EPCIA gave good estimates of the peak width down to a resolution of 0.1.

The ability of the three methods to estimate retention time was found to be very similar in this study (Figure 5.3c). All the methods tended to overestimate the retention time of the second component and underestimate that of the first. As a result, the starting profiles suggest a larger separation between the two peaks than is actually present. This implies that these methods emphasize the pure component regions on the 'outsides' of the peak cluster.



**Figure 5.3.** The effect of chromatographic resolution on the starting profiles. Profiles generated by EPCIA (●), varimax rotation (○) and needle search (▲) compared to the true concentration profiles: (a) sum of squared residuals, (b) systematic error in peak width, (c) systematic error in retention time.

**Transformed Profiles.** The more important results are those for the transformed profiles produced by ITT of the starting profiles. The profiles from all three methods converged with 10 iterations or less. The EPCIA profiles usually converged fastest, often taking only three or four iterations. The quality of these transformed profiles is indicated in Table 5.1. The method with the smallest SSR for the initial profiles gave the smallest SSR for predicted concentration profiles in every case. The differences among the transformed profiles from the three methods were less dramatic than the differences evident in the starting profiles. Notably, the results from the needle search and varimax rotation were nearly identical despite the differences in their starting profiles.

**Table 5.1** *The effect of chromatographic resolution on the concentration profiles predicted by ITT is given for each type of starting profile.*

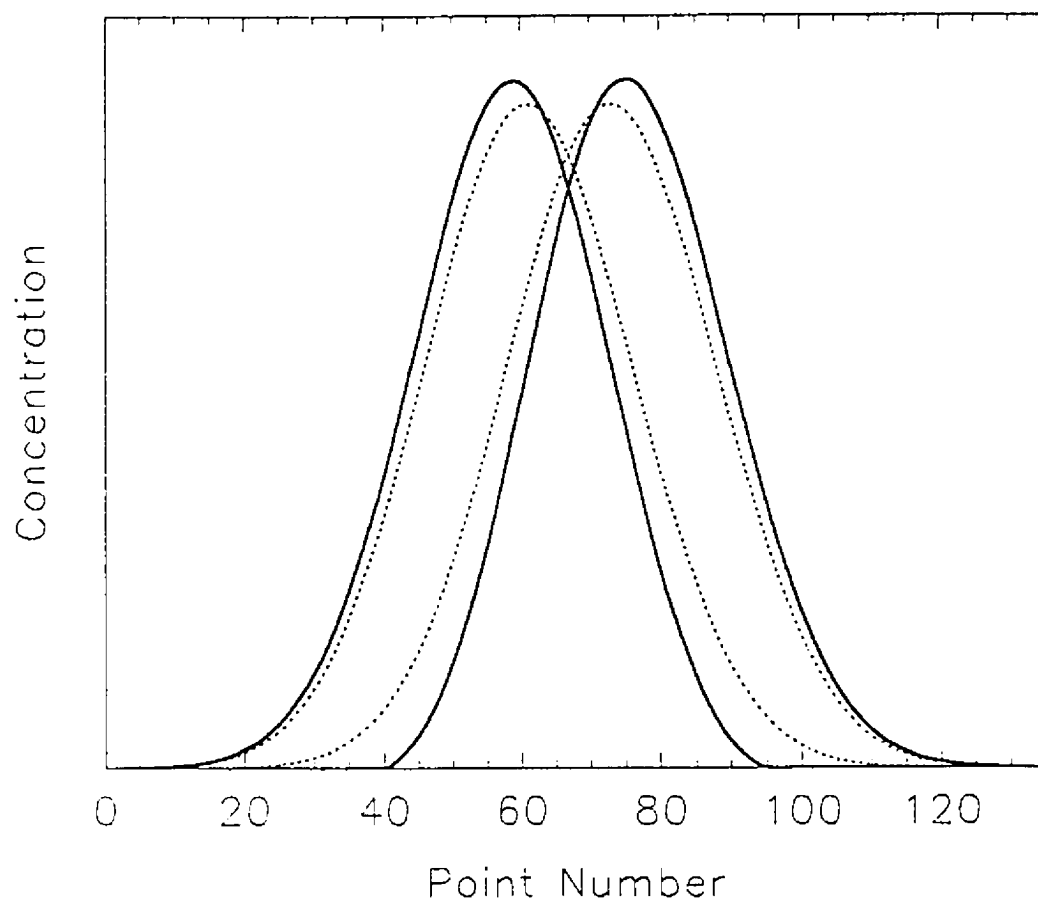
<i>Rs</i>	Sum of the Square Residuals		
	<i>EPCIA</i>	<i>Needle Search</i>	<i>Varimax</i>
1.0	0.001	0.000	<b>0.000</b>
0.8	0.002	0.001	<b>0.001</b>
0.6	<b>0.003</b>	0.011	0.010
0.5	<b>0.017</b>	0.028	0.025
0.4	<b>0.023</b>	0.062	0.061
0.35	<b>0.056</b>	0.087	0.085
0.2	<b>0.083</b>	0.119	0.119
0.1	<b>0.146</b>	0.200	0.200

ITT improved the peak width estimates of the varimax and needle profiles, but EPCIA still gave more accurate results. The transformed profiles had correct retention times when the chromatographic resolution was larger than 0.4. For

more overlapped peaks, the predicted retention times still overestimated the separation between the two components, but these errors were reduced in magnitude. For example, with  $R_s = 0.1$  the predicted retention time from EPCIA had an error of 4 s, while the errors from the varimax and needle search were 6 s. For comparison, all three methods had an error of 8 s for their initial profiles. As the ITT results from the varimax and needle search were practically identical the following sections will report varimax results only, unless the two results differ.

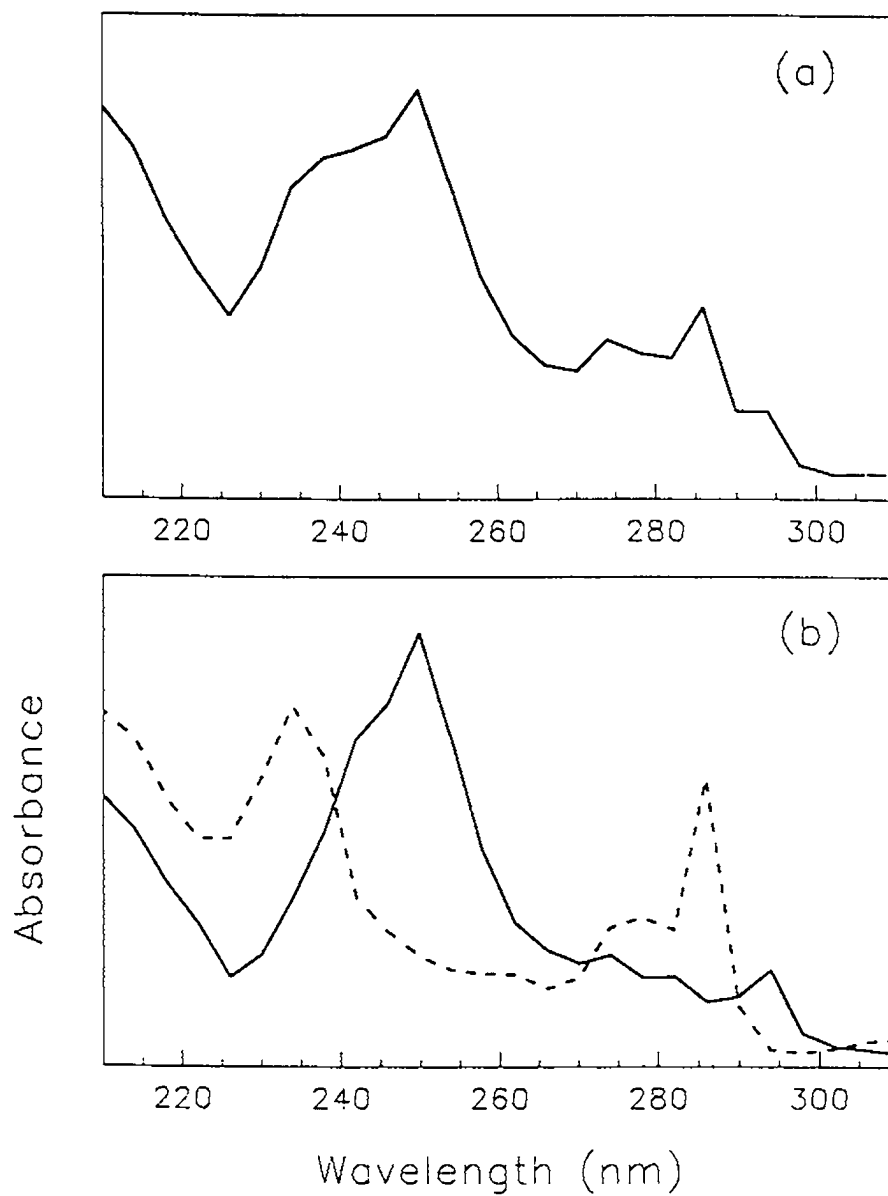
Figure 5.4 compares two concentration profiles with  $R_s = 0.2$  to the predicted profiles from the ITT of EPCIA starting profiles. The systematic errors in peak width and retention time are apparent at this degree of overlap, but otherwise the profiles are reasonable. These profiles have been normalized to unit area. In the absence of calibration standards, SMCR cannot provide absolute concentration estimates. The predicted profiles and the true concentration profiles are related through unknown scaling factors. The *relative* concentrations of the two components can be estimated by assuming the compounds give the same detector response. One way of using this assumption is to estimate the relative concentrations from the lengths of their predicted spectral vectors.

**Spectral Profiles.** Once the concentration profiles have been estimated the spectral profiles can also be extracted with Equation 5.8. Table 5.2 compares the estimated and true spectral profiles for a range of chromatographic resolutions. The accuracy of these results follow the same trends as those seen in the concentration profiles, which is not surprising since both estimates are based on the same transformation matrix. Figure 5.5 illustrates the systematic errors in a spectrum extracted from a highly overlapped ( $R_s = 0.1$ ) two-



**Figure 5.4.** Concentration profiles predicted by ITT (solid line) compared with the true profiles (dashed line) for a two-component peak with  $R_S = 0.2$ . EPCIA was used to generate the starting profiles.





**Figure 5.5.** Results from ITT a two-component peak with  $R_s = 0.1$ : (a) the extracted spectrum of phenanthrene, (b) the true spectra of phenanthrene (solid line) and triphenylene (dashed line).

component peak. The estimated spectrum of the first component (phenanthrene) is recognizable, but it shows some features of the second component (fluoranthene) at this degree of overlap. This spectral "leakage" is due to errors in the concentration profiles being propagated into the spectral profiles.

**Table 5.2** *The effect of chromatographic resolution on the pure spectral profiles predicted by ITT is given for each type of starting profile. The SSR is the sum for both components.*

<i>Rs</i>	Sum of the Square Residuals	
	<i>EPCIA</i>	<i>Varimax</i>
1.0	0.001	<b>0.000</b>
0.8	0.001	<b>0.000</b>
0.6	<b>0.001</b>	0.005
0.5	<b>0.010</b>	0.015
0.4	<b>0.014</b>	0.034
0.35	<b>0.035</b>	0.050
0.2	<b>0.104</b>	0.129
0.1	<b>0.194</b>	0.211

In summary, the starting profiles obtained from EPCIA gave the best results for two-component mixtures of equal concentration. The starting profiles and predicted concentration profiles from ITT were closer to the actual elution profiles in most cases, particularly those with low chromatographic resolution. This result is probably due to their more realistic shape, since the retention times of the starting profiles from all three methods were similar. All three methods tended to overestimate the separation between the two components, which limits their performance for peaks with chromatographic resolution of less than 0.4.

### 5.4.3 Effect of Spectral Correlation

The effect of spectral correlation was assessed for six different pairs of spectra (Table 5.3). The starting profiles from varimax rotations were identical in each case. In contrast, the starting profiles from EPCIA improved with increasing spectral correlation. Recall that in studying the limitations of EPCIA (Chapter 4), the maximum of the innovation was found to decrease with increasing spectral overlap. The results from this study suggest that the EPCIA approach also gives better starting profiles with highly correlated spectra, perhaps because under these conditions the innovations are similar to what would be obtained with adaptive filtering.

**Table 5.3** *The effect of spectral resolution on the concentration profiles predicted by ITT is given for EPCIA and varimax starting profiles.*

<i>Components</i>	<i>Spectral Angle</i>	<u>Sum of the Square Residuals</u>	
		<i>EPCIA</i>	<i>Varimax</i>
Anth / Phen	17.4	<b>0.076</b>	0.656
Phen / Tri	32.9	<b>0.147</b>	0.656
Anth / Tri	38.0	<b>0.171</b>	0.656
Phen / Flu	39.6	<b>0.175</b>	0.656
Anth / Flu	51.7	<b>0.189</b>	0.656
Flu / Tri	55.9	<b>0.212</b>	0.656

Unlike the innovations, which are calculated from the raw data, the varimax method is used to generate starting profiles from abstract factors that are normalized to unit area. The general appearance of these abstract

chromatograms is independent of spectral correlation, long as the two components are spectroscopically different. The only property of the abstract chromatograms which depends on the spectral correlation is their signal-to-noise ratio. As the differences between the spectra decrease they will eventually become comparable to the experimental noise. At this extreme, the quality of the varimax starting profiles would be degraded. This behavior was not observed for the varimax starting profiles in Table 5.3, because the S/N ratio of these data is very high.

#### **5.4.4 Effect of Concentration Ratio**

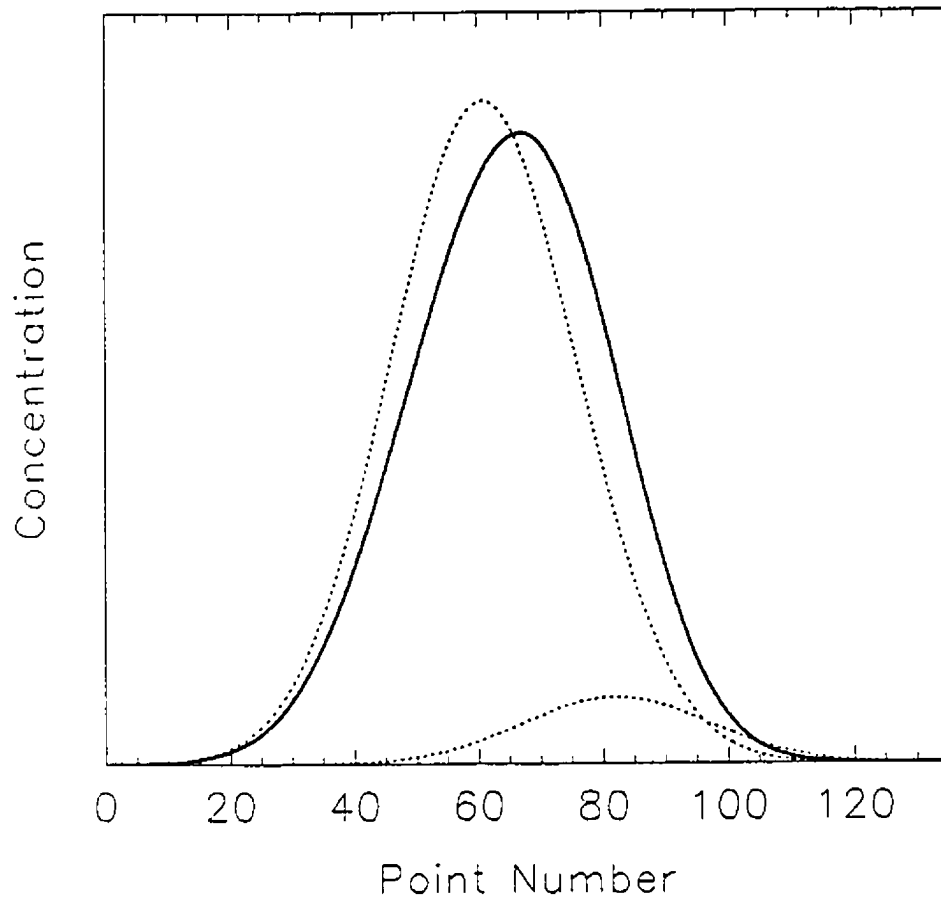
From the studies in Chapter 4, the effective concentration ratio was known to affect the appearance of EPCIA starting profiles. For instance, EPCIA gave better estimates of the peak shape for the minor component of a two-component peak. The rationalization was that the results for the minor component are closer to what would be produced by adaptive filtering in these cases. This argument also seems consistent with the effect of increased spectral correlation, which decreased the magnitude of the innovations and made them better starting profiles. The results in Table 5.4 show that while the starting profile of the minor component changed slightly with the changes in concentration ratio, the concentration profile predicted by ITT was relatively unaffected. In all cases, the EPCIA-generated starting profiles gave better results than those obtained from a varimax rotation.

In this study, the EPCIA profiles were found to be less suitable for the major component, particularly for systems with an ECR below 0.1. As the ECR decreased, the following changes were observed in the EPCIA starting profiles:

(1) the width of the major peak increased, (2) the predicted retention time moved towards the center of the peak cluster, (3) the width of the minor peak decreased. In some ways, these changes are consistent with the data as the major component does dominate the chromatographic peak for a wider region when the contribution of the minor component is decreased. Unfortunately, these changes in the major peak make it a poor starting profile for ITT. Note that the ITT does not improve the major peak significantly (Table 5.4). As seen in Figure 5.6, the major peak converges to a combination of the real concentration profiles. The constraints do not prevent this, because the composite peak still has positive concentration profiles. Thus the EPCIA starting profiles are unsuitable for a major component overlapped with a minor component.

**Table 5.4** *The effect of concentration ratio on elution profiles predicted by ITT of EPCIA starting profiles. Sum of the square residuals is given for each peak. Bold type indicates that EPCIA outperforms the varimax method.*

<i>Effective Conc. Ratio</i>	<u>Major Component</u>		<u>Minor Component</u>	
	<i>Starting Profile</i>	<i>Transformed Profile</i>	<i>Starting Profile</i>	<i>Transformed Profile</i>
1.0	<b>0.088</b>	<b>0.058</b>	<b>0.096</b>	<b>0.060</b>
0.5	<b>0.017</b>	<b>0.015</b>	<b>0.133</b>	<b>0.068</b>
0.2	<b>0.128</b>	0.108	<b>0.153</b>	<b>0.071</b>
0.1	<b>0.224</b>	0.206	<b>0.159</b>	<b>0.072</b>
0.05	<b>0.294</b>	0.282	<b>0.161</b>	<b>0.072</b>
0.02	0.345	0.339	<b>0.162</b>	<b>0.073</b>
0.01	0.370	0.364	<b>0.170</b>	<b>0.073</b>

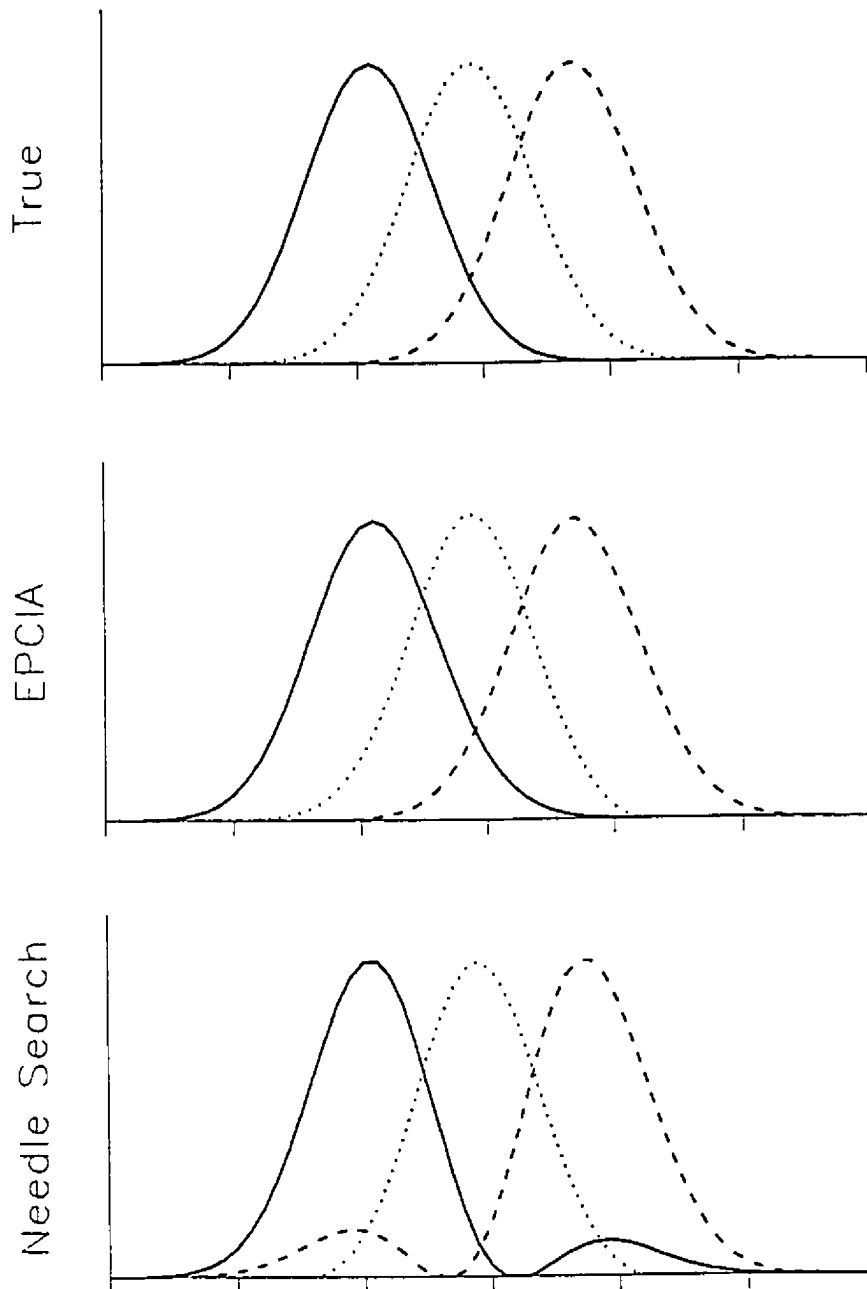


**Figure 5.6.** The concentration profile predicted by ITT for the major component (solid line) compared to the true profiles (dashed lines). See text for details.

### 5.4.5 Three-Component Systems

To demonstrate that the EPCIA approach can be applied to more complicated systems, an example of a three-component mixture is shown in Figure 5.7. Both the one- and two- component models would be expected to show regions of model failure in this case. Consider the Kalman filter applied in the forward direction. The one-component model would show an increase in its innovations upon encountering the second chemical component. These innovations would remain above the baseline until after the third component elutes, since the spectra in this region have contributions from more than one component. The two-component model would indicate the region in which the third component elutes. Therefore, the two-component innovations can be used as a starting profile for the last chemical component. Similarly, the reverse two-component model can be used for the first component. What remains to be achieved is a reasonable starting profile for the middle component. Although neither of the one-component models accurately describe its concentration profile, the leading edge of this peak is identified by the forward model and the trailing edge by the reverse model. Thus a reasonable starting profile can be obtained by splicing these two innovations sequences together. This synthesis of the third component is similar to the approach used in evolving factor analysis, where the forward and reverse sequences of eigenvalues are joined to identify components in the middle of a set.

This method was applied to a three-component peak generated with phenanthrene, fluoranthene and triphenylene with  $R_s = 0.4$  (between adjacent peaks) and ECR = 1:1:1. The result from this study were very interesting, in that



**Figure 5.7.** The concentration profiles predicted by ITT compared to the true profiles. Results from EPCIA and needle search starting profiles are shown.



the EPCIA, varimax, and needle search methods gave different results. The results from the ITT are shown on Figure 5.7 for needle search and EPCIA starting profiles. Starting profiles from the needle search converged to bimodal distributions which were combinations of the true concentration profiles. Results from the ITT of varimax starting profiles were also bimodal, but not to as great an extent. This problem has been noted in other work. Additional constraints are usually applied during the refinement of ITT profiles to eliminate these minor peaks. Vandeginste<sup>185</sup> removed any peak that was separated from the major peak by a region of zeros. Gemperline<sup>184</sup> applied linear constraints, that forced the concentration profile to decrease on either side of the maximum. These additional constraints were not needed with the EPCIA starting profiles used in this example to obtain good estimates of the concentration profiles. These results suggest that prior knowledge of peak shapes may become important as the complexity of the problem increases, however.

## 5.5 CONCLUSIONS

The advantage of self-modeling curve resolution over traditional methods of mixture analysis is that it can be applied without a knowledge of the pure component spectra. In this work, the iterative target transform (ITT) was applied to simulated data of overlapped chromatographic peaks with UV-visible detection. The performance of ITT was compared for three different methods of generating starting profiles: EPCIA, varimax rotation and needle search. The innovation traces from EPCIA were found to be good starting profiles since they accurately estimated peak widths and contain only positive concentrations. For two-component mixtures of equal concentration, the concentration profiles

predicted by ITT of EPCIA profiles were more accurate than the other methods, particularly for cases of low chromatographic resolution and high spectral overlap. The major limitation of all three types of starting profiles was that they overestimated the chromatographic separation between the two peaks. The ITT could not correct for this at resolutions of less than 0.4, and thus, the accuracy of the predicted concentration profiles suffered. When the concentrations of the two components were dissimilar, the EPCIA profiles used to start ITT gave poor results for the major component. If this problem can be overcome, EPCIA-based approaches for mixture analysis might have significant advantages over existing methods, particularly for more complex systems.

## CONCLUSIONS AND FUTURE WORK

---

### 6.1 CONCLUSIONS

To deal with complex mixtures, analytical chemists often use multidimensional methods, in which two instruments that provide complementary information about a sample are linked in series. The advantages of multidimensional methods include their efficiency, precision and accuracy. This work has demonstrated how digital filters, particularly the Kalman filter, enhance these properties of multidimensional methods.

In Chapter 2 the concept of parallel Kalman filter networks was introduced. These networks take advantage of the diagnostic properties of the Kalman filter, namely its ability to detect the extent and nature of modeling errors in real-time. The network can be implemented in a quasi-continuous or discrete form. The quasi-continuous form was used in kinetics determinations to compensate for variations in the pseudo-first-order rate constant. The Kalman filter models predicted the changes in analyte absorbance based on a range of rate constants. The performance of each model was evaluated by examining the innovations, the differences between the measured value of a signal and the value predicted by the Kalman filter. The best model was chosen as the one with the smallest sum of squared innovations. Experimental data from the molybdenum blue method for the determination of phosphate was used to demonstrate the insensitivity of this algorithm to variations in the rate constant. The results were found to be more accurate than those from a fixed-time method, furthermore, the estimated concentrations were more precise because the

Kalman filter uses all the experimental points. The efficiency of this algorithm made it suitable for real-time applications on a laboratory computer.

In Chapter 3, the discrete form of the network was used to perform recursive principal components analysis. This network contained models for describing one- and two-component bilinear responses, that were used to elucidate the rank of data from chromatography with multisensor detection. The performance of each model was evaluated by plotting the rms innovations as a function of time. For pure chromatographic peaks the innovations gave a flat baseline, that reflected the level of random noise. The elution of additional chemical components produced local regions of model failure, indicated by innovations that exceeded the baseline value. This algorithm for peak purity analysis in real-time was called evolving principal components analysis (EPCIA).

In Chapter 4, the fundamental and experimental limitations of EPCIA were examined for unresolved mixtures in liquid chromatography with UV-visible detection. It was shown that the innovation sequence from the EPCIA model approximates the elution profile of the minor component, and that the maximum of this sequence determines the detectability of the impurity. This maximum decreases with increasing spectral correlation and decreasing chromatographic resolution. When these two variables are fixed, the maximum innovation is proportional to the concentration of the minor component over a wide range. Experimentally, the validity of the EPCIA model may be limited by a variety of effects including heteroscedastic noise, spectral scan time, baseline perturbation, and nonlinear instrumental response. At high absorbance values (>1.0 AU) the one-component innovations were inflated by the nonlinear response of the diode-array spectrometer, attributed to the effects of stray light and polychromatic radiation.

In Chapter 5, self-modeling curve resolution was used to extract pure component elution profiles and spectra from overlapped chromatographic peaks. In particular, the iterative target transform was applied to concentrations profiles generated by EPCIA. These starting profiles were found to give better results than established methods for two-component mixtures of roughly equal concentrations. The disadvantage to the EPCIA profiles was the poor results obtained for the major peak when the two components had dissimilar contributions to the observed spectral region.

## 6.2 FUTURE WORK

The work presented here has demonstrated the potential of Kalman filter networks and laid the groundwork for future applications. While the application of quasi-continuous methods was demonstrated for reaction-rate methods with first-order kinetics, many other possibilities can be envisioned. For example, the measurement model could be expanded to include kinetic interferences, such as those due to silicate in the determination of phosphate. The Kalman filter models could also include absorbances at multiple wavelengths, thus allowing spectral as well as kinetic differences to be observed. In addition, there are many other types of experimental results that would benefit from the application of these Kalman filter networks. In general, these are cases where the Kalman filter alone is unsuitable, due to nonlinear models or the presence of systematic errors that cause global failures of a linear model. Possible applications include real-time correction of wavelength shifts due to instrumental drift in emission spectroscopy and environment-dependent changes in fluorophores.

EPCIA was demonstrated as an application of discrete networks. It was shown to be a powerful technique for examining spectrochromatograms. The

ability of this algorithm to detect minor impurities could be improved, particularly for samples with large absorbances, with models that compensate for the nonlinear response of the spectrometer. More sophisticated models for background correction should also be pursued. In this work only one- and two-component models were explored, but it would be useful to extend the algorithm to include more than two-component models. This would allow EPCIA to be applied to more complex systems in chromatography and to results from spectrophotometric titrations.

In this work the results from EPCIA were usually shown as an rms innovation trace, plotted as a function of elution time. The disadvantage to using the rms values is that the direction and magnitude of innovations from individual models are lost. This information may be useful in distinguishing among random errors, systematic errors, and real chemical information. For example, the 'raw' innovations could indicate the wavelengths where different models fail. There are other properties of the Kalman filter that should be considered in future work, including its ability to include prior knowledge in the state vector. This may prove useful for chemical systems where there is a partial knowledge of the members of a mixture. Another possibility is the use of nondeterministic models that consider errors caused by drifting systems. These improvements to both forms of Kalman filter networks should allow them to extract information from a wide range of existing and future experimental techniques.

# APPENDIX A

## PROGRAM LISTING FOR HP-PCA.BAS

---

```

' HP-PCA.BAS
' Version 4.0
' Written by Stephen Vanslyke
' Dalhousie University
' Sept./90
'
' HP-PCA loads data files that are created by the operating software of the
' Hewlett Packard diode array. The absorbances at selected wavelengths are
' stored into an array which is then passed to the Kalman filter routine for
' principal component analysis. The results should indicate the presence of
' coeluting peaks.
'
' Revision 2.0 the Kalman filtering subroutine BigKal3 is used.
'           The major difference being the use of orthogonal innovations
'
' Revision 3.0 the filter is run through the data in both directions.
'
' Revision 4.0 modified to load the variance data from the *.TIM file.
'           (note that the variance is not used by the KF routine)
'=====
'
'*** Declarations and such ***
DECLARE SUB Dispinn (Innov!(), SpecData#(), NPnts%, NWaves%)
DECLARE SUB Display (Array2D#(), Index%, FirstPnt%, LastPnt%, Min, Max!)
DECLARE SUB HGen (IWX%, IWY%, IPT%, AB#(), H#(), NWaves%)
DECLARE SUB Kalman (Profile#(), NWaves%, NPnts%, Thresh!, StateP!(), Innov!())
DECLARE SUB KeyPress (K%, K$)
DECLARE SUB Label8 (X!, Label$)
DECLARE SUB ListDir (Wild$)
DECLARE SUB LoadData (SpecData#(), VarData(), NPnts%, NWaves%, Min, Max, MaxVar, SpecLabel$(),
MaxWaves%)
DECLARE SUB SaveFile (Array!(), NRows%, NCol%, Trans%)
DECLARE SUB Wind (SpecData#(), NWaves%, NPnts%, Thresh)
DECLARE SUB VGADisplay (Array2D!(), Index%, FirstPnt%, LastPnt%, Max!, Col%)

CONST Black% = 0, Blue% = 1, Green% = 2, Aqua% = 3, Red% = 4, Purple% = 5
CONST Orange% = 6, White% = 7, Grey% = 8, BBlue = 9, BGreen% = 10
CONST BAqua% = 11, BOrange% = 12, Violet% = 13, Yellow% = 14, BWhite% = 15
CONST Path$ = "\\SJV\SIM\"

'*** Dynamic Arrays ***
MaxPnts% = 100
MaxWaves% = 51 ' 26 corresponds to 4nm resolution over 100nm
DIM SpecData#(MaxWaves% + 1, MaxPnts%) 'One extra to store the maximum
DIM VarData(MaxWaves% + 1, MaxPnts%) 'One extra to store the maximum
DIM StateP(2 * MaxWaves% - 3, 3)
DIM Innov(5, MaxPnts%)
DIM SpecLabel$(MaxWaves%)

'*** Defaults for Variables ***
NWaves% = 26
CONST StdDev = .0005
Thresh = 6 * StdDev

```

```

***** Main program starts here *****

SCREEN 0
CLS
DO
' *** Load the data into the array SpecData#()***

    CALL LoadData(SpecData#(), VarData(), NPnts%, NWaves%, Min, Max, MaxVar, SpecLabel$(), MaxWaves%)

' *** Display the data ***
    DO
        CLS
        PRINT " Column# Wave Length"
        FOR I% = 1 TO NWaves%
            PRINT " "; I%;
            PRINT " "; SpecLabel$(I%)
        NEXT I%
        PRINT " "; NWaves% + 1;
        PRINT " "; "MaxPlot"
        PRINT " "; NWaves% + 2;
        PRINT " "; "VarPlot"

        DO
            INPUT "Chromatogram"; Chrom%
            LOOP UNTIL (Chrom% >= 0) AND (Chrom% <= NWaves% + 2)
            IF Chrom% <> 0 THEN
                SCREEN 12
                IF Chrom% <= NWaves% THEN
                    CALL Display(SpecData#(), Chrom%, 1, NPnts%, Min, Max)
                    TempMax = 0
                    FOR I% = 1 TO NPnts%
                        IF VarData(Chrom%, I%) > TempMax THEN TempMax = VarData(Chrom%, I%)
                    NEXT I%
                    CALL VGADisplay(VarData(), Chrom%, 1, NPnts%, TempMax, Purple%)
                ELSEIF Chrom% = (NWaves% + 1) THEN
                    CALL Display(SpecData#(), Chrom%, 1, NPnts%, Min, Max)
                ELSEIF Chrom% = NWaves% + 2 THEN
                    CALL Display(SpecData#(), NWaves% + 1, 1, NPnts%, Min, Max)
                    CALL VGADisplay(VarData(), NWaves% + 1, 1, NPnts%, MaxVar, Purple%)
                END IF
                CALL KeyPress(K%, K$)
                SCREEN 0
            END IF
        LOOP UNTIL Chrom% = 0

' ** Window the data ***
        CALL Wind(SpecData#(), NWaves%, NPnts%, Thresh)

' *** Filter the Data ***
        CALL Kalman(SpecData#(), NWaves%, NPnts%, Thresh, StateP(), Innov())
        CALL DispInn(Innov(), SpecData#(), NPnts%, NWaves%)
        CALL SaveFile(Innov(), 5, NPnts%, 1)
        VIEW
        WINDOW
        SCREEN 0
        'LOCATE 3, 15
        'PRINT "Press 'Q' to quit"
        'CALL KeyPress(K%, K$)
        LOOP UNTIL K$ = "Q"

```



END

```

=====
'This section traps errors during file loading
ErrorTrap:
  PRINT ERR
  FileError% = 1
  RESUME NEXT
=====

SUB DispInn (Innov!(), SpecData#(), NPnts%, NWaves%)
'
'Subroutine to display the innovations traces
'
' *** Find the maxima ***
' The array Max() stores the following maxima
' Max(1) = forward innovation for one component
' Max(2) = forward innovation for two component
' Max(3) = reverse innovation for one component
' Max(4) = revrse innovation for two component
' Max(5) = absorbance

  DIM Max(5)
  FOR I% = 1 TO NPnts%
    IF Innov!(1, I%) > Max(1) THEN Max(1) = Innov!(1, I%)
    IF Innov!(2, I%) > Max(2) THEN Max(2) = Innov!(2, I%)
    IF Innov!(3, I%) > Max(3) THEN Max(3) = Innov!(3, I%)
    IF Innov!(4, I%) > Max(4) THEN Max(4) = Innov!(4, I%)
    Innov!(5, I%) = SpecData#(NWaves% + 1, I%)
    IF Innov!(5, I%) > Max(5) THEN Max(5) = Innov(5, I%)
  NEXT I%

'*** Display the Innovations and the Max Absorbance ***
  VIEW PRINT
  SCREEN 12
  CLS
  LOCATE 2, 10

' ** Results from the forward filter **
  COLOR BBlue%
  PRINT "One Component Model"
  COLOR BWhite%
  CALL VGADisplay(Innov!(), 1, 1, NPnts%, Max(1) * 2, BBlue%)

  LOCATE 2, 35
  COLOR Green%
  PRINT "Two Component Model"
  COLOR BWhite%
  CALL VGADisplay(Innov!(), 2, 1, NPnts%, Max(1) * 2, Green%)
  CALL Label8(Max(1), Label$)
  LOCATE 16, 1
  PRINT Label$;

  LOCATE 2, 59
  PRINT "Maximum Absorbance"
  CALL VGADisplay(Innov!(), 5, 1, NPnts%, Max(5), BWhite%)

```

```

CALL KeyPress(K%, K$)
CLS

' ** Results from the reverse filter **
LOCATE 2, 10
COLOR BBlue%
PRINT "One Component Model"
COLOR BWhite%
CALL VGADisplay(Innovl(), 3, 1, NPnts%, Max(3) * 2, BBlue%)

LOCATE 2, 35
COLOR Green%
PRINT "Two Component Model"
COLOR BWhite%
CALL VGADisplay(Innovl(), 4, 1, NPnts%, Max(3) * 2, Green%)
CALL Label8(Max(3), Label$)
LOCATE 16, 1
PRINT Label$;

LOCATE 2, 59
PRINT "Maximum Absorbance"
CALL VGADisplay(Innovl(), 5, 1, NPnts%, Max(5), BWhite%)

LOCATE 4
FOR I% = 1 TO 5
  LOCATE , 65
  PRINT Max(I%)
NEXT I%

END SUB

SUB Display (Array2D#(), Index%, FirstPnt%, LastPnt%, Min, Max) STATIC

' ** The subroutine Display3 uses screen mode 12 to display the collected
' data in the array Array2D#() as an X-Y graph. The X-axis is scaled to
' the number of data points and the Y-axis displays the value
' of that array element.
'
' Array2D#(i,j) = the array containing the data to be plotted.
' Index%       = the row of the array to be plotted
' FirstPnt%, LastPnt% = first and last j values to be plotted
' Min,Max     = Values to scale the screen to
'
' ** Label axis **
CALL Label8(Max, Label$)
LOCATE 4, 1
PRINT Label$;
CALL Label8(Min, Label$)
LOCATE 28, 1
PRINT Label$;

' ** Draw a frame around the graph **
LINE (69, 40)-(626, 445), Aqua%, B
LINE (71, 42)-(624, 443), Aqua%, B

' ** Set up a view port inside this box **

```

```

IF LastPnt% < 2 THEN EXIT SUB
VIEW (75, 48)-(620, 437)

' ** Scale this graphics window **
WINDOW (FirstPnt%, Min)-(LastPnt%, Max)
LINE (FirstPnt%, 0)-(LastPnt%, 0), White%
' ** Plot the data in Array2D#( ) **
LINE (FirstPnt%, Array2D#(Index%, FirstPnt%))-(FirstPnt% + 1, Array2D#(Index%, FirstPnt% + 1)),
Yellow%
FOR I% = (FirstPnt% + 2) TO LastPnt%
    LINE -(I%, Array2D#(Index%, I%)), Yellow%
NEXT I%

' ** Reset to normal screen **
WINDOW
VIEW

END SUB

SUB HGen (IWX%, IWY%, IPT%, AB#( ), H#( ), NWaves%)
'
'Subroutine to generate observation matrices.
'
'
'**** WARNING ****
' When the number of nonzero elements is changed, the values of
' the NCOEFFS% matrix (in the Kalman subroutine) should be modified.
'*****

' ** One-component model **
FOR I% = 1 TO NWaves% - 1
    H#(I%, 1) = AB#(IWX%, IPT%)
    H#(I%, 2) = 1
    H#(I%, 3) = 0
NEXT I%

' ** Two-component model **
FOR I% = NWaves% TO (2 * NWaves% - 3)
    H#(I%, 1) = AB#(IWX%, IPT%)
    H#(I%, 2) = AB#(IWY%, IPT%)
    H#(I%, 3) = 1
NEXT I%
'
END SUB

DEFDBL A-Z
SUB Kalman (Profile(), NWaves%, NPnts%, Thresh!, StateP!(), Innov!())
'
'Subroutine to perform parallel Kalman filtering of multiwavelength
'elution profiles (recursive PCA).
'
'
' Profile!() - the elution absorbance matrix, dimensioned at least
' (NWaves%,NPnts%), which contains absorbances at each wavelength
' at each point in time. Maximum number of wavelengths which
' can currently be handled by this subroutine is 20.
' Thresh! - Threshold for turning on the filter
' StateP() - is an array which returns the state parameter estimates.
' Innov!() - array containing the rms of the innovations at each point
'
'
' XFor%,XBac% - indices of the wavelengths used for one-component models.
' YFor%,YBac% - indices of the wavelengths used for two-component models.

```

```

'
'   IStart%   - point number where the KF starts
'
' Shared variables
'   NWaves%   - is the number of wavelengths recorded (should be <=20).
'   NPnts%    - is the number of spectral readings made (should be <=300).
'=====
'
' Initialization.
'
'   NModel% = 2 * NWaves% - 3
'   DIM AB(1 TO NWaves%, 1 TO NPnts%) 'Double precision array for the absorb.
'   DIM P(NModel%, 3, 3)
'   DIM H(NModel%, 3), X(NModel%, 3)
'   DIM Gain(3)
'   DIM NCOEFFS%(NModel%)
'   DIM TMP1(3), TMP2(3, 3), TMP3(3, 3)
'
'   Var = CDBL(StdDev * StdDev)      'Variance
'   PINIT = Var * 1D+59              'Approximation to infinity

*** Initialize matrices **
REDIM Innov!(5, NPnts%)
FOR Dir% = 1 TO -1 STEP -2   *** Filters the data in both directions
  CLS
  LOCATE 18, 10
  IF Dir% = 1 THEN
    PRINT "Forward Filter"
    FOR I% = 1 TO NPnts%
      FOR J% = 1 TO NWaves%
        AB(J%, I%) = (Profile(J%, I%))
      NEXT J%
    NEXT I%
  ELSE
    PRINT "Reverse Filter"
    FOR I% = 1 TO NPnts%
      FOR J% = 1 TO NWaves%
        AB(J%, I%) = (Profile(J%, NPnts% - I% + 1))
      NEXT J%
    NEXT I%
  END IF

  FOR I% = 1 TO NModel%
    IF I% <= NWaves% - 1 THEN
      NCOEFFS%(I%) = 1      ' <=====
    ELSE
      NCOEFFS%(I%) = 2      ' <=====
    END IF
  NEXT I%

  FOR I% = 1 TO NModel%
    FOR J% = 1 TO NCOEFFS%(I%)
      FOR K% = 1 TO NCOEFFS%(I%)
        P(I%, J%, K%) = 0
      NEXT K%
      P(I%, J%, J%) = PINIT
      X(I%, J%) = 0
    NEXT J%
  NEXT I%
'
' Wait for a peak to be detected, then take two largest signals for model.
' IStart% = 0

```

```

DO
  IStart% = IStart% + 1
  IWx% = 1           'Assume these are the greatest absorbances
  IWy% = 2
  IF AB(1, IStart%) < AB(2, IStart%) THEN SWAP IWx%, IWy%
  FOR I% = 3 TO NWaves%
    IF AB(I%, IStart%) > AB(IWx%, IStart%) THEN
      IWy% = IWx%
      IWx% = I%
    ELSEIF AB(I%, IStart%) > AB(IWy%, IStart%) THEN
      IWy% = I%
    END IF
  NEXT I%
LOOP UNTIL (AB(IWx%, IStart%) > Thresh!) OR (IStart% = NPnts%)

IF IStart% = NPnts% THEN
  PRINT "          ERROR -- No peak detected."
  BEEP
  CALL KeyPress(K%, K$)
  EXIT SUB
ELSE
  PRINT "          Peak detected at point"; IStart%
  IF (Dir% = 1) THEN
    IF (XFor% <> 0) THEN
      PRINT "          Using input from menu":
      IWx% = XFor%
    ELSE
      XFor% = IWx%
      PRINT "          Using the maximum absorbances:"
    END IF
    YFor% = IWy%
  ELSEIF (Dir% = -1) THEN
    IF (XBac% <> 0) THEN
      PRINT "          Using input from menu":
      IWx% = XBac%
    ELSE
      XBac% = IWx%
      PRINT "          Using the maximum absorbances:"
    END IF
    YBac% = IWy%
  END IF
END IF

*** Temp ***
  IWx% = 15
  IWy% = 10
  *****

  PRINT "          Wavelength X ="; IWx%; " Wavelength Y = "; IWy%
END IF
:
*** Loop to filter each point for each model **
PRINT "          Kalman Filtering Point:"
InCnt% = 1
FOR I% = IStart% TO NPnts%
  LOCATE 22, 32
  PRINT I%
  CALL HGen(IWx%, IWy%, I%, AB(), R(), NWaves%)
  RSum1! = 0
  RSum2! = 0
  FOR M% = 1 TO NModel%
    *** Calculate Kalman gain **

```

```

IF M% < NWaves% THEN          'One component model
  y = AB(M%, I%)
  IF M% >= IW% THEN y = AB(M% + 1, I%)
ELSE                            'Two component model
  IDX% = M% - NWaves% + 1
  IF IW% > IWY% THEN
    IF IDX% >= IWY% THEN IDX% = IDX% + 1
    IF IDX% >= IW% THEN IDX% = IDX% + 1
  ELSE
    IF IDX% >= IW% THEN IDX% = IDX% + 1
    IF IDX% >= IWY% THEN IDX% = IDX% + 1
  END IF
  y = AB(IDX%, I%)
END IF

Temp = 0
FOR J% = 1 TO NCOEFFS%(M%)
  TMP1(J%) = 0
  FOR K% = 1 TO NCOEFFS%(M%)
    TMP1(J%) = TMP1(J%) + P(M%, J%, K%) * H(M%, K%)
  NEXT K%
  Gain(J%) = TMP1(J%)
  Temp = Temp + H(M%, J%) * TMP1(J%)
NEXT J%
Temp = Temp + Var
FOR J% = 1 TO NCOEFFS%(M%)
  Gain(J%) = Gain(J%) / Temp
NEXT J%

,
'Got Kalman gain, now update estimate of the state vector.
YCalc = 0
FOR J% = 1 TO NCOEFFS%(M%)
  YCalc = YCalc + X(M%, J%) * H(M%, J%)
NEXT J%
Diff = y - YCalc
'Calculate the orthogonal residuals.

P1 = AB(IW%, I%)
P2 = AB(IWY%, I%)
IF M% < NWaves% THEN
  ODiff = -X(M%, 1) * P1 + y - X(M%, 2)
  ODiff = ODiff / SQRT(X(M%, 1) ^ 2 + 1)
ELSE
  P3 = AB(IDX%, I%)
  top = -X(M%, 1) * P1 - X(M%, 2) * P2 + P3 - X(M%, 3)
  ODiff = top / SQRT(X(M%, 1) ^ 2 + X(M%, 2) ^ 2 + 1)
END IF

FOR J% = 1 TO NCOEFFS%(M%)
  X(M%, J%) = X(M%, J%) + Gain(J%) * Diff
NEXT J%

Inn1 = CSNG(ABS(ODiff))
IF M% < NWaves% THEN
  RSum1! = RSum1! + Inn1 ^ 2
ELSE
  RSum2! = RSum2! + Inn1 ^ 2
END :F

,
'Now update the covariance matrix.
FOR J% = 1 TO NCOEFFS%(M%)

```

```

    FOR K% = 1 TO NCOEFFS%(M%)
        TMP2(J%, K%) = -Gain(J%) * H(M%, K%)
    NEXT K%
    TMP2(J%, J%) = 1 + TMP2(J%, J%)
NEXT J%

FOR J% = 1 TO NCOEFFS%(M%)
    FOR K% = 1 TO NCOEFFS%(M%)
        TMP3(J%, K%) = 0
        FOR LX = 1 TO NCOEFFS%(M%)
            TMP3(J%, K%) = TMP3(J%, K%) + TMP2(J%, LX) * P(M%, LX, K%)
        NEXT LX
    NEXT K%
NEXT J%

FOR J% = 1 TO NCOEFFS%(M%)
    FOR K% = 1 TO NCOEFFS%(M%)
        P(M%, J%, K%) = 0
        FOR LX = 1 TO NCOEFFS%(M%)
            P(M%, J%, K%) = P(M%, J%, K%) + TMP3(J%, LX) * TMP2(K%, LX)
        NEXT LX
        P(M%, J%, K%) = P(M%, J%, K%) + Gain(J%) * Var * Gain(K%)
    NEXT K%
NEXT J%
NEXT M%

**** Store all but the first innovation ***
IF (Dir% = -1) AND (InCnt% > 1) THEN
    Innov!(3, NPnts% - I% + 1) = SQR(RSum1! / (NModel% - 1))
    Innov!(4, NPnts% - I% + 1) = SQR(RSum2! / (NModel% - 2))
ELSEIF (Dir% = 1) AND (InCnt% > 1) THEN
    Innov!(1, I%) = SQR(RSum1! / (NModel% - 1))
    Innov!(2, I%) = SQR(RSum2! / (NModel% - 2))
END IF

    InCnt% = InCnt% + 1
NEXT I%
NEXT Dir%

*** Transfer the state parameters into their array **
FOR I% = 1 TO NModel%
    FOR J% = 1 TO NCOEFFS%(I%)
        StateP!(I%, J%) = CSNG(X(I%, J%))
    NEXT J%
NEXT I%

END SUB

DEFSNG A-Z
SUB KeyPress (K%, K$)

' stores the value of a non-null inkey$ into the variable >K$<
' stores its ascii value into the variable >K%<
'
' if inkey$ returns a two-byte string (extended ascii), then KeyPress strips
' the first byte and sets >K$< equal to the 2nd byte. >K%< is then set to the
' ascii value of >K$< + 300.

' *** First clear the input buffer ***

```

```

DO
    K$ = INKEY$
    LOOP UNTIL K$ = ""

' *** Then get the next keypress ***
DO
    K$ = INKEY$
    LOOP WHILE K$ = ""

' *** This conditional deals with the extended characters ***
IF LEN(K$) = 1 THEN
    K$ = UCASE$(K$)
    K% = ASC(K$)
ELSE
    K$ = RIGHT$(K$, 1)
    K% = ASC(K$) + 300
END IF
'
END SUB

SUB Label8 (X, Label$) STATIC
'
'This subroutine converts the real number X into a character string
'not greater than eight characters long and returns it in LABEL$. It
'is intended to fix the length of tick mark labels in graphics applications
'to a maximum of 8 characters.
'
'First convert to string and strip off excess spaces.
    ORIG$ = STR$(X)
    STRIP$ = ""
    FOR I% = 1 TO LEN(ORIG$)
        Temp$ = MID$(ORIG$, I%, 1)
        IF Temp$ <> " " THEN STRIP$ = STRIP$ + Temp$
    NEXT I%
    IF LEN(STRIP$) < 9 THEN
        Label$ = SPACE$(8 - LEN(STRIP$)) + STRIP$
        EXIT SUB
    END IF
'That didn't do the trick. Next look for exponential notation. If
'this is the case, we can take the last four characters (exponent)
'and the first four characters (mantissa). Round if necessary.
    EPOS = INSTR(STRIP$, "E")
    IF EPOS <> 0 THEN
        CHAR1$ = MID$(STRIP$, 4, 1) 'Last digit in mantissa
        CHAR2$ = MID$(STRIP$, 5, 1) 'Truncated digit in mantissa
        IF CHAR2$ > "4" THEN CHAR1$ = CHR$(ASC(CHAR1$) + 1) 'Round it
        Label$ = LEFT$(STRIP$, 3) + CHAR1$ + RIGHT$(STRIP$, 4)
        EXIT SUB
    END IF
'Not exponential notation. The only exception I have found to this
'is of the form "-1.234567". This is assumed for anything which has made
'it this far and the leftmost 8 characters will be taken, with rounding.
    CHAR1$ = MID$(STRIP$, 8, 1) 'Last retained character
    CHAR2$ = MID$(STRIP$, 9, 1) 'First truncated character
    IF CHAR2$ > "4" THEN CHAR1$ = CHR$(ASC(CHAR1$) + 1) 'Round it
    Label$ = LEFT$(STRIP$, 7) + CHAR1$
    END SUB

SUB ListDir (Wild$)

Cmd$ = "DIR/W " + Wild$ + ">DIRFILE"

```



```

LOCATE 3, 15
PRINT "Directory "; Wild$
PRINT
SHELL Cmd$
OPEN "DIRFILE" FOR INPUT AS #1
  FOR I% = 1 TO 4
    LINE INPUT #1, Line$
  NEXT I%
  DO
    LINE INPUT #1, Line$
    PRINT Line$
  LOOP UNTIL EOF(1)
CLOSE #1
KILL "DIRFILE"

END SUB

SUB LoadData (SpecData#(), VarData(), NPnts%, NWaves%, Min, Max, MaxVar, SpecLabel$(), MaxWaves%)
'
' The subroutine LoadData is used to load in the information stored in the
' file "FileName.TIM", a file created by the Hewlett Packard diode array
' software. The structure of this file is documented in appendix B of the
' UV/VIS Software handbook. These files consist of a header section and a
' data section. This subroutine checks the header section to see if the
' acquisition mode of the selected file is 5, wavelength range spectra.
' It also checks that the std.dev = 0 (data stored without variance
' estimates). Files that do not meet these requirements will not be loaded.
'
' The std.dev feature of the diode array could be added in a latter version
' of this program to provide a noise estimate for the Kalman filter.
'
' *** Open the Data File ***
  SHARED FileError%      'A shared variable needed for the error trap

  DO
    ErrorFlag% = 0
    DO
      FileError% = 0          'Clear the error flag
      ON ERROR GOTO ErrorTrap 'Turn on the error trap
      CLS
      LOCATE 3, 15
      CALL ListDir(Path$ + "*.TIM")
      PRINT
      INPUT "File to be loaded (no extension)"; FileName$
      IF UCASE$(FileName$) = "Q" THEN END
      FileName$ = Path$ + UCASE$(FileName$) + ".TIM"      'add the path & ext.
      OPEN FileName$ FOR INPUT AS #1      'Open the file
      CLS
      LOCATE 3, 15
      IF FileError% = 1 THEN      'Has the error trap been set?
        PRINT " - Unable to find "; FileName$;
        SOUND 300, 4
        CALL KeyPress(K%, X%)
      END IF
    LOOP WHILE FileError%

    ON ERROR GOTO 0      'Turn off error trap

    CLS
    LOCATE 3, 15

```

```

PRINT FileName$;

' *** Get out the data for the Kalman filter ***
INPUT #1, Junk, Junk$, TimeData$, DateData$
INPUT #1, Junk, IntTime, StdDevOn%
INPUT #1, Junk
LINE INPUT #1, Junk$
LINE INPUT #1, Junk$
LINE INPUT #1, Junk$
INPUT #1, WaveMode%, Junk, NWavesStored%, Junk
INPUT #1, StartWave%, EndWave%, Junk, Junk, Junk, Junk
INPUT #1, AcqMode%, RunTime, CycleTime, StartTime

' Check the file's storage mode
IF StdDevOn% <> 0 THEN
  'ErrorFlag% = 1
  'CLOSE #1
  PRINT " - Standard deviation was set"
END IF
IF AcqMode% <> 5 THEN
  ErrorFlag% = 1
  CLOSE #1
  PRINT " - ERROR - wrong acquisition mode"
  SOUND 400, 3
  CALL KeyPress(K%, K$)
END IF
LOOP WHILE ErrorFlag%
NPnts% = CINT((RunTime - StartTime) / CycleTime) + 1

PRINT " "; TimeData$, DateData$
PRINT TAB(14); NWavesStored%; "Wavelengths from"; StartWave%; "to"; EndWave%
PRINT TAB(15); "Cycle Time = "; CycleTime; "; Integration Time; "; IntTime; ""
PRINT TAB(15); "Run time ="; RunTime; ", ", Start; "Time = "; StartTime
PRINT TAB(15); "Number of Readings ="; NPnts%
PRINT

'*** Now the file is open, get out the goodies ***
DO
  PRINT "          Number of wavelengths to use <" ; NWaves% ; ">";
  INPUT Temp%
  LOOP UNTIL (Temp% >= 3) AND (Temp% <= MaxWaves%) AND (Temp% <= NWavesStored%) OR Temp% = 0
  IF Temp% <> 0 THEN NWaves% = Temp%

' After the first wavelength has been read, there will be (NWavesStored% - 1)
' wavelengths left to choose from. We want to divide this range up evenly and
' select the remaining (NWaves% - 1) wavelengths from it. The distance between
' these wavelengths is StepSize.

  StepSize = CSNG(NWavesStored% - 1) / CSNG(NWaves% - 1)

'Fill in the nm wavelength values for the Waves

  FOR IX = 1 TO NWaves%
    SpecLabel$(IX) = STR$(StartWave% + CINT(CSNG(IX - 1) * StepSize * 2))
  NEXT IX

  LOCATE 12, 15
  PRINT "Loading point#"
  Min = 0
  Max = 0
  FOR IX = 1 TO NPnts%

```

```

LOCATE 12, 30           'Update display
PRINT I%
INPUT #1, Temp         'Input abs at the first wavelength
IF Temp > Max THEN Max = Temp
IF Temp < Min THEN Min = Temp
SpecData#(1, I%) = CDBL(Temp) 'Store it in the array
WavePointer% = 1
,
' At this point we have to do some careful bookkeeping to ensure we know
' where we are in the sequential file. WavePointer% stores the position of
' the last absorbance stored to the array. The number of absorbances to read
' before we get to the next one to store in the array is stored in Skip%.
FOR J% = 1 TO (NWaves% - 1)
  Skip% = 1 + INT(CSNG(J%) * StepSize) - WavePointer%
  WavePointer% = WavePointer% + Skip%
  FOR K% = 1 TO Skip%
    INPUT #1, Temp
  NEXT K%
  IF Temp > Max THEN Max = Temp
  IF Temp < Min THEN Min = Temp
  SpecData#(J% + 1, I%) = CDBL(Temp)
NEXT J%

' If the last absorbance read was not the last one on that line of the file,
' then we read the rest of the line.
IF WavePointer% < NWavesStored% THEN LINE INPUT #1, Junk$

' If the Standard Deviation were recorded they will be on the next line

IF StdDevOn% <> 0 THEN
  INPUT #1, VarData(1, I%) 'Input Var at the first wavelength
  WavePointer% = 1
  FOR J% = 1 TO (NWaves% - 1)
    Skip% = 1 + INT(CSNG(J%) * StepSize) - WavePointer%
    WavePointer% = WavePointer% + Skip%
    FOR K% = 1 TO Skip%
      INPUT #1, Temp
    NEXT K%
    VarData(J% + 1, I%) = Temp
  NEXT J%
  IF WavePointer% < NWavesStored% THEN LINE INPUT #1, Junk$

END IF
NEXT I%
CLOSE #1

' The last trick is to store the maximum absorbance of each sample
MaxVar = 0
FOR I% = 1 TO NPnts%
  RowMax = -10
  MeanVar = 0
  FOR J% = 1 TO NWaves%
    IF SpecData#(J%, I%) > RowMax THEN RowMax = SpecData#(J%, I%)
    MeanVar = MeanVar + VarData(J%, I%)
  NEXT J%
  SpecData#(NWaves% + 1, I%) = RowMax
  Temp = MeanVar / CSNG(NWaves%)
  VarData(NWaves% + 1, I%) = Temp
  IF MaxVar < Temp THEN MaxVar = Temp
NEXT I%

```

END SUB

SUB SaveFile (Array(), NRows%, NCol%, Trans%)

```

LOCATE 30, 25
PRINT "Output the data to a file (Y/N) ?";
CALL KeyPress(K%, K$)
IF K$ = "Y" THEN
  LOCATE 30, 25
  PRINT "                                     ";
  LOCATE 30, 25
  INPUT ; "File Name (No Extension)"; FileName$
  FileName$ = FileName$ + ".DAT"
  OPEN FileName$ FOR OUTPUT AS #1

  IF Trans% = 1 THEN
    FOR J% = 1 TO NCol%
      FOR I% = 1 TO NRows%
        PRINT #1, Array(I%, J%); ", ";
      NEXT I%
      PRINT #1,
    NEXT J%
  ELSE
    FOR J% = 1 TO NRows%
      FOR I% = 1 TO NCol%
        PRINT #1, Array(J%, I%); ", ";
      NEXT I%
      PRINT #1,
    NEXT J%
  END IF
  CLOSE #1
END IF
LOCATE 30, 25
PRINT "                                     ";

```

END SUB

SUB VGADisplay (Array2D(), Index%, FirstPnt%, LastPnt%, Max, Col%) STATIC

```

' ** The subroutine VGADISPLAY uses screen mode 12 to display the collected
' data in the array Array() as an X-Y graph. The X-axis is scaled to
' the number of data points and the Y-axis displays the value
' of that array element.

```

```

' ** Label axis **
IF LastPnt% - FirstPnt% < 1 THEN EXIT SUB
IF Max <= 0 THEN Max = 1

```

```

CALL Label8(Max, Label$)
LOCATE 4, 1
PRINT Label$
CALL Label8(0, Label$)
LOCATE 28, 1
PRINT Label$

```

```

' ** Draw a frame around the graph **
LINE (69, 40)-(626, 445), Aqua%, B
LINE (71, 42)-(624, 443), Aqua%, B

```

```

' ** Set up a view port inside this box **
VIEW (75, 48)-(620, 437)

```

```

' ** Scale this graphics window **
    WINDOW (FirstPnt%, -.1 * Max)-(LastPnt%, Max * 1.03)

' ** Plot the data in Array2D() **

    LINE (FirstPnt%, Array2D(Index%, FirstPnt%)-(FirstPnt% + 1, Array2D(Index%, FirstPnt% + 1)),
Col%
    FOR IX = (FirstPnt% + 2) TO LastPnt%
        LINE -(IX, Array2D(Index%, IX)), Col%
    NEXT IX

' ** Reset to normal screen **
    WINDOW
    VIEW

END SUB

SUB Wind (SpecData#(), NWaves%, NPnts%, Thresh)
    Start% = 1
    Stops% = NPnts%

    SCREEN 12
' ** Draw a frame around the graph **
    LINE (69, 40)-(626, 445), Aqua%, B
    LINE (71, 42)-(624, 443), Aqua%, B

' ** Set up a view port inside this box **
    VIEW (75, 48)-(620, 437)
    LOCATE 1, 10
    PRINT "Use the cursor/ tab keys to select a window of data. S to save it."
    Toggle% = 1
    DO
        CLS
        Min = 0
        Max = 0
        FOR IX = Start% TO Stops%
            Temp = SpecData#(NWaves% + 1, IX)
            IF Temp < Min THEN Min = Temp
            IF Temp > Max THEN Max = Temp
        NEXT IX

' ** Scale this graphics window **
        WINDOW (Start%, Min)-(Stops%, Max)
        LINE (Start%, 0)-(Stops%, 0), White%

' ** Plot the data **
        Temp = SpecData#(NWaves% + 1, Start%)
        IF Temp > Thresh THEN
            LINE (Start%, SpecData#(NWaves% + 1, Start%)-(Start% + 1, SpecData#(NWaves% + 1, Start% +
1)), Yellow%
        ELSE
            LINE (Start%, SpecData#(NWaves% + 1, Start%)-(Start% + 1, SpecData#(NWaves% + 1, Start% +
1)), BBlue%
        END IF
        FOR IX = (Start% + 2) TO Stops%
            Temp = SpecData#(NWaves% + 1, IX)
            IF Temp > Thresh THEN
                LINE -(IX, Temp), Yellow%
            ELSE
                LINE -(IX, Temp), BBlue%
        NEXT IX
    LOOP UNTIL Toggle% = 0

```

```

    END IF
NEXT I%

IF Toggle% = 1 THEN
    LOCATE 29, 9
    COLOR Red%
    PRINT Start%; " ";
    COLOR White%
    LOCATE 29, 75
    PRINT Stops%; " ";
ELSE
    LOCATE 29, 9
    COLOR White%
    PRINT Start%; " ";
    COLOR Red%
    LOCATE 29, 75
    PRINT Stops%; " ";
    COLOR White%
END IF
LOCATE 29, 40
PRINT "Threshold = "; Thresh; "      ";

CALL KeyPress(K%, K$)
PRINT K%
IF Toggle% = 1 THEN
    IF (K% = 377) AND (Start% < (Stops% - 2)) THEN Start% = Start% + 1
    IF (K% = 375) AND (Start% > 1) THEN Start% = Start% - 1
ELSE
    IF (K% = 377) AND (Stops% < (NPnts%)) THEN Stops% = Stops% + 1
    IF (K% = 375) AND (Stops% > (Start% + 2)) THEN Stops% = Stops% - 1
END IF
IF K% = 9 THEN Toggle% = Toggle% * -1
IF (K% = 372) THEN
    IF (Thresh > StdDev) THEN
        Thresh = Thresh * 1.5
    ELSE
        Thresh = Thresh + StdDev
    END IF
END IF
IF (K% = 380) THEN
    IF (Thresh > StdDev) THEN
        Thresh = Thresh / 1.5
    ELSE
        Thresh = Thresh - StdDev
    END IF
END IF
IF K$ = "S" THEN

    LOCATE 30, 20
    PRINT "File name (no extention)";
    INPUT ; FileName$
    FileName$ = FileName$ + ".DAT"
    OPEN FileName$ FOR OUTPUT AS #1
    FOR J% = 1 TO NPnts%
        FOR I% = 1 TO (NWaves%)
            PRINT #1, CSNG(SpecData#(I%, J%)); ", ";
        NEXT I%
    PRINT #1,
    NEXT J%
    CLOSE #1

```

```
        LOCATE 30, 20
        PRINT SPACE$(40);
    END IF

    LOOP UNTIL K% = 13
    ' ** Reset to normal screen **
    WINDOW
    VIEW
    SCREEN 0

    ' ***
    FOR I% = Start% TO Stops%
        FOR J% = 1 TO (NWaves% + 1)
            SpecData#(J%, (I% - Start% + 1)) = SpecData#(J%, I%)
        NEXT J%
    NEXT I%
    NPnts% = Stops% - Start% + 1

END SUB
```

## APPENDIX B

### PROGRAM LISTING FOR TARGET.M

---

```
% Target.M
% Performs Iterative Target Factor Analysis

% Part 1. Load in the data and results to the Kalman filter
% Output:
%   C = Concentration
%   S = Spectra
%   D = data = C*S + noise
%   Tk = target vectors (Kalman innovations)

clear all
clear global
fname = 'test';
fname = input('File Name (no ext) >', 's');

% Concentration matrix *.CON
temp = [fname, '.con'];
eval(['load ', temp]);
C = eval([fname, '(:,1:3)']);
ns = length(C(:,1));

% Spectral Matrix *.SPC
temp = [fname, '.spc'];
eval(['load ', temp]);
S = eval([fname, '(:,1:2)']);
S = S';
nw = length(S(1,:));

% Data matrix *.DAT
temp = [fname, '.dat'];
eval(['load ', temp]);
D = eval(fname);

% Innovations Matrix *.INN
temp = [fname, '.inn'];
eval(['load ', temp]);
Tktemp = eval(fname);

clear fname temp test;

% Part 2. factor analysis on data matrix (D*D')
%
% Output:
%   V = Eigenvector matrix
%   A = Scores matrix
%   L = eigenvalues

fa
C = C(:,1:nc);
Cn = C./((ones(length(C(:,1))),1) * (sum(C.^2)).^0.5));

% Part 3. Kalman filter
% Output: Tk
% For a two-component system the innovations from the one-component
```



```

% Kalman filter are used as target vectors.
if nc==2
    Tk = Tktemp(:,[3,1]);
elseif nc==3
    Tk = zeros(ns,nc);
    Tk(:,1) = Tktemp(:,4);
    Tk(:,3) = Tktemp(:,2);
    cross = max(find(Tktemp(:,1) < Tktemp(:,3)))
    Tk(:,2) = [Tktemp(1:cross,1)', Tktemp((cross+1):ns,3)']';
end
Tk = Tk./(ones(length(Tk(:,1)),1) * (sum(Tk.^2)).^0.5);
clear cross Tktemp

```

```

% Part 4. Needle Search
%
% Outputs:
%   Tn = Target vector for needle search

```

```

    needle

```

```

% Part 5. Varimax rotation
% Output: Tv = target vector

```

```

varimax

```

```

% part 6

```

```

% plot(Te)
% title('EFA target vectors');
% hold on
% plot(Cn,'r:')
% hold off
% pause

```

```

plot(Tk)
title('Kalman filter target vectors');
hold on
plot(Cn,'r:')
hold off
pause

```

```

plot(Tn)
title('needle search target vectors');
hold on
plot(Cn,'r:')
hold off
pause

```

```

plot(Tv)
title('varimax target vectors');
hold on
plot(Cn,'r:')
hold off

```

```

% F = [diag(Te' * Cn)',

```

```

% diag(Tk' * Cn)',
% diag(Tn' * Cn)',
% diag(Tv' * Cn)'];

F = [diag(Tk' * Cn)',
     diag(Tn' * Cn)',
     diag(Tv' * Cn)'];
F = acos(F).*(180/pi);

% Dist = [(sum((Te - Cn).^2)).^0.5,
%         (sum((Tk - Cn).^2)).^0.5,
%         (sum((Tn - Cn).^2)).^0.5,
%         (sum((Tv - Cn).^2)).^0.5];
Dist = [(sum((Tk - Cn).^2)).^0.5,
        (sum((Tn - Cn).^2)).^0.5,
        (sum((Tv - Cn).^2)).^0.5];

Fit1 = [F,Dist]
%T = [Te,Tk,Tn,Tv];
T = [Tk,Tn,Tv];

Part 7
' ITT (separate script file)

% FA.M
% FA performs factor analysis on data matrix D*D'
%
% Output:
%   V = Eigenvector matrix
%   A = Scores matrix
%   L = eigenvalues
%
[V,L] = eig(D*D');
L = flipud(fliplr(L));
V = fliplr(V);
A = V'* D;
echo off

subplot(211),plot (diag(L(1:5,1:5)));
title ('Eigenvalues')
subplot(212),plot(V(:,1:3));
title('Abstract Chromatograms')

nc = input('Number of Components? ');
%nc = 2;

V = V(:,1:nc);
A = A(1:nc,:);
L = L(1:nc,1:nc);
% L = diag(L);

% subplot(211),mesh (D);
% title('Raw Data Matrix')
% subplot(212),mesh(V * A);
% title('Reconstruction from Principle Components')
% pause

```

```

subplot(111)
% global D V A L

% Needle.M
% Needle Search
%
% Inputs:
%     V = Eigenvectors of D*D'

l = size(V(:,1));
t = [1:l];
for i = 1:l,
    T = zeros(l,1);
    T(i) = 1;           %Make test vector
    R = V' * T;        %Calc. rotation
    Tp = (V * R);      %Calc. predicted vector
    c(i) = T' * Tp;    %Compare vectors
end

% Plot the results
plot(max(max(C))*c/max(c),'g+')
hold on
plot(max(max(C))*c/max(c),'g-')
plot(C,'r-')
hold off
pause

% Produce a target vector

Tn = zeros(l,nc);
for i=1:nc
    p = find(Cn(:,i)==max(Cn(:,i)))
    plot(t((p-10):(p+10)) , c((p-10):(p+10)),'g+')
    j = input('Target Vector? ');
    Tn(j,i) = 1;
end

clear i j R c T Tp l p

% Varimax.M
% This procedure follows algorithm as spelled out in
% Harman (1960) in Chapter 14, section 4. To run the
% program - the loadings are put in an array called
% lding. Type return to continue processing.
% The notation follows Harman. The routine vfunct.m is
% called to compute the variance of the loadings
% squared.

lding = V;
b=lding;
[n,nf]=size(lding);

V0 = vfunct(lding) ;           % variances of loadings^2
for it=1:10; % Never seems to need very many iterations
for i=1:nf-1 % Program cycles through 2 factors
    jl=i+1; % at a time.
    for j=jl:nf

```

```

xj=lding(:,i); % notation here closely
yj=lding(:,j); % follows harman
uj=xj.*xj-yj.*yj;
vj=2*xj.*yj;
Ao=sum(uj);
Bo=sum(vj);
Co=uj.*uj-vj.*vj;
Do=2*uj.*vj;
num=Do-2*Ao*Bo/n;
den=Co-(Ao^2-Bo^2)/n;
tan4p=num/den;
phi=atan2(num,den)/4;
angle=phi*180/pi;

if abs(phi)>.0000001;
    Xj=cos(phi)*xj+sin(phi)*yj;
    Yj=-sin(phi)*xj+cos(phi)*yj;
    bj1=Xj;
    bj2=Yj;
    b(:,i)=bj1;
    b(:,j)=bj2;
    lding(:,i)=b(:,i);
    lding(:,j)=b(:,j);
end
end;
lding=b;
Vm=vfunct(lding);
if abs(Vm-V0)<.000001;break;else V0=Vm;end;
end;

Tv = lding;
rot = ones(1,nf) - 2*(max(Tv) < abs(min(Tv)));
Tv = Tv * diag(rot);

peakpos = zeros(0,1);
for i=1:nc
    peakpos(i) = find(Tv(:,i)==max(max(Tv(:,i)))));
end
[peakpos,j] = sort(peakpos);
Tv = Tv(:,j);

clear Ao Bo Co Do V0 Vm Xj Yj angle lding rot peakpos
clear b bj1 bj2 den i it j jl n nf num phi tan4p uj vj xj yj

% IT.M
% Iterative Target Transform Factor Analysis
%
% Inputs:
%     T = target vector
%     V = Eigenvectors of D*D'
m = length(T(1,:));
t = [1:ns];
clear F1 F2

% Target testing

%ni = input('number of iterations ');
ni = 10;
for i = 1:ni      % Number of iterations

```

```

R = V' * T;      % Rotation vector
Tp = V * R;     % Tp = target projected into the factor space
%plot(Tp)
%hold on
T = Tp;

% Target transform
% Set the negative elements to zero
z = find(T < 0.000);
T(z) = zeros(size(T(z)));

% Normalize the target profiles to unit length
Sum = (sum(T.^2)).^0.5;
Sum = ones(length(T(:,1)),1) * Sum;
T = T./Sum;

% Compare the targets to the true concentration profiles
F = T' * Cn;
f1 = acos(F(:,1)).*(180/pi) ;
F1 = [F1',f1]';
f2 = acos(F(:,2)).*(180/pi);
F2 = [F2',f2]';
if nc==3
    f3 = acos(F(:,3)).*(180/pi);
    F3 = [F3',f3]';
end
end
hold off

% Show the results
%pause
subplot(211)
plot(F1)
title('peak one')
subplot(212)
plot(Cn(:,1), 'b+')
hold on
plot(T(:, [1:nc:m]))
hold off
pause

subplot(211)
plot(F2)
title('peak two')
subplot(212)
plot(Cn(:,2), 'b+')
hold on
plot(T(:, [2:nc:m]))
hold off
pause

if nc==3
    subplot(211)
    plot(F3)
    title('peak three')
    subplot(212)
    plot(Cn(:,3), 'b+')
    hold on
    plot(T(:, [3:nc:m]))
    hold off
end
pause

```

```
end
subplot

%Dist = [(sum((T(:,1:2) - Cn).^2)).^0.5,
%       (sum((T(:,3:4) - Cn).^2)).^0.5,
%       (sum((T(:,5:6) - Cn).^2)).^0.5)];

Fit2 = [f1,f2]

clear R Sum i j z Tp m Dist f1 f2
```

## REFERENCES

---

1. Shenk, J. S. In *Near-Infrared Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications*; Ellis Horwood: New York, 1990; pp 235-240.
2. Lorber, A.; Harel, A.; Goldbart, Z.; Brenner, I. B. *Anal. Chem.* **1987**, *59*, 1260-1266.
3. Beebe, K. R.; Kowalski, B. R. *Anal. Chem.* **1987**, *59*, 1007A-1017A.
4. Meglen, R. R. *J. Chemom.* **1991**, *5*, 163-179.
5. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics; Chemical Analysis Series 82*; Wiley-Interscience: New York, 1986.
6. *Practical Guide to Chemometrics*; Haswell, S. J., Ed.; Dekker: New York, 1992.
7. Meloun, M.; Militky, J.; Forina, M. *Chemometrics for Analytical Chemistry*; Ellis Horwood: New York, 1992.
8. Brown, S. D.; Bear, R. S. Jr.; Blank, T. B. *Anal. Chem.* **1992**, *64*, 22R-49R.
9. Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: a textbook*; Elsevier: New York, 1988; pp 5-9.
10. Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*; Wiley: New York, 1978.
11. Haaland, D. M. In *Practical Fourier Transform Infrared Spectroscopy*; Ferraro, J. R.; Krishnan, K., Eds.; Academic: San Diego, 1990; pp 402-405.
12. Windig, W. J. *Chemom.* **1993**, *7*, 213-222.
13. Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 247-264.
14. Sanchez, E.; Kowalski, B. R. *J. Chemom.* **1988**, *2*, 265-280.
15. Skoog, D. A.; Leary, J. J. *Principles of Instrumental Analysis 4<sup>th</sup> ed.*; Saunders: Fort Worth, 1992; p 592.
16. Giddings, J. C. *Anal. Chem.* **1967**, *39*, 1027-1028.

17. Crozier, A.; Reeve, D. R. *Anal. Proc.* **1992**, *29*, 422-425.
18. Hirschfeld, T. *Anal. Chem.* **1980**, *52*, 297A-312A.
19. Giddings, J. C. *Anal. Chem.* **1984**, *56*, 1258A-1270A.
20. Davis, J. M.; Giddings, J. C. *Anal. Chem.* **1983**, *55*, 418-424.
21. Rosenthal, D. *Anal. Chem.* **1982**, *54*, 63-66.
22. Miller, J. C.; Miller, J. N. *Statistics for Analytical Chemistry*, Ellis Horwood: London, 1984; pp 35-40.
23. Hieftje, G. M. *Anal. Chem.* **1972**, *44*, 69A-78A.
24. Hamming, R. W. *Digital Filters 2<sup>nd</sup> ed.*; Prentice-Hall: Toronto, 1983.
25. Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge University: Cambridge, 1986; Chapter 14.
26. Gauss, K. G. *Theory of Motion of the Heavenly Bodies*; Dover: New York, 1963.
27. Savitsky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627-1639.
28. Bialkowski, S. E. *Anal. Chem.* **1988**, *60*, 355A-361A.
29. Larivee, R. J.; Brown, S. D. *Anal. Chem.* **1992**, *64*, 2057-2066.
30. Rzhetskii, A. M.; Mardilovich, P. P. *Appl. Spectrosc.* **1994**, *48*, 13-20.
31. Tahboub, Y. R.; Pardue, H. L. *Anal. Chem.* **1985**, *57*, 38-41.
32. Erickson, C. L.; Lysaght, M. J.; Callis, J. B. *Anal. Chem.* **1992**, *64*, 1155A-1163A.
33. Brown, S. D. *Anal. Chim. Acta* **1986**, *181*, 1-26.
34. Rutan, S. C. *J. Chemom.* **1987**, *1*, 7-18.
35. Tranter, R. L. *Anal. Proc.* **1990**, *27*, 134-136.
36. Cooper, W. S. *Rev. Sci. Instrum.* **1986**, *57*, 2862-2869.



37. Hamming, R. W. *Digital Filters 2<sup>nd</sup> ed.*; Prentice-Hall: Toronto, 1983; pp 1-7.
38. Sorenson, H. W. *IEEE Spectrum* **1970**, 7, 63-68.
39. Wiener, N. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*; Wiley: New York, 1949.
40. Kalman, R. E. *J. Basic Eng.* **1960**, 82D, 35-45.
41. Wentzell, P. D.; Karayannis, M. I.; Crouch, S. R. *Anal. Chim. Acta* **1989**, 224, 263-274.
42. Alberty, R. A. *Physical Chemistry, Sixth ed.*; Wiley: New York, 1983; Chapter 17.6.
43. Brown, R. G. *Introduction to Random Signal Analysis and Kalman Filtering*; Wiley: New York, 1983; Chapter 5.
44. Wentzell, P. D.; Wade, A. P.; Crouch, S. R. *Anal. Chem.* **1988**, 60, 905-911.
45. Rutan, S. C.; Brown, S. D. *Anal. Chim. Acta* **1984**, 160, 99-119.
46. Rutan, S. C. *Anal. Chem.* **1991**, 1103A-1109A.
47. *Applied Optimal Estimation*; Gelb, A. , Ed.; M.I.T. Press: Cambridge Massachusetts, 1974; pp 115-119.
48. San, K. Y.; Stephanopoulos, G. *Biotechnol. Bioeng.* **1984**, 26, 1189-1197.
49. Seelig, P. F.; Blount, H. N. *Anal. Chem.* **1976**, 48, 252-258.
50. Poulisse, H. N. J. *Anal. Chim. Acta* **1979**, 112, 361-374.
51. Biakowski, S. E. *Anal. Chem.* **1988**, 60, 403A-413A.
52. Lavagnini, I.; Pastore, P.; Mango, F. *Anal. Chim. Acta* **1990**, 239, 95-106.
53. Poulisse, H. N. J.; Jansen, R. T. P. *Anal. Chim. Acta* **1983**, 151, 433-439.
54. Jansen, R. T. P.; Poulisse, H. N. J. *Anal. Chim. Acta* **1983**, 151, 441-446.
55. Lilley, T. *Anal. Proc.* **1984**, 21, 147-148.

56. Rutan, S. C.; Brown, S. D. *Anal. Chim. Acta* **1985**, 167, 23-37.
57. Velasco, A.; Rui, X.; Silva, M.; Perez-Bendito, D. *Talanta* **1993**, 40, 1505-1510.
58. Jimenez-Prieto, R.; Velasco, A.; Silva, M.; Perez-Bendito, D. *Talanta* **1993**, 40, 1731-1739.
59. Quencer, B. M.; Crouch, S. R. *Anal. Chem.* **1994**, 66, 458-463.
60. Rutan, S. C.; Brown, S. D. *Anal. Chim. Acta* **1985**, 175, 219-229.
61. Didden, C. B. M.; Poullisse, H. N. L. *Anal. Lett.* **1980**, 13, 921-935.
62. Brown, T. F.; Brown, S. D. *Anal. Chem.* **1981**, 53, 1410-1417.
63. Scolari, C. A.; Brown, S. D. *Anal. Chim. Acta* **1984**, 166, 253-260.
64. Shi, L.; Li, Z.; Xu, Z.; Pan, Z.; Wang, L. *J. Chemom.* **1991**, 5, 193-199.
65. van Loosbroek, A.; Debets, H. J. G.; Coengracht, P. M. J. *Anal. Lett.* **1984**, 17, 779-792.
66. Yongnian, N.; Selby, M.; Kokot, S.; Hodgkinson, M. *Analyst* **1993**, 1049.
67. van Veen, E. H.; de Loos-Vollebregt M. T. C. *Anal. Chem.* **1991**, 63, 1441-1448.
68. Gerow, D. D.; Rutan, S. C. *Anal. Chem.* **1988**, 60, 847-852.
69. Brown, S. D. *J. Chemom.* **1991**, 5, 147-161.
70. Thijssen, P. C.; Prop, L. T. M.; Kateman, G.; Smit, H.C. *Anal. Chim. Acta* **1985**, 174, 27-40.
71. Wienke, O.; Vijn, T.; Buydens, L. *Anal. Chem.* **1994**, 66, 841-849.
72. Rutan, S. C.; Carr, P. W. *Anal. Chim. Acta* **1988**, 215, 131-142.
73. Webster, G. H.; Cecil, T. L.; Rutan, S. C. *J. Chemom.* **1988**, 3, 21-32.
74. Xie, Y.; Wang, J.; Liang, Y.; Yu, R. *Anal. Chim. Acta* **1992**, 307-316.

75. Hayashi, Y; Helburn, R. S.; Rutan, S. C. In *Computer-Enhanced Analytical Spectroscopy Vol. 4*; Wilkins, C. L. Ed.; Plenum: New York, 1993; Chapter 11.
76. Rutan, S. C.; Brown, S. D. *Anal. Chim. Acta* **1985**, 167, 39-50.
77. Wilk, H. R.; Brown, S. D. *Anal. Chim. Acta* **1989**, 225, 37-52.
78. van Veen E. H.; de Loos-Vollebregt, M. T. C. *Spectrochim. Acta* **1990**, 45B, 313-328.
79. Wentzell, P. D.; Vanslyke, S. J. *Anal. Chim. Acta* **1992**, 257, 173-181.
80. Azimi-Sadjadi, M. R.; Lu, T.; Nebot, E. M. *IEEE Trans. Signal Process.* **1991**, 39, 137-147.
81. Mottola, H. A. *Kinetic Aspects of Analytical Chemistry*; Wiley: New York, 1988.
82. Pardue, H. L. *Anal. Chim. Acta* **1989**, 216, 69-107.
83. Perez-Bendito, D.; Silva, M.; Gomez-Hens, A. *Trends in Anal. Chem.* **1989**, 8, 302.
84. Carr, P. W. *Anal. Chem.* **1978**, 50, 1602-1607.
85. Davis, J. E.; Renoe, B. *Anal. Chem.* **1979**, 51, 526-528.
86. Atwood, J. G.; DiCesare, J. L. *Clin. Chem.* **1975**, 21, 1263-1269.
87. Landis, J. B.; Rebec, M.; Pardue, H. L. *Anal. Chem.*, **1977**, 49, 785-788.
88. Holler, F. J.; Calhoun, R. K.; McClanahan, S. F. *Anal. Chem.* **1982**, 54, 755-761.
89. Mieling, G. E.; Pardue, H. L. *Anal. Chem.* **1978**, 50, 1611-1618.
90. Harris, R. C.; Hultman, E. *Clin. Chem.* **1983**, 29, 2079-2081.
91. Wentzell, P. D.; Crouch, S. R. *Anal. Chem.* **1986**, 58, 2851-2855.
92. Corcoran, C. A.; Rutan, S. C. *Anal. Chem.* **1988**, 60, 1146-1153.
93. Larsson, J. A.; Pardue, H. L. *Anal. Chim. Acta* **1989**, 224, 289-303.

94. Wentzell, P. D.; Crouch, S. R. *Anal. Chem.* **1986**, *58*, 2855-2857.
95. Corcoran, C. A.; Rutan, S. C. *Anal. Chem.* **1988**, *60*, 2450-2454.
96. Ingle, J. D. Jr.; Crouch, S. R. *Anal. Chem.* **1971**, *43*, 697-701.
97. Campi, G. L.; Ingle, J. D. Jr.; *Anal. Chim. Acta* **1989**, *224*, 275-287.
98. Harmon, H. H. *Modern Factor Analysis*; University of Chicago: Chicago, 1967.
99. Weiner, P. H. *Chem. Tech.* **1977**, 321-328.
100. Wold, S. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37-52.
101. Windig, W. *Chemom. Intell. Lab. Syst.* **1988**, *4*, 201-213.
102. Aries, R. E.; Lidiard, D. P.; Spragg, R. A. *Chem. in Brit.* **1991**, 821-824.
103. Malinowski, E. R. *Factor Analysis in Chemistry, Second ed.*; Wiley: New York, 1991.
104. Balke, S. T. *Quantitative Column Liquid Chromatography*; Journal of Chromatography Library Vol. 29; Elsevier: Amsterdam, 1984; Chapter 4.
105. Alfreson, T.; Sheedan, T. *J. Chromatogr. Sci.* **1986**, *24*, 473-482.
106. Jones, D. G. *Anal. Chem.* **1985**, *57*, 1057A-1073A.
107. Fetzer, J. C.; Biggs, W. R. *J. Chromatogr.* **1993**, *642*, 319-327.
108. Namba, R. In *Practical Fourier Transform Infrared Spectroscopy*, Ferraro, J. R.; Krishnan, K., Eds.; Academic: San Diego, 1990; Chapter 9.
109. Mason, P. B.; Zhang, L.; Carnahan, J. W.; Winans, R. E. *Anal. Chem.* **1993**, *65*, 2596-2600.
110. Covey, T. R.; Lee, E. D.; Bruins, A. P.; Henion, J. D. *Anal. Chem.* **1986**, *58*, 1451A-1461A.
111. Hamilton, H. J.; Gemperline, P. J. *J. Chemom.* **1990**, *4*, 1-13.
112. Vandeginste, B. G. M. *Topics Curr. Chem.* **1987**, *141*, 24-32.
113. Malinowski, E. R. *Anal. Chem.* **1977**, *49*, 606-612.

114. Gillette, P. C.; Koenig, J. L. *Appl. Spectrosc.* **1982**, 36, 535-539.
115. Healy, M. J. R. *Matrices for Statistics*; Oxford: New York, 1986; Chapter 4.
116. Gerritson, M. J. P.; Faber, N. M.; van Rijn, M.; Vandeginste, B. G. M.; Kateman, G. *Chemom. Intell. Lab. Syst.* **1992**, 12, 257-268.
117. Kowalski, B. R.; Bender, C. F. *J. Am. Chem Soc.* **1972**, 94, 5632-5639.
118. Lavine, B. K. in *Practical Guide to Chemometrics*; Haswell, S. J., Ed.; Dekker: New York, 1992; Chapter 7.
119. Grahn, H.; Delagio, F.; Delusuc, M. A.; Levy, G. C. *J. Magn. Reson.* **1988**, 77, 294-307.
120. Haaland, D. M. In *Practical Fourier Transform Infrared Spectroscopy*; Ferraro, J. R.; Krishnan, K., Eds.; Academic: San Diego, 1990; Chapter 8.
121. Kowalski, B. R.; Seasholtz, M. B. *J. Chemom.* **1991**, 5, 129-145.
122. Laatikainen, R.; Tuppurainen, K. *Comput. Chem.* **1990**, 14, 109-126.
123. Malinowski, E. R. *Anal. Chem.* **1977**, 49, 612-617.
124. Rasmussen, G. T.; Isenhour, T. L.; Lowry, S. R.; Ritter, G. L. *Anal. Chim. Acta* **1978**, 103, 213-221.
125. Malinowski, E. R. *Anal. Chim. Acta* **1978**, 103, 339-354.
126. Malinowski, E. R. *J. Chemom.* **1988**, 1, 33-40.
127. Tauler, R.; Casassas, E.; Izquierdo-Ridorsa, A. *Anal. Chim. Acta* **1991**, 447-458.
128. Kormos, D. W.; Waugh, J. S. *Anal. Chem.* **1983**, 55, 633-638.
129. Hirsch, R. F.; Wu, G. L.; Tway, P. C. *J. Chemom.* **1987**, 265-272.
130. Rossi, T. M.; Warner, I. M. *Anal. Chem.* **1986**, 58, 810-815.
131. Fay, M. J.; Proctor, A.; Hoffman, D. P.; Hercules, D. M. *Anal. Chem* **1991**, 63, 1058-1063.

132. Deuwer, D. L.; Kowalski, B. R.; Fasching, J. L. *Anal. Chem.* **1976**, *48*, 2002-2010.
133. Eastment, H. T.; Krzanowski, W. J. *Technometrics* **1982**, *24*, 73-77.
134. Geladi, P.; Wold, S. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 273-281.
135. Gampp, H.; Maeder, M.; Meyer, C. J.; Zuberbühler, A. D. *Talanta* **1985**, *32*, 1133-1139.
136. Gampp, H.; Maeder, M.; Meyer, C. J.; Zuberbühler, A. D. *Talanta* **1986**, *33*, 943-951.
137. Maeder, M.; Zuberbuehler, A. D. *Anal. Chim. Acta* **1986**, *181*, 287-291.
138. Maeder, M. *Anal. Chem.* **1987**, *59*, 527-530.
139. Maeder M.; Zilian, A. *Chemom. Intell. Lab. Syst.* **1988**, *3*, 205-213.
140. Gemperline, P. J.; Hamilton, J. C. *J. Chemom.* **1989**, *3*, 455-461.
141. E.R. Malinowski in *Computer Enhanced Analytical Spectroscopy Vol. 1*; Meuzelaar H. L. C.; Isenhour, T. L., Eds; Plenum: New York, 1987; pp. 55-102.
142. Keller H. R.; Massart D. L. *Anal. Chim. Acta* **1991**, *246*, 379-390.
143. Kvalheim O. M.; Liang, Y. *Anal. Chem.* **1992**, *64*, 936-946.
144. Malinowski, E. R. *J. Chemom.* **1992**, *6*, 29-40.
145. Yost, R.; Stoveken, J.; MacLean, W. *J. Chromatogr.* **1977**, *134*, 73-82.
- 146 Vanslyke, S. J.; Wentzell, P. D. *Anal. Chem.* **1991**, *63*, 2512-2519.
147. Barker, T.; Brown, S. D. *J. Chromatogr.* **1989**, *469*, 77-90.
148. Redmond, M.; Brown, S. D.; Wilk, H. R. *Anal. Lett.* **1989**, *22*, 963-979.
149. Hayashi, Y.; Yoshioka, S.; Takeda, Y. *Anal. Chim. Acta* **1988**, *212*, 81-94.
150. McCue, M.; Malinowski, E. R. *Appl. Spectrosc.* **1983**, *37*, 463-469.
151. Gemperline, P. J. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 206-212.

152. Williamson, J. H. *Can. J. Phys.* **1968**, *46*, 1845-1847.
153. McCue, M.; Malinowski, E. R. *Anal. Chem.* **1981**, *133*, 125-136.
154. Vaidya, R. A.; Hester, R. D. *J. Chromatogr.* **1984**, *287*, 231-244.
155. Gerritsen, M. J. P.; Tanis, H.; Vandeginste, B. G. M.; Kateman, G. *Anal. Chem.* **1992**, *64*, 2042-2056.
156. Keller, H. R.; Massart, D. L.; Kiechle, P.; Erni, F. *Anal. Chim. Acta* **1992**, *256*, 125-131.
157. Ingle, J. D. Jr.; Crouch, S. R. *Spectrochemical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 1988; pp 34-35.
158. Hargis, L. G.; Howell, J. A. in *Physical Methods of Chemistry*, 2nd ed.; Rossiter, B. W.; Baetzold, R. C., Eds.; Wiley: New York, 1986; Vol. 8, Chapter 1.
159. Milano, M. J.; Lam, S.; Savonis, M.; Pautler, D. B.; Pav, J. W.; Grushka, E. *J. Chromatogr.* **1978**, *149*, 599-614.
160. Talmi, Y.; Simpson, R. W. *Appl. Opt.* **1980**, *19*, 1401-1414.
161. Talmi, Y. *Appl. Spectrosc.* **1982**, *36*, 1-18.
162. Lobinski, R.; Marczenko, Z. *Crit. Rev. Anal. Chem.* **1992**, *23*, 55-111.
163. Slavin, W. *Anal. Chem.* **1963**, *35*, 561-566.
164. Dose, E. V.; Guiochon, G. *Anal. Chem.* **1989**, *61*, 2571-2579.
165. Keller, H. R.; Massart, D. L. *Anal. Chim. Acta* **1992**, *263*, 21-28.
166. Skoog, D. A.; Leary, J. J. *Principles of Instrumental Analysis 4<sup>th</sup> ed.*; Saunders: Fort Worth, 1992; pp 131-134.
167. Voigtman, E. *Anal. Instrum.* **1993**, *21*, 43-62.
168. Keller, H. R.; Massart, D. L.; Liang, Y. Z.; Kvalheim, O. M. *Anal. Chim. Acta* **1992**, *263*, 29-36.
169. Keller, H. R.; Massart, D. L.; Liang, Y. Z.; Kvalheim, O. M. *Anal. Chim. Acta* **1992**, *267*, 63-71.

170. Keller, H. R.; Massart, D. L.; De Beer, J. O. *Anal. Chem.* **1993**, *65*, 471-475.
171. Sanchez, F. C.; Khots, M. S.; Massart, D. L.; De Beer, J. O. *Anal. Chim. Acta* **1994**, 181-192.
172. Thomas, E. V.; Haaland, D. M. *Anal. Chem.* **1990**, *62*, 1091-1099.
173. Otto, M.; Wegscheider, W. *Anal. Chem.* **1985**, *57*, 63-69.
174. Brown, C. W.; Lynch, P. F.; Obremski, R. J.; Lavery, D. S. *Anal. Chem.* **1982**, *54*, 1472-1479.
175. Strasters, J. K.; Billiet, H. A. H.; De Galan, L.; Vandeginste, B. G. M.; Kateman, G. E. *Anal. Chim. Acta* **1987**, *385*, 181-200.
176. Kankare, J. J. *Anal. Chem.* **1970**, *42*, 1322.
177. Schostack, K. J.; Malinowski, E. R. *Chemom. and Intell. Lab. Syst.* **1993**, *20*, 173-182.
178. Delaney, M. F. *Anal. Chem.* **1984**, *56*, 261R-277R.
179. Lawton, W. H.; Sylvestre, E. A. *Technometrics* **1971**, *13*, 617-633.
180. Knorr, F. J.; Futrell, J. H. *Anal. Chem.* **1979**, *51*, 1236-1241.
181. Gillette, P. C.; Lando, J. B.; Koenig, J. L. *Anal. Chem.* **1983**, *55*, 630-633.
182. Sasaki, K.; Kawata, S.; Miniama, S. *Appl. Opt.* **1983**, *22*, 3599-3603.
183. Borgen, O. S.; Kowalski, B. R. *Anal. Chim. Acta* **1985**, *174*, 1-26.
184. Gemperline, P. J. *Anal. Chem.* **1986**, *58*, 2656-2663.
185. Vandeginste, B. G. M.; Derks, W.; Kateman, G. *Anal. Chim. Acta* **1985**, *173*, 253-264
186. Stasters, J. K.; Billiet, H. A. H.; de Galan, L.; Vandeginste, B. G. M.; Kateman, G. *Anal. Chem.* **1988**, *60*, 2745-2751.
187. Gerritsen, M. J. P.; Tanis, H.; Vandeginste, B. G. M.; Kateman, G. *Anal. Chem.* **1992**, *64*, 2042-2056.



188. Vandeginste, B.; Essers, R.; Bosman, T.; Reijnen, J.; Kateman, G. *Anal. Chem.* **1985**, *57*, 971-985.
189. Vandeginste, B. G. M.; Leyten, F.; Gerritsen, M.; Noor, J. W.; Kateman, G.; Frank, J. *J. Chemom.* **1987**, *1*, 57-71.
190. Karjalainen, E. J.; Karjalainen, U. P. *Anal. Chim. Acta* **1991**, *250*, 169-179.
191. Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425-1432.
192. Kvalheim, O. M.; Liang, Y. *Anal. Chem.* **1992**, *64*, 936-946.
193. Tauler, R.; Casassas, E. *J. Chemom.* **1988**, *3*, 151-161.
194. Brown, S. D.; Harper, A. M. In *Computer-Enhanced Analytical Spectroscopy Vol. 4*; Wilkins, C. L. Ed.; Plenum: New York, 1993; Chapter 6.