MAXIMUM LIKELIHOOD MULTIVARIATE METHODS IN ANALYTICAL CHEMISTRY

Darren Thomas Andrews

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
at
Dalhousie University
Halifax, Nova Scotia
May, 1997

0-612-24729-5

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 "Maximum Likelihood Multivariate Methods in Analytical Chemistry"

by Darren Thomas Andrews

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

Dated: May 26, 1997

External Examiner

Research Supervisor

Examining Committee

[redacted]

DALHOUSIE UNIVERSITY

May 29, 1997

Author: **Darren Thomas Andrews**

Title: **Maximum Likelihood Multivariate Methods in Analytical Chemistry**

Department: Chemistry
Degree: Ph.D.
Convocation: Fall, 1997

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 in individuals or institutions.

[redacted]

Signature of Author

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

3 Modeling in Higher Dimensions

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The development of analytical methods demands a reliable method for modeling the instrumental response function. This is particularly true for multivariate measurements and is closely related to the measurement error characteristics of the method. This work presents new methods for extracting information from univariate and multivariate data sets based on the principle of maximum likelihood. These maximum likelihood techniques allow for the incorporation of measurement errors in the modeling process and therefore yield a more reliable representation of the true underlying model.

The potential of these techniques is first evaluated in two-dimensions where it was shown that these methods have properties that make them statistically desirable. A maximum likelihood analog to principal component analysis (PCA) is then developed for the multivariate case. The theoretical foundations of maximum likelihood principal component analysis (MLPCA) are initially established using a regression model and then extended to the framework of PCA and singular value decomposition (SVD). The proposed technique also allows for the incorporation of correlated errors and intercept terms. Simulated and experimental data are used to evaluate the performance of the new algorithm. In all cases, models determined by MLPCA are found to be superior to those obtained by PCA when non-uniform error distributions are present, although the level of improvement depends on the error structure of the particular data set.

To demonstrate the practical implications of MLPCA, this technique was applied to problems in multivariate calibration, the modeling of incomplete data sets, and calibration transfer. Two new calibration methods, maximum likelihood principal component regression (MLPCR) and maximum likelihood latent root regression (MLLRR), are developed which exhibit superior performance over conventional multivariate calibration methods when there is a non-uniform error structure. MLPCA is also shown to be useful in handling incomplete data sets in a reliable and simple manner by assigning large uncertainties to missing measurements. This approach is extended to the general problem of multivariate calibration transfer.

ABBREVIATIONS AND SYMBOLS

In general, the conventions used in this paper is are follows. Matrices are represented as upper case bold letters and column vectors are represented as bold lower case letters. Normal face fonts (upper and lower ease) are used for scalars. Normal, Greek and script fonts are used with no particular pattern, but where possible, we have tried to adhere to symbols commonly used in the literature. Symbols which represent estimates of unknown quantities are designated with a caret ("^"). Symbols which represent truncated matrices are designated with a tilde ("~"). A matrix transpose is indicated by a superscript "T" and the Euclidean norm of a vector by "‖•‖". The Kronecker product of two matrices is indicated by "⊗ ".

A list of important abbreviations in the paper follows:

ANN	artificial neural networks
CLS	classical least squares
CR	continuum regression
DAD	diode array detector
EVM	effective variance method
GRAM	generalized rank annihilation method
HE	homoscedastic, equal
HU	homoscedastic, unequal
iid	independent and identically distributed
ILS	inverse least squares
LRR	latent root regression
LWR	locally weighted regression
MLCFA	maximum likelihood common factor analysis

MLLRR	maximum likelihood latent root regression
MLPCA	maximum likelihood principal component analysis
MLPCR	maximum likelihood principal component regression
MLR	multiple linear regression
MSE	mean squared error
MWR	multiply weighted regression
NAS	net analyte signal
NAS*	net analyte signal projected into PCA or MLPCA subspace
NPS	noise power spectrum
OLS	ordinary least squares
PCA	principal component analysis
PCR	principal component regression
PE	proportional, equal
PLS	partial least squares
PMF	positive matrix factorization
PPR	projection pursuit regression
PU	proportional, unequal
RA	random
RMSE	root-mean-square-error
RMSECV	root-mean-square-error of cross-validation
RMSEP	root-mean-square-error of prediction
RR	ridge regression
SEN	sensitivity
SVD	singular value decomposition
TLS	total least squares
UPCA	unweighted principal component analysis
WLS	weighted least squares
WPCA1	weighted principal component analysis (using standard deviations)
WPCA1*	WPCA1 parameters scaled by R

WPCA2	weighted principal component analysis (using statistical weights)
WPCR	weighted principal component analysis
WPLS	weighted partial least squares

A list of important symbols in the paper follows:

$\mathbf{0}_\mathbf{p}$	an px1 vector of zeros
$\mathbf{1}_\mathbf{m}$	an mx1 vector of ones
a	(i) model coefficient (slope) (ii) slope of double sigmoid mask for data set 8 (Chapter 5)
a_{ij}	a model coefficient (element of **A**)
$\mathbf{a}_i$	a left hand vector of the true model
A_{ij}	absorbance of sample i at wavelength j
$\mathbf{A}_2$	lower $(m$-$p)$xp submatrix of **A**
$\mathbf{A}$	(i) matrix of absorbances (Chapter 1) (ii) matrix of model coefficients (iii) left-hand matrix of p-dimensional model (Equation 4.2)
$\hat{\mathbf{A}}$	matrix of estimated model coeffieients
b	an intercept term (offset of regression vector)
$\mathbf{b}$	a column vector of offsets (intercept term)
$\mathbf{b}_j$	a column vector of **B** (offsets)
$\mathbf{b}_2$	lower $(m$-$p)$x1 subvector of **B**
$\hat{\mathbf{b}}$	overall vector of regression coefficients for PCR
$\mathbf{B}$	(i) matrix of model offsets (ii) right hand matrix of p-dimensional model (Equation 4.2)
c_{ik}	concentration of component k in solution i
c	offsets for MWR model
$\mathbf{c}$	vector of column offsets for MLPCA model
$\mathbf{C}$	matrix of concentrations
d_{ij}'	element of **D'**
$\mathbf{d}$	vector of row offsets for MLPCA model
$\mathbf{D'}$	matrix of measurements weighted by corresponding measurement errors

e_{ij}	element of $\mathbf{E}$(i)
$\mathbf{E}$	(i) matrix of measurement residuals (ii) unfiltered measurement errors for $\mathbf{X}$
$\mathbf{E_A}$, $\mathbf{E_C}$	matrices of measurement residuals for CLS and ILS models
$\mathbf{f}$	residual matrix for PCR
$\mathbf{F}$	matrix of filter coefficients (*mnxmn*) for calculating Ω of filtered data
$\mathbf{G}_i$	an intermediate matrix in the calculation of $d\bar{S}^2/d\alpha_i$
$\mathbf{H}_j$	a substitution equal to $\left(\mathbf{U}^T\Psi_i^{-1}\mathbf{U}\right)^{-1}\mathbf{U}^T\Psi_i^{-1}$
$\mathbf{I}_n$, $\mathbf{I}_p$	*nxn* and *pxp* identity matrices
$\mathbf{J}_i$	derivative of rotation matrix, $\mathbf{T}$, with respect to angle α_i
k	normalization constant for normal distribution
$\mathbf{K}$	(i) commutation matrix for $\mathbf{X}$ (Chapter 4) (ii) matrix of coefficients for CLS model (Chapter 1)
$\hat{\mathbf{K}}$	matrix of estimated coefficients for CLS model
l	function minimized for maximum likelihood projection (log likelihood function)
L	probability density function for measurement vector $\mathbf{x}$
$\mathbf{L}$	matrix of PCA loadings
$\mathbf{L}_i$	an exchange matrix with the property that $\mathbf{F}_i = \mathbf{L}_i\,\mathbf{T}_i$
m	number of rows in $\mathbf{X}$
n	number of columns in $\mathbf{X}$
N	number of experimentally observed points
N_{cal}, N_{pred}	number of calibration and prediction samples
p	(i) rank of data matrix (pseudomatrix) (ii) amount of proportional error added (data set 9)
P	probability of observing a χ^2 value less than a given value
$\mathbf{P}, \hat{\mathbf{P}}$	matrix of theoretical and estimated coefficients for ILS model
$\mathbf{P}_j$	maximum likelihood projection matrix for $\mathbf{X}_j$
$\mathbf{q}$, $\hat{\mathbf{q}}$	regression vector for PCR
$\mathbf{q}_i$	projection of $\mathbf{r}_i$ into PCA or MLPCA subspace

$\mathbf{Q}$	*mxn* matrix of measurement error variances
$\mathbf{Q}_i$	projection of $\mathbf{R}_i$ into PCA or MLPCA subspace
$\mathbf{r}_i$	column vector of pure spectrum for component *i*
r_{max}	amplitude of double sigmoid error mask for data set 8
$\mathbf{R}$	ratio of σ_y to σ_x
$\mathbf{R}_i$	matrix of column vectors of all pure spectra excluding spectrum for component *i*
s_θ, σ_θ	standard deviations of estimated angles
S^2	objective function
$\hat{\mathbf{t}}_{unk}, \tilde{\mathbf{t}}_{unk}$	scores for unknown sample by MLPCA and PCA
$\mathbf{T}_1, \mathbf{T}_2, \ldots$	individual rotation matrix
$\mathbf{T}$	(i) matrix of PCA scores (ii) overall rotation matrix (Chapter 4)
$\hat{\mathbf{T}}$	matrix of maximum likelihood scores
$\tilde{\mathbf{T}}$	matrix of truncated scores from PCA
u,s,v	elements of $\hat{\mathbf{U}}$, $\hat{\mathbf{S}}$ and $\hat{\mathbf{V}}$
$\mathbf{U}, \mathbf{S}, \mathbf{V}^T$	matrices returned by SVD
$\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}^T$	matrices returned by MLPCA (SVD form)
$\tilde{\mathbf{U}}, \tilde{\mathbf{S}}, \tilde{\mathbf{V}}^T$	truncated $\mathbf{U}, \mathbf{S}, \mathbf{V}^T$ matrices (obtained by PCA)
$\hat{\mathbf{U}}_1, \hat{\mathbf{U}}_2$	upper and lower submatrices of $\hat{\mathbf{U}}$ (*p*x*p* and (*m-p*)x*p*)
$\hat{\mathbf{U}}_o, \hat{\mathbf{V}}_o$	initial estimates for $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$
$\hat{\mathsf{U}}, \hat{\mathsf{V}}$	super-matrices for $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$
V_N	retention volume
$\mathbf{V}_1, \mathbf{V}_2$	upper *pxp* and (*n-p*)x*p* submatrices of $\mathbf{V}$
$\tilde{\mathbf{V}}_p$	upper *pxp* and (*n-p*)x*p* submatrices of $\tilde{\mathbf{V}}$
W	weight of absorbent
W_i	weights for WPCA2
x_i	measured data point (independent variable)

x_i^{o}	true value of x_i
$\hat{x}_i$	predicted vaue of x_i^{o} by maximum likelihood
$x_{ij}, \hat{x}_{ij}, x_{ij}^{o}$	elements of $\mathbf{X}$, $\hat{\mathbf{X}}$ and $\mathbf{X}^{o}$
x_i^{PCA}	estimated value of x_i^{o} by PCA
$\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}$	column vectors of $\mathbf{X}$
$\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j, \hat{\mathbf{x}}$	column vectors of $\hat{\mathbf{X}}$
$\mathbf{x}^{o}, \mathbf{x}_i^{o}, \mathbf{x}_j^{o}$	column vectors of $\mathbf{X}^{o}$
$\hat{\mathbf{x}}_{unk}$	predicted value of $\mathbf{x}^{o}$ by MLLRR
$\mathbf{x}_p, \hat{\mathbf{x}}_p, \mathbf{x}_p^{o}$	upper px1 vectors of $\mathbf{X}$, $\hat{\mathbf{X}}$ and $\mathbf{X}^{o}$
$\Delta\mathbf{x}, \Delta\mathbf{x}_j$	residual vector for $\mathbf{x}$ $(\mathbf{x}-\hat{\mathbf{x}})$
$\mathbf{X}$	measurement data matrix
$\mathbf{X}_1$	upper p elements of $\mathbf{X}_j$
$\hat{\mathbf{X}}$	matrix of maximum likelihood estimates of $\mathbf{X}^{o}$
$\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2$	upper and lower submatrices of $\hat{\mathbf{X}}$ (pxn and $(m$-$p)$xn)
$\mathbf{X}^{o}$	matrix of true measurements
$\tilde{\mathbf{X}}$	matrix of truncated PCA estimates of $\mathbf{X}^{o}$
X^2	goodness of fit
y	(i) dependent variable in classical regression (ii) transformed retention data
y_i, y_i^{obs}	observed dependent data matrix
y_i^{OLS}, y_i^{PCA}	estimated values of y_i^{o} by OLS and PCA
y^{pred}	predicted value of y_i^{o} (general)
y^{o}	true value of observed y
$\tilde{y}$	estimated value of y_i^{o} by LRR
$\hat{y}, \hat{y}_i$	maximum likelihood estimate of y^{o}
$\mathbf{y}$	vector of dependent variables in classical regression

$\mathbf{y}^{ref}$	vector of reference concentrations
$\Delta\mathbf{y}$	residual vector for $\mathbf{y}$ $(=\mathbf{y}-\hat{\mathbf{y}})$
$\hat{\mathbf{y}}_{unk}$	vector of maximum likelihood estimates of $\mathbf{y}^{o}_{unk}$ by MLLRR
α	significance level for hypothesis testing
α_i	rotation angle about axis i
$\boldsymbol{\beta}$	vector of regression coefficients in classical regression
δ	(i) bias component of error (Chapter 1) (ii) vector of errors in $\mathbf{y}$ in classical regression
ε	random error (Chapter 1)
ε_j	column vector of measurement errors for column j of $\mathbf{X}$
$\varepsilon_i^{PCA}, \varepsilon_i^{OLS}$	model residuals associated with PCA and OLS
ε_{ij}	measurement error for x_{ij} after filtering
ε_{kj}	molar absorptivity of component k at wavelength j
$\boldsymbol{\Phi}_{ij}$	matrix of filter coefficients for errors corresponding to measurement x_{ij}
γ	vector of row offsets for Mandel model
Γ^{inc}	incomplete gamma function
λ	(i) convergence criterion for MLPCA algorithm (Chapter 4) (ii) wavelength (Chapter 5)
μ	grand mean for $\mathbf{X}$
ν	degrees of freedom
θ	angle of model parameter with respect to axis
$\hat{\theta}$	angle of estimated model parameter with respect to axis
$\bar{\theta}$	average of estimated angles
θ_i	angular deviation of eigenvector i from true model space
θ_{true}	angle of true model parameter with respect to axis
ρ	vector of column offsets for Mandel model
Σ_i, Σ_{unk}	error covariance matrix for row i of $\mathbf{X}$ (or column i of $\mathbf{X}^T$) or unknown sample

Σ_i^{pred}	predicted covariance matrix of sample *i* by maximum likelihood projection
Ω	full error covariance matrix of *vec*(**X**)
Ξ	full covariance matrix of *vec*($\mathbf{X}^T$)
Ψ_i	scores covariance matrix for sample *i*
Ψ_j	error covariance matrix for column vector *j* of **X**
σ	standard deviation of the Gaussian distribution
σ_x, σ_y	standard deviation of *x* and *y* variables
σ_o	baseline noise for data set 8
σ_{ij}^2	measurement variance for x_{ij}
$\hat{\sigma}_{jj}$	jj^{th} element of Σ_i^{pred}
$(\sigma_y')^2$	propagated error containing σ_x^2 and σ_y^2

ACKNOWLEDGMENTS

My research would never have 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, Mitch Lohnes, Chris Brown and Stephanie Mehlman.

Dr. David Hamilton, Dr. Louis Ramaley, and Dr. Neil Burford are thanked for their insights on matters academic and otherwise.

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

Finally, heartfelt thanks go out to my family (for their unconditional support), especially to Jeanie, Daniel and Max. I couldn't have done it without you.

1
INTRODUCTION

1.1 THE NEED FOR STATISTICAL ANALYSIS

The demand for chemical information in virtually all fields of technology is now greater than it has ever been before. Furthermore, innovative advances in chemical instrumentation allow more data to be obtained for a system under study. Unfortunately, situations may arise in which potentially useful information is lost amid the complexity of the data. The field of chemometrics has evolved from a need to address this problem. The role of chemometrics is to provide a variety of mathematical and statistical tools to aid in the extraction of information from measurements on chemical systems [1-7]. This work introduces a generalized approach to the treatment of analytical data from a modern generation of analytical instruments.

Over the span of only a few decades, chemometrics has revolutionized many areas of interest in the scientific community. The importance of chemometrics can be illustrated with a few examples [8-12]. In the petroleum industry, chemometric tools permit the determination of octane numbers in gasoline through simple spectroscopic techniques, as opposed to a long and costly procedure involving a specially designed engine and a skilled technician. Likewise in the food industry, methods have been developed for the measurement of various properties (*e.g.* protein, fat and moisture content) which previously required tedious wet-chemical procedures. In medicine, chemometric methods have allowed for the early diagnosis of cancers through simple spectroscopic techniques, and are leading to the development of non-invasive sensors used for the determination of glucose in blood. In industry, chemometric procedures are routinely employed in process monitoring and improving the quality and output of process streams. All of these advances are only possible because of the ability of chemometrics to extract information out of seemingly-useless data.

Chemometric techniques range in complexity from the very simple to the very complex, and which approach is appropriate for a particular analysis depends very much on the data being scrutinized. The remainder of this chapter will examine the nature of experimental data and discuss some general approaches to modeling these data.

1.2 CLASSICAL AND MULTIDIMENSIONAL MODELING

As noted in the preceding section, the required complexity of a mathematical model is tied to the nature of the data under analysis which, in turn, is dictated by the instrumental means used to acquire the data. The type and amount of data an instrument can produce depends on the characteristics of the instrument. In general, instruments can be classified by the *order* of the data they produce (zero-order, first-order, etc.) [13]. This classification refers to the order of tensor an instrument produces. The relationship between the form of the data and the order of instrumentation is presented in Figure 1.1. The remainder of this section will briefly consider the characteristics of zero, first and

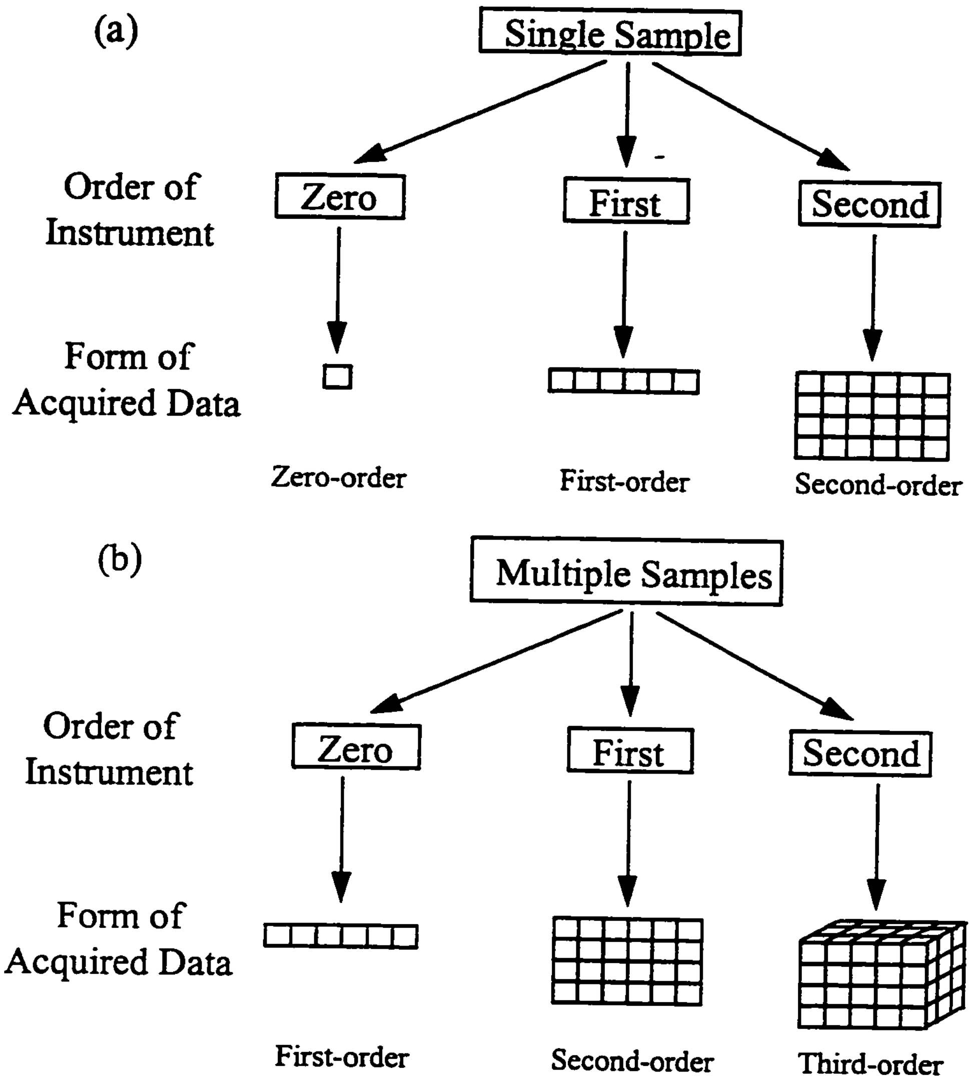


Figure 1.1 Illustration of the relationship between the nature of data and instrumental characteristics

second-order instruments, but the emphasis will be placed on zero and first-order instruments as they have direct applicability to this work.

Zero-order instruments are those which yield a single datum per sample analyzed, such as pH meters and single wavelength spectrometers. When modeling data of this type, a regression model is often proposed that relates a single response to some known property for a particular sample which is given by Equation 1.1:

$$r = f(x) \tag{1.1}$$

where $f(x)$ is the function relating the property, x, such as concentration, to the response, r. It should be noted from Figure 1.1 that when a single sample is analyzed by a zero-order instrument, the form of the data is also zero-order, but when a series of samples are considered, the result is a first-order data set. In general, the order of the data to be analyzed is one greater than the order of the instrument used for the acquisition. When a chemist refers to modeling, two steps are implied: (1) the determination of a mathematical model from known data (calibration step), and (2) the prediction of some property for an unknown sample from new responses (prediction step). This latter step may be accomplished using the inverse form of Equation 1.1:

$$x_{unk} = f^{-1}(r) \tag{1.2}$$

Of course, this implies that the inverse is uniquely defined, which is not always the case.

The widespread popularity of univariate regression arises from the fact that the model building and property estimation are fairly straightforward. Also, models are usually parsimonious (*i.e.* simple) and can be related to fundamental principles. However, the choice of the response to be measured and its relationship to the property of interest is critical. For example, if one were to consider a typical absorbance-calibration experiment, an investigator would choose a single wavelength at which to measure the response for a sample. Obviously, this wavelength must be related to the concentration of the analyte and be free of any interferences from other species in solution. A statistical framework for linear models that may be used for univariate calibration, such as ordinary least-squares (OLS), is given in Chapter 2.

The development of new and innovative instrumentation over the years has led to the availability of more chemically relevant data to be analyzed. Instruments that yield a vector of zero-order data for one sample are called first-order instruments. One example of a first-order technique is chromatography, in which a single channel (zero-order) detector reading is acquired at set time intervals. The spectrometer is another common example of a first-order instrument. Innovations such as diode array technology for UV-visible absorption instruments, charge coupled devices for fluorescence instruments, Fourier transform techniques for infrared spectroscopy, and quadrupoles for mass spectrometry have made first-order data more readily available with such instruments. In the case of chromatography, the vector of data is the chromatogram, while in spectroscopy, the vector is the spectrum. The growing popularity of these methods have led to the development of a number of data analysis techniques (known as multivariate methods) to extract as much information as possible from the multichannel responses. Two traditional methods available to model multivariate data from linear systems such as those described above are classical least-squares (CLS) and inverse least-squares (ILS). Although these methods have some limitations in modeling complex data sets, they are described here because of their historical significance.

Classical least-squares, also known as the K-matrix method, assumes that the responses are related to a proposed model where all the components are known. To illustrate this we will consider a Beer's Law model applied to m solutions containing p components for which measurements are made at n wavelengths. The absorbance for solution i at wavelength j is given by:

$$A_{ij} = \sum c_{ik} \varepsilon_{kj} + e_{ij} \tag{1.3}$$

where ε_{kj} is the molar absorptivity of component k at wavelength j (a path length of 1 cm is assumed), c_{ik} is the concentration of component k in solution i, and e_{ij} is the error associated with the measurement of A_{ij} (this will be addressed in Section 1.4). In matrix form, this can be written as:

$$\begin{bmatrix} A_{11} & \cdots & A_{1n} \\ A_{21} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix} = \begin{bmatrix} c_{11} & \cdots & c_{1p} \\ c_{21} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{bmatrix} \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1n} \\ \varepsilon_{21} & \cdots & \varepsilon_{2n} \\ \vdots & \vdots & \vdots \\ \varepsilon_{p1} & \cdots & \varepsilon_{pn} \end{bmatrix} + \begin{bmatrix} e_{11} & \cdots & e_{1n} \\ e_{21} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots \\ e_{m1} & \cdots & e_{mn} \end{bmatrix} \tag{1.4}$$

or by:

$$\mathbf{A} = \mathbf{CK} + \mathbf{E}_{\mathbf{A}} \tag{1.5}$$

where **K** is a matrix of constants estimated by the direct measurement of pure component spectra or by employing a calibration data set of known concentrations and solving for **K**: $\hat{\mathbf{K}} = \left(\mathbf{C}_{cal}^{\mathrm{T}}\mathbf{C}_{cal}\right)^{-1}\mathbf{C}_{cal}^{\mathrm{T}}\mathbf{A}_{cal}$. Here $\hat{\mathbf{K}}$ is the classical least squares estimate of **K** determined by assuming uniform, normally-distributed, uncorrelated measurement errors in $\mathbf{E}_{\underline{\mathrm{A}}}$. The concentration of unknowns may then be determined by:

$$\hat{\mathbf{C}}_{unk} = \mathbf{A}_{unk}\hat{\mathbf{K}}^{\mathrm{T}}\left(\hat{\mathbf{K}}\hat{\mathbf{K}}^{\mathrm{T}}\right)^{-1} \qquad (1.6)$$

The principal advantages of using first-order data for calibration are: (1) it allows for the simultaneous determination of multiple components, (2) potential interferences can be included in the calibration model (even if they themselves are not of interest), (3) the precision of the analytical result is often improved, and (4) the presence of interferences not included in the calibration set and outliers can be detected [14].

CLS is normally the best approach to use for multivariate calibration when all components within the samples are known, but tends to fail when dealing with complex mixtures containing one or more unknown components. An alternative model to CLS is inverse least-squares, also known as the P-matrix model or multiple linear regression (MLR), which assumes the model:

$$\mathbf{C} = \mathbf{AP} + \mathbf{E}_{\mathrm{c}} \qquad (1.7)$$

where **C** is the (*m*x*p*) concentration matrix and **A** is the (*m*x*n*) absorbance matrix. The matrix **P** is a (*n*x*p*) matrix of regression coefficients to be determined using a calibration set of *m* solutions and is given by: $\hat{\mathbf{P}} = \left(\mathbf{A}_{cal}^{\mathrm{T}}\mathbf{A}_{cal}\right)^{-1}\mathbf{A}_{cal}^{\mathrm{T}}\mathbf{C}_{cal}$. Equation 1.7 is the inverse relationship of the Beer's Law model given in Equation 1.5. This approach has an advantage over CLS in that all of the chemical constituents in the sample need not be known. However, there are a number of drawbacks inherent in ILS [14-15], with the first being a restriction on the number of wavelengths used in the calibration. Because the calculation of **P** requires the inversion of $\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)$, the number of wavelengths must be less than or equal to the total number of samples ($n \leq m$) or else singularity will result. A second problem is the potential for a high degree of collinearity among the chosen wavelengths (*i.e.* absorbances may be linearly related at certain wavelengths) resulting in a matrix which is nearly singular and numerically ill-conditioned. Therefore, application of ILS typically requires the selection of a few wavelength channels. Optimization of this selection is tedious and the reduction in the number of wavelength channels can counteract precision advantages gained in making multichannel measurements.

The central weakness of ILS is in the computation of the so-called pseudoinverse of A, designated as $\mathbf{A}^{+} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}$. A wide range of methods have been developed to address this problem and this list includes techniques such as principal component regression (PCR), ridge regression (RR), partial least squares (PLS) and continuum regression (CR). Some of these techniques are discussed in more detail in Chapter 5.

At present, methods to deal with first-order data sets are more refined than for second-order and higher, but significant advantages in second-order calibration have been made. Examples of second-order bilinear data sets include those obtained from chromatography with multivariate detection (*e.g.* GC-MS, LC-DAD) and fluorescence emission/excitation spectra. Methods applied to these data sets include the generalized rank annihilation method (GRAM) and trilinear decomposition [16-17]. A potential advantage of these approaches is that they can accoũnt for interferences not included in the calibration data. Although the methods presented in this thesis have potential application to second-order instruments, they are currently restricted to the first-order case and higher order methods have been described only for completeness.

1.3 PRINCIPAL COMPONENT ANALYSIS

In the previous section, the difficulty in calculating the pseudoinverse of the matrix A in ILS was introduced. This problem can be resolved by using methods based on principal component analysis (PCA). PCA is a member of a broader group of methods known as factor analysis, originally developed by social scientists, and is a powerful and widely used tool for the analysis of multivariate data sets. The differences between PCA and factor analysis have been discussed by Lawley and Maxwell [18]. PCA has important chemical applications in mixture analysis, exploratory data analysis/pattern recognition, modeling and multivariate calibration [19-22]. The first of these applications attempts to determine the number of components present in an unknown mixture. For exploratory data analysis, PCA attempts to detect distinct classes among the samples. When used in modeling, PCA can determine whether the system being analyzed has a simple underlying physical model. Finally, when applied to multivariate calibration, PCA is used to solve some of the problems associated with CLS and ILS. To achieve these various goals, each of these applications make use of the dimensionality-reducing aspect of PCA.

PCA is usually carried out by a method known as singular value decomposition (SVD). For an *m*x*n* data matrix, **X**, the application of SVD gives:

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} = \mathbf{TL} \tag{1.8}$$

where **U** is *m*x*n*, and **S** and **V** are *n*x*n* (**S** is diagonal). The matrix **T** $\left(= \mathbf{US}\right)$ is called the *m*x*n* scores matrix and the matrix **L** $\left(= \mathbf{V}^{\mathrm{T}}\right)$ is the *n*x*n* loadings matrix, also known as the matrix of eigenvectors of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ or principal components. The scores describe the contribution of each principal component to each sample, while the loadings for a given eigenvector (rows of **L** or columns of **V**) indicate the importance of each of the original variables in defining that eigenvector. The decomposition is carried out so that the first eigenvector will account for most of the variance in the data set, the second will account for the most of the remaining variance, and so on. The eigenvalues (squared elements of the matrix **S**) give the amount of variance that is accounted for by each eigenvector. In general, there is no real advantage to the decomposition if the full matrices are retained. Therefore, a reduction in dimensionality to *p*-dimensions occurs when **U** is reduced to *m*x*p*, **S** to *p*x*p* and **V** to *n*x*p* (**T** is *m*x*p* and **L** is *p*x*n*). It is hoped that by removing the eigenvectors associated with lesser amounts of variance (i.e. the last (*n*-*p*) "factors" or principal components) a minimum amount of information will be lost. The question regarding how many principal components should be included in the truncated model will not be discussed here, but will be addressed in the following chapters.

1.4 ERROR, NOISE AND SOURCES OF VARIATION

In modeling data from chemical systems, the chemometrician generally assumes that there is some true underlying mathematical model (ultimately connected to some chemical principle), and it is the goal of chemometrics to identify this model as completely as possible. In chemistry, the term "model" most often refers to a deterministic (or "functional") model, while in statistics the model description is often a probabilistic (or "structural") one. In attempting to develop the functional model underlying a set of chemical data, it is necessary to realize that all chemical measurements are corrupted with measurement errors. If the true measurement (unknown to the observer) is x°, and the observed measurement is x, then we have:

$$x = x^{0} + e \tag{1.9}$$

where e is the error in the measurement. It is apparent that since the inference of a model is based on measured observations, errors in these observations will lead to errors in the inferred model. Thus, an understanding of the characteristics of the errors is essential in developing methods to estimate models.

The error term, e, in Equation 1.9 is a random variable and by definition cannot be predicted, only characterized. One way to characterize the error is in terms of a non-random bias component, δ, and a random error, or "noise" component, ε, given by:

$$e = \delta + \varepsilon \qquad (1.10)$$

The distinction between bias and noise is often blurred and typically depends on the definition of the measurement. For example, if we consider an absorbance measurement in UV-visible spectroscopy, this can be represented as

$$\mathrm{A} = \mathrm{A}^{\circ} + e = \mathrm{A}^{\circ} + \delta + \varepsilon \qquad (1.11)$$

If a large number of replicate measurements are made it is expected that

$$E(\mathrm{A}) = \mathrm{A}^{\circ} + \delta \qquad (1.12)$$

because the expectation value of e is:

$$E(e) = \delta \qquad (1.13)$$

Obvious sources of random noise in this example include photon shot noise and Johnson noise in the electrical circuits. Another source of random noise is the random variations in the source intensity, known as flicker noise. The positioning of the sample cell is also a potential source of error, since small changes in the orientation can modify the optical characteristics of the instrument. Whether this is considered bias or noise, however, depends on how the replicate measurement is defined. If the replicate involves the removal and replacement of the sample cell, the variation in the cell positioning can be considered to be noise. However, if a replicate measurement is performed by leaving the sample cell stationary, then this can be considered to be bias. This applies to other sources of error as well (*e.g.* blank measurement). Analytical chemists often distinguish bias and noise using the terms accuracy and precision. In principle, bias can be eliminated by a careful experimental design and, in tfiis work, will be assumed to be zero.

Random errors can be characterized by a number of properties, the most common being variance. The variance of *e* is given by:

$$\sigma^2 = E(e^2) = \lim_{N\to\infty}\left[\sum_{i=1}^{\infty} e_i^2 \Big/ N\right] \qquad (1.14)$$

In practice, σ^2 is usually not known but can be estimated from a limited number of measurements and is represented as s^2. Another way to describe random errors is by looking at the distribution of these errors which is represented by a probability density function. The most common distribution assumed for experimental measurements is the normal or Gaussian distribution, which (in the absence of bias) is given by Equation 1.15.

$$f(e) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-e^2/2\sigma^2\right] \qquad (1.15)$$

Other distributions are also found in experimental systems (log normal, Poisson) but in the absence of other information, a normal distribution is usually assumed.

For single measurements, the preceding descriptions are generally sufficient. However, for multiple measurements, such as a series of measurements from a zero-order instrument or a vector of measurements from a first-order instrument, additional descriptions are often used. One such description classifies the noise as homoscedastic or heteroscedastic. Homoscedastic means that all measurements have the same variance while heteroscedastic implies different variances among the measurements, but how these terms apply has to be considered in context. For example, for data obtained on a spectrometer, errors could be homoscedastic within a particular wavelength channel (*i.e.* all samples measured give the same error variance at that wavelength), but heteroscedastic among different channels. If heteroscedastic noise is present, the characteristics of the noise may be further described in terms of the form of the heteroscedasticity. For example, the term "proportional errors" is often used to describe cases where the standard deviation is proportional to the magnitude of the measurement.

Another way to describe the relationship among multiple measurements is through their *covariance*. For measurements x_1 and x_2 (with corresponding errors e_1 and e_2), the error covariance is given by:

$$\sigma_{12} = E(e_1 e_2) = \rho_{12}\sigma_1\sigma_2 \qquad (1.16)$$

where ρ_{12} is the correlation coefficient between e_1 and e_2. For multiple measurements, a general description of errors is provided by the error covariance matrix. For example, if three measurements are considered the error covariance matrix is:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix} \qquad (1.17)$$

Note that this is a symmetric matrix with the diagonal elements being the measurement variances. For uncorrelated measurement errors, the form of the error covariance matrix will be diagonal.

If multiple measurements are related by time, the correlation among adjacent measurements is often described by the noise power spectrum (NPS), which is the expectation value of the Fourier transform of the sequence of errors. Errors which are uncorrelated in time give rise to *white noise* which has an NPS that is constant with frequency. For errors correlated in time, the NPS is not flat and the most common example of this is *pink noise* or 1/*f* noise, which has an NPS that decreases in magnitude with the inverse of the noise frequency. Some examples of pink noise are source flicker noise in spectroscopy and signal drift. Interference noise, such as 60 Hz noise, usually gives regular peaks in the NPS and is another example of correlated noise. Examples of noise power spectra for different types of noise are shown in Figure 1.2. The NPS is also used to describe correlations among errors that are related by a variable other than time (*e.g.* among adjacent wavelength channels in a spectrometer). In this case, the abscissa of the NPS will not correspond to the frequency.

As for a single measurement error, errors in multiple measurements can be described by a multivariate probability distribution. In practice, multivariate probability density functions are difficult to obtain, and a multivariate normal distribution is usually assumed. For a vector of measurement errors, **e**, this is given by:

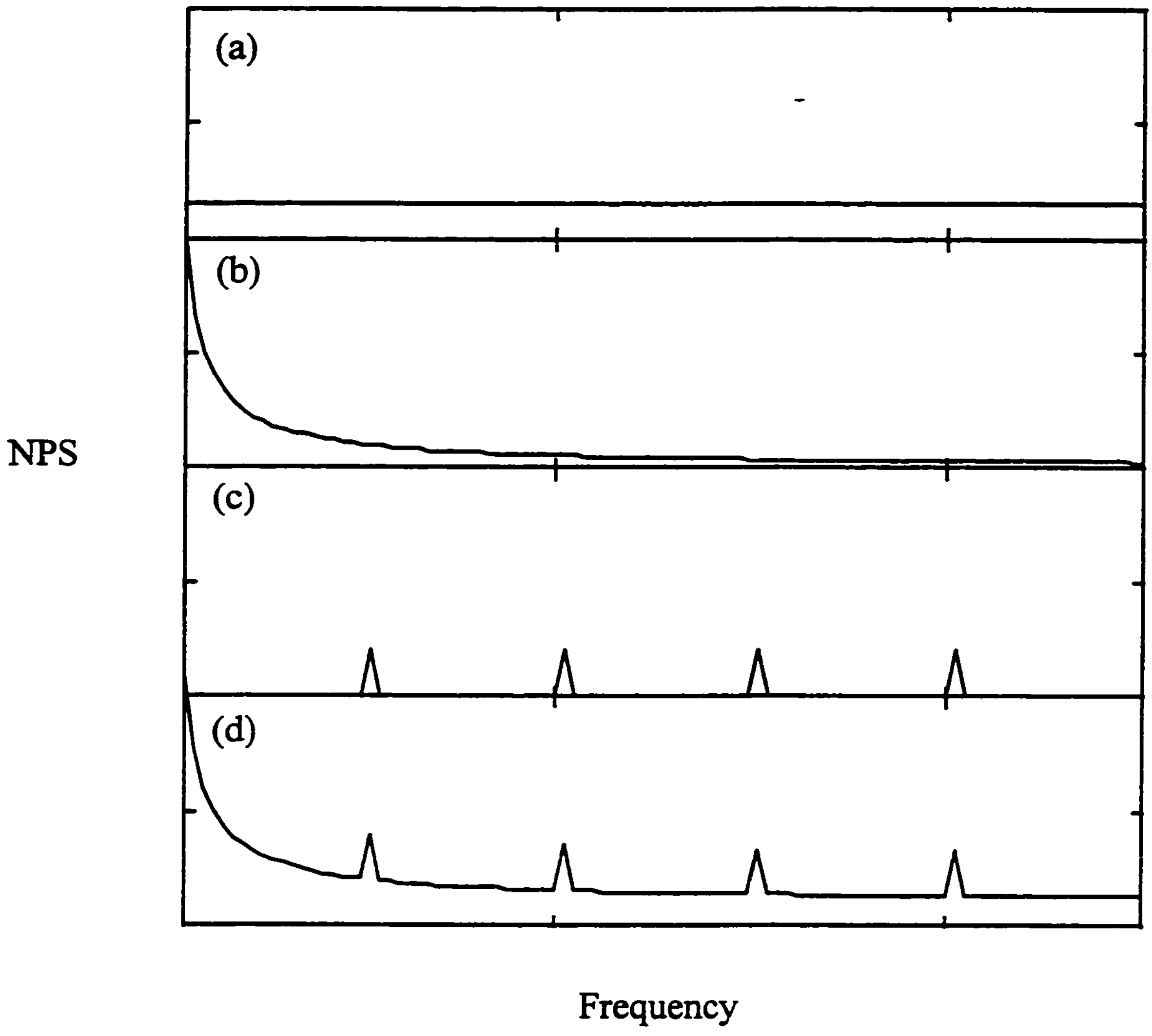


Figure 1.2 Noise power spectra for (a) white noise, (b) 1/*f* noise, (c) interference noise and (d) a combination of all three.

$$f(\mathbf{e}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}\mathbf{e}^{\mathrm{T}}\Sigma^{-1}\mathbf{e}\right] \tag{1.18}$$

where n is the number of response channels (*i.e.* the length of **e**) and Σ is the error covariance matrix. This is the distribution generally assumed in this work.

1.5 RESEARCH OBJECTIVES

Traditional methods of data analysis from univariate and multivariate data sets usually make very little use of the information about measurement errors. Typically, these methods make assumptions about error structures (*e.g.* homoscedastic, independent errors) that are frequently violated. While these assumptions often have only minor effects on the development of a model, this is not always true. The consequences of making invalid error assumptions is particularly important in multivariate data analysis, where the effects of minor variations can be exaggerated and the model can be rendered useless.

What is needed, from both a practical and theoretical perspective, is a method to deal with a variety of error structures in an optimal or near-optimal manner. The objective of this work is to describe new approaches to data analysis based on principles of maximum likelihood estimation. Chapter 2 begins with the simple case of bivariate data sets. Chapters 3 and 4 extend these theoretical principles into higher dimensions for multivariate measurements. Chapters 5 and 6 examine practical applications of these new methods by comparing them with traditional methods and using them to solve new kinds of problems. Finally, Chapter 7 considers some of the implications for future work.

2
MODELING IN TWO DIMENSIONS

2.1 INTRODUCTION

Perhaps no tools are more widely used when analyzing chemical data than regression and principal component analysis (PCA). Least-squares regression methods, including univariate and multivariate methods in both linear (such as CLS and ILS, described in Section 1.2) and nonlinear forms, have been used extensively by chemists for many years. PCA has had a shorter history, but nevertheless has become an indispensable tool for modern multivariate analysis in areas such as exploratory data analysis, modeling, mixture analysis, and calibration. Although regression and PCA have developed from different origins and are generally viewed as serving different purposes, some similarities between the two methods can be drawn, particularly in cases where PCA is used for modeling applications. Throughout this chapter, these similarities will be examined and, perhaps more importantly, differences that are relevant to modeling applications will be identified. In accomplishing this task, the emphasis in this chapter has been placed on bivariate data sets, primarily because of the relative simplicity of the treatment and the ease of visualizing the results in two dimensions.

As a typical example, consider a series of *x,y* pairs that are linearly related, but corrupted by measurement errors in one or both dimensions. To further simplify the problem (and facilitate a direct comparison of least squares and PCA) we will assume that the intercept for the true linear relationship is zero and therefore can be ignored in any modeling approach. For such a data set, two methods that can be used to estimate the true linear relationship are least squares regression (hereinafter referred to as ordinary least squares, or OLS) and PCA. For OLS, the model is described by the estimated slope parameter, while for PCA it is the direction of the first eigenvector that describes the linear model. Generally speaking, these two model estimates will be similar but not identical. For example, the PCA solution is invariant under a reassignment of the axes, whereas the OLS solution is not. On the other hand, the OLS solution is invariant under a scaling of the axes, but the PCA solution is not. The question: "Which method best describes the true model?" can now be posed. In attempting to answer this question, one quickly becomes entangled in the issue of measurement errors and a variety of other modeling methods that have been developed over the years.

For the purpose of demonstration, seven different methods for modeling bivariate linear data are considered: ordinary least squares, weighted least squares, the effective variance method, multiply weighted regression, principal component analysis, and two forms of weighted PCA. It will be shown that these methods can be unified by considering their relationship to maximum likelihood estimation. Furthermore, the equivalence of certain methods under a variety of error conditions will be demonstrated. While some of these results will have immediate utility for simple modeling applications, the principal objective of this chapter is to present a more unified view of these methods and suggest guidelines for their use.

2.2 TECHNIQUES FOR MODELING IN TWO DIMENSIONS

2.2.1 Principal Component Analysis

In its early applications, PCA was primarily used to describe the relationships among random variables in the social sciences. Consider, for example, samples drawn from a bivariate normal distribution in which there is a correlation between the variables. An example might be, say, the height and shoe size of individuals. If all of the sample pairs are plotted in a two-dimensional space, the distribution of measurements might appear as shown in Figure 2.1. In this example, the data have been mean-centered in both variables, designated as **V1** and **V2**. PCA will describe the major axes of the ellipse representing the distribution, both in terms of their direction (as indicated by the eigenvectors or loadings, **E1** and **E2**) and their magnitude (as reflected in the eigenvalues). In this way, PCA characterizes the variance and covariance of the variables in a useful way. Later, applications in the physical sciences became more widespread, but the objectives of such applications were often considerably different. Generally, in the social sciences, the question of measurement errors does not arise, since the variance in the population can be considered to be the dominant factor. In chemistry, where PCA has served a multitude of purposes, there are two principal objectives: dimensionality reduction and modeling.

Dimensionality reduction means that data which were originally represented as objects (or samples) in an n-dimensional variable (or feature) space are represented as samples in a redefined m-dimensional space ($m<n$) with minimal loss of information.

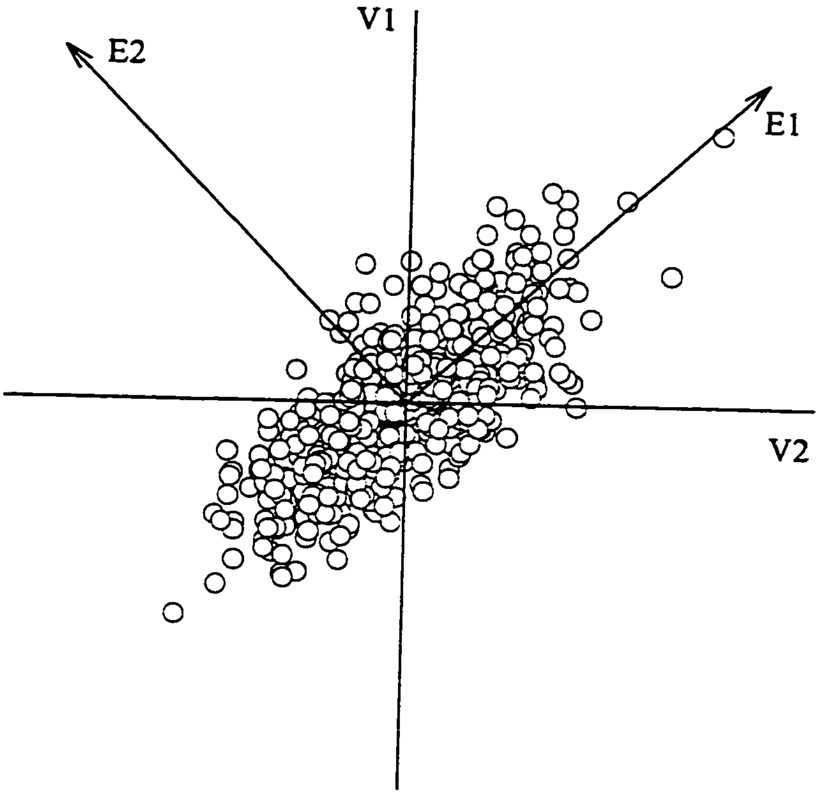


Figure 2.1 Bivariate normal distribution with a correlation between variables V1 and V2. E1 and E2 represent the eigenvectors.

PCA is a very useful tool in this regard, since the *m* eigenvectors, or principal components, account for the largest amount of variance in the data set and in that sense provide the optimum linear combination of the original variables. For this reason, PCA has been widely used for exploratory data analysis in chemistry.

Modeling is also a form of dimensionality reduction, but in this case the objective is not only to describe the data in a space of lower dimensionality, but also to determine the minimum dimensionality needed to reproduce the information within experimental measurement error. Thus, modeling involves two steps: rank estimation and model determination. In this process, the objective is to separate the chemically meaningful variance in the data set from that which is associated with measurement uncertainty. Perhaps the most dominant applications of this type in chemistry are methods for the analysis of mixtures based on second-order data.

The simple case of using PCA to produce a one-dimensional model in a two-dimensional space is analyzed within this chapter. Consider an example in which the absorbance of a solution containing one chromophore is measured at two different wavelengths for ten different concentrations. The result is a 10 x 2 (or a 2 x 10) data matrix. This data set can be visualized either as ten points plotted in a two-dimensional absorbance space, or as two points plotted in a ten-dimensional sample space. Because these two perspectives are ultimately equivalent from the standpoint of PCA, only the former will be considered here. A typical data set is plotted in the absorbance space in Figure 2.2a. As a consequence of Beer's Law, and barring any effects of nonlinearities or interferences, we expect the ratio of the two absorbances to be fixed. The data should

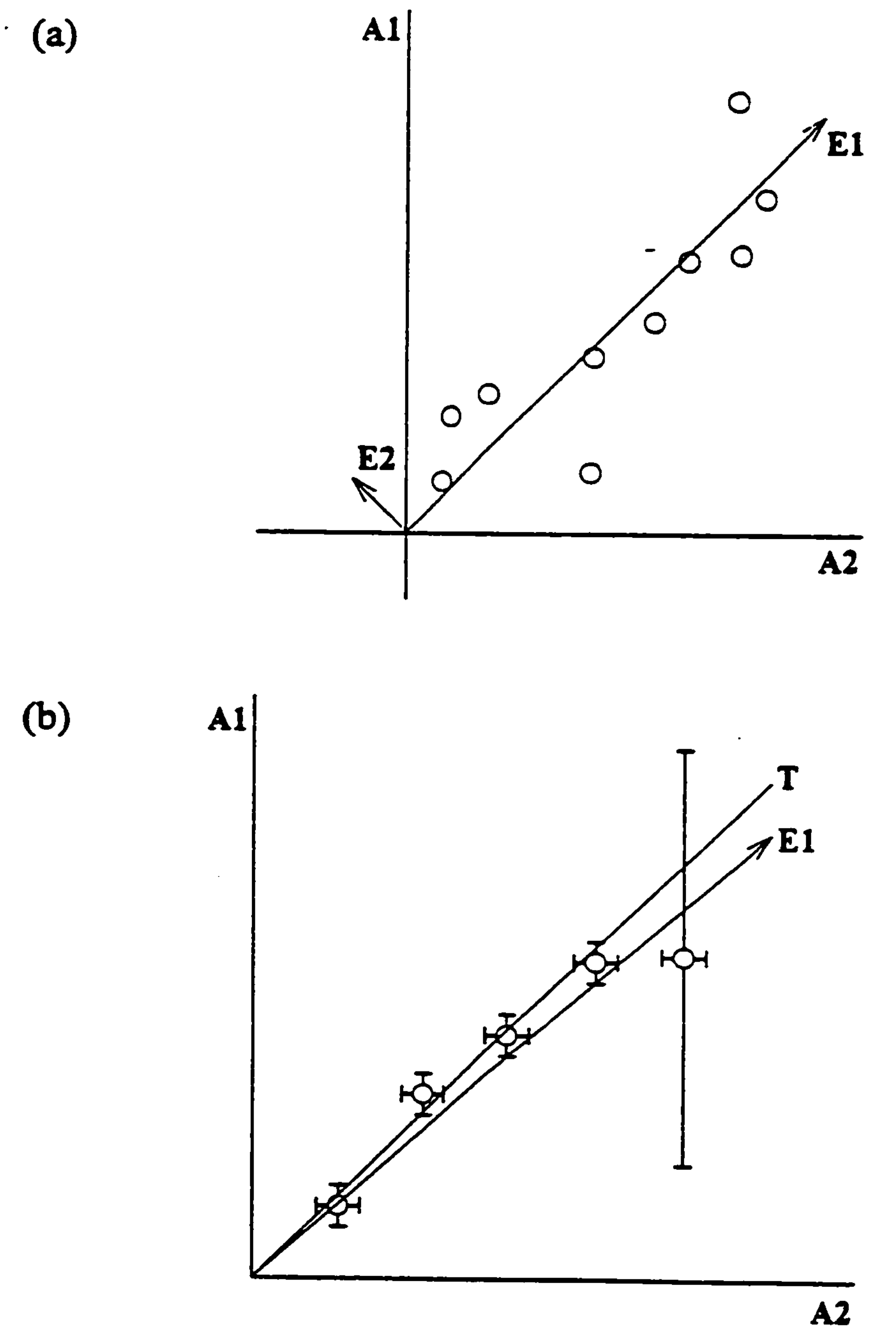


Figure 2.2 (a) PCA results for two-dimensional linear data showing the direction of the eigenvectors, E1 and E2. (b) Effect of large uncertainties on linear modeling by PCA: T=true model, E1=direction of first eigenvector.

therefore fall on a straight line, although there will be some deviations from this due to experimental errors. The directions of the first and second eigenvectors calculated by PCA (**E1** and **E2**) are also shown in Figure 2.2a. The eigenvectors themselves should both be unit length, but for the purposes of illustration, they have been scaled. Note that the first eigenvector accounts for the largest amount of the variance and is predominantly affected by changes in concentration. The slope of this eigenvector will ideally be equal to the ratio of the molar absorptivities at the two wavelengths, but the influence of measurement noise will cause some deviation from this. The second eigenvector will be orthogonal to the first and will account for the remaining variance. Ideally, the residual variance would be entirely attributed to measurement errors and therefore would not contain chemically relevant information, but in practice this is not the case. For this situation, we would say that the data set has an intrinsic dimensionality or a pseudorank of one, since the important information is described by the first eigenvector. In many chemical applications, such as mixture analysis, estimation of the pseudorank (or chemical rank) of a data set is a principal objective. In such applications, the usual procedure is to reproduce the data matrix with progressively more factors until it is reconstructed within experimental error. Obviously, the role of measurement errors is of critical importance in such applications. For this chapter, the concern is not with the problem of rank estimation as with finding the best parameters for the rank one model, although the two are ultimately related. A procedure to determine the true rank of a multidimensional data set, while accounting for experimental error, will be discussed in Chapter 4.

The data in Figure 2.2a have not been mean-centered, but an intercept of zero has been assumed. The eigenvectors always extend from the origin and a non-zero intercept will generally increase the rank estimate by one. Often the intercept can be forced to zero by mean-centering, but this will not always be the case if measurement errors are not uniform. This will be further discussed in Section 3.4.2. For this reason, results reported here are for non-mean-centered data.

It should also be noted that the arguments presented here for two-dimensions also extend to higher dimensions. If more wavelengths were used for one-component mixtures, the first eigenvector would still retain essentially all of the significant information. Alternatively, if two-component mixtures were used, the plane described by the first two eigenvectors would contain the relevant information.

Measurement errors can play a very important role in the application of PCA to chemical data sets, both in problems of dimensionality reduction and modeling. This is illustrated with the following two examples.

In modeling applications, the failure of PCA to develop accurate models under certain measurement error conditions is readily illustrated. Figure 2.2b demonstrates this for the case where absorbance measurements are made at two wavelengths for several samples and the error associated with one measurement is inordinately large. Because of the leverage associated with this sample, the error in the first eigenvector compared with the true model (*i.e.* the ratio of absorbances at the two wavelengths) is quite large. Under these circumstances, recourse to some sort of weighted regression method is suggested.

In exploratory data analysis, one is often interested in visually determining if classes can be distinguished when samples are projected into a space of lower dimensionality. PCA is a way to do this, since projection onto the first two or three eigenvectors accounts for the largest amount of variance and this often produces the best separation. One of the contentious issues in this application is whether or not variables should be scaled prior to PCA. The argument for scaling is that variables with radically different ranges (sometimes orders of magnitude) will lead to eigenvectors skewed towards the variables with largest ranges. On the other hand, scaling can also lead to problems, as illustrated by the following hypothetical case. Consider two measurements made on a series of samples which belong to two classes. Measurement one is capable of distinguishing the two classes and is about two orders of magnitude larger than measurement two, which consists of pure noise. Figure 2.3a shows the data (mean-centered) in the two-dimensional measurement space, and Figure 2.3b shows the projection of the data onto the measurement one axis. It is obvious that the two classes are separated in the one-dimensional space. Figure 2.3c shows the projection onto the first-eigenvector when no scaling of the original data is used and it is clear that the two classes are still separated. Figure 2.3d shows the corresponding projection when both variables are autoscaled (scaled to unit variance). In this case, the random contribution of the second variable means that the two classes are no longer completely separated when the samples are projected into one dimension. This weakness of autoscaling can be extrapolated into higher dimensions, but there are many counter-examples which show

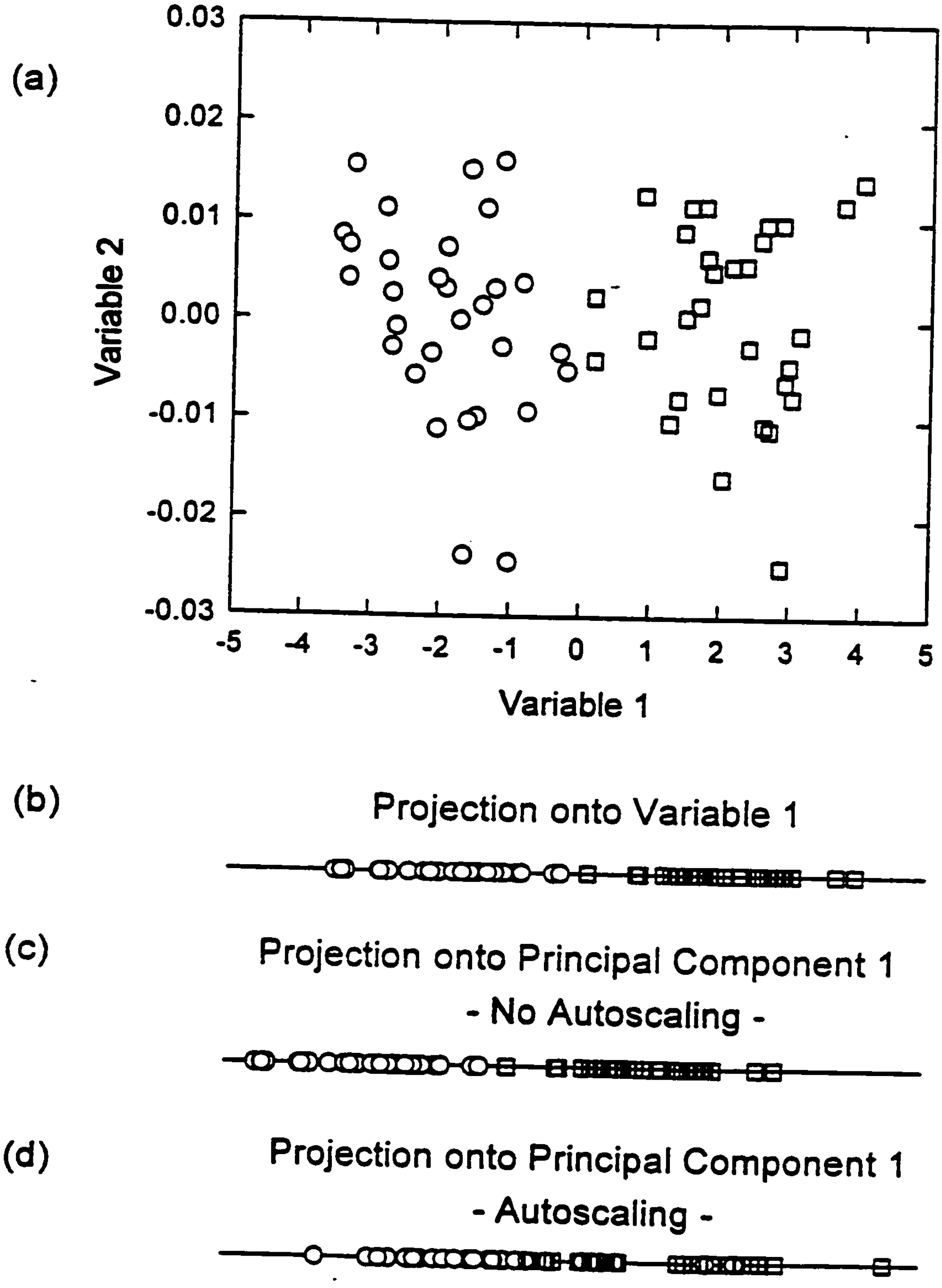


Figure 2.3 A two-dimensional pattern recognition example. (a) Original data plotted in measurement space, (b) projection of data onto first measurement axis, (c) projection onto first eigenvector obtained without autoscaling, (d) projection onto first eigenvector after autoscaling.

the benefit of autoscaling. It is apparent that the key to correct scaling lies in estimates of the measurement error and a more general model is necessary to take this into account.

In summary, the objective of PCA in most chemical applications is to efficiently extract information from the variance in multiple dimensions. A major weakness of this approach, however, is that it makes implicit assumptions about measurement errors which are often incorrect. This corrupts the quality of information provided and may lead to erroneous results. It is clear, then, that a more general method of dealing with measurement errors is required.

2.2.2 Linear Regression

Returning to Figure 2.2a (or Figure 2.1), it might be asserted that an approximation to the first eigenvector could be obtained by a least-squares fit to a straight line with the equation,

$$y = a\,x \tag{2.1}$$

followed by appropriate normalization. Note that there is no intercept included here because the eigenvector is defined to extend from the origin. A question then arises regarding the distinction between PCA and unweighted linear regression (or ordinary least squares, OLS). These methods produce similar but usually not identical results, and it is known that regression of y on x does not generally produce the same result as regression of x on y, whereas the first eigenvector is invariant (relative to the original frame of reference) under such an exchange of axes. Both methods minimize a sum of squared residuals, but the definition of the residual changes as shown in Figure 2.4. For OLS, the residual, ε_i, is calculated in the vertical direction:

$$\varepsilon_i^{OLS} = y_i - y_i^{OLS} \tag{2.2}$$

where y_i is the observed response and y_i^{OLS} is the pǫint on the line corresponding to x_i. In contrast, PCA seeks to minimize the sum of the squares of the residuals orthogonal to the first eigenvector:

$$\varepsilon_i^{PCA} = \sqrt{\left(x_i - x_i^{PCA}\right)^2 + \left(y_i - y_i^{PCA}\right)^2} \tag{2.3}$$

where x_i and y_i are the observed data and x_i^{PCA} and y_i^{PCA} are the coordinates of the orthogonal projection of the point onto the first eigenvector.

2.2.3 Maximum Likelihood Estimation

For the general case, we imagine that we have *N* experimental (observed) data points (*x,y* pairs), each of which has an associated standard deviation (σ_x, σ_y). For the development of the method, we will assume that the measurement errors are random, uncorrelated, and normally distributed. Although the assumption of independent, normally distributed errors is not universally valid, it is commonly made for the analysis of chemical data sets and will be used throughout this work to simplify the discussion. Further, it will be assumed that there is a true (unknown) linear relationship of the form:

$$\hat{y} = ax + b \tag{2.4}$$

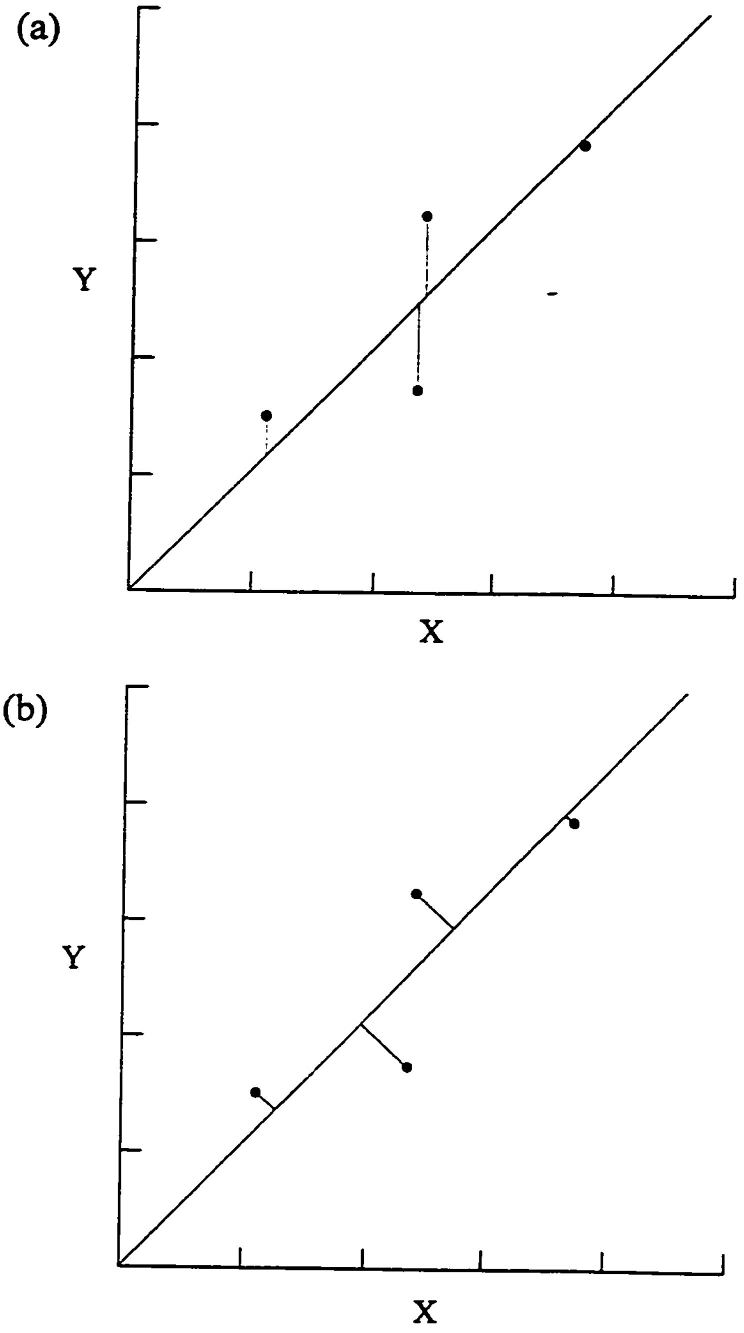


Figure 2.4 Graphical representation of residuals for (a) OLS and (b) PCA.

where $\hat{y}$ is the y value estimated from the model for a given x, and a and b are the slope and intercept parameters estimated for the model.

Using these assumptions, the principle of maximum likelihood estimation is illustrated in Figure 2.5. The line through the points represents a trial solution for the desired model. For each experimentally observed pair of measurements (represented by the X's in the lower part of the figure) we will assume that there is a corresponding "true" point (x_i°, y_i°) that lies on the line defining the trial model. Of course, we do not know where these "true" points are, or even if we have the correct model, but we use the maximum likelihood principle to guide our estimates of them. Centered around each of the "true" points on the line is a bivariate probability distribution associated with the errors in both dimensions, as indicated by the mesh plots in the top part of the figure and the contour lines in the lower part of the figure. For the maximum likelihood solution, the positions of the "true" points along the line are first adjusted so that the observed values are at the highest point on the bivariate probability density functions. The points on the line are then called the maximum likelihood estimates of $(x_i^\circ,\ y_i^\circ)$ for each pair of x_i and y_i, and their coordinates will be designated $\hat{x}_i$ and $\hat{y}_i$. The maximum likelihood solution is obtained by changing the slope of the trial solution to minimize the objective function, S^2, which is given by:

$$S^2 = \sum_{i=1}^{N}\left[\frac{\left(x_i - \hat{x}_i\right)^2}{\sigma_{x_i}^2} + \frac{\left(y_i - \hat{y}_i\right)^2}{\sigma_{y_i}^2}\right] \tag{2.5}$$

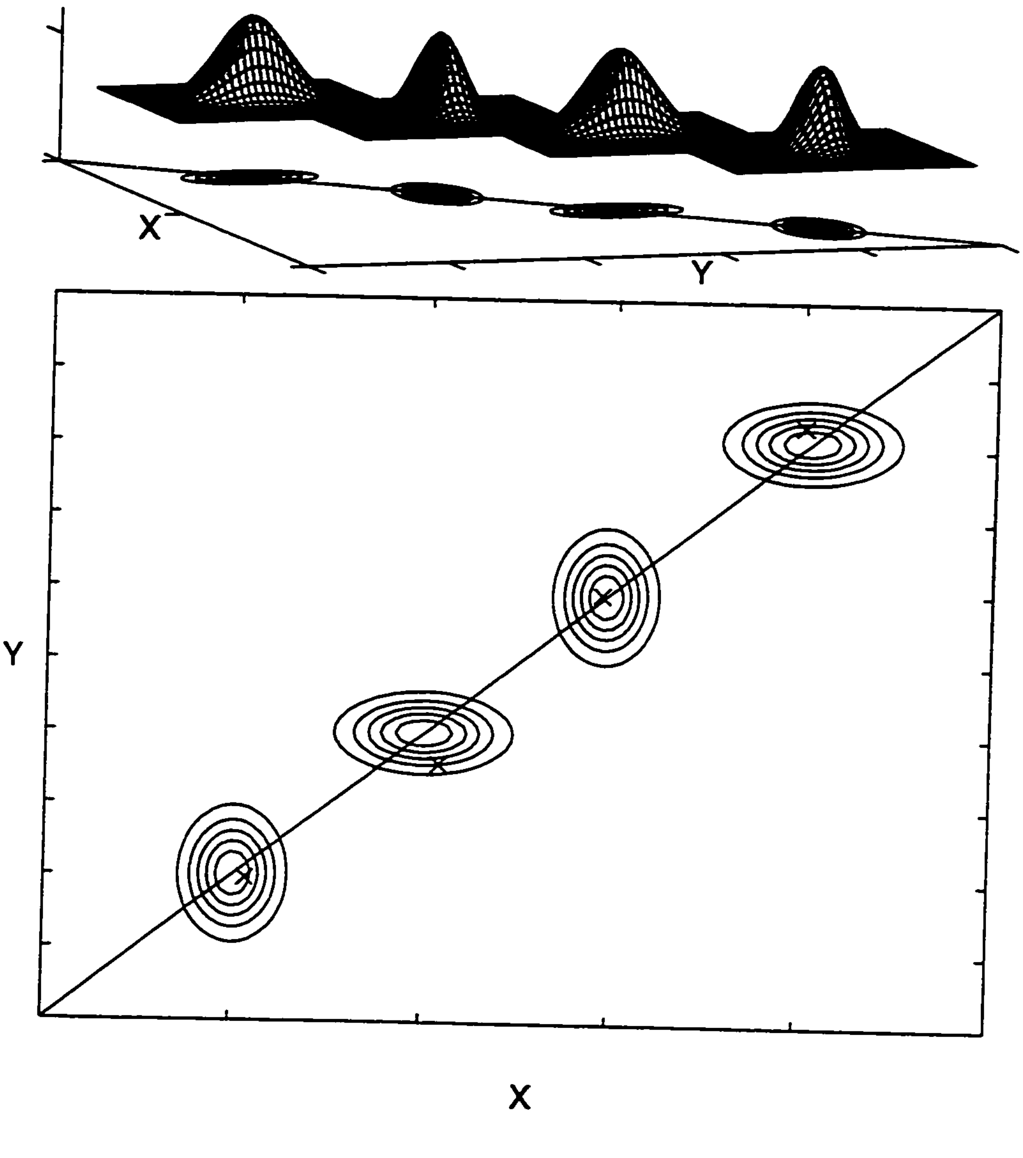


Figure 2.5 Illustration of maximum likelihood estimation. The surfaces and contours show the bivariate probability density functions around the maximum likelihood estimates. The X's show the observed measurements.

Minimizing this function maximizes the joint probability density function for the observed points and therefore is called the maximum likelihood solution. Note that this equation is similar to the goodness of fit (X^2) that is normally minimized in regression problems, except that it considers errors in both variables and also requires maximum likelihood estimates of the "true" points. Mathematically, the procedure is as follows. First, we define the coordinate system to have its origin at ($\hat{x}$, $\hat{y}$) as shown in Figure 2.6. With this definition, moving the maximum likelihood point along the line described by the model (l_1 in Figure 2.6) is equivalent to moving the experimental point along a parallel line (l_2), which is offset by an amount equal to ($y-\hat{y}$) when $\hat{x}$ $= x$. In this new coordinate system, we are actually determining Δx and Δy directly, where $\Delta x = (x\text{-}\hat{x})$ and $\Delta y = (y-\hat{y})$. The equation of the line containing the experimental point is:

$$\Delta y = a\,\Delta x + c \tag{2.6}$$

Note that the slope of this line is the same as for the trial solution, but the intercept is different in the new coordinate system and will be given by:

$$c = y - (ax + b) \tag{2.7}$$

The bivariate probability density (likelihood) function around ($\hat{x}$, $\hat{y}$) is given by:

$$L = k\,e^{-\frac{1}{2}\left[\frac{(x-\hat{x})^2}{\sigma_x^2} + \frac{(y-\hat{y})^2}{\sigma_y^2}\right]} \tag{2.8}$$

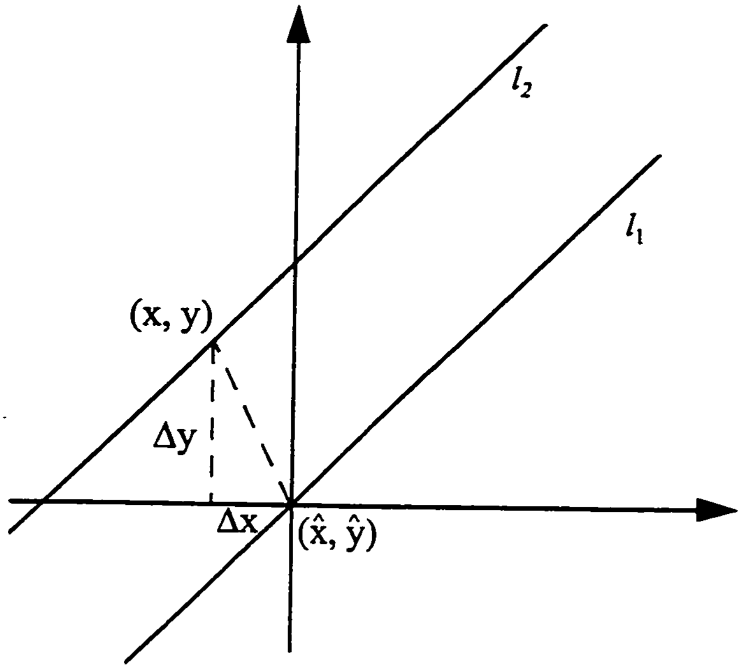


Figure 2.6 The geometry of maximum likelihood estimation for a single measurement. See text for details.

where k is a normalization constant. Our objective is to find the $\hat{x}$ and $\hat{y}$ that maximize L subject to the constraint of Equation 2.4. Normally, the negative logarithm of the likelihood function is used for simplification. Substituting for $\hat{y}$, taking the negative logarithm and dropping the constant term gives:

$$l = \frac{1}{2}\left\{\frac{(x-\hat{x})^2}{\sigma_x^2} + \frac{(y - a\hat{x} - b)^2}{\sigma_y^2}\right\} \tag{2.9}$$

Minimizing Equation 2.5 is the same as maximizing L. Differentiating with respect to $\hat{x}$ and setting the result to zero gives:

$$\frac{(x-\hat{x})}{\sigma_x^2} + \frac{a(y - a\hat{x} - b)}{\sigma_y^2} = \frac{\Delta x}{\sigma_x^2} + \frac{a\Delta y}{\sigma_y^2} = 0 \tag{2.10}$$

Substitution of Equation 2.6 and rearrangement provides the results:

$$\Delta x = -\frac{ac\sigma_x^2}{\sigma_y^2 + a^2\sigma_x^2} \tag{2.11}$$

$$\Delta y = \frac{c\sigma_y^2}{\sigma_y^2 + a^2\sigma_x^2} \tag{2.12}$$

Equations 2.11 and 2.12 allow calculation of $\hat{x}$ $(= x - \Delta x)$ and $\hat{y}$ $(= y - \Delta y)$ for this point given the trial model parameters. The procedure is repeated for each of the N points and the objective function is calculated. When the objective function is minimized, the trial model is the maximum likelihood solution. Although maximum likelihood model estimates are often preferred, the method of solution is not as straightforward as for PCA and other regression methods, so they tend to be underused.

It is important to note that OLS and PCA do yield maximum likelihood estimates when certain conditions arise. OLS is consistent with a maximum likelihood solution if the errors in x are negligible with respect to the errors in y, and the errors in y are homoscedastic and normally distributed. ("Homoscedastic" indicates that the errors for all measurements of a given variable have the same standard deviation). PCA produces the maximum likelihood solution if the errors for all measurements are independent and identically distributed (*iid*); that is, they are homoscedastic in x and y, and $\sigma_x=\sigma_y$ for all measurement pairs. This means that the contours in Figure 2.5 are vertical lines for OLS and circles for PCA. While the assumptions made for both PCA and OLS may be valid in many cases, it is clear that they lack generality for dealing with measurement errors.

2.2.4 Weighting Methods in PCA and Regression

A number of methods have been developed to provide a more reliable treatment of errors in both regression and PCA. The simplest approaches scale or weight the data in an attempt to change the error structure (without altering the form of the underlying model) and reduce the influence of measurements with large errors. In general, the goal is to turn conventional solutions into maximum likelihood solutions by transforming the measurements. Some of the more common weighting techniques will be summarized in this section. In order to distinguish these, the unweighted methods will be referred to as UPCA (for unweighted PCA) and OLS.

From a regression perspective, most scientists are familiar with weighted least squares (WLS) [23]. This method minimizes the goodness of fit, given by,

$$X^2 = \sum_{i=1}^{N} \frac{\left(y_i^{calc} - y_i^{obs}\right)^2}{\sigma_{y_i}^2} \tag{2.13}$$

This procedure produces a maximum likelihood estimate when errors in y are not homoscedastic, but still requires that the errors in x are negligibly small. For cases where there are errors in both variables, Riu and Rius provide a thorough discussion of univariate regression methods in a recent review [24]. One of these methods is the effective variance method (EVM) [25], which is a form of iteratively reweighted least squares. With this method, errors in x are propagated to y using the relation,

$$\left(\sigma_y'\right)^2 = \sigma_y^2 + a^2\sigma_x^2 \tag{2.14}$$

In this equation, σ_x and σ_y are the standard deviations in x_i and y_i, a is the slope of the straight line model, and σ_y' is the total error propagated to y_i. Once the error has been propagated this way, WLS can be used. Of course, an initial estimate of a, a_o, is required, so this method is carried out iteratively, and continues until convergence is achieved. Although EVM does incorporate the errors in x and is relatively simple to implement, the results are not consistent with a maximum likelihood solution [26].

For bivariate data sets, maximum likelihood estimation of the desired parameters is accomplished through a technique which will be referred to here as multiply weighted regression (MWR). This approach minimizes the objective function, S^2, described in the previous section. Implementation of the method was outlined by York in 1966 [27], with subsequent improvements by Williamson [28] and others [29-30]. This study of the bivariate case utilizes the regression procedure detailed by Lybanon [31].

A variety of weighting methods have also been used for PCA. Generally these involve scaling of the data prior to analysis by PCA. In an excellent analysis of these scaling procedures, Paatero and Tapper [32] showed that, in order for such scaling to be optimal in a maximum likelihood sense, the matrix of standard deviations associated with the measurements needs to be of rank one. While this will not be true in the general case, it will be realistic in certain situations. Paatero and Tapper demonstrate, for example, that scaling by the norm for each variable (*i.e.* autoscaling) will be optimal only if the uncertainty in each variable is a constant proportion of the variable norm. For this study, two approaches to scaling will be examined. One of the simplest and most popular approaches is to divide each measurement in the data matrix by its associated standard deviation [33]:

$$d'_{ij} = \frac{d_{ij}}{\sigma_{ij}} \tag{2.15}$$

PCA is then carried out on **D'**. For this study, this method will be referred to as WPCA1. Because this approach does not impose any structure on the matrix of standard deviations, it is easy to see that in the general case it will destroy the inherent structure in the data matrix. However, for certain error structures, it will be shown that this pretreatment can be effective.

Another method for weighting data matrices prior to PCA was described by Simeon and Pavkovic in 1992 [34]. With this method, each row of the data matrix is divided by a weight value, W_i, calculated according to

$$W_i = \left(\frac{w_i}{\sum_{i=1}^{r} w_i} \right)^{\frac{1}{2}} \tag{2.16}$$

where,

$$w_i = \left(\prod_{j=1}^{c} \sigma_{ij}^2 \right)^{-\frac{1}{c}} \tag{2.17}$$

In these equations, r and c are the number of rows and columns in the data matrix, respectively. If each row of the data matrix is considered to represent a point in the original data space, that point is weighted in inverse proportion to the geometric mean of the variances of the coordinates of that point. Therefore, the greater the error, the smaller the influence of the point. This method will be referred to as WPCA2.

Other scaling methods have also been described in the literature. One of these was employed by Cochran and Horne in the analysis of spectroscopic data from rapid scanning kinetics experiments [35]. Their method assumed that the error structure could be represented by a linear relationship of the form:

$$\sigma_{ij}^2 = x_i z_j \tag{2.18}$$

where x_i is a function of the wavelength and z_i is a function of the spectrum number. Another approach is the “balanced scaling” method of Paatero and Tapper [32]. This is a heuristic technique which attempts to find the best scaling for cases where the rank of the matrix of standard deviations is greater than unity. This method produces optimal scaling (*i.e.* the maximum likelihood solution) when this matrix is of rank one. Since the conditions for maximum likelihood estimation are clearly defined for both of these methods, they were not included in this study.

In higher dimensions, analogs to MWR have been developed that produce maximum likelihood solutions for PCA. These include the criss-cross regression method of Gabriel and Zamir [36], the positive matrix factorization method of Juntto and Paatero [37], and the maximum likelihood PCA method that will be introduced in the next chapter. Because these methods will produce results equivalent to those from MWR for bivariate data sets, they are not considered separately here.

Under particular error conditions, many of the methods described here can be considered equivalent to one another. It is the objective of this chapter to compare these methods for two dimensional data sets and to make recommendations regarding methods of data pretreatment and analysis.

2.3 EXPERIMENTAL

All of the results presented in this chapter were obtained using simulated data. This is necessary because the large number of data sets required to make statistically valid conclusions precludes the use of experimental data. Simulations were carried out using Matlab v. 4.2c.1 for Windows (Mathworks, Natick, MA) on a P5-based personal computer.

A variety of simulated data sets were examined, but in all cases the true model was linear with a zero intercept in accordance with Equation 2.1. There are a number of parameters that may influence the fit of a line to a given set of data. The slope, number of points used, and the magnitude and nature of the errors were varied in the studies carried out, but it was primarily the effect of experimental errors that was the subject of this study. To minimize the effect of the slope, the line being estimated was kept at a constant length.

There were five error structures examined in this study. In all cases, experimental errors were simulated using random numbers drawn from a normal distribution. In the simplest case, homoscedastic equal errors (HE), the standard deviations in x and y were taken to be equal to each other and the same for all points. The magnitude of σ is specified as a percentage of the length of the line being fit. For the homoscedastic unequal case (HU), all of the standard deviations in x and y are the same, but σ_x does not equal σ_y. The level of error is specified by a percentage of the length of the line being fit for x and by the desired ratio of $\sigma_y/\sigma_x=R$ for y. For proportional equal errors (PE), σ_x and σ_y are given as a constant percentage of the displacement of the true points along the vector from the origin. For proportional unequal errors (PU), σ_x is calculated by the procedure outlined for PE errors, and σ_y is determined by again using the specified ratio of the standard deviations, R. Finally, for random errors (RA), σ_x and σ_y for each point are determined by drawing a random number from a uniform distribution between zero and a percentage of the length of the line being fit. These five cases are intended to represent cases which might arise for a typical chemical data set, with the last encompassing the most general situation. The five cases are summarized pictorially in Figure 2.7. The known standard deviations (as opposed to measured values) for the data sets were used in the data analysis. Experimentally measured standard deviations can themselves possess a high level of uncertainty [38],

$$RSD(s) \approx \frac{1}{\sqrt{2(N-1)}} \tag{2.19}$$

but this source of variability was not explored in this study.

Obviously, any study like this cannot possibly encompass all possible scenarios, so the simulations were carried out to represent a large range of reasonable circumstances. The slopes employed were 0.1, 0.5, 2, and 10, and ratios of standard deviations (R) were 0.01, 0.2, 5, and 100. Error levels, as defined above, were 4% and 10% of the line length. In all cases, 10 simulated data points were generated, with the

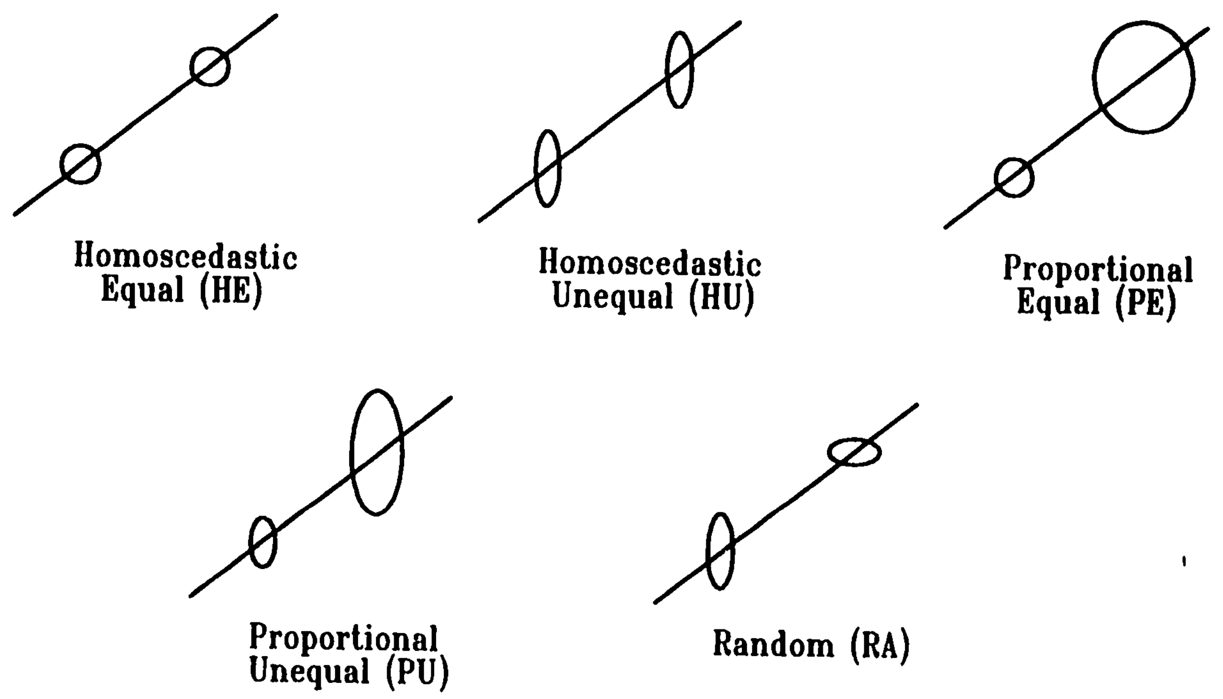


Figure 2.7 Pictorial representation of the error conditions examined in this work.

noise-free values located at equal intervals along the length of the line and no point at the origin. Because the objective of this study is to compare regression methods and PCA, no intercept term is included.

2.4 MATHEMATICAL EQUIVALENCE OF METHODS

In this section, the seven methods and five error structures under consideration are compared in several ways. To simplify the study, the equivalence of various methods is examined from a mathematical perspective under the different error conditions.

Each of the methods treated here tries to minimize some criterion that is generally known as the objective function. Table 2.1 presents the form of the objective function for each of the methods used in this study when different noise types are imposed. For the general case, the expressions for OLS and WLS can be found in standard introductory textbooks on regression, and the equation for EVM can be derived from these with the weights adapted accordingly. The expressions for MWR and UPCA can be readily derived and have been given in the literature [39-40]. The weighted PCA equations can be obtained from appropriate extensions of the UPCA equation.

Table 2.1 Forms of the objective function under specific error conditions.

	General (Random)	Homoscedastic Equal	Homoscedastic Unequal	Proportional Equal	Proportional Unequal
OLS	$\sum_i (y_i - ax_i)^2$	$\sum_i (y_i - ax_i)^2$	$\sum_i (y_i - ax_i)^2$	$\sum_i (y_i - ax_i)^2$	$\sum_i (y_i - ax_i)^2$
WLS	$\sum_i \frac{(y_i - ax_i)^2}{\sigma^2_{y_i}}$	$\sum_i (y_i - ax_i)^2$	$\sum_i (y_i - ax_i)^2$	$\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$	$\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$
EVM	$\sum_i \frac{(y_i - ax_i)^2}{\sigma^2_{y_i} + a_0^2\sigma^2_{x_i}}$	$\sum_i (y_i - ax_i)^2$	$\sum_i (y_i - ax_i)^2$	$\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$	$\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$
MWR	$\sum_i \frac{(y_i - ax_i)^2}{\sigma^2_{y_i} + a^2\sigma^2_{x_i}}$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{R^2+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$	$\frac{1}{R^2+a^2}\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$
UPCA	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$
WPCA1	$\frac{1}{1+a^2}\sum_i \left(\frac{y_i}{\sigma_{y_i}} - a\frac{x_i}{\sigma_{x_i}}\right)^2$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{R^2+(aR)^2}\sum_i (y_i - aRx_i)^2$	$\frac{1}{1+a^2}\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$	$\frac{1}{R^2+(aR)^2}\sum_i \frac{(y_i - aRx_i)^2}{x_i^2}$
WPCA2	$\frac{1}{1+a^2}\sum_i \frac{(y_i - ax_i)^2}{\sigma_{x_i}\sigma_{y_i}}$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i (y_i - ax_i)^2$	$\frac{1}{1+a^2}\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$	$\frac{1}{1+a^2}\sum_i \frac{(y_i - ax_i)^2}{x_i^2}$
Mathematically Equivalent Methods	None	OLS=WLS=EVM MWR=UPCA =WPCA1=WPCA2	OLS=WLS=EVM UPCA=WPCA2 MWR=WPCA1*	WLS=OLS MWR=WPCA1 =WPCA2	WLS=EVM MWR=WPCA1*

It can be seen from Table 2.1 that when certain noise structures arise, methods that generally differ can become mathematically equivalent. For example, when the errors are homoscedastic equal (HE), all of the linear regression methods minimize the same objective function. This can also be seen for MWR and the principal component methods. For other noise structures, the equivalence among methods is not as extensive, but is useful nonetheless. It should be noted that for the cases of HU and PU errors, the minimization criteria show that the MWR estimate of the slope can be obtained from the WPCA1 estimate of slope by multiplying the latter by *R*. (We denote this method WPCA1*). This equivalence is expected, since the scaling introduced by WPCA1 effectively reduces the problem to HE and PE conditions, but also alters the slope of the line. These results are also consistent with the analysis of Paatero and Tapper [32] and Cochran and Horne [35]. For all of the error conditions other than RA, the matrix of standard deviations can be represented as the outer product of two vectors and will have a rank of unity. Therefore, it should be possible to provide the maximum likelihood solution through simple scaling. Note, however, that WPCA2 only provides results equivalent to MWR for the cases where $\sigma_x = \sigma_y$, and for the fairly common case of HU errors, it is identical to unweighted PCA.

The equivalence of methods expressed in Table 2.1 simplifies the task of comparison since redundant methods may be omitted when certain noise cases are examined. Furthermore, the cases where the results of MWR are identical to other methods are particularly important since they simplify extension of the maximum likelihood method to higher dimensions. This extension will be the focus of subsequent chapters.

2.5 COMPARISON OF METHODS

A study similar to the current investigation was carried out recently by Kalantar *et al* [41], but with several important differences. First, these authors compared only OLS and MWR in their study and did not include PCA methods or WLS. Second, because the cited study focused on implications for regression, both slope and intercept terms were included. Because the objective of the present study is to compare regression methods and PCA, no intercept term is included. Finally, Kalantar *et al* concentrated on examining the bias of the methods, which is not necessarily the most important criterion from a chemist's point of view.

The utility of a particular modeling method relates to how closely the estimated model compares to the true model. This can be evaluated by Monte-Carlo methods employing a large number of data sets with the same noise characteristics and examining the distribution of estimated slopes around the true slopes. To assess performance, measures of central tendency and dispersion can be used, but, as indicated by Kalantar *et al* [41], care must be taken in the parameter to be assessed. For this study, the estimated slope would be considered a poor parameter choice because it is a biased estimator of the true model slope, even when there is a geometrically symmetric distribution of estimated models around the true model. For example, an angular change of, say $+10°$, around the true line will not correspond to an equal change in slope in both directions unless the line is coincident with one of the axes. Thus the distribution of slopes around the true value will be skewed. This is because there is a nonlinear relationship between the slope and the angle that the model makes with the axes ($\theta = \tan^{-1} a$). It is expected that a symmetric distribution of models about the true model would be reflected by a symmetric distribution of angles. Therefore, it is the angular distribution about the true model ($\hat{\theta} - \theta_{true}$) that is evaluated in this study.

Initial studies were carried out to examine the angular distributions about the true model for the various cases of noise and modeling methods used in this chapter. Figure 2.8 shows typical results for the case of purely random error (RA) at a noise level of 4%. In this example, ten coordinate pairs (equidistant along a line of length 25) and a slope of 0.5 were used to generate the noise free data. Results from other cases were similar. Distributions, based on 10,000 data sets, were generated with each of the seven modeling methods included in this study (OLS, WLS, MWR, EVM, UPCA, WPCA1, and WPCA2). The distributions given in Figure 2.8 are shown relative to the true slope, which therefore corresponds to an angular deviation of zero. In general the distributions appear fairly symmetric about the true model, with the exception of WPCA1. (It should be noted that under other error conditions examined (HE, HU, PE, PU), the distributions for WPCA1 were in fact symmetric.) The angular deviations were therefore considered to be useful for comparison of the methods.

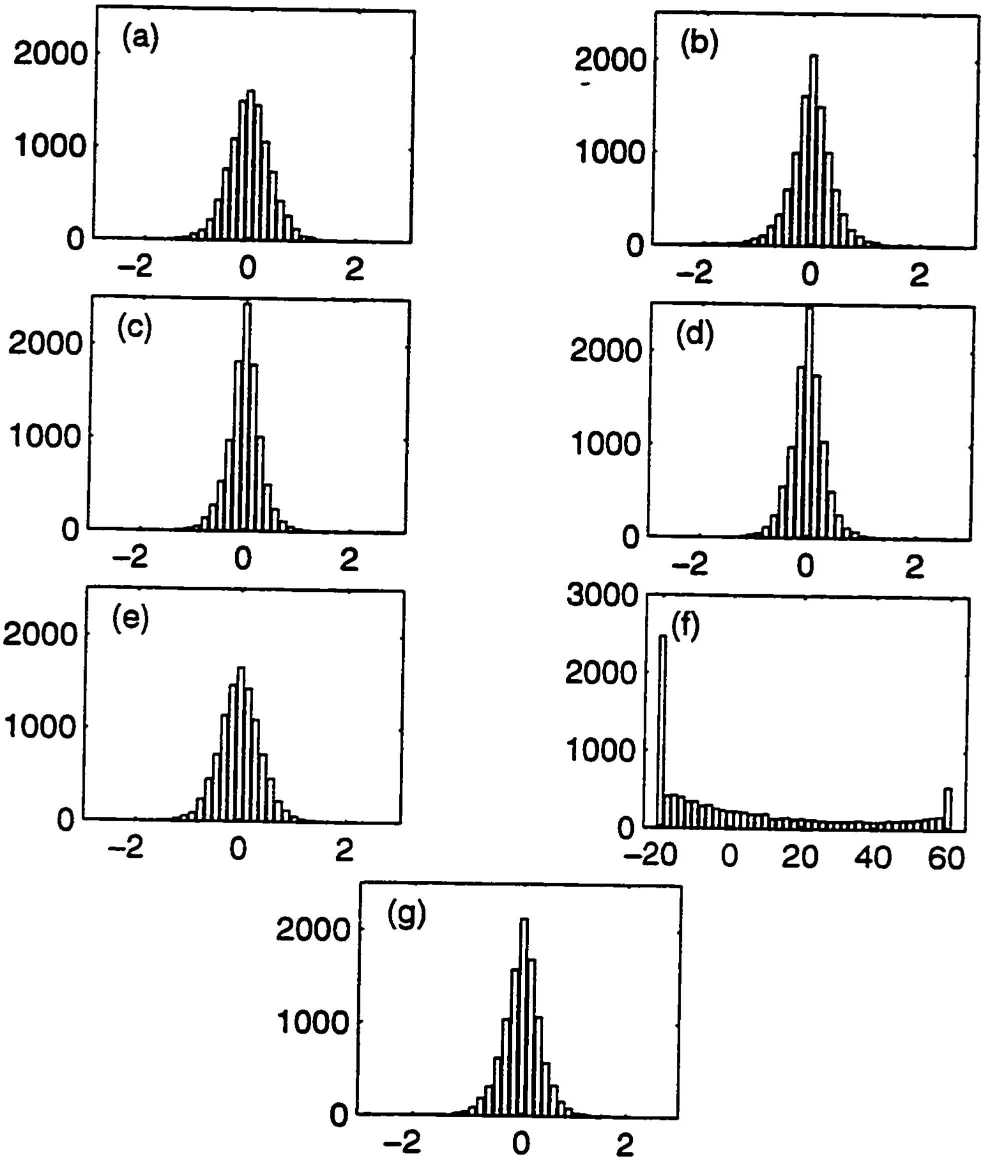


Figure 2.8 Histograms showing the distributions of centered angular deviations (4% RA error with a=0.5) for (a) OLS, (b) WLS, (c) EVM, (d) MWR, (e) UPCA, (f) WPCA1 and (g) WPCA2.

2.5.1 Bias Study

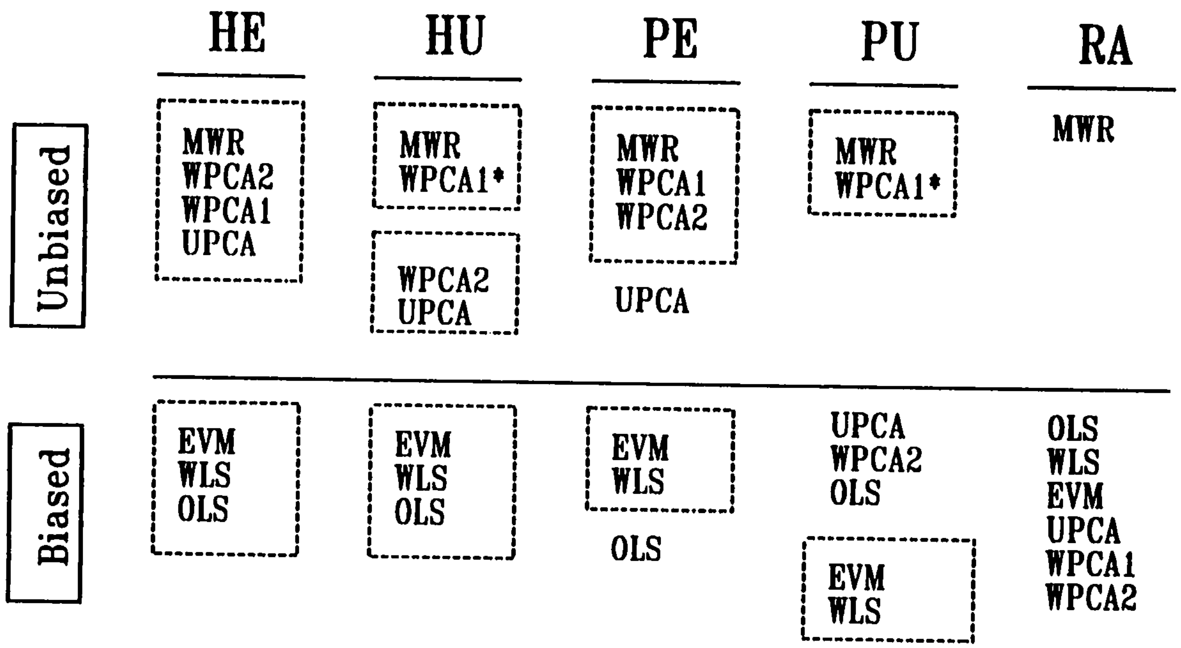
One method of evaluating the different modeling techniques is according to their bias, which is defined as the difference between the expectation value for the estimated angle of the linear model, $\hat{\theta}$, and the angle for the true model:

$$bias = E\left(\hat{\theta}\right) - \theta_{true} \tag{2.20}$$

For this study, the mean of the angle estimates obtained from the 10,000 simulations was compared to the true model angle. The large number of simulations ensures, by the Central Limit Theorem, that the mean estimated angle is normally distributed, and that a Z-test of the bias can be used. The test statistic has the form:

$$Z = \frac{\left(\overline{\theta} - \theta_{true}\right)\sqrt{N}}{s_{\theta}} \tag{2.21}$$

where $\overline{\theta}$ and s_{θ} are the average and standard deviation of the 10,000 estimated angles. Values of Z greater than 1.96 in magnitude were taken to indicate bias which is statistically significant. A summary of typical results for the case of 10 evenly distributed points is shown in Figure 2.9. In this figure, the methods are ranked according to the magnitude of the Z statistic and classified as either biased or unbiased based on the critical value. Also, methods which are mathematically identical under various error conditions are indicated by being enclosed in the same dashed box. Several observations can be made for this particular study. For homoscedastic equal (HE) errors, MWR and the equivalent principal component methods produce unbiased results while the



	HE	HU	PE	PU	RA
Unbiased	MWR WPCA2 WPCA1 UPCA	MWR WPCA1* WPCA2 UPCA	MWR WPCA1 WPCA2 UPCA	MWR WPCA1*	MWR
Biased	EVM WLS OLS	EVM WLS OLS	EVM WLS OLS	UPCA WPCA2 OLS EVM WLS	OLS WLS EVM UPCA WPCA1 WPCA2

□ indicates methods that are mathematically identical under the conditions.

Evaluation of bias based on Z-test at .05 level.

Figure 2.9 Summary of typical trends observed for bias with different methods and error conditions.

regression methods become significantly biased as the orientation of the line becomes more vertical. For the HU and PU cases, these same methods were again found to be unbiased (but were not equivalent as in the HE case), although these results were sensitive to the particular conditions used and under the right circumstances all of the methods exhibited bias. For the remaining error types (PU and RA), only MWR (and its equivalent for the PU case, WPCA1*) was found to give unbiased results.

One should be careful not to read too much into this particular study, since different levels of bias can be observed as the conditions of the simulations are changed. However, some general conclusions can be drawn. First, for the majority of the error conditions studied here MWR and the PCA methods appeared to be less biased than the linear regression methods. (As the errors in x approach zero, however, the regression methods are expected to perform better than the PCA methods.) Second, and perhaps more importantly, in all of the cases that were studied in this chapter, MWR consistently provided results with a smaller bias than the other methods. Finally, it should be noted that, while bias is a useful means of comparison, it is not necessarily the most important parameter to consider in most chemical applications, as discussed below.

2.5.2 Mean-Squared Error Study

From a chemist's point of view, the utility of a modeling method should be based on how closely the results from a single set of data approximate the true model underlying those data. This will depend not only on bias, but also on the dispersion of results from the true model under a given set of error conditions. The total error of an estimator, $\hat{\theta}$, arising from both of these sources is referred to as the mean-squared error (MSE) [42] and is given by:

$$MSE = E\left(\hat{\theta} - \theta_{true}\right)^2 \tag{2.22}$$

where E is the expectation operator, $\hat{\theta}$ is the angle of the linear model obtained by a particular method for a given data set, and θ_{true} is the angle for the true model. It can be shown that:

$$MSE = \sigma_{\hat{\theta}}^2 + \left(bias\right)_{\hat{\theta}}^2 \tag{2.23}$$

where $\sigma_{\hat{\theta}}^2$ is the variance of the estimated angles about their mean, and the bias is defined in Equation 2.20. In other words, the uncertainty in the analysis of a single data set will be determined not only by the bias, but also by the distribution of the estimated parameter about the mean. This is illustrated in Figure 2.10. When the bias of a method is small relative to the spread of the distribution (as in the cases examined here), it will be the latter that is the dominant factor in the determination of uncertainty for a given data set.

To examine the performance of the various methods included in this chapter in terms of MSE, a study similar to that used to examine bias was carried out. The results of the MSE study will vary with the conditions used (number of points, error level, slope), but trends emerge and some general observations can be made. As expected, the performance of the various methods with respect to MSE depends on the error structure and other conditions used. Typical results of the MSE study are summarized in Figure

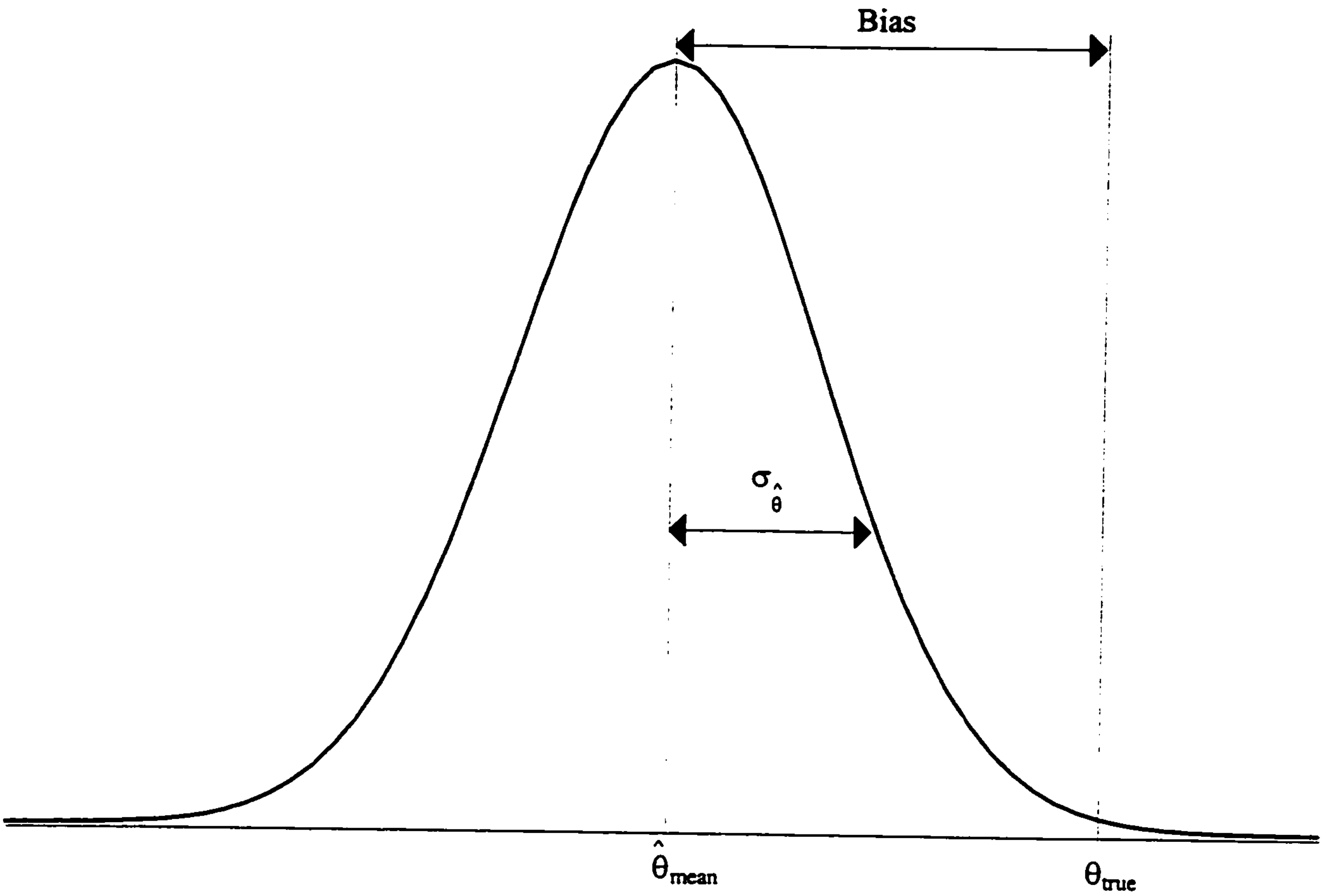


Figure 2.10 Graphical representation of mean-squared error.

2.11. In generating these results, paired Z-tests were used to compare the MSEs from 10,000 data sets. In the figure, methods are ranked according to their MSE. Because of the sensitivity of these tests (due to the large number of data sets) statistically significant differences were found for virtually every pair of methods, but such differences were not always meaningful. For example, while a 5% difference in variances may be statistically significant, from a practical point of view it is of little importance. For this reason, the solid boxes in the figure encompass methods which were the same from a practical perspective. A practically significant difference is somewhat arbitrarily defined as a case where the difference in MSEs is greater than 10% of the average MSE for the two methods. As in Figure 2.9, the dashed boxes indicate methods that are mathematically identical.

Figure 2.11 shows that there is considerable variation in the performance of the methods under different error conditions, as was observed for the bias study. For many cases, it is found that several methods produce results that are not practically distinguishable, at least for the rather conservative conditions used here. An exception is the most general case (RA), where more meaningful differences in MSE were observed. Overall, however, MWR consistently produced the smallest MSE, which means that it should give the smallest uncertainty in any single experiment. This conclusion reinforces the results of the bias study in the last section and suggests that MWR is the preferred method among the seven techniques examined.

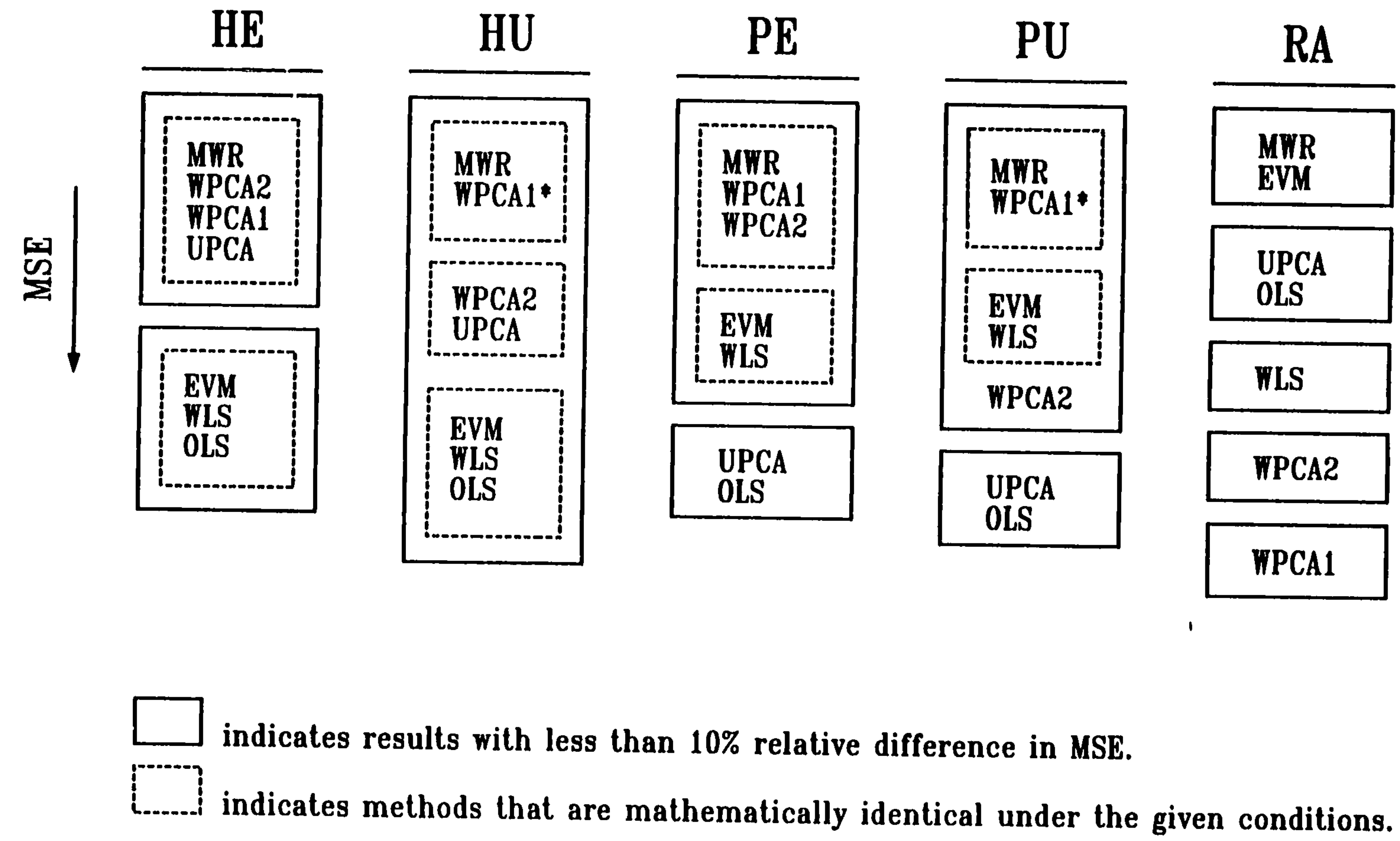


Figure 2.11 Summary of typical trends observed for mean-squared error with different methods and error conditions.

2.6 CONCLUSIONS

It was the primary objective of this chapter to clearly elucidate the relationships among regression and PCA methods for modeling linear bivariate data with normally distributed, uncorrelated errors. It has been shown that these relationships can be understood from the perspective of maximum likelihood estimation. MWR can be regarded as providing the maximum likelihood estimate for the model parameters in the general case, while each of the other methods are special cases which will provide the maximum likelihood solution when the appropriate error structures hold. It is the implicit assumptions about errors that are made for each modeling method that distinguishes it from the others and determines its strengths and weaknesses in a given situation. It was found that, under certain conditions, all of the various methods studied here will reduce to the maximum likelihood method. For ease of reference, equivalent methods under the various conditions are shown in Table 2.1.

A secondary objective of this chapter was to demonstrate that, at least for some common cases, MWR produces results which are superior in terms of both bias and mean-squared error. While this conclusion was not rigorously proven or generalized through an exhaustive study of all possible conditions, the results are consistent with the widely held belief that maximum likelihood methods perform best for well-behaved data sets.

Of the error structures considered here, it will be noted that WPCA1 is equivalent to MWR in all cases but RA. Therefore, this approach to scaling should be useful as a pretreatment for PCA in higher dimensions when the appropriate error structure is valid. However, it should be noted with some caution that this method of scaling performs badly when general unstructured errors are present and can destroy the structure of the data in those instances. This observation is consistent with the conclusions of Paatero and Tapper [32]. Furthermore, no particular advantages to using WPCA2 were found in this study, so no further results for WPCA2 will be reported for the remainder of this work.

It can be concluded that the maximum likelihood method, manifested as MWR for the two-dimensional case, serves as a general approach to modeling, and that other methods can be unified by their relationship to MWR. Additionally, the use of maximum likelihood renders redundant issues of scaling in PCA, since appropriate weighting of the data is automatically incorporated into the method. The conclusions presented here for simple two-dimensional linear modeling have implications for multivariate modeling in higher dimensional spaces as well. Thus, the use of maximum likelihood modeling as a substitute for PCA in higher dimensions will eliminate the need for scaling and produce results with less bias and smaller uncertainties. In the next chapter, a method for maximum likelihood modeling in dimensions greater than two will be developed.

3
MODELING IN HIGHER DIMENSIONS

3.1 INTRODUCTION

In the previous chapter, the use of PCA in two dimensions was described. Although shown to have utility in this case, the strength of PCA arises in its ability to readily handle multivariate data sets. Because of the many applications of PCA, some of which were outlined in Section 1.3, it has become a tool of choice for chemometricians.

One important aspect of PCA which has been largely ignored (despite its widespread use) is the role of measurement errors in the decomposition procedure. Problems arising from these errors have been loosely acknowledged through the introduction of various scaling techniques, and this has led to numerous interesting debates over when such techniques should be used. For example, autoscaling, in which data columns are first mean-centered and then scaled to unit variance, is commonly used, but can lead to problems for certain types of data.

Despite the fact that PCA has been described (albeit briefly) in the first chapter, a more in-depth treatment is required for application in higher dimensions. In this chapter, a novel approach to PCA is described that inherently accounts for measurement errors (if their variances are known) and removes the need for any kind of preprocessing to scale or offset the data. This new technique will be referred to as maximum likelihood principal component analysis (MLPCA) because it is based on the principle of maximum likelihood model estimation. This name is somewhat inaccurate, since PCA is described by a specific definition, and to alter this means that one is no longer performing PCA, but we will use the description to emphasize the close relationship of the new method to the objectives of PCA. It is demonstrated that MLPCA produces results identical to traditional PCA when conditions of uniform measurement error are assumed, but provides results that better estimate the true model when non-uniform noise is present. MLPCA also renders mean-centering and scaling obsolete as data pretreatment steps and therefore avoids the problems that these techniques can introduce.

3.2 MODELING IN HIGHER DIMENSIONS

3.2.1 PCA

For multivariate analytical measurements, we may consider the *m*x*n* matrix **X** which consists of *m* samples measured at *n* sensors. Whatever the application, the general objective of PCA in chemistry is to map such multivariate data into a space of lower dimensionality. This is done by first determining the direction of greatest variance in the data set and assigning a unit vector to this direction. The unit vector is called the first eigenvector or the first principal component. The projections onto this vector are called the principal component scores and represent the linear mapping of the data into a one-dimensional space. The second eigenvector accounts for the largest amount of residual variance (*i.e.* that not accounted for by the first eigenvector) and is orthogonal to the first eigenvector. The scores on the first two eigenvectors map the data into a two-dimensional space defined by the plane of these two vectors. This process continues until the number of eigenvectors equals the dimensionality of the original space. By selecting the appropriate number of eigenvectors (normally in order of decreasing variance) projection of the data into a space of arbitrary dimensionality can be achieved.

In practice, PCA is now most often implemented through singular value decomposition (SVD). This decomposes the original data matrix into the product of three matrices:

$$\mathbf{X} = \mathbf{USV}^{\mathrm{T}} \tag{3.1}$$

where the superscript "T" denotes the transpose of a matrix. In this application, **U** is *mxn*, **S** is a diagonal *nxn* matrix of singular values, and **V** is *nxn*. To describe the data in a *p*-dimensional subspace, **U** is truncated to *mxp*, **V** to *nxp*, and **S** to *pxp*. This gives,

$$\tilde{\mathbf{X}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^{\mathrm{T}} \tag{3.2}$$

where $\tilde{\mathbf{X}}$ represents the orthogonal projection of the measurements onto the model. The product $\tilde{\mathbf{U}}\tilde{\mathbf{S}}$ is often referred to as the scores matrix and $\tilde{\mathbf{V}}^{\mathrm{T}}$ as the loadings matrix. Various methods can be used to determine whether the *p*-dimensional space is sufficient to represent the data within experimental error, but these will not be addressed here. Instead, we will consider the question of whether or not PCA is the best method to model the *p*-dimensional subspace.

One of the weaknesses of PCA is that, while it is sensitive to variance in a data set, it ignores the source of the variance. Chemists are interested in systematic variance, that which arises from sources other than purely random measurement error, but PCA does not distinguish between these two sources. If some information about the magnitude of the measurement errors is known, it seems logical that a better model could be obtained by taking this into consideration. To see how this could be done, it is helpful to view PCA from a geometric perspective.

Each of the rows (samples) in the data matrix **X** can be considered to represent a point in an n-dimensional sensor or column space. Alternatively, we can view each of the columns as a point in the m-dimensional row or sample space. In either case, our objective is to fit these points to a p-dimensional model (hyperplane) in the corresponding higher dimensional space. As demonstrated in Chapter 2, both PCA and classical regression methods can be considered to be maximum likelihood estimation methods under the right conditions. PCA will provide a maximum likelihood estimate of the p-dimensional model when the errors in the measurements are independent and identically distributed (*iid*) with a normal distribution.

The requirement for *iid* measurement errors for maximum likelihood estimation by PCA has been one of the most cumbersome aspects of its use. It has long been known that PCA is scale sensitive and this is a direct consequence of the implicit assumptions regarding measurement errors. Cases of non-uniform noise abound in analytical chemistry. For example, many applications involve measurement variables that have inherently different uncertainties because of different ranges or different units. As well, measurement errors in areas such as spectroscopy typically vary with signal intensity (often referred to as heteroscedasticity) and may also be correlated. Numerous methods have been developed to reduce these cases to uniform uncertainty prior to PCA, but no universal method exists. Methods such as WPCA1 and WPCA2 previously discussed in Section 2.2.4 may be adequate for certain error structures, but are ineffective in general situations. It would therefore be useful to develop a true maximum likelihood estimation procedure which can take measurement errors into account.

3.2.2 Maximum Likelihood PCA

For each measurement, x_{ij}, we will assume there is an associated random measurement error, e_{ij}, which is characterized by its variance, σ_{ij}^2, or standard deviation, σ_{ij}. Furthermore, if we consider each sample as a point in *n*-dimensional sensor space and that the measurement errors are normally distributed, then each of these points can be considered to be bounded by an *n*-dimensional hyper-ellipsoid which represents a region of specified probability (e.g. 95%) for the true measurement. This ellipsoid is characterized by the *nxn* error covariance matrix for the point, given by:

$$\Sigma_i = E\left[\left(\mathbf{x}_i - \mathbf{x}_i^\circ\right)\left(\mathbf{x}_i - \mathbf{x}_i^\circ\right)^{\mathrm{T}}\right] \tag{3.3}$$

where $\mathbf{x}_i$ is the column vector of measurements (*n*x1) for sample *i* (*i.e.* the transpose of the *i*th row of **X**) and $\mathbf{x}_i^\circ$ is the column vector of expectation values for those measurements. If the measurement errors are uncorrelated, Σ_i is a diagonal matrix whose elements are σ_{ij}^2. In this case, the major axes of the ellipsoid will be parallel to the sensor axes. If the errors in the measurements are correlated (as is often the case in spectral measurements, for example), the off-diagonal elements of Σ_i will not be zero and the ellipsoid can be viewed as tilted away from the sensor axes. Similarly, an *mxm* covariance matrix can be calculated for each point in the sample space.

In maximum likelihood estimation, the error ellipsoid for each point in the space is considered to be associated with some "true" point which lies on the hyperplane described by a trial model. Of course, the actual error-free measurement is not known, so a best guess is required, and this is the maximum likelihood estimate for the point. For both classical regression and PCA, the location of this point is defined by a particular projection onto the hyperplane, although it will be a true maximum likelihood estimate only if the error assumptions made for those two methods are valid. For the general case, the maximum likelihood estimate is obtained by finding the point in the plane of the model for which the experimentally measured point is at the highest point on the multivariate probability density function for that measurement. Thus, the maximum likelihood estimate of point $\boldsymbol{x}_{\mathrm{i}}$ (= $[x_{\mathrm{i1}} \;\; x_{\mathrm{i2}} \;\; \ldots \; x_{\mathrm{in}}]^{\mathrm{T}}$) is the one for which the observed measurement is "most probable". This process is repeated for all of the points in the space until a matrix of maximum likelihood estimates, designated $\hat{\mathbf{X}}$, is obtained for the trial model parameters.

This is only the first step in the maximum likelihood estimation, since it finds the most likely true values of the measurements for a set of trial model parameters, but does not assess those parameters. The maximum likelihood fit is obtained by adjusting the parameters in the trial solution to minimize the goodness-of-fit objective function, S^2, given by:

$$S^2 = \sum_{i=1}^{m} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \Sigma_i^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \qquad (3.4)$$

where $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ represent the observed and maximum likelihood estimates of the measurement vector for sample i. In effect, this is a sum of squares of residuals for all measurements weighted by the appropriate error covariance matrix. For uncorrelated measurement errors, this function reduces to:

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(x_{ij} - \hat{x}_{ij}\right)^2}{\sigma_{ij}^2} \tag{3.5}$$

For the maximum likelihood fit, this sum of squared residuals will approximate a χ^2 distribution with $(m\text{-}p)$x$(n\text{-}p)$ degrees of freedom (for the case of no intercept terms), where p is the dimensionality of the model. The minimization of S^2 corresponds to the maximum likelihood estimation of the p-dimensional hyperplane.

As seen in Section 2.2.3, the maximum likelihood estimation procedure therefore consists of two nested algorithms: one which determines the maximum likelihood estimates of the true measurements and the corresponding objective function in conjunction with a set of trial model parameters, and the other which updates the parameters to minimize S^2. Further details of these algorithms are treated in subsequent sections, but first the method outlined above is considered in view of current literature.

The concept of maximum likelihood estimation is quite general in nature and therefore has been employed extensively in all fields of science, including chemistry. However, an extensive search of the literature has indicated that no reference to maximum likelihood implementations of PCA have appeared in the literature, at least in the manner considered here. The closest work of this kind describes "maximum likelihood common factor analysis" (MLCFA) and was reported in the chemical literature quite recently by De Volder and coworkers [43-44]. Although the terms are often used interchangeably by chemists, PCA and factor analysis are distinctly different approaches to multivariate analysis [18,p.109]. MLCFA is based on an approach described by Lawley and Maxwell [18,Ch. 4] and later employed in programs such as LISREL [45-46], but does not include estimated measurement errors in the determination of common factors. The MLCFA method assumes that the measurements are random variables, an assumption that does not generally hold for chemical measurements. Nevertheless, the chemical reports claim better results with MLCFA than with PCA. The reason for this is unclear, but it is suspected that the superior results are a consequence of the fact that autoscaling is used and factor analysis, which minimizes residual covariance, should perform better under these conditions than PCA, which minimizes residual variance. However, MLCFA does not generally use information about measurement errors. Elsewhere, Thomas discussed maximum likelihood prediction based on a calibration model, but does not use the principles to develop the calibration model itself [47]. Several authors [36,48] discuss the development of models by minimizing an objective function of the form of Equation 3.5, but do not incorporate maximum likelihood estimates of the measurements in doing so.

Another errors-in-variables method that has become popular recently is total-least-squares (TLS) [40]. This method uses SVD for the purpose of developing a regression model and is similar to MLPCA in some ways. However, it is less general in its ability to obtain maximum likelihood estimates of model parameters. To our knowledge, the method described in this work is unique in its approach to PCA model estimation.

The remainder of this chapter is devoted to the specific aspects of the MLPCA algorithm. Extensions of maximum likelihood parameters into higher dimensions will be made, building on the bivariate principles set down in Section 2.2.3.

3.2.3 Extension of Model Parameters into Higher Dimensions.

The procedure for the estimation of a rank one model in a two-dimensional space is relatively straightforward and has been addressed in detail in Chapter 2. To develop an analogue to PCA based on maximum likelihood, the principles described in two dimensions need to be extended to higher dimensional models in higher dimensional spaces and incorporate a non-diagonal error covariance matrix. If a point in an n-dimensional space is given by $(x_1, x_2, \dots, x_n)$, then the objective function for a trial model is calculated in a manner described by Equation 3.4. First, however, we must develop equations to calculate the maximum likelihood estimates of the points in the multidimensional space. This requires a set of model equations, which will be defined by the dimensionality of the space and the dimensionality of the model. For example, a one-dimensional model in three dimensions would be given by a set of parametric equations:

$$\begin{aligned} x_2 &= a_{21}x_1 + b_2 \\ x_3 &= a_{31}x_1 + b_3 \end{aligned} \qquad (3.6)$$

where the a's and b's are the model parameters to be estimated (the subscript notation will become evident shortly) and x_1 is the (arbitrarily chosen) independent variable. Note that

$n-1$ equations are necessary to describe a line in n dimensions. When the objective function is optimized with respect to the parameters, the solution will be analogous to the first eigenvector produced by PCA for this three-dimensional data set, but should possess better properties when the homoscedastic error assumption for PCA does not hold.

Models of higher dimensionality can also be used, generally requiring $n-p$ equations, where p is the dimensionality of the model. For example, a two-dimensional model in four dimensions is described by the two equations:

$$\begin{aligned} x_3 &= a_{31}x_1 + a_{32}x_2 + b_3 \\ x_4 &= a_{41}x_1 + a_{42}x_2 + b_4 \end{aligned} \tag{3.7}$$

These equations will define a two dimensional plane in a four-dimensional space in the same way that the equations in Equation 3.6 define a line in three dimensions. (In Chapter 4, the dimensionality of the subspace will be more apparent from the PCA formulation.) Again, with maximum likelihood estimation, the assignment of "independent" and "dependent" variables is arbitrary, but we will normally consider the first p variables to be "independent".

Determination of maximum likelihood estimates of points in higher dimensions proceeds in the same manner as for the two-dimensional case, except that now we are dealing with matrices of independent and dependent variables in the equation for the log likelihood and a covariance matrix that is more complicated if the assumption of independence is no longer valid. Suppose we define the measurements for a particular point in the n dimensional space to be given by the vector $\mathbf{x} = (x_1, x_2, x_3, \ldots\ x_n)^{\mathrm{T}}$. A p-dimensional model in this space is given by the matrix equation,

$$\mathbf{x} = \mathbf{A}\mathbf{x}_\mathrm{p} + \mathbf{b} \tag{3.8}$$

where,

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_\mathrm{p} \\ \mathbf{A}_2 \end{bmatrix} = \begin{bmatrix} & \mathbf{I}_\mathrm{p} & & \\ a_{p+1,1} & a_{p+1,2} & \cdots & a_{p+1,p} \\ a_{p+2,1} & a_{p+2,2} & \cdots & a_{p+2,p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,p} \end{bmatrix}, \; \underset{-}{\mathbf{b}} = \begin{bmatrix} \mathbf{0}_\mathrm{p} \\ \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{0}_\mathrm{p} \\ b_{p+1} \\ b_{p+2} \\ \vdots \\ b_n \end{bmatrix}, \; \mathbf{x}_\mathrm{p} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \tag{3.9}$$

Here $\mathbf{I}_\mathrm{p}$ indicates a *pxp* identity matrix and $\mathbf{0}_\mathrm{p}$ indicates a px1 vector of zeros. **A** is the *nxp* matrix of regression coefficients, with the coefficients for the p "independent" variables set to unity in the corresponding variable. The nx1 vector **b** contains the intercept terms for the model. The point **x** will have an associated error covariance matrix Σ and the multivariate probability density function is described by Equation 3.10.

$$L = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\, exp\left[-\tfrac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^\mathrm{T}\Sigma^{-1}(\mathbf{x}-\hat{\mathbf{x}})\right] \tag{3.10}$$

The vector of maximum likelihood estimates, $\hat{\mathbf{x}}$, is obtained by minimizing the function:

$$l = \tfrac{1}{2}(\mathbf{x}-\hat{\mathbf{x}})^\mathrm{T}\Sigma^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \tag{3.11}$$

subject to Equation 3.8. It can be shown in a manner analogous to the two dimensional case that the solution is,

$$\Delta\mathbf{x} = \mathbf{x} - \hat{\mathbf{x}} = -\mathbf{A}\left(\mathbf{A}^\mathrm{T}\Sigma^{-1}\mathbf{A}\right)^{-1}\mathbf{A}^\mathrm{T}\Sigma^{-1}\mathbf{c} + \mathbf{c} \tag{3.12}$$

where,

$$\mathbf{c} = \mathbf{x} - \mathbf{A}\mathbf{x}_\mathrm{p} - \mathbf{b} \tag{3.13}$$

From this result, $\hat{\mathbf{x}}$ can be calculated, for the given point ($\hat{\mathbf{x}} = \mathbf{x} - \Delta\mathbf{x}$), but usually it is $\Delta\mathbf{x}$ that is required for substitution into Equation 3.4. This procedure is repeated for each point in turn prior to evaluation of the objective function given by Equation 3.4. The model coefficients (**A** and **b**) are then updated and the process continues until S^2 is minimized, at which point the coefficients represent the maximum likelihood model.

The maximum likelihood model obtained in this way will describe a *p*-dimensional hyperplane in the *n*-dimensional space, analogous to the hyperplane represented by the first *p* eigenvectors found by PCA. Note that the MLPCA approach does not define individual eigenvectors as in the case of conventional PCA, but in most applications it is the subspace defined by the eigenvectors and not the vectors themselves that are important. Where individual vectors are required, it is a simple matter to calculate them once the subspace has been defined.

3.2.4 General Procedure

The MLPCA algorithm begins with an *mxn* data matrix, **X**, where the rows of **X** represent the points to be modeled. An estimate of the error covariance matrix for each row of **X** is also required. Generally, unless the noise is very well characterized, the errors in **X** will be assumed to be independent. In this case, the standard deviations in each measurement can be placed in a matrix that is the same size as **X** and the rows of this used to generate the diagonal elements of the error covariance matrix. After these matrices have been defined, the first step is to obtain an initial estimate of the coefficients. If an intercept is not included in the model, perhaps the simplest way to do this is to perform SVD on **X**. If this is done and the results are truncated to the form given in Equation 3.2, the coefficients are calculated as:

$$\mathbf{A} = \tilde{\mathbf{V}}\tilde{\mathbf{V}}_p^{-1} \tag{3.14}$$

where $\tilde{\mathbf{V}}$ is the loadings matrix from SVD truncated to rank *p*, and $\tilde{\mathbf{V}}_p$ is the upper *pxp* submatrix of $\tilde{\mathbf{V}}$. This method becomes more complicated when an intercept is involved and in this case we have used the effective variance method (EVM) [25] to determine initial estimates. This method is an adaptation of weighted regression which propagates the errors in the independent variables to the dependent variables in an iterative fashion. EVM was found in general to give the best initial estimates of the coefficients, even in the absence of an intercept.

Once initial estimates of **A** and **b** have been obtained, Equation 3.12 is used to evaluate the maximum likelihood estimates for each row of **X** and S^2 is calculated from Equation 3.4. The process of searching for a minimum then begins. At each iteration in this process, a new trial solution is generated and a new value for S^2 calculated until the optimum solution has been found. A variety of optimization procedures could be employed, including gradient algorithms, genetic algorithms and simulated annealing, but in this study we used simplex optimization [49]. This is not necessarily the most efficient method, but is robust and was convenient to implement for these initial studies. A more efficient solution will be introduced in the next chapter. It should be noted that to enhance the performance of the optimization procedure, it is not carried out in the original space of the model parameters, but rather the space of angles that correspond to the **A** matrix (*i.e.* $\theta_{ij} = \tan^{-1} a_{ij}$). The reason for this is the nonlinear behavior of the slope parameters. For example, in two dimensions, a step change in the slope causes a much greater adjustment for a line with a slope of 1 than for a line with a slope of 1000. Because of this, the rate of movement of the simplex over the search space would depend on the value of the parameter and it would be difficult to choose an optimum step size. In angle space, the surface is more homogeneous, leading to a more effective search and an easier determination of convergence. Of course, the intercept terms cannot be treated in this way.

One of the major differences between conventional PCA and MLPCA is that the former produces eigenvectors and scores, whereas the latter produces model parameters and maximum likelihood estimates (confined to a p-dimensional hyperplane in the original n-dimensional space). As already noted, it is generally the subspace that is of interest rather than the way it is defined, but the availability of eigenvectors is comforting to some and useful in many respects (*e.g.* for visualization of the samples in the subspace). Because of this, a simple method has been developed for generating what will be termed "pseudo-eigenvectors". This is accomplished by carrying out singular value decomposition (SVD) on the maximum likelihood estimate of **X** and retaining the first p eigenvectors (note that because these estimates are confined to a p-dimensional hyperplane, additional eigenvectors are meaningless). In PCA, it is assumed that the maximum likelihood estimates are the orthogonal projections of the data onto the first p eigenvectors (*i.e.* the scores). Unfortunately, this does not necessarily hold true for MLPCA and an orthogonal projection may result in a loss of any extra information gained by incorporation of the error estimates. Instead, "pseudo-scores" are obtained by multiplication of the maximum likelihood estimate of **X** by the p pseudo-eigenvectors. These procedures produce results that are analogous to conventional PCA. Likewise, new measurement vectors **x** can be projected into the principal component space by first computing their maximum likelihood values in accordance with the model obtained.

Another difference between PCA and MLPCA is that the former produces estimates of all eigenvectors in a single treatment. Therefore, the dimensionality of the subspace is specified simply by selecting the corresponding number of eigenvectors. In contrast, for MLPCA, the dimensionality of the desired model must be determined in advance, and the process must be restarted if a model of higher or lower dimensionality is desired. While this, when coupled with the potentially long calculation times for MLPCA, is an inconvenience, it will be shown that in certain cases the advantages of better model estimation warrant its use.

3.2.5 Equivalence of Methods

As noted above, the optimization procedure for MLPCA can be quite time consuming when compared to traditional PCA carried out via SVD. Because of this, it is worthwhile to point out cases where more conventional methods will serve as well. In the case of *iid* errors, MLPCA and conventional PCA produce identical results. The case of homoscedastic, unequal errors (*i.e.* same variance for all measurements of a given variable, but different variances among variables) can also be readily treated. In such an instance, the method previously described in Section 2.2.4, referred to as WPCA1 can be used. With WPCA1, each variable is scaled by dividing by its corresponding standard deviation prior to analysis by PCA. This has the effect of reducing the problem to the homoscedastic, equal error case. This produces a projection which is equivalent to MLPCA in terms of the spatial relationships of the samples, although the scale and orientation of the axes are different from MLPCA. For most other error cases it is necessary to use the MLPCA approach. Using the WPCA1 method in cases where errors are not homoscedastic within a variable can seriously distort the data.

3.3 EXPERIMENTAL

All simulations in this study were carried out using Matlab v.4.0 for Windows (Mathworks, Natick, MA) on a 486-based personal computer. Simplex optimization of maximum likelihood solution was based on the modified simplex of Nelder and Mead [49]. Further details of the simulations employed are given in the following sections.

3.4 ADVANTAGES OF MLPCA

There are four principal advantages of MLPCA over conventional PCA for multivariate data analysis: (1) it eliminates the need for scaling the data, (2) it eliminates the need for mean-centering the data, (3) it provides a reliable statistic for rank estimation, and (4) it handles missing data in a simple and statistically valid manner. The first two advantages are illustrated with some simple examples in the following sections.

The latter points will be explored further, in Chapters 4 and 6 respectively, following an examination of the theory behind MLPCA in the next chapter.

3.4.1 Scaling

Over the years, many methods of preprocessing data prior to implementing PCA have been devised. The most common techniques involve mean-centering and scaling. We will consider each of these steps separately. Scaling of data has always been a subject of debate among practitioners of PCA, but it is clear that in some cases it is very beneficial, while in others it can be detrimental. Deming has considered the effect of some common scaling methods from a geometric perspective and illustrated how these can distort the original data [50]. Perhaps the most common type of scaling is variance scaling, which consists of dividing the rows or columns of a data matrix by the standard deviation of that row or column so that it has unit variance. When used in conjunction with mean-centering, this technique is known as autoscaling. It is normally done in such a way that the row or column treated consists of measurements of one type. Range scaling, in which each row or column vector is scaled by the range of its elements, is also used. In practice, both of these methods are attempts to reduce the measurements to uniform measurement error by assuming that the errors for a given variable are homoscedastic and are a fixed percentage of the observed range. This is a rather simplistic view and fails dramatically, for example, if one or more measurements consists of pure error. Another technique is to scale each measurement by the reciprocal of its standard deviation. This approach, which we have referred to as WPCA1, is somewhat better, but seriously distorts the data if measurement errors for a given variable are not the same. Paatero *et al* give an excellent summary of different scaling methods and discuss their assumptions [32]. They show that optimal scaling requires a rank one matrix of standard deviations, which is often not observed in practice. A "balanced scaling" approach is shown to perform well, but is still sub-optimal. MLPCA can provide an optimal model for the data with no preprocessing.

To illustrate the effects of scaling on data, we will examine an artificial rank two data set with three variables. In practice, most data sets will contain many more variables, but this small data set is more easily visualized. Consider an equilateral triangle with sides of unit length centered at the origin of a three-dimensional coordinate system and lying in the *xy* plane. Now, ten points, representing samples of different classes, are placed at each corner of the triangle, and normally distributed noise representing measurement variance, is added to each sample for all three variables. This noise has a standard deviation of 0.1 in the x and y directions, and 1 in the z direction. The projection of the samples onto the *xy* plane is shown in Figure 3.1a and exhibits a good separation of the three classes. The z axis is not shown, but has no value in class separation, although it does have a variance (from measurement error) roughly equivalent to the systematic variance along the x and y axes. Because of this, the projection of the samples into the two-dimensional space defined by the first two eigenvectors obtained from conventional PCA produces a poor separation of the clusters compared to the original *xy* projection and is shown in Figure 3.1b. The significant random variance in the z direction leads to a major contribution of that axis to the first two eigenvectors, thus contaminating the projection with noise. In this case, variance scaling does not help, as

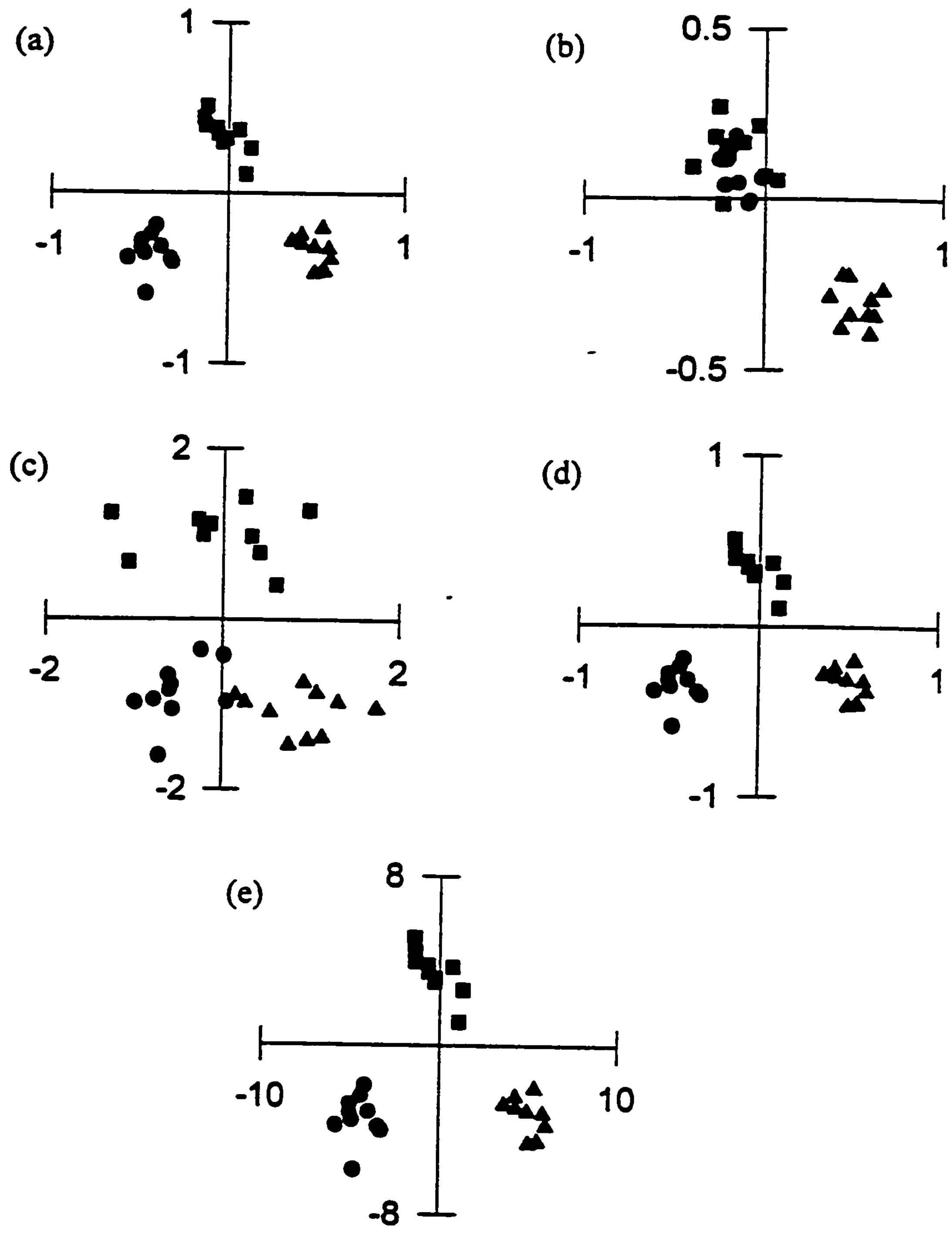


Figure 3.1 (a) Projection of three-dimensional data ($\sigma_x = \sigma_y = 0.1$, $\sigma_z = 1$) simulating three classes onto *xy* plane. (b) Projection of data onto first two eigenvectors following PCA.. (c) Projection of scaled data onto first two eigenvectors following PCA. (d) Projection of data onto first two pseudo-eigenvectors following MLPCA. (e) Projection of data onto first two eigenvectors following WPCA1.

shown in Figure 3.1c. This is because the variance in each variable is comparable. Although the variance in x and y is largely due to class differences and the variance in z is due to random noise, variance scaling does not distinguish between the two, so the projection is once again contaminated by noise. In contrast, Figure 3.1d shows that the projection of the maximum likelihood values onto the first two pseudo-eigenvectors obtained by MLPCA produces a class separation which is comparable to the ideal projection into the xy plane. In this case, the known measurement standard deviations were used in the calculation, so systematic variance could be more effectively distinguished from random variance by the algorithm. As a final comparison, WPCA1 was carried out on the data matrix and the projections are shown in Figure 3.1e. As previously noted, the spatial relationships of the samples in this space is equivalent to that produced by MLPCA (and the separation is therefore as good), but the orientation and scaling of the axes are different.

Given the equivalent performance of MLPCA and WPCA1, one might question the necessity of using MLPCA, especially since it is much more involved computationally. This is best illustrated with an example in which the noise is no longer homoscedastic within the variables. The same procedure was used to generate the data as for Figure 3.1, but in this case half of the samples in each group had a standard deviation of 0.02 in x and y, and the other half had a standard deviation of 0.2. As before, the standard deviation of z is unity. The projection of these samples onto the xy plane is shown in Figure 3.2a, along with the projection onto the first two eigenvectors by conventional PCA in Figure 3.2b. Again, PCA provides a rather poor separation of

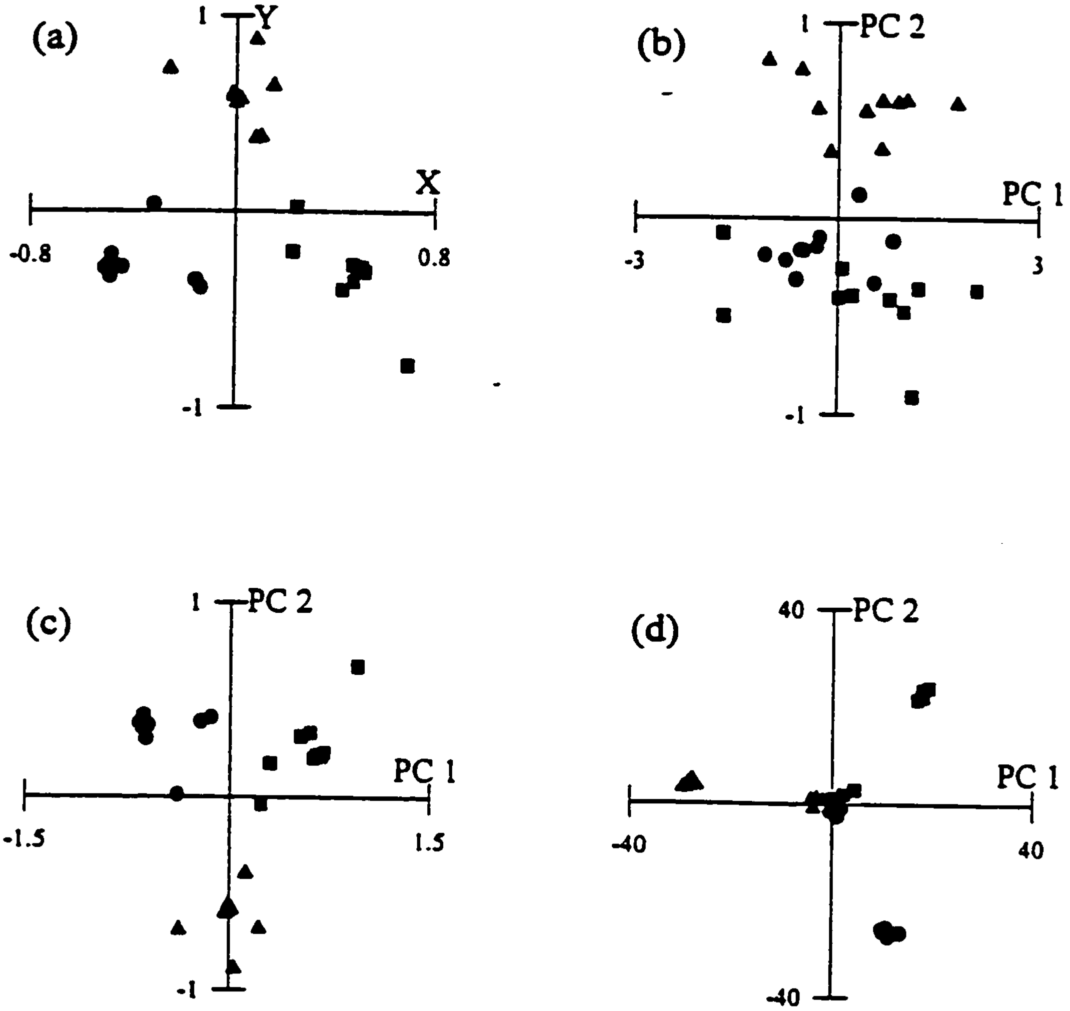


Figure 3.2 (a) Projection of simulated three-dimensional data with non-uniform errors onto the *xy* plane. (b) Projection resulting from PCA. (c) Projection from MLPCA. (d) Projection from WPCA1.

classes. In contrast, MLPCA provides quite a good separation as shown in Figure 3.2c. For comparison, WPCA1 results are shown in Figure 3.2d. Not only does this method provide a poor separation for half of the samples, but the results are misleading. Four groups are clearly evident rather than three, a direct consequence of the scaling used. For errors which are heteroscedastic within the variables, WPCA1 can seriously distort the original data because of the scaling used. Although these simple examples are simulated, they indicate some of the pitfalls that can be encountered when traditional scaling methods are used with PCA. Such problems will likely be compounded as the dimensionality of the measurement space increases.

3.4.2 Mean-centering

The subject of mean-centering as a preprocessing tool for PCA is almost as contentious as the issue of scaling. Mean-centering involves subtracting the mean measurement for each variable from the row or column corresponding to that variable. From a modeling perspective, the purpose of mean-centering in conventional PCA is to eliminate the contribution of intercept terms in describing the subspace. These intercepts may arise, for example, from a constant baseline contribution or the existence of closure in a data set. Generally speaking, variance due to the intercept is usually not considered important in evaluating systematic variance in a chemical system. For the case of homoscedastic errors in all variables, mean-centering eliminates the variance due to the intercepts in a manner consistent with maximum likelihood estimation. Because of this, if intercept terms are non-zero, mean-centering prior to PCA reduces the pseudo-rank of the data matrix by one. A number of authors have discussed the effects of mean-centering on multivariate analysis [20,42,51].

Although the sample mean can be considered to be an unbiased estimator of the population mean, the variance in this estimation will be related to the variance in the measurements. Because of this, an imprecise estimate of the mean can cause the rank of a data set to be overestimated by PCA. This can be especially problematic in the case of heteroscedastic errors. In MLPCA, however, the intercept can be estimated using principles of maximum likelihood, so each measurement is given its proper weight. This is illustrated with the following simple example in two dimensions.

Consider the two-dimensional data set shown in Figure 3.3a. These data were generated from the relationship:

$$y = x + 1 \tag{3.15}$$

where $x = 1,\ldots,10$. Normally distributed noise with a standard deviation of 0.1 was added to all of the x values and to all of the y values except the last two. The last two y values were assigned noise with a standard deviation of 10. In this particular realization of the data, both of these measurements are high, but they cannot be considered outliers since they fall within an acceptable region of the known distribution. The dashed line in the figure (T) shows the true model and the solid line (E1) shows the direction of the first eigenvector obtained by applying MLPCA (with an intercept term) to these data. Note that no preprocessing was performed in this case and the slope of the first eigenvector (1.070) is close to the true value of unity, and the intercept value found by MLPCA was 0.9594. In contrast, Figure 3.3b shows the mean-centered data and the first eigenvector

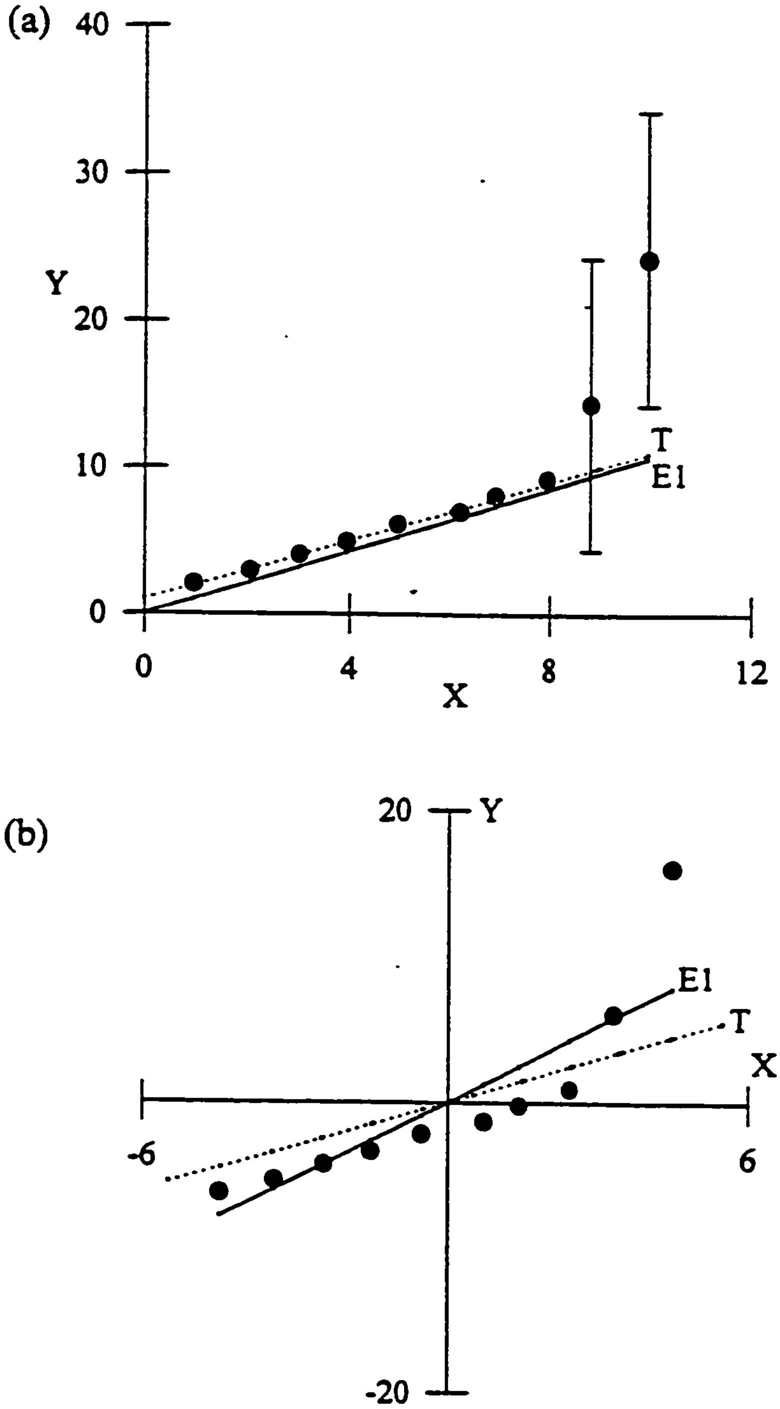


Figure 3.3 (a) Simulated two-dimensional data with non-uniform noise showing the true model (T) and the direction of the first pseudo-eigenvector found by MLPCA (E1). (b) Mean-centered data showing the direction of the first eigenvector from PCA (E1).

resulting from the application of conventional PCA. It can be seen that, because the errors in the last two points are larger and occur in the same direction, mean centering does not compensate for the true intercept of the model. In this case, the slope of the first eigenvector is 1.733, significantly different from the true value. This error is due in part to the fact that measurement precision was not taken into account, but is also a consequence of mean-centering.

Although the use of MLPCA with intercept terms makes mean-centering of the data unnecessary, it is still left to the analyst to address the question of whether intercept terms should be included in the MLPCA model or forced to zero. This is analogous to the case of fitting or forcing an intercept in a simple linear calibration plot and can only be answered with a knowledge of the data set. In order to remain consistent with PCA when an intercept is fit, the pseudo-scores and pseudo-loadings in MLPCA are obtained by conducting SVD after subtracting the intercept terms from the maximum likelihood estimates. This means that the reproduced data matrix will have an offset, but this is easily corrected by adding the intercept terms. Also, when performing rank estimation, the degrees of freedom will have to be reduced by $n-p$ if intercept terms are included. Further discussion on the role of intercepts in general can be found in the following chapter.

3.5 CONCLUSIONS

The purpose of this chapter was to introduce the principles of MLPCA and to illustrate its features with some relatively simple examples. It has been shown that MLPCA is a reliable method for dealing with measurement errors in PCA problems. By incorporating measurement errors into the model estimation, a better model should be obtained. Furthermore, it has also been shown that MLPCA removes questions of mean-centering and scaling which have plagued applications of multivariate analysis to chemistry from the beginning. These features support the further development of this technique which will be addressed in the next chapter.

The largest drawback to the use of the MLPCA algorithm described here is that it is more time consuming and cumbersome to implement than conventional PCA. This problem will be addressed by exploring the theory behind the method in the next chapter. From this treatment, an improved, efficient algorithm will be proposed to make implementation of MLPCA easier. In addition, a procedure for rank estimation via the S^2 statistic will be described.

4
THEORETICAL FOUNDATION OF MLPCA AND ALGORITHMIC IMPROVEMENTS

4.1 INTRODUCTION

In the previous chapter, a procedure was introduced to incorporate measurement errors into the multivariate modeling process. This is normally done by minimizing the usual weighted residual sum of squares in accordance with some *p*-dimensional model. Mathematically, this corresponds to the minimization of Equation 4.1,

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(x_{ij}-\hat{x}_{ij}\right)^2}{\sigma_{ij}^2} \qquad (4.1)$$

where $\hat{x}_{ij}$ corresponds to the estimated value of the measurement. In the general case, where there are no offsets in the model, this is given by:

$$\hat{\mathbf{X}} = \mathbf{AB} \qquad (4.2)$$

where **A** is *mxp* and **B** is *pxn*. By analogy to PCA, **A** and **B** correspond to scores and loadings matrices, but in Equation 4.2 the individual vectors which make up the columns of **A** and rows of **B** are not required to be orthogonal. A variety of methods have been devised to obtain **A** and **B** through minimization of Equation 4.1 and these differ largely in their representation of the problem, the constraints applied to the solution, and their approach to the nonlinear optimization. Gabriel and Zamir [36] describe a method based on "criss-cross regressions" as a means to obtain lower rank approximations of the matrix **X**. Paatero *et al* [48] have described what they call "positive matrix factorization" (PMF) and have applied this to environmental problems. In addition to satisfying the minimization criterion, PMF also requires **A** and **B** to have only positive entries. The MLPCA solution developed in the previous chapter is somewhat different from these earlier approaches in several respects. First, rather than alternately optimizing **A** and **B**, one of the matrices is determined through a maximum likelihood projection onto the other. This leaves only one matrix to be optimized, simplifying the procedure. This is important, since these are multiparameter problems, often with more than one solution and prone to local minima. A second difference is that the parametric equations allow for the inclusion of intercept terms, an important consideration in chemistry where offsets are commonplace and cannot generally be remedied by mean centering in maximum likelihood estimation. Finally, this approach explicitly includes the error covariance matrix in the estimation of the model parameters.

Although the utility of MLPCA was demonstrated in Chapter 3 with several examples, the method was considered to have several drawbacks. These are best illustrated by considering the matrix form of the parametric equations for a *p*-dimensional hyperplane in an *m*-dimensional space:

$$\mathbf{x} = \mathbf{A}\mathbf{x}_{\mathrm{p}} + \mathbf{b} \qquad (4.3)$$

In this equation, **x** represents a column vector of **X** (**X** has been transposed from the previous chapter to facilitate the development of the MLPCA theory), $\mathbf{x}_{\mathrm{p}}$ is a vector of the upper *p* elements of **x**, **A** is an *mxp* matrix of model coefficients (slope parameters) and **b** is a *p*x1 vector of intercept terms. Note that the upper *pxp* matrix of **A** is the identity matrix and the upper *p* elements of **b** are zeros. If we assume that the intercepts are zero as in traditional PCA, this leaves the coefficients of **A** to be estimated - a total of *(m-p)*x*p* parameters. However, finding the optimum model should be just a matter of finding the optimal rotation of the p-dimensional hyperplane, so only m-1 angles should be required. Therefore, the problem is "over-parameterized". Furthermore, the number of parameters to be estimated can change dramatically depending on whether $\mathbf{X}$ or $\mathbf{X}^T$ is being analyzed. Another problem is the practical implementation of this model. In Equation 4.3, $\mathbf{x}_p$ represents the p "independent variables" for the parametric equations, which are arbitrarily chosen to be the first p rows of $\mathbf{x}$. In practice, the maximum likelihood estimates of $\mathbf{x}_p$ are obtained from the observed measurements and the trial value of $\mathbf{A}$, and then used to estimate the remaining rows of $\mathbf{x}$. In principle, the maximum likelihood approach is independent of which rows are chosen as the independent variables, but in reality computational instabilities can arise if they are highly collinear. A final drawback of this approach is that the problem is not presented in terms of scores and loadings which are so familiar to practitioners of PCA.

In this chapter, the theory of MLPCA will be formulated in terms of singular value decomposition (SVD), which is a very common method for implementing PCA. Additionally an easy-to-implement MLPCA algorithm will be described, consisting of an alternating least squares procedure which is robust and very efficient compared to conventional gradient search methods.

The objective of this chapter is to develop the MLPCA approach in a manner consistent with the PCA formulation and present algorithms which are computationally practical. A complete analysis of the statistical properties of the method is beyond the scope of this treatment, but examples are presented to validate the method and demonstrate some of its features. Additional applications will be presented in Chapters 5 and 6.

4.2 THEORY OF MLPCA

The development of the theoretical aspects of MLPCA is presented here in four sections. First, the parametric models discussed in Chapter 3 are extended to a PCA framework, and a strategy for gradient optimization of the model parameters is discussed. In the second section, a more efficient optimization procedure based on an alternating least squares approach is described. This procedure assumes that the model contains no intercept terms and the measurements have uncorrelated errors. This algorithm will be referred to as the "standard" MLPCA algorithm since it represents the simplest case. The more general case which accommodates correlated errors is discussed in Section 4.2.3. Finally, as an analog to mean centering in traditional PCA, the incorporation of intercept terms into the MLPCA procedure is treated in Section 4.2.4.

4.2.1 MLPCA with No Intercept Terms

As mentioned in the previous chapter, starting with the *mxn* matrix of measurements, **X**, the MLPCA problem can be regarded simply as one of finding the equation for the optimum *p*-dimensional hyperplane to fit *n* points in the *m*-dimensional row space or, alternatively, *m* points in the *n*-dimensional column space. In the analysis presented here, the former approach is used, but it will become apparent that this is not important. As previously mentioned, maximum likelihood model estimation is an iterative two step procedure. First, for a set of given hyperplanar model parameters (*i.e.* slopes), the maximum likelihood estimates for the points (the column vectors of the observed data matrix, **X**) are found. These are then used to calculate the objective function in Equation 4.1 (or an analogous equation). In the second step, the model parameters are adjusted in an attempt to minimize S^2. The new model parameters are used to calculate new maximum likelihood estimates of the points and a new S^2, and the process continues until the objective function is minimized. Thus, there are two problems to address: (1) how to calculate the maximum likelihood estimates for a given set of model parameters, and (2) how to optimize the model parameters. The former was addressed in Chapter 3, but these results will be reiterated to make comparison with PCA easier. The latter problem was initially addressed with simplex optimization, but due to the points raised in Section 4.1, a more elegant optimization is described.

In accordance with the assumptions stated earlier, each column of the data matrix **X** can be considered to represent a point in the *m*-dimensional row space, with the true measurements corrupted by normally distributed errors:

$$\mathbf{x} = \mathbf{x}^{\circ} + \boldsymbol{\varepsilon} \qquad (4.4)$$

Here **x** is a column vector of **X**, $\mathbf{x}^{\circ}$ represents the error-free column vector, and $\boldsymbol{\varepsilon}$ is the vector of measurement errors, which has an error covariance matrix Ψ:

$$\Psi = cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}^{\mathrm{T}}) \qquad (4.5)$$

where "$E(\bullet)$" denotes an expectation value. Note that each column of **X** can have a different error covariance matrix. In the development of an MLPCA model, it is assumed that the error-free measurements lie on a *p*-dimensional hyperplane that can be modeled by a set of parametric equations with *p* independent variables. The independent variables for these parametric equations will be arbitrarily chosen to be the first p rows of **X**. As before, the general equation to be solved is:

$$\mathbf{x} = \mathbf{A}\mathbf{x}_p^\circ + \varepsilon \tag{4.6}$$

where $\mathbf{x}_p^\circ$ is the vector containing the first p elements of $\mathbf{x}^\circ$. In this equation, **A** is the *mxp* matrix of model coefficients (slope parameters), and the upper *pxp* submatrix of **A** is the identity matrix. Our problem here is to find the best estimate for $\mathbf{x}_p^o$ given the vector of observations (**x**), a matrix of estimated model coefficients ($\hat{\mathbf{A}}$), and an error covariance matrix (Ψ). The form of the solution is analogous to that for generalized least squares regression and yields:

$$\hat{\mathbf{x}}_p = \left(\hat{\mathbf{A}}^T\Psi^{-1}\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}^T\Psi^{-1}\mathbf{x} \tag{4.7}$$

The derivation of Equation 4.7 is presented in Appendix A. Here $\hat{\mathbf{x}}_p$ is the maximum likelihood estimate of $\mathbf{x}_p$. Substitution back into the model equation gives the maximum likelihood estimates for the remaining elements of **x**.

$$\hat{\mathbf{x}} = \hat{\mathbf{A}}\hat{\mathbf{x}}_p = \hat{\mathbf{A}}\left(\hat{\mathbf{A}}^T\Psi^{-1}\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}^T\Psi^{-1}\mathbf{x} \tag{4.8}$$

Equation 4.8 solves the first of the problems posed, allowing maximum likelihood estimates of the measurements to be easily obtained for a given set of model parameters. However, to obtain the maximum likelihood fit, it is necessary to adjust the model coefficients ($\hat{\mathbf{A}}$) to minimize the objective function, S^2. In the case of uncorrelated errors, this objective function is given by Equation 4.1. In the case where errors are correlated among the rows, a more general form of Equation 4.1 is minimized:

$$\begin{aligned} S^2 &= \sum_{j=1}^{n} \left(\mathbf{x}_j - \hat{\mathbf{x}}_j\right)^{\mathrm{T}} \Psi_j^{-1} \left(\mathbf{x}_j - \hat{\mathbf{x}}_j\right) \\ &= \sum_{j=1}^{n} \Delta\mathbf{x}_j^{\mathrm{T}} \Psi_j^{-1} \Delta\mathbf{x}_j \end{aligned} \tag{4.9}$$

where, as before, $\mathbf{x}_j$ represents a column vector of **X**. Equation 4.9 reduces to Equation 4.1 for a diagonal error covariance matrix. For the case where the error covariance matrix, Ψ, is the same for each column of **X**, Fuller has given a closed form solution for $\hat{\mathbf{A}}$ that minimizes S^2 [39, p. 292]. If Ψ is also diagonal, this is equivalent to the solution obtained by SVD if appropriate row scaling is used. However, in the general case, when the error covariance matrix varies with the columns of **X**, there is no closed form solution for $\hat{\mathbf{A}}$. Fuller suggests an iterative solution in this case [39, p. 217]. Simplex and gradient based algorithms have been successfully employed to optimize the coefficients of $\hat{\mathbf{A}}$, but in general convergence is slow and prone to local minima. Furthermore, depending on which rows are used for the "independent" variables, the numerical stability of the solution algorithm is questionable. Another drawback to this approach is that the equations developed thus far are in the form of a regression model rather than the PCA model which is sought. For these reasons, it would be more convenient to represent Equation 4.8 in terms of a PCA decomposition, *i.e.* in terms of scores and loadings. To do this, consider the form of the PCA model normally arrived at through SVD:

$$\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} \tag{4.10}$$

where $\hat{\mathbf{X}}$ is *mxn*, $\hat{\mathbf{U}}$ is *mxp*, $\hat{\mathbf{S}}$ is *pxp*, and $\hat{\mathbf{V}}$ is *nxp*. The caret on **X** denotes that these are the maximum likelihood estimates of the measurements in accordance with the *p*-dimensional model, and $\hat{\mathbf{U}}$, $\hat{\mathbf{S}}$, and $\hat{\mathbf{V}}$ are obtained from the singular value decomposition of $\hat{\mathbf{X}}$, which is constrained to be rank p. Now $\hat{\mathbf{X}}$ and $\hat{\mathbf{U}}$ are partitioned into the upper p rows ($\hat{\mathbf{X}}_1$ and $\hat{\mathbf{U}}_1$) and the lower m-p rows ($\hat{\mathbf{X}}_2$ and $\hat{\mathbf{U}}_2$) to give:

$$\hat{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{X}}_2 \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{U}}_1 \\ \hat{\mathbf{U}}_2 \end{bmatrix} \hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} = \left(\begin{bmatrix} \hat{\mathbf{U}}_1 \\ \hat{\mathbf{U}}_2 \end{bmatrix} \hat{\mathbf{U}}_1^{-1} \right) \hat{\mathbf{U}}_1 \hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} = \begin{bmatrix} \mathbf{I}_p \\ \hat{\mathbf{U}}_2 \hat{\mathbf{U}}_1^{-1} \end{bmatrix} \hat{\mathbf{X}}_1 \tag{4.11}$$

or, with reference to Equation 4.8:

$$\hat{\mathbf{A}} = \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \tag{4.12}$$

which is similar to the first step in the general procedure introduced in Chapter 3. Thus, there is a direct relationship between the parametric equations and SVD form of the model in the absence of intercepts. Substituting this into Equation 4.8:

$$\begin{aligned}
\hat{\mathbf{x}}_j &= \hat{\mathbf{A}}\left(\hat{\mathbf{A}}^{\mathrm{T}} \Psi_j^{-1} \hat{\mathbf{A}}\right)^{-1} \hat{\mathbf{A}}^{\mathrm{T}} \Psi_j^{-1} \mathbf{x}_j \\
&= \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \left(\left(\hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1}\right)^{\mathrm{T}} \Psi_j^{-1} \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \right)^{-1} \left(\hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1}\right)^{\mathrm{T}} \Psi_j^{-1} \mathbf{x}_j \\
&= \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \left(\left(\hat{\mathbf{U}}_1^{-1}\right)^{\mathrm{T}} \hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \right)^{-1} \left(\hat{\mathbf{U}}_1^{-1}\right)^{\mathrm{T}} \hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \mathbf{x}_j \\
&= \hat{\mathbf{U}} \hat{\mathbf{U}}_1^{-1} \hat{\mathbf{U}}_1 \left(\hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \hat{\mathbf{U}}\right)^{-1} \hat{\mathbf{U}}_1^{\mathrm{T}} \left(\hat{\mathbf{U}}_1^{\mathrm{T}}\right)^{-1} \hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \mathbf{x}_j \\
&= \hat{\mathbf{U}} \left(\hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \hat{\mathbf{U}}\right)^{-1} \hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \mathbf{x}_j \\
&= \mathbf{P}_j \mathbf{x}_j
\end{aligned} \tag{4.13}$$

where the projection matrix, $\mathbf{P}_j$, is given by:

$$\mathbf{P}_j = \hat{\mathbf{U}} \left(\hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \hat{\mathbf{U}}\right)^{-1} \hat{\mathbf{U}}^{\mathrm{T}} \Psi_j^{-1} \tag{4.14}$$

Like Equation 4.8, Equation 4.13 allows the maximum likelihood values for the matrix of measurements to be calculated, but in accordance with a given SVD model rather than a regression model. It is also similar to an equation developed by Bartlett in the psychometrics literature as early as 1937 [52-53] and discussed later by Lawley and Maxwell [18, Ch. 4], who describe it as an unbiased method for estimating factor scores. It should be noted that because $\hat{\mathbf{S}}$ is diagonal, the scores matrix, ($=\hat{\mathbf{U}}\hat{\mathbf{S}}$), can be substituted for $\hat{\mathbf{U}}$ in Equations 4.12-4.14. In order for Equation 4.13 to provide maximum likelihood estimates, the measurement errors should be normally distributed and the error covariance matrix needs to be available. In the case of uncorrelated errors, Ψ will be a diagonal matrix with the diagonal elements equal to the measurement variances.

As before, Equation 4.13 is used to optimize the elements of $\hat{\mathbf{U}}$ in accordance with the objective function given in Equation 4.9. In this case, however, there should be fewer parameters to optimize. While the matrix $\hat{\mathbf{A}}$ has $p(m\text{-}p)$ variable coefficients, the columns of $\hat{\mathbf{U}}$ define an orthonormal set of vectors in the row space and it is only necessary to optimize (m-1) angles in this space to define the optimum hyperplane.

To optimize the SVD model, an initial estimate for $\hat{\mathbf{U}}$, designated $\hat{\mathbf{U}}_0$, is first obtained. The column vectors of $\hat{\mathbf{U}}_0$ are then rotated in the m-dimensional space by applying an mxm rotation matrix, **T**. This gives a new estimate for $\hat{\mathbf{U}}$:

$$\hat{\mathbf{U}} = \mathbf{T}\hat{\mathbf{U}}_0 \qquad (4.15)$$

One easy way to define the rotation matrix is in terms of successive rotations about each axis. In an m-dimensional space, there are m-1 rotation angles to be specified, so we have:

$$\mathbf{T} = \mathbf{T}_1\mathbf{T}_2\cdots\mathbf{T}_{m-1} = \prod_{i=1}^{m-1}\mathbf{T}_i \qquad (4.16)$$

where,

$$\mathbf{T}_1 = \begin{bmatrix} \cos\alpha_1 & -\sin\alpha_1 & 0 & \cdots & 0 \\ \sin\alpha_1 & \cos\alpha_1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \; \mathbf{T}_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \cos\alpha_2 & -\sin\alpha_2 & \cdots & 0 \\ 0 & \sin\alpha_2 & \cos\alpha_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \text{ etc.} \qquad (4.17)$$

The problem now is one of optimizing the rotation angles (α_1 ... α_{m-1}) to minimize the objective function in Equation 4.9.

The optimization of the rotation angles can be carried out by a number of methods, but gradient methods are generally regarded as being the most efficient. These require the calculation of derivatives of S^2 with respect to the rotation angles. Since this is a non-trivial calculation, the derivation is included as Appendix B. The result is given in Equation 4.18:

$$\begin{aligned}\frac{\partial S^2}{\partial \alpha_i} = -2\sum_{j=1}^{n} \big[&\Delta\mathbf{x}_j^T\Psi_j^{-1}\mathbf{G}_i\mathbf{P}_j\mathbf{x}_j - \mathbf{x}_j^T\Psi_j^{-1}\mathbf{P}_j\mathbf{G}_i\mathbf{P}_j\Delta\mathbf{x}_j \\ &-\Delta\mathbf{x}_j^T\Psi_j^{-1}\mathbf{P}_j\mathbf{G}_i\mathbf{P}_j\mathbf{x}_j + \mathbf{x}_j^T\Psi_j^{-1}\mathbf{G}_i\mathbf{P}_j\Delta\mathbf{x}_j\big]\end{aligned} \qquad (4.18)$$

where the matrix $\mathbf{G}_i$ is defined in the appendix. This equation has been checked against numerically calculated derivatives and found to be correct. It can be used in conjunction with standard gradient techniques to find the optimum rotation of eigenvectors to minimize the objective function in Equation 4.9. In practice, this procedure is faster and more reliable than using the regression form of the equation, but it is still relatively slow and susceptible to local minima. Therefore, an alternative approach was sought. This is described in the next section.

4.2.2 An Efficient MLPCA Algorithm

In order to be useful with large data sets, an MLPCA procedure is needed which converges relatively quickly. Among the most efficient methods in this regard are iterative procedures, such as alternating regression approaches. Such a solution was developed for the MLPCA problem and was based on the following rationale.

It will be assumed for the moment that all measurement errors are independent so that error covariance matrices in both the row and column spaces are diagonal. If this is true, the *p*-dimensional model obtained by maximum likelihood estimation must be equivalent in both spaces. This follows because the objective function in both cases reduces to the same summation given by Equation 4.1. Mathematically,

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(x_{ij}-\hat{x}_{ij}\right)^2}{\sigma_{ij}^2} = \sum_{j=1}^{n}\Delta\mathbf{x}_j^{\mathrm{T}}\Psi_j^{-1}\Delta\mathbf{x}_j = \sum_{i=1}^{m}\Delta\mathbf{x}_i^{\mathrm{T}}\Sigma_i^{-1}\Delta\mathbf{x}_i \qquad (4.19)$$

Here, $\Delta\mathbf{x}_j$ is a column vector of $\Delta\mathbf{X}$, $\Delta\mathbf{x}_i$ is a column vector of $(\Delta\mathbf{X})^{\mathrm{T}}$, and Ψ_j and Σ_i are the corresponding column and row error covariance matrices for **X**, both of which are diagonal. For ease of visualization, some of the matrices are represented pictorially in Figure 4.1a. In order to develop the alternating regression algorithm, Equation 4.10 will be rewritten as:

$$\hat{\mathbf{X}}^{\mathrm{T}} = \hat{\mathbf{V}}\hat{\mathbf{S}}\hat{\mathbf{U}}^{\mathrm{T}} \qquad (4.20)$$

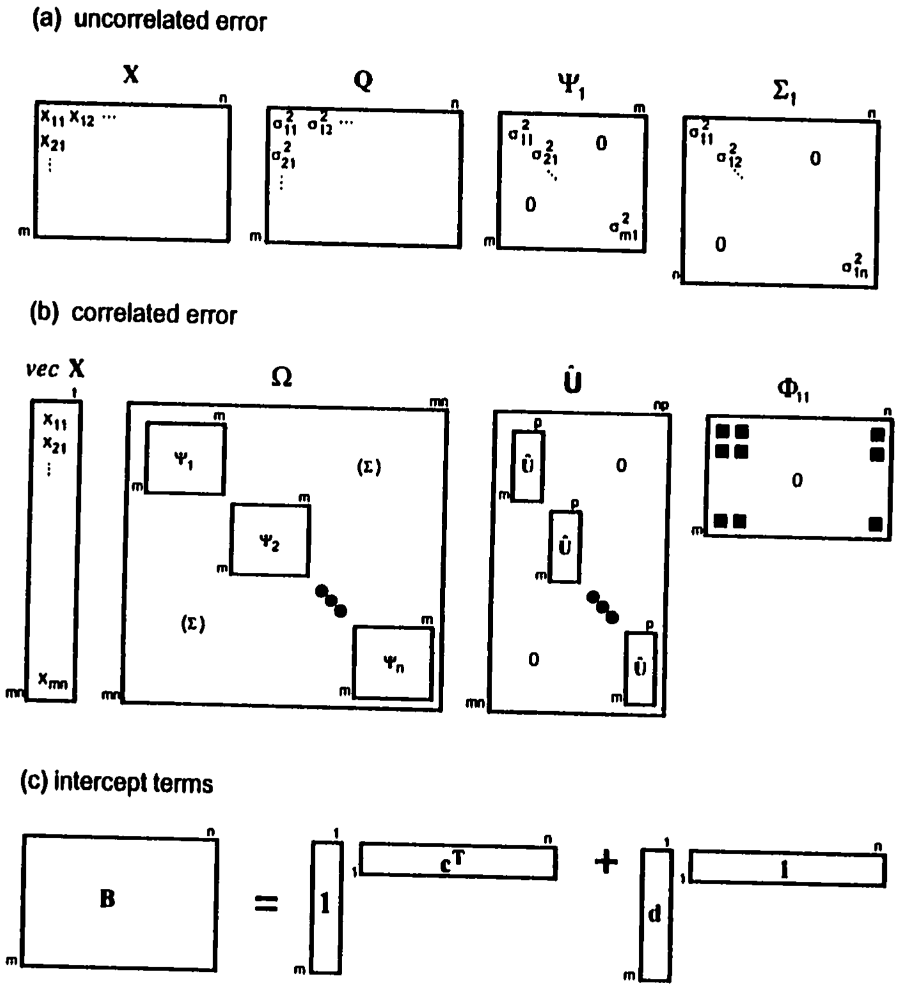


Figure 4.1 Pictorial representation of some MLPCA matrices: (a) matrices used in the standard algorithm, (b) matrices used in the algorithm which incorporates error covariance, (c) composition of the background matrix in the algorithm incorporating intercept terms.

This suggests that the maximum likelihood estimates of the measurements in column space are given by an equation which is analogous to Equation 4.13:

$$\hat{\mathbf{x}}_i = \hat{\mathbf{V}}\left(\hat{\mathbf{V}}^T\Sigma_i^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^T\Sigma_i^{-1}\mathbf{x}_i \tag{4.21}$$

where, as before, $\mathbf{x}_i$ is a column vector of $\mathbf{X}^T$ and Σ_i is the corresponding error covariance matrix. When the maximum likelihood solution has been obtained, the estimates of **X** in the row and column spaces will be identical. This implies that an alternating regression approach can be developed by alternately transposing the maximum likelihood estimates and performing SVD.

The algorithm for the alternating regression procedure is given in Table 4.1 (with the Matlab code presented in Appendix C). It should be noted that the algorithm has been expanded to show a full iteration for clarity, but there are some redundancies in the procedure that can be exploited to make the actual code more compact. The algorithm alternately uses the maximum likelihood estimates in the original row space to update the estimates in the column space (*i.e.* the row space of the transposed matrix), and vice versa. This procedure has been found to be simple, fast and reliable. It does not appear to be susceptible to local minima, as is the case for gradient methods. Convergence time will depend on the dimensionality of the problem, the accuracy of the initial SVD estimate, and the structure of the errors. The algorithm is easily applied to cases where there are missing data simply by incorporating large variances for the missing measurements. Convergence is somewhat slower in these cases due to the poor initial estimates obtained when the missing measurements are replaced with zeros, but is still reliable. Some comparative data on convergence times are given in the Section 4.4.1.

Table 4.1 Standard MLPCA algorithm (uncorrelated errors, no intercepts)

1. Given an *mxn* data matrix **X** and a corresponding *mxn* matrix **Q** of measurement error variances, use SVD to obtain an initial approximation to the MLPCA solution. The SVD solution is truncated to rank *p* as indicated by the notation *svd*(**X**,*p*). This means that **U**, **S**, and **V** are truncated to *mxp*, *pxp*, and *nxp*, respectively.

$$\left[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}\right] \Leftarrow svd(\mathbf{X}, p) \qquad \text{(T-1)}$$

2. Transpose **X** and **Q** and calculate the maximum likelihood estimates in the alternate space using $\hat{\mathbf{V}}$.

$$\mathbf{X} \Leftarrow \mathbf{X}^{\mathrm{T}}, \qquad \mathbf{Q} \Leftarrow \mathbf{Q}^{\mathrm{T}}, \quad \Sigma_{\mathrm{i}} \Leftarrow diag(\mathbf{q}_{\mathrm{i}}) \qquad \text{(T-2)}$$

$$\hat{\mathbf{x}}_{\mathrm{i}} = \hat{\mathbf{V}}\left(\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_{\mathrm{i}}^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_{\mathrm{i}}^{-1}\mathbf{x}_{\mathrm{i}} \qquad \text{(T-3)}$$

Here $\mathbf{x}_{\mathrm{i}}$ is a column vector of the now transposed **X**. From this result, the objective function can be calculated using eqn T-4.

$$S_1^2 = \sum_{\mathrm{i}=1}^{\mathrm{m}}(\mathbf{x}_{\mathrm{i}} - \hat{\mathbf{x}}_{\mathrm{i}})^{\mathrm{T}}\Sigma_{\mathrm{i}}^{-1}(\mathbf{x}_{\mathrm{i}} - \hat{\mathbf{x}}_{\mathrm{i}}) = \sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\frac{\left(x_{\mathrm{ji}} - \hat{x}_{\mathrm{ji}}\right)^2}{\sigma_{\mathrm{ji}}^2} \qquad \text{(T-4)}$$

3. Compute the SVD of $\hat{\mathbf{X}}$ from step 2 and, as before, truncate the results to obtain a new $\hat{\mathbf{V}}$.

$$\left[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}\right] \Leftarrow svd\left(\hat{\mathbf{X}}, p\right) \qquad \text{(T-5)}$$

4. Repeat step 2 to estimate the model in the original space.

$$\mathbf{X} \Leftarrow \mathbf{X}^{\mathrm{T}}, \qquad \mathbf{Q} \Leftarrow \mathbf{Q}^{\mathrm{T}}, \quad \Psi_{\mathrm{j}} \Leftarrow diag(\mathbf{q}_{\mathrm{j}}) \qquad \text{(T-6)}$$

$$\hat{\mathbf{x}}_{\mathrm{j}} = \hat{\mathbf{V}}\left(\hat{\mathbf{V}}^{\mathrm{T}}\Psi_{\mathrm{j}}^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^{\mathrm{T}}\Psi_{\mathrm{j}}^{-1}\mathbf{x}_{\mathrm{j}} \qquad \text{(T-7)}$$

$$S_2^2 = \sum_{\mathrm{j}=1}^{\mathrm{n}}\left(\mathbf{x}_{\mathrm{j}} - \hat{\mathbf{x}}_{\mathrm{j}}\right)^{\mathrm{T}}\Psi_{\mathrm{j}}^{-1}\left(\mathbf{x}_{\mathrm{j}} - \hat{\mathbf{x}}_{\mathrm{j}}\right) = \sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\frac{\left(x_{\mathrm{ij}} - \hat{x}_{\mathrm{ij}}\right)^2}{\sigma_{\mathrm{ij}}^2} \qquad \text{(T-8)}$$

5. Compute the SVD of $\hat{\mathbf{X}}$ to obtain a new estimate of the MLPCA solution in the original space.

$$\left[\hat{\mathbf{U}}, \hat{\mathbf{S}}, \hat{\mathbf{V}}\right] \Leftarrow svd\left(\hat{\mathbf{X}}, p\right) \qquad \text{(T-9)}$$

6. Calculate the convergence parameter, λ.

$$\lambda = (S_1^2 - S_2^2)\,/\,S_2^2 \qquad \text{(T-10)}$$

If λ is less than the convergence limit (typically 10^{-10} in this work), terminate. Otherwise return to step 2.

The algorithm presented in Table 4.1 does impose certain restrictions. First, it is assumed that there are no offsets in the row or column space. Normally, this would be equivalent to saying that the data have been mean-centered, but in the case of non-uniform measurement errors, mean-centering is not generally equivalent to eliminating offsets. The topic of row and column offsets is discussed in Section 4.2.4.

Another restriction to the algorithm presented here is that it assumes uncorrelated errors in the row and column spaces. The algorithm will not converge to a common solution if the covariance matrices used are not diagonal in both spaces. This raises an interesting question. Suppose, for example, that one is dealing with a series of *m* samples whose spectra are measured at *n* different wavelengths. Also imagine that, because of instrumental characteristics, errors are correlated in the wavelength direction, but there is no correlation in the errors among the samples. Under these conditions, the *m*x*m* error covariance matrices in the row space, Ψ_j, are diagonal and minimization of S^2 should lead to the same solution regardless of whether the *n*x*n* error covariance matrices in the column space, Σ_i , are diagonal or not, since there is no information about wavelength correlation in the Ψ_j. However, it is apparent that the maximum likelihood estimates of **X** obtained by Equation 4.21 depend on whether or not the Σ_i are diagonal and will not be the same as the maximum likelihood estimates obtained by Equation 4.13 if there are correlated errors. Therefore, it seems that the maximum likelihood solution found in one space is not generally equivalent to that found in the alternate space in the presence of correlated errors. The reason for this apparent paradox is that the points in the row space are assumed to be independent, which will not be true if errors are correlated in the wavelength direction, so the model is invalid. The subject of correlated measurement errors is addressed in the next section.

4.2.3 Error Covariance

When measurements are made during the course of an experiment, there is a realistic possibility that random errors in-these measurements will be correlated with one another because of the design of the experiment or the nature of the samples. Even if the original measurement errors are not correlated, it is possible that pre-processing methods such as digital filtering can introduce such correlation. To our knowledge, no one has attempted to develop PCA models which deal with correlated measurement errors, although there has been recognition of the importance of correlated errors in the literature [54]. Earlier works cited [36,48] attempt to develop algorithms to minimize Equation 4.1, which assumes uncorrelated errors for maximum likelihood estimation. In the more general case, we wish to minimize Equation 4.9, which incorporates the non-diagonal error covariance matrix. Furthermore, the model developed should be consistent with an SVD formulation such that the maximum likelihood estimates obtained in either the row or the column spaces will be the same. In practice, one is fortunate to have individual measurement standard deviations available, and information on error covariance is rare. Nevertheless, it is useful to develop a theoretical framework for using such information if for no other reason than to assess its value.

There are essentially three common cases of error correlation that can be distinguished: (1) all measurement errors are uncorrelated, (2) correlations among errors exist along either the rows or columns of the data matrix, but are uncorrelated in the other direction, and (3) there is some degree of possible correlation among all of the measurement errors. The first case was dealt with in the preceding two sections and the third is the completely general case which has yet to be addressed. To begin, however, it is helpful to examine the second case which is more restricted.

An example of the second case was presented earlier and will be considered again here. Consider a series of spectra whose errors are correlated in the wavelength direction (*e.g.* by source fluctuations) but not correlated among samples. If this were true, the error covariance matrix for column j of $\mathbf{X}$ (m samples by n wavelengths), Ψ_j, would be diagonal, but that for row i, Σ_i, would not. Variance information is carried in both spaces, but covariance information is only carried in the column space in this case. Therefore it would seem logical to compute the maximum likelihood estimates using Equation 4.21 and minimize the objective function by rotating the columns of $\mathbf{V}_0$. In principle, this can be done and will lead to the correct result. However, it would have to be done using the gradient methods described in Section 4.2.1 rather than the much more efficient algorithm described in Section 4.2.2, since we can no longer interchange the row and column spaces. Furthermore, when the final solution is obtained, the maximum likelihood estimates of $\mathbf{X}$ computed by Equation 4.21 in the column space will not be the same as those calculated in the row space using Equation 4.13. This is an apparent contradiction, since there should only be one set of maximum likelihood projections which are the same in either space. The reason for this paradox is that there is no information about wavelength correlation in the row space, so the maximum likelihood estimates generated there are wrong. Realizing this, one could simply use the estimates obtained from Equation 4.21, but this does not address the more general problem of incorporating the error covariance information in both spaces.

To arrive at a more general solution for correlated errors, it is necessary to realize that any pair of measurement errors could be correlated and redefine the problem accordingly. Rather than considering it as modeling *n* points in an *m*-dimensional space, or *m* points in an *n*-dimensional space, it will be viewed as modeling a single point in an *mn*-dimensional space. To do this, **X** is vectorized by applying the *vec* operator and the equations are adapted as necessary. The generalizations of Equations 4.13 and 4.9 are:

$$vec\left(\hat{\mathbf{X}}\right) = \hat{\mathsf{U}}\left(\hat{\mathsf{U}}^{T}\Omega^{-1}\hat{\mathsf{U}}\right)^{-1}\hat{\mathsf{U}}^{T}\Omega^{-1}\, vec(\mathbf{X}) \tag{4.22}$$

and

$$S^{2} = vec(\Delta\mathbf{X})^{T}\,\Omega^{-1}\, vec(\Delta\mathbf{X}) \tag{4.23}$$

where,

$$\hat{\mathsf{U}} = \mathbf{I}_{n} \otimes \hat{\mathbf{U}} \tag{4.24}$$

and

$$\Omega = E\left[\left(vec\left(\mathbf{X}-\mathbf{X}^{\circ}\right)\right)\cdot\left(vec\left(\mathbf{X}-\mathbf{X}^{\circ}\right)\right)^{T}\right] \tag{4.25}$$

Here the *vec* operator gives an *mn*x1 vector with the column vectors of **X** arranged in sequence [55]. The symbol "⊗ " indicates the Kronecker product such that each element of $\mathbf{I}_n$ is multiplied by $\hat{\mathbf{U}}$ [55]. Thus, $\hat{\mathsf{U}}$ is an *mn*x*np* matrix with $\hat{\mathbf{U}}$ (*m*x*p*) repeating along the diagonal. Ω is the full covariance matrix for *vec*(**X**), providing the error covariance among all of the measurements. $\mathbf{X}^{\circ}$ represents the true (or expectation) values for **X**. For greater clarity, some of these matrices are shown pictorially in Figure 4.1b. Note that the column covariance matrices of **X**, represented as Ψ, fall along the diagonal of Ω. The remainder of Ω is made up with the row covariance information (Σ) and other covariances.

With these definitions, an alternating regression algorithm similar to the one in the preceding section can be developed, and is given in Table 4.2. As before, the algorithm above uses the maximum likelihood estimates in one space to estimate the solution in the alternate space. As the solutions are exchanged, the error covariance matrix for *vec*(**X**) (given by Ω) needs to be modified to give the covariance matrix for *vec*($\mathbf{X}^{\mathrm{T}}$) (given by Ξ). This can be done on an element-by-element basis, but it is easier to use the commutation matrix, **K** [55]. The commutation matrix is an orthonormal matrix that has the property:

$$vec\left(\mathbf{A}^{\mathrm{T}}\right) = \mathbf{K}\, vec\left(\mathbf{A}\right) \qquad (4.26)$$

When combined with the definition in Equation 4.23, this leads to the use in Equation B-12 to transform the error covariance matrix into the alternate space. In practice, the commutation matrix can be computed as follows. Begin with a *mn*x1 vector **a**, such that $a_i = i$. Reshape **a** so that it forms the *m*x*n* matrix **A** and then set **b** = *vec*($\mathbf{A}^{\mathrm{T}}$). Now the corresponding elements of **a** and **b** are the row and column indices, respectively, of the elements of the *mn*x*mn* commutation matrix, **K**, that should be set to 1. The remaining elements of **K** should be set to zero, making it a sparse matrix with *mn* non-zero elements.

The algorithm in Table 4.2 represents a completely general treatment for the case of correlated measurement errors and therefore is a significant advance in multivariate modeling. The algorithm (written in Matlab code) is presented in Appendix D. It converges rapidly to an optimal solution (unless the matrices involved are numerically unstable) and yields results identical to the earlier algorithm in the presence of uncorrelated errors. In practice, use of the algorithm is currently limited to some extent by the size and stability of the matrices. In the completely general case, the covariance

Table 4.2 MLPCA algorithm for correlated measurement errors.

1. Given an *mxn* matrix **X**, a corresponding *mnxmn* matrix $\boldsymbol{\Omega}$ of measurement error covariances for *vec*(**X**), and a commutation matrix, **K**, for **X**, use a truncated SVD to obtain an initial approximation to the MLPCA solution.

$$\left[\hat{\mathbf{U}},\hat{\mathbf{S}},\hat{\mathbf{V}}\right] \Leftarrow svd(\mathbf{X},p) \qquad \text{(T-11)}$$

2. Transpose **X** and calculate the maximum likelihood estimates in the alternate space using $\hat{\mathbf{V}}$.

$$\mathbf{X} \Leftarrow \mathbf{X}^{\mathrm{T}}, \qquad \boldsymbol{\Xi}^{-1} \Leftarrow \mathbf{K}\boldsymbol{\Omega}^{-1}\mathbf{K}^{\mathrm{T}} \qquad \text{(T-12)}$$

$$\hat{\boldsymbol{\mathcal{V}}} = \mathbf{I}_{\mathrm{m}} \otimes \hat{\mathbf{V}} \qquad \text{(T-14)}$$

$$vec\left(\hat{\mathbf{X}}\right) = \hat{\boldsymbol{\mathcal{V}}}\left(\hat{\boldsymbol{\mathcal{V}}}^{\mathrm{T}}\boldsymbol{\Xi}^{-1}\hat{\boldsymbol{\mathcal{V}}}\right)^{-1}\hat{\boldsymbol{\mathcal{V}}}^{\mathrm{T}}\boldsymbol{\Xi}^{-1}\,vec(\mathbf{X}) \qquad \text{(T-15)}$$

From this result, the objective function can be calculated using eqn T-16.

$$S_1^2 = vec(\Delta\mathbf{X})^{\mathrm{T}}\,\boldsymbol{\Xi}^{-1}\,vec(\Delta\mathbf{X}) \qquad \text{(T-16)}$$

3. Reconstruct $\hat{\mathbf{X}}$ from *vec*($\hat{\mathbf{X}}$) and compute the truncated SVD of $\hat{\mathbf{X}}$.

$$\left[\hat{\mathbf{U}},\hat{\mathbf{S}},\hat{\mathbf{V}}\right] \Leftarrow svd\left(\hat{\mathbf{X}},p\right) \qquad \text{(T-17)}$$

4. Repeat step 2 to estimate the model in the original space.

$$\mathbf{X} \Leftarrow \mathbf{X}^{\mathrm{T}}, \qquad \boldsymbol{\Omega}^{-1} \Leftarrow \mathbf{K}^{\mathrm{T}}\boldsymbol{\Xi}^{-1}\mathbf{K} \qquad \text{(T-18)}$$

$$\hat{\boldsymbol{\mathcal{V}}} = \mathbf{I}_{\mathrm{n}} \otimes \hat{\mathbf{V}} \qquad \text{(T-19)}$$

$$vec\left(\hat{\mathbf{X}}\right) = \hat{\boldsymbol{\mathcal{V}}}\left(\hat{\boldsymbol{\mathcal{V}}}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\mathcal{V}}}\right)^{-1}\hat{\boldsymbol{\mathcal{V}}}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\,vec(\mathbf{X}) \qquad \text{(T-20)}$$

$$S_2^2 = vec(\Delta\mathbf{X})^{\mathrm{T}}\,\boldsymbol{\Omega}^{-1}\,vec(\Delta\mathbf{X}) \qquad \text{(T-21)}$$

5. Reconstruct $\hat{\mathbf{X}}$ (original dimensions) and compute the truncated SVD of $\hat{\mathbf{X}}$ in the original space.

$$\left[\hat{\mathbf{U}},\hat{\mathbf{S}},\hat{\mathbf{V}}\right] \Leftarrow svd\left(\hat{\mathbf{X}},p\right) \qquad \text{(T-22)}$$

6. Compute the convergence parameter (eqn T-10) and terminate if is less than the convergence limit. Otherwise, return to step 2.

matrix will have m^2n^2 elements and easily exceeds the storage capacity of most machines for large matrices unless special measures are used. The matrices also tend to become ill-conditioned as **X** becomes large, causing convergence problems. However, for many chemical problems, error covariance is limited to either the row or column directions. In these cases either Ω or Ξ will be block diagonal and can be stored as sparse matrices. The diagonal blocks of these matrices (Ψ or Σ) can be inverted individually, and the covariance matrix in the alternate space can be calculated with the commutation matrix. In this way, the algorithm can be extended to a much wider set of problems.

4.2.4 MLPCA with Intercepts

In models for chemical systems, it is common for row and column offsets to be present for the matrix of measurements. Returning to the earlier example from spectroscopy, one can imagine a situation in which a constant background spectrum is present for all of the samples. If **X** is *m* samples by *n* wavelengths, this can be considered a vector of column offsets. Since this is invariant for all samples, it is often desirable to subtract it from each sample spectrum prior to decomposition of the data matrix, thereby achieving a reduction in rank. Likewise, one can imagine a vector of row offsets that arises from, say, variations in cell position or sample preparation. This can also be removed. Models which include such effects in chemistry are the same as that developed by Mandel for analysis of variance [56], namely:

$$x_{\mathrm{ij}} = \mu + \rho_{\mathrm{i}} + \gamma_{\mathrm{j}} + \sum_{\mathrm{k=1}}^{\mathrm{p}} u_{\mathrm{ik}} s_{\mathrm{kk}} v_{\mathrm{jk}} \qquad (4.27)$$

Here μ is the grand mean of **X**, ρ_i and γ_j represent row and column offsets, respectively, and *u*, *s*, and *v* are individual elements of the SVD of the matrix with the offsets removed. The elements of the vectors ρ and γ are often taken to be the means of the rows and columns after the grand mean is subtracted. Note that Equation 4.27 is a general formulation. In a given application, the row and/or column offsets could be set to zero (as they often are) or could even be constrained, for example, to each have identical elements (a situation rarely imposed in chemistry). Also note that the grand mean, since it is constant, can be incorporated into either or both of the offset terms, so can be excluded from Equation 4.27. For the purposes of this discussion, an alternate form of Equation 4.27 will be used:

$$\begin{aligned}\hat{\mathbf{X}} &= \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} + \mathbf{1}_{\mathrm{m}}\mathbf{c}^{\mathrm{T}} + \mathbf{d}\,\mathbf{1}_{\mathrm{n}}^{\mathrm{T}} \\ &= \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} + \mathbf{B}\end{aligned} \qquad (4.28)$$

In this equation, **c** and **d** are column vectors of the row and column offsets, and $\mathbf{1}_{\mathrm{m}}$ and $\mathbf{1}_{\mathrm{n}}$ are column vectors of 1's of length *m* and *n*. This representation of the matrix of offsets, **B**, is shown pictorially in Figure 4.1c. It is clear from the figure that the presence of row or column offsets will increase the rank of an untreated data matrix by one, while the presence of both will increase the rank by two.

In chemistry, realization of a model of the form of Equation 4.28 is normally accomplished by column and/or row mean centering to determine **c** and **d**. If only one of these is to be used, the means of the columns or rows can be used directly as **c** or **d**. If both are used, the grand mean must first be subtracted from the data matrix, or one set of offsets needs to be calculated after the other has been subtracted from **X**.

It should be pointed out that there are an infinite number of row and column offset vectors which will provide equivalent models in terms of quality of fit. This is illustrated in Figure 4.2 for the case of rank one data (with an offset) in a two-dimensional space. Note that a zero intercept for the model can be obtained in at least three ways, and there are infinitely more. In fact, the illustration shows that any one of the offset terms (x or y) can be set to zero without changing the quality of the fit. In general, any p offset terms can be set to zero for both **c** and **d**, which is why the degrees of freedom is reduced by (n-p) and/or (m-p) when mean-centering is used.

When all of the measurements in **X** have normal *iid* errors, mean-centering to remove offsets is a convenient approach to use because the characteristics of PCA *guarantee* that the mean will fall on the optimum model, so forcing the mean to zero ensures that all of the intercept terms will also be zero for the centered data. However, for MLPCA, the presence of non-uniform and/or correlated error distributions means that this is no longer generally true, although it may be a good approximation. For this reason, the row and column offset vectors need to be optimized along with the scores and loadings in order to obtain a true maximum likelihood solution. Attempts to include these parameters into the alternating regression algorithms already presented have not been successful thus far, and generally result in convergence on a suboptimal solution. As an alternative, more traditional gradient methods have been coupled with the alternating regression procedure to yield the MLPCA solution. Although this is slower than the standard MLPCA algorithm, it converges reliably. The algorithm (written in Matlab code) is presented in Appendix E.

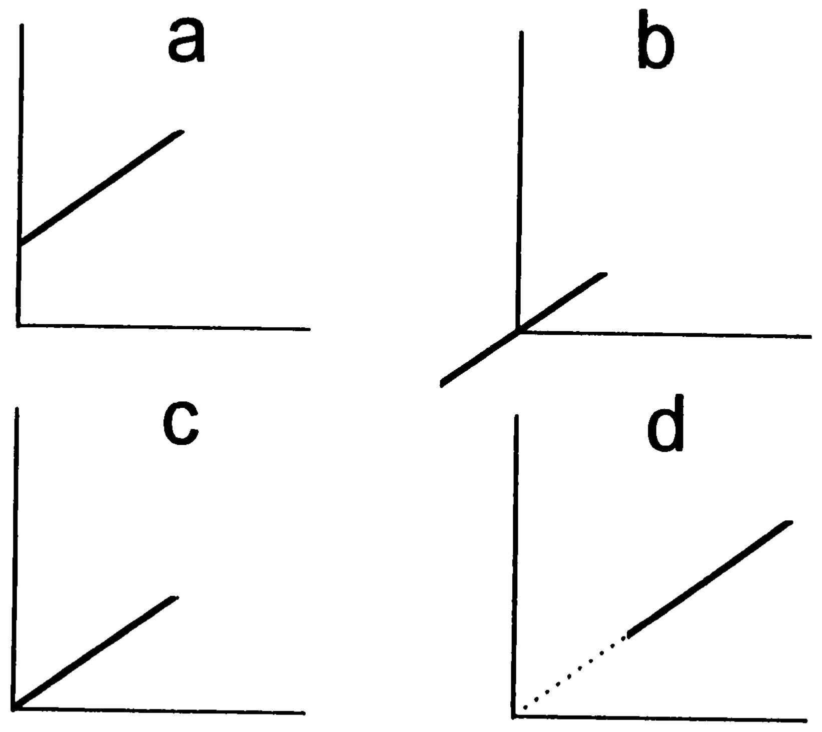


Figure 4.2 Representation of equivalent translations for removing model intercepts: (a) original data, (b) mean-centered in x and y, (c) offset of zero in x (d) offset of zero in y.

The procedure begins by finding initial estimates for the row and/or column offset vectors. One way to do this would be to use the corresponding means, but an alternate procedure is chosen here. The data are first analyzed using the algorithm with no intercepts, but increasing the model rank by 1 or 2 to account for the offset vectors. The row and/or column means of the maximum likelihood estimates are then used as a starting point for the offset parameters. As noted above, the procedure should require the optimization of only (*n-p*) row offsets and/or (*m-p*) column offsets. However, the full vectors in each direction (*n* and/or *m* parameters) have been used in this work to simplify the conversion from and comparison with the row and column means. This will lead to degenerate solutions, but does not seem to affect convergence.

Once initial estimates for **c** and **d** have been obtained, these are used to calculate **B** (see Equation 4.28) and this is subtracted from **X**. The alternating regression algorithm is then applied to the adjusted matrix. As soon as the convergence criterion has fallen below an acceptable value, a gradient search is implemented to optimize **c** and/or **d**. The results of this are used to calculate a new **B**, which is then subtracted from the original **X**, and the process is repeated until the change in the objective function is acceptably small. In order to carry out the gradient optimization, the derivatives of S^2 with respect to the intercept parameters are needed. For uncorrelated errors, these can be obtained by using the equation for $\Delta\mathbf{x}$ in the presence of intercepts:

$$\Delta\mathbf{x}_{\mathrm{j}} = \left(\mathbf{I} - \hat{\mathbf{U}}\left(\hat{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{j}}^{-1}\hat{\mathbf{U}}\right)^{-1}\hat{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\Psi}_{\mathrm{j}}^{-1}\right)\left(\mathbf{x}_{\mathrm{j}} - \mathbf{b}_{\mathrm{j}}\right) \qquad (4.29)$$

Here $\mathbf{b}_{\mathrm{j}}$ is a column vector of **B**. This gives:

$$\frac{\partial S^2}{\partial \mathrm{c}_j} = \Delta \mathbf{x}_j^T \Psi_j^{-1}\left(\mathbf{I} - \hat{\mathbf{U}}\left(\hat{\mathbf{U}}^T \Psi_j^{-1} \hat{\mathbf{U}}\right)^{-1} \hat{\mathbf{U}}^T \Psi_j^{-1}\right) \mathbf{1}_m \tag{4.30}$$

$$\frac{\partial S^2}{\partial \mathbf{d}} = \sum_{j=1}^{n}\left[\Delta \mathbf{x}_j^T \Psi_j^{-1}\left(\mathbf{I} - \hat{\mathbf{U}}\left(\hat{\mathbf{U}}^T \Psi_j^{-1} \hat{\mathbf{U}}\right)^{-1} \hat{\mathbf{U}}^T \Psi_j^{-1}\right)\right]^T \tag{4.31}$$

These equations were employed for the gradient optimization.

4.3 EXPERIMENTAL

All of the calculations performed in this study were carried out on a DEC 3000/300X UNIX-based workstation with a clock speed of 175 MHz and 96 MB of memory (Digital Equipment Corp., Maynard, MA). Programs were written in Matlab v. 4.2a (The Math Works Inc., Natick, MA). Seven data sets were used to evaluate the algorithms.

Data set 1 was a simulated rank two data set of dimensions 10x20. The error-free data matrix was generated by multiplying a 10x2 matrix of elements from a uniform distribution of random numbers between 0 and 1 (U(0,1)) by a 2x20 matrix that was also drawn from U(0,1). Measurement standard deviations corresponding to this 10x20 matrix were determined by generating a 10x20 matrix of random numbers from U(0,0.01). This ensured that there was no pattern in the standard deviations. Finally, a 10x20 matrix of measurement errors was generated by taking a 10x20 matrix of normally distributed random numbers (mean=0, standard deviation=1, or N(0,1)) and multiplying this on an element-by-element basis by the matrix of standard deviations. The result was added to the error free matrix to give the noisy data, **X**. The matrix of variances, **Q**, was obtained by squaring the elements of the standard deviation matrix. The matrices **X** and **Q** were passed to the MLPCA algorithm for uncorrelated errors.

Data sets 2 and 3 were also rank two matrices generated in the same manner as data set 1, except that their dimensions were 20x20 and 20x100.

Data set 4 was simulated rank three spectral data. Pure component spectra were simulated as three Gaussian profiles spaçed 20 nm apart, each with a standard deviation of 20 nm and a maximum height of unity. The maximum of the center profile was at 500 nm, and 41 equally spaced points were calculated for each spectrum in the range of 400 to 600 nm. A 20x3 concentration matrix was generated by drawing random numbers from a U(0,1) distribution. The 20x41 error-free matrix was the product of the concentration matrix and the 3x41 matrix of spectral profiles. To provide a matrix of standard deviations that was unstructured (*i.e.* rank>one) but still realistic, constant and proportional errors were used. The constant part was taken to be 1% of the maximum value of the noise-free data matrix. The 20x41 matrix of proportional standard deviations was calculated as 5% of the elements in the error-free data matrix. The overall matrix of standard deviations was the square root of the sum of the squares of the proportional part and the constant part. Finally, random numbers from an N(0,1) distribution were multiplied by each element of the standard deviation matrix to give the error matrix, which was added to the error-free data to give the noisy data matrix, **X**.

Data set 5 was generated in exactly the same manner as data set 4, except that random offsets, drawn from an N(0,0.1) distribution, were added to each row and column of the final **X**. This was intended to test the version of the MLPCA algorithm designed to fit intercept terms.

Data set 6 consisted of near-infrared spectroscopic data for three-component mixtures containing toluene, chlorobenzene, and heptane. The mixtures were prepared as part of a calibration transfer study [57] by Scott Specialty Gases (Houston, TX) and consisted of 31 samples from an augmented three-level, three-factor factorial design. The concentrations varied between 20% and 70% by weight for toluene and chlorobenzene, and 2% and 10% by weight for heptane: The mixtures were sealed into standard 1 cm pathlength cuvettes and spectra were obtained over the range 400-2500 nm on an NIRSystems Model 6500 (NIRSystems, Silver Spring, MD) grating spectrometer at intervals of 2 nm and were the average result of 32 scans. The spectrometer employed a Si detector in the range 400-1100 nm and a PbS detector at longer wavelengths. A typical spectrum is shown in Figure 4.3a. Clearly there are some regions above 1600 nm which are essentially opaque and therefore of little utility for analysis. Consequently, standard deviations in this region are high. However, these wavelengths were retained in this study for the purpose of illustrating the features of MLPCA. Unfortunately, replicate data were only available for the first sample, for which 400 spectra had been obtained, so a complete matrix of standard deviations could not be constructed. Instead, the standard deviation data for the first sample, shown in Figure 4.3b, was used for all of the samples. Although this is not completely accurate, it should serve as a reasonable approximation, especially for regions where the standard deviation is very large due to high absorbance.

Data set 7 was a 5x10 matrix constructed in the same manner a data set 1, except that correlated errors were introduced. To produce error covariance, a 3x3 moving average filter (coefficients = 1/9) was applied to the 5x10 matrix of errors before it was added to the error-free measurements. At the boundaries of the error matrix, the filter was wrapped around to the opposite side in order to eliminate edge effects. Although this approach is not particularly realistic, it represents a general case for which the covariance structure could be easily predicted. The covariance matrix for this data set was calculated using the following definitions:

$$\mathbf{F} = \left[vec(\Phi_{11}) \quad vec(\Phi_{21}) \quad \cdots \quad vec(\Phi_{mn}) \right] \tag{4.32}$$

$$\Omega = \mathbf{F}^{\mathrm{T}}\, diag\big(vec(\mathbf{Q})\big)\ \mathbf{F} \tag{4.33}$$

Here **Q** is the 5x10 matrix of error variances prior to the application of the moving average filter. The 5x10 matrix Φ_{ij} contains the nine filter coefficients applied to the error matrix to give the filtered error corresponding to measurement *ij*. This is illustrated in Figure 4.1b for Φ_{11}, where the filled squares show the positions of the filter coefficients. For Φ_{12}, the squares shift right, and for Φ_{21}, they shift down. Expressed another way, if **E** represents the 5x10 matrix of uncorrelated errors generated in accordance with the variances in **Q**, and ε_{ij} represents the error added to element *ij* of the pure data matrix, we have:

$$\varepsilon_{ij} = \left[vec(\Phi_{ij}) \right]^{\mathrm{T}} vec(\mathbf{E}) \tag{4.34}$$

The errors generated in this way were added to the pure data matrix. The noisy data matrix, **X**, and the error covariance matrix, Ω, were passed to the MLPCA algorithm.

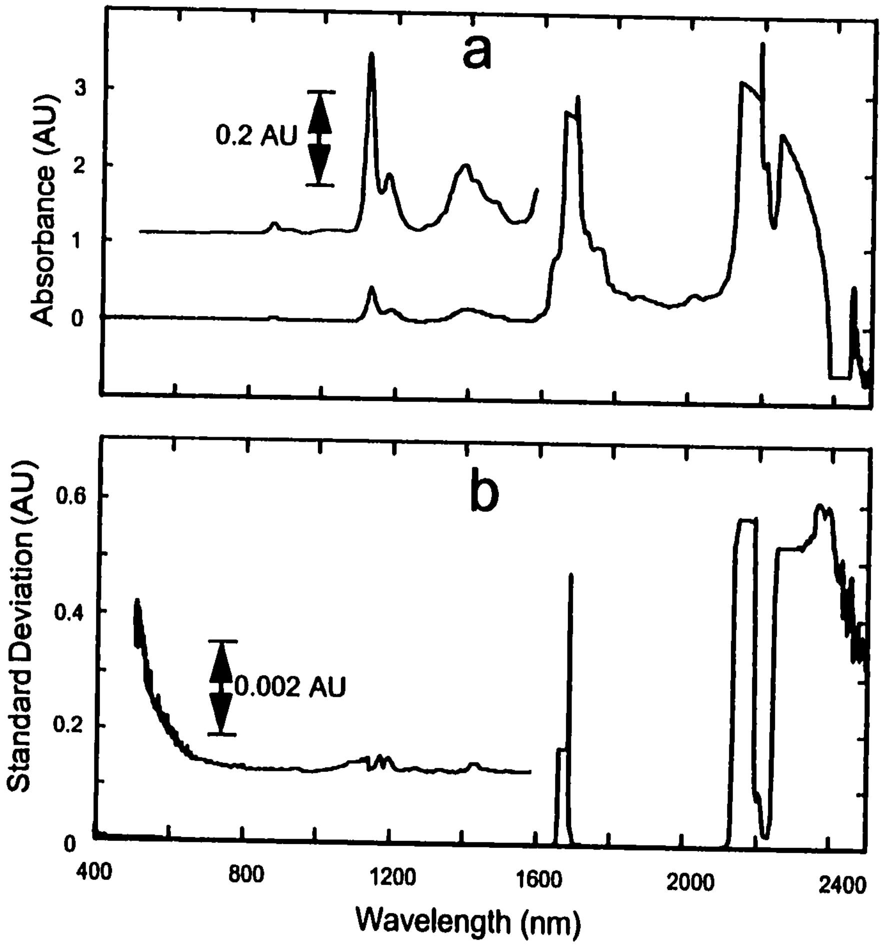


Figure 4.3 (a) Typical near infrared spectrum of three-component (toluene-chlorobenzene-heptane) mixture and (b) Standard deviation of replicate scans. The region between 500 and 1600 nm has been enlarged (inset) for greater clarity.

4.4 RESULTS

4.4.1 Algorithm Performance

Table 4.3 summarizes the results of applying the various MLPCA algorithms to the data sets described in the preceding section. For data sets 1-4 and 6, the standard algorithm (Table 4.1) was used. The version which incorporates intercept terms (Section 4.2.4) was used for data set 5, and the routine for correlated errors (Table 4.2) was employed for data set 7. In all cases, the MLPCA model rank was varied from 1 to p+1, where p is the true rank of the data set (*i.e.* the rank in the absence of measurement errors). All of the data sets were also analyzed by PCA, with mean centering used as a pretreatment step for data set 5. The models generated by PCA and MLPCA were used to estimate the error-free measurements ($\hat{\mathbf{X}}$) by orthogonal and maximum likelihood projections of the measurements, respectively. These were used to calculate the objective function, S^2, in each instance. For this purpose, Equation 4.1 was used for data sets 1-6 and Equation 4.23 was used for data set 7. For cases with no intercept terms, S^2 for the model with correct rank should approximate a χ^2 distribution with $(m\text{-}p)(n\text{-}p)$ degrees of freedom. In accordance with this, the last two columns of Table 4.3 give the probability of realizing a value of S^2 below that observed if the model were correct. In other words, values of P below 0.025 or above 0.975 would constitute rejection of the null hypothesis that the model is correct for a two-sided test (α=0.05). For data set 5, the same test was done using $(m\text{-}p\text{-}1)(n\text{-}p\text{-}1)$ degrees of freedom to account for row and column intercepts.

The convergence times given in Table 4.3 are the result of single runs that were carried out with no competing tasks running on the computer. The results are "typical" in

Table 4.3 Results of MLPCA on test data sets.

Data Set Number	Size	Error Type	Offset?	True Rank	Model Rank	Convergence Time (min)	S^2		P^*	
							PCA	MLPCA	PCA	MLPCA
1	10x20	Random	N	2	1	3.00	3.90×10^6	4.53×10^5	1.00	1.00
					2	1.46	3.29×10^4	123.04	1.00	0.10
					3	2.12	5209.0	88.48	1.00	0.02
2	20x20	Random	N	2	1	1.21	7.17×10^6	4.07×10^5	1.00	1.00
					2	1.65	1.51×10^4	311.82	1.00	0.32
					3	2.56	1.86×10^4	267.29	1.00	0.18
3	20x100	Random	N	2	1	700.50	8.71×10^9	7.13×10^6	1.00	1.00
					2	48.46	2.83×10^7	1757.3	1.00	0.46
					3	40.10	2.67×10^8	1548.3	1.00	0.04
4	20x41	Proportional + Constant	N	3	1	0.03	8419.2	7100.9	1.00	1.00
					2	0.03	1240.2	1217.0	1.00	1.00
					3	0.03	724.90	683.43	0.98	0.85
					4	0.24	654.99	571.85	0.96	0.28
	20x41	Proportional + Constant	Y	3	1	13.73	8138.4	6592.5	1.00	1.00
					2	5.35	1094.4	953.02	1.00	1.00
					3	4.21	677.40	622.28	0.99	0.81
					4	12.78	603.87	512.46	0.97	0.20
6	31x1050	?	N	3	1	3.35	1.72×10^8	8.88×10^7	1.00	1.00
					2	8.13	1.63×10^8	1.44×10^7	1.00	1.00
					3	5.45	1.27×10^8	3.38×10^5	1.00	1.00
					4	10.84	5.39×10^7	1.43×10^5	1.00	1.00
7	5x10	Random, Correlated	N	2	1	0.02	2.59×10^6	1.53×10^6	1.00	1.00
					2	0.11	199.92	23.01	1.00	0.48
					3	0.06	129.45	9.97	1.00	0.24

* P = probability that a value less than the corresponding S^2 would be observed for a χ^2 distribution with the appropriate degrees of freedom.

the sense that no attempt was made to adjust the random number seeds to improve performance. Note that there is a separate convergence time listed for each rank analyzed by MLPCA. This is because, unlike conventional PCA, the MLPCA solutions are not generally nested (*i.e.* the rank p model does not contain the rank (p-1) model). Convergence times listed are generally reasonable, with most cases requiring no more than a few minutes and all but one case requiring less than an hour. Although size seems to play some role in convergence time, a much more important factor is the structure of the errors. Experience has indicated that the totally random error structure, such as that in data sets 1-3, is the most difficult case, and this observation is supported by the results in Table 4.3. This case is therefore useful in estimating upper limits for the convergence time. The rank one model for data set 3 (the largest of the "random" error models) proved to be unusually difficult to solve, requiring nearly 12 hours. It is not clear at this point whether the slow convergence is the result of the error structure itself or poor initial estimates that arise from it. However, convergence seems to be considerably faster for other error structures, such as data set 4, where the convergence time is typically a few seconds. This case is likely to be much more typical of experimental data than the random structure. Data set 5, which is the same as data set 4 except for the presence of row and column offsets, required considerably more time because of the gradient optimization of intercept vectors, as described in section 4.2.4. Data set 6 represents a typical experimental data set and demonstrates that MLPCA is a practical alternative to PCA for such cases. Relatively slow convergence in this instance was probably due to poor initial estimates resulting from the inclusion of very noisy measurements in the data set. Finally, data set 7 shows that convergence time is not a problem for correlated errors.

Analysis of the objective function values in Table 4.3 shows that MLPCA always produces a lower S^2 than the corresponding PCA model. This is expected, since PCA does not optimize the same criterion. It is interesting to note, however, that for data sets 2 and 3, the value of S^2 obtained by PCA actually *increases* in going from a rank 2 to a rank 3 model. At first this may seem contradictory, since the general rule is that increasing the rank of a model improves the fit to the data. However, accounting for a greater amount of variance in the data set by increasing the number of factors in PCA does not necessarily decrease the value of S^2. This seems to be especially true for the case of unstructured errors.

With reference to the last two columns of Table 4.3, it is expected that P should drop significantly below unity when a model of the correct rank is found, since that is the point at which S^2 should follow a χ^2 distribution. For MLPCA, this is true for all cases but one. The one exception is the experimental data, data set 6. In this case, it is not surprising that P remains at unity for several reasons: (a) the matrix of standard deviations was approximated using only information from the first sample, and therefore is incorrect for the remaining samples, (b) the variance estimates are the results of replicate scans, and do not account for other sources of variance, such as cell positioning, (c) the variance estimates appear to reflect truncation of the signal in some places, (d) the noise is known to be correlated, and (e) although there are only three known components in the system, there is a real possibility of row and/or column offsets. This example highlights some of the difficulties in using this sensitive statistical measure to estimate the rank in practical cases, but does not diminish the utility of MLPCA for model estimation. (Note that the objective function for the rank three model is more than two orders of magnitude smaller for MLPCA.) In contrast to MLPCA, the PCA models almost always give P values of unity in the table, indicating an incorrect model even when the rank is overestimated. The only exceptions are data sets 4 and 5, where the error structure is closer to uniform, but even in those cases P does not fall below 0.98 for models of the correct rank.

4.4.2 Statistical Validation

Although the cases in the preceding section indicate that MLPCA produces smaller values of the objective function than PCA, it does not guarantee that the procedure converges on the optimum solution, since local optima are always possible. One way to test for a global optimum is to use a different initial estimate and compare the final solutions, but this method is not foolproof. A better method in this case is to exploit the statistical characteristics of S^2 for the correct model. This is done by analyzing replicate data sets, each with the same matrix of error-free data and standard deviations, but with different errors. If the distribution of the S^2 values for these replicates follows a χ^2 distribution with the appropriate degrees of freedom, then it can be concluded that the method is finding the maximum likelihood solution.

A convenient way to make this comparison is to use probability plots. First, the replicate data sets (100 in this case) are analyzed and the S^2 values are stored. The S^2 values are then sorted and assigned a cumulative probability according to their position in the list (the observed probability). For example, the second element in the list would be assigned an observed probability of $2/n$, where n is the number of replicates. Then an expected probability is calculated using the χ^2 distribution. The cumulative probability density function for χ^2 can be calculated using the incomplete gamma function [58]:

$$P(S^2|\nu) = \Gamma_{\text{inc}}\left(\frac{\nu}{2}, \frac{S^2}{2}\right) \tag{4.35}$$

where ν is the degrees of freedom. If the two distributions are the same, a plot of the expected probabilities against observed probabilities should yield a straight line with a slope of unity. If the model is insufficient to account for the systematic variance, either because the form of the model is incorrect or the parameters are suboptimal, then the points of the plot will lie above the ideal line. This means that the distribution of S^2 is shifted right from the χ^2 distribution. If the model accounts for an excessive amount of variance (*i.e.* the estimated rank is too high and measurement variance is modeled), the points will lie below the ideal line.

Figure 4.4 shows probability plots for four of the data sets used in this study: 2, 4, 5, and 7. These data sets were chosen to reflect the different error structures and algorithms used. It is clear from the figure that, in all cases, the results from MLPCA follow the expected distribution, with only minor deviations attributable to the statistical limitations of this study. Therefore, it can be concluded that each of the algorithms is converging on a global optimum (the quality of the optimum is discussed in Section 4.4.5). Furthermore, to varying degrees, the models generated by PCA do not adequately account for the systematic variance in the data sets and are therefore likely to be inferior. Additional comments on these plots are made in the following two sections.

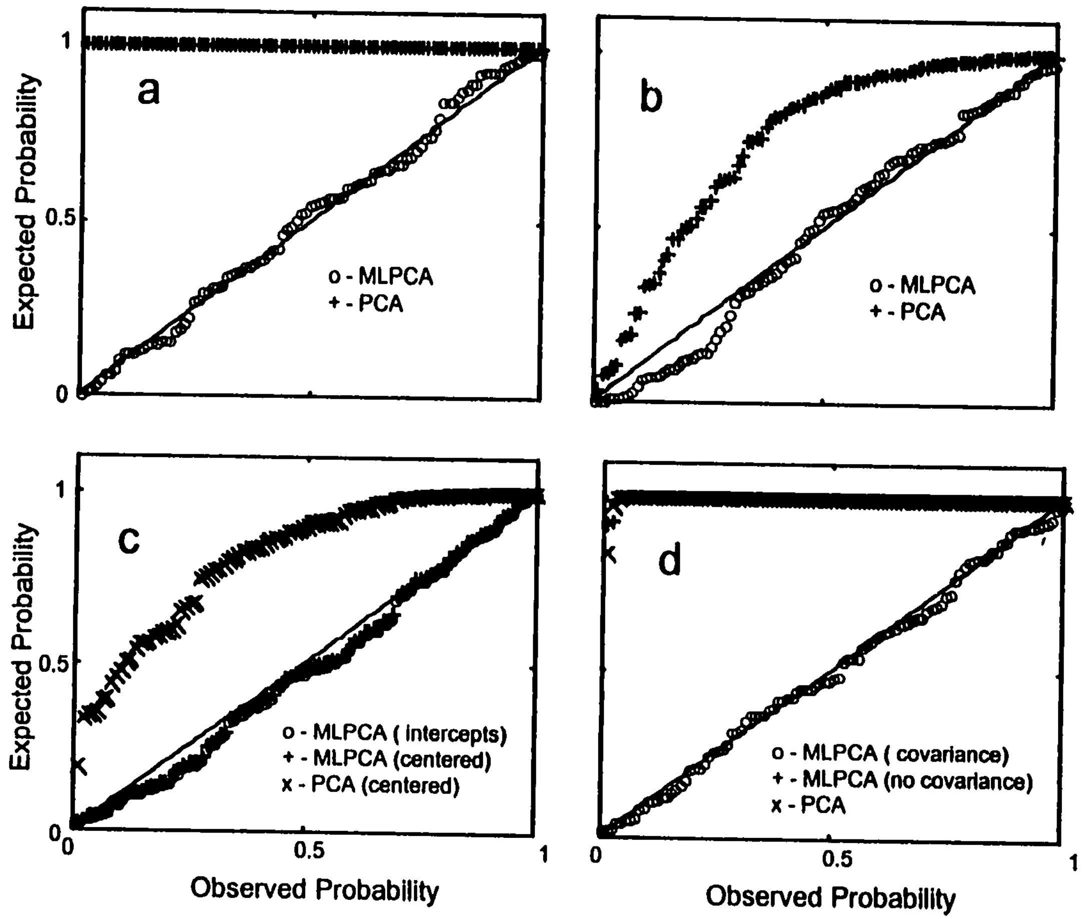


Figure 4.4 Probability plots of S^2 for replicate runs of simulated data sets: (a) data set 2, (b) data set 4, (c) data set 5, (d) data set 7.

4.4.3 MLPCA with Intercepts

The results shown in Figure 4.4c confirm that the gradient optimization method used to determine the intercept vectors for MLPCA performed as expected. However, an additional analysis was done using the standard MLPCA algorithm after the data were column and row mean centered. While this approach always produced S^2 values which were higher than those for which the intercepts were optimized, the difference is barely distinguishable in the figure. This implies that the intercepts determined are very close to the mean values for this particular error structure. Therefore, although the use of mean-centering may mean that the models obtained by the standard MLPCA algorithm are suboptimal, the differences may be negligibly small in many cases and justify the use of the faster algorithm. Larger differences are likely to be observed for cases where the standard deviations become very large or where the data matrix is small. Such cases justify efforts to improve the efficiency of the modified algorithm.

4.4.4 Correlated Errors

Figure 4.4d reveals some interesting characteristics of MLPCA when applied to cases of correlated errors. It is clear from the figure that the version of the MLPCA algorithm that incorporates error covariance provides the optimum model according to the maximum likelihood criterion, and that PCA models are inferior in this regard. An additional analysis was also carried out using the standard MLPCA algorithm by assuming no correlation among the errors and using the diagonal elements of the full covariance matrix (Ω) for the variances. The models generated by this approach were not visibly any better than the PCA models, although the plot does not allow a direct comparison of these two sets of results. Further studies have shown that PCA and the standard MLPCA algorithm produce inferior results for correlated errors even when the standard deviations are the same for all measurements (in this case, these two algorithms are equivalent). This indicates the importance of the error covariance information. Since many chemical measurements and data preprocessing methods give rise to correlated errors, future studies need to be carried oüt to assess the importance of this contribution to model estimation and to improve the numerical reliability of the algorithm incorporating covariance.

4.4.5 Model Quality

Although MLPCA generates models with smaller values of the objective function than does PCA, the key questions have yet to be answered: "Are the MLPCA models closer to the true model and do they offer significant advantages over the PCA models?". The second part of this question cannot be answered outside the context of particular applications, since the advantages gained by MLPCA will undoubtedly depend on the type and magnitude of errors involved, as well as the intended use of the model (regression, mixture analysis, etc.). This aspect will be addressed in Chapter 5. However, the first part of the question is readily answered using simulated data.

One way to compare the MLPCA and PCA models is to project the original vectors used to generate the error free data onto the row space (U) or column space (V) of the model. The angle between the projected vector and the original vector then gives an indication of the agreement between the true model and the fitted model. Mathematically, the angular deviations for the left-hand vectors are given by:

$$\theta_i = cos^{-1}\left(\frac{\mathbf{a}_i^T\mathbf{U}\mathbf{U}^T\mathbf{a}_i}{\left\|\mathbf{a}_i^T\right\|\cdot\left\|\mathbf{U}\mathbf{U}^T\mathbf{a}_i\right\|}\right) \tag{4.36}$$

where θ_i is the angular deviation of the left-hand vector $\mathbf{a}_i$ from the space of the model. A similar expression can be used for the right-hand vectors. In order to be able to draw statistically valid conclusions, 100 replicates were run for data sets 2, 4, and 7. The results are reported in Table 4.4. Note that, since Equation 4.36 always gives positive values, the results in the table are the mean values of the absolute angular deviations. Also given are the standard deviations of the distributions.

Application of the *t*-test to the results in Table 4.4 clearly indicates that the MLPCA algorithm produces models with smaller angular deviations for all three data sets. The extent of improvement varies considerably with the nature of the data and the errors, however. Although these differences are statistically significant, the practical significance of the differences remains as a subject for future research.

To further assess the quality of MLPCA model estimation, the experimental data set, data set 6, was used. Of course, in this case, replicate data sets were unavailable, as were pure component spectra. However, the deviations of the concentration vectors from the model space could be measured. The concentrations for this data set were determined gravimetrically and so were accurately known. Table 4.5 shows that, when the full wavelength range is used, the model obtained by MLPCA is far superior to that obtained by PCA. This is not surprising, since PCA attempts to model the variance due to measurement errors in the high absorbance regions. Normally, in a situation like this, one would preselect wavelengths prior to analysis by PCA. If the wavelength region used is 700-1600 nm, the variance is essentially uniform and PCA and MLPCA produce

Table 4.4 Comparison of vector angle accuracies for PCA and MLPCA. Results are based on 100 replicates and uncertainties are given as standard deviations.

Data Set Number	Rank	Mean Angular Deviations (PCA) (degrees)		Mean Angular Deviations (MLPCA) (degrees)	
		U	V	U	V
2	2	0.41 ± 0.08	0.39 ± 0.06	0.22 ± 0.04	0.21 ± 0.05
		0.42 ± 0.08	0.41 ± 0.06	0.22 ± 0.04	0.23 ± 0.04
4	3	3.1 ± 0.6	3.4 ± 0.6	2.3 ± 0.4	3.0 ± 0.5
		4.7 ± 1.0	3.8 ± 0.7	3.8 ± 0.7	3.1 ± 0.5
		3.0 ± 0.7	3.5 ± 0.6	2.3 ± 0.4	2.9 ± 0.4
7	2	0.10 ± 0.05	0.16 ± 0.05	0.023 ± 0.012	0.10 ± 0.04
		0.11 ± 0.06	0.16 ± 0.06	0.027 ± 0.016	0.12 ± 0.04

Table 4.5 Angular deviations of concentration vectors from PCA space for data set 6.

Wavelength Range (nm)	Component	Angular Deviations (degrees)	
		PCA	MLPCA
400-2500	toluene	13.7	0.127
	chlorobenzene	13.6	0.135
	heptane	25.8	0.688
700-1600	toluene	0.281	0.281
	chlorobenzene	0.359	0.359
	heptane	0.792	0.792

equivalent results. However, it will be noted that the angular deviation of the concentration vectors is smaller when MLPCA is used with the full wavelength region as opposed to the truncated data set, even though the error estimates in this case are only approximate. This illustrates that information can be lost when data are excluded from the modeling process. MLPCA uses the measurement error variance to optimize the amount of information extracted from the data and in that sense represents a significant advance in multivariate analysis.

4.5 CONCLUSIONS

In this chapter, the theoretical foundations of MLPCA have been established using the framework of PCA (and SVD). By incorporating information about the measurement errors, the procedure has been shown to be optimal for principal components modeling in accordance with a maximum likelihood criterion. The algorithm presented here is particularly efficient in its use of alternating regression to achieve rapid convergence. Modifications to the algorithm also permit the incorporation of intercept terms consistent with a maximum likelihood model. Furthermore, generalization of the method allows the incorporation of correlated measurement errors. This represents the first time that a PCA procedure has been developed that has the capability of dealing with measurement error covariance. Results using simulated data show that the objective function minimized by the algorithm approximates a χ^2 distribution with $(m\text{-}p)(n\text{-}p)$ degrees of freedom (in the absence of intercept terms) provided that the measurement error covariance matrix is known and the form of the model is correct. Results using simulated and experimental data also demonstrate that model estimation by MLPCA is superior to models produced by PCA in cases where non-uniform or correlated errors are present.

Practical implications of the theoretical aspects of MLPCA put forward here will be the focus of the next two chapters. This work has clearly demonstrated the positive features of MLPCA and answered many questions related to optimal scaling for PCA models. The advantages of MLPCA over PCA are balanced to some degree by the greater computational efficiency of the latter. As long as measurement errors are approximately uniform or are very small in the context of the intended application, PCA results may be sufficient, if suboptimal. However, there are many cases in chemistry where these conditions do not hold. This study has demonstrated the importance of measurement error data for maximizing the information available from chemical data sets and MLPCA should serve as an important archetype for optimal modeling by PCA. The area of calibration will benefit from the optimal modeling features of MLPCA and will be addressed in the following chapter.

5
MAXIMUM LIKELIHOOD MULTIVARIATE CALIBRATION

5.1 INTRODUCTION

Over the past several decades, advances in chemometrics have led to the development of a multitude of multivariate calibration methods for the analysis of chemical mixtures [59-61]. As a result, such methods are now routinely applied and are indispensable tools for solving many "real-world" problems. At times, the proliferation of multivariate calibration techniques seems unending and includes such methods as multiple linear regression (MLR) (described in Section 1.2), principal components regression (PCR), partial least squares regression (PLS), continuum regression (CR), projection pursuit regression (PPR), locally weighted regression (LWR), and artificial neural networks (ANNs), among others. Each of these methods possesses its own strengths and weaknesses and which works best for a given problem depends on the characteristics of the data and the objectives of the analysis. However, as research produces a clearer distillation of the similarities and differences among methods, a number of techniques, such as PLS and PCR, have established themselves as the practical workhorses of multivariate calibration. PCR is one of the oldest and most well-studied methods currently in use and this chapter describes two fundamental enhancements to the methodology involved which will extend its utility and reliability even further. Although the techniques described in this chapter are general in their application, the focus will be on spectroscopic data sets.

Traditional univariate calibration, which assumes no interference with the measured response variable, typically applies weighted or unweighted least squares regression to a series of standards to develop the calibration model. This was discussed in Chapters 1 and 2. Under the right conditions, the model developed in this manner will be optimal in a maximum likelihood sense. Maximum likelihood parameter estimation methods are widely used because of their desirable statistical characteristics [62]. In the present context, maximum likelihood estimation means that the parameters determined for the model are the ones most likely to give rise to the observed data given the statistical characteristics of the measurement uncertainties. The conditions necessary for ordinary least squares (weighted or unweighted, as appropriate) to provide maximum likelihood parameter estimates are: (1) the form of the model needs to be correct (*e.g.* straight line, quadratic, intercept if necessary), (2) the measurement uncertainties in the response variable need to be uncorrelated and normally distributed, and (in the case of weighted regression) have known variances, and (3) the measurement uncertainties in the concentrations (x-variable) need to be negligibly small relative to the uncertainties in the response variable (y). In practice, these ideal conditions are rarely met exactly, but maximum likelihood methods are often still regarded as the best alternative if the conditions are approximately valid.

In contrast to traditional univariate calibration, techniques such as PCR are known as *inverse* calibration methods because the concentrations are regressed on the responses (factor scores in PCR) rather than the other way around. Accordingly, PCR can only qualify as a maximum likelihood method if the uncertainties in the responses (scores) are negligible compared to those in the concentrations. While this is often true when the reference method for concentrations is relatively imprecise, there are many cases where the assumption is somewhat tenuous. Furthermore, PCR ignores the uncertainty in the spectroscopic data when it performs the initial decomposition by principal component analysis (PCA). As pointed out in Chapter 2, PCA yields a maximum likelihood decomposition only when the measurement uncertainties are independent and identically distributed with a normal distribution ("*iid* normal"). It has long been known that spectroscopic measurements inherently possess non-uniform measurement standard deviations which can vary as a function of both signal amplitude and wavelength. Furthermore, instrument characteristics often lead to correlated noise characteristics.

Although the noise characteristics for most common spectroscopic methods have been well-studied by Ingle and Crouch [63], this information is generally ignored in establishing multivariate calibration models. It should be apparent that, since each spectral measurement can possess a different uncertainty, each can also carry a different amount of information into the calibration procedure. (A summary of important noise sources for some common spectroscopic techniques is presented in Table 5.1.) In PCR, for example, principal component analysis (PCA) is first used to determine the subspace of the component spectra of a mixture. The spectrum of each calibration sample is then projected into this subspace to give a set of scores, or latent variables. These scores are used in the regression procedure to produce the PCR calibration model. This projection has the effect of combining the spectral measurements to reduce the overall error and also makes the regression step more mathematically tractable. Obviously, the quality of results obtained by PCR will depend on the quality of the estimation of the spectral subspace by PCA. Unfortunately, as mentioned earlier, PCA tries to maximize the

Table 5.1 Summary of limiting noise sources in selected spectroscopic techniques.

Type of Spectroscopy	Concentration Range	Important Noise Sources
Atomic Absorption (flame)	Low	Signal shot noise, Source flicker, Flame transmission
	Moderate	Absorption flicker
	High	0% T noise
Atomic Emission (flame, plasma)	Low	Background emission flicker and shot noise
	Moderate/High	Analyte emission shot and flicker noise
Molecular Absorption (visible)	Low	Signal flicker (source, cell transmission)
	Moderate	Signal flicker, Signal shot noise
	High	Signal shot noise, 0% T noise
Molecular Absorption (infrared)	All	Detector noise
Molecular Fluorescence	Low	Background emission flicker and shot noise
	Moderate/High	Analyte emission shot and flicker noise

variance accounted for by the extracted latent variables, regardless of whether the variance is due to chemical effects (*i.e.* changes in chemical concentrations) or simply measurement uncertainty. Because of this, including measurements with a large uncertainty can degrade the quality of the calibration model developed by PCR. While this problem has been addressed informally through approaches such as scaling and wavelength selection, these pretreatments are generally suboptimal in a maximum likelihood sense.

In this chapter, two new methods are described to account for measurement uncertainty in multivariate calibration. These new methods, which will be referred to as maximum likelihood principal components regression (MLPCR) and maximum likelihood latent root regression (MLLRR), are actually more general forms of PCR and latent root regression (LRR) and will produce solutions identical to these methods under the right conditions. However, the new techniques, based on maximum likelihood principal component analysis (MLPCA), are better-suited to providing optimal solutions in the maximum likelihood sense when there are non-uniform uncertainties in the data. It will be shown using both simulated and experimental data that MLPCR and MLLRR can provide significantly better predictive ability than conventional methods in realistic situations. Perhaps more importantly, the maximum likelihood methods provide a general unifying framework from which multivariate calibration methods can be examined.

Throughout this chapter, a number of assumptions and simplifications have been made that should be clarified from the outset. First, it has been assumed that measurement errors are normally distributed. While the principles of maximum likelihood estimation are general in nature, mathematical tractability in the development of MLPCA demanded that this restriction be imposed. Although this assumption may not be strictly valid in all cases, it is viewed as reasonable and, unless the violation is severe, should not greatly diminish the general utility of the methods, just as simple regression is often used without strict adherence of the underlying assumptions. A second assumption made by MLPCA for maximum likelihood estimation (and by weighted regression, for that matter) is that measurement error variances are exactly known. In practice, however, this is rarely the case, so true maximum likelihood estimates are technically unattainable for real experimental data. Nevertheless, it will be shown that variance estimates are sufficient to obtain significant improvement in results; *i.e.* that some knowledge of measurement uncertainty is often better than an implicit assumption of uniform variance. Finally, throughout this chapter, it has been assumed that measurement errors are uncorrelated; *i.e.* the error covariance matrix is diagonal. While such a condition can be controlled in simulations, it is almost certainly invalid for experimental measurements. It is demonstrated, however, that significant improvement in predictive ability can be achieved even when the assumption of uncorrelated errors is tenuous. There are two main reasons for excluding error covariance in this work. First, while estimates of measurement error variance are often available, knowledge of the covariance matrix in practice is still quite rare, so we wished to demonstrate the utility of these methods when the covariance matrix is unavailable. Second, although the theory of MLPCA is capable of dealing with correlated errors, there are several practical problems that need to be addressed. These include rank deficiency in the estimated error covariance matrix and the computational burden of large matrices. These are subjects of ongoing research and will not be treated in this thesis.

5.2 CALIBRATION METHODS

5.2.1 Principal Components Regression

For the purposes of this discussion, it will be assumed that we are trying to develop a calibration model for a single analyte in the presence of multiple unknown interferences, and that the measurements consist of spectroscopic data (although other analytical techniques could also be employed). Conventional PCR begins with a set of calibration samples for which the concentration of the analyte has been obtained by some independent means. The first step in the procedure is to apply PCA to the spectra of the calibration samples. This is usually done by way of singular value decomposition (SVD) to give:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{5.1}$$

Here **X** is the matrix of spectra in the calibration set (*m* samples by *n* wavelengths). The component concentrations in the calibration set should reflect the distribution of those concentrations for future samples (*i.e.* the calibration set should span the space of samples to be predicted), and the number of samples and wavelength channels should be greater than the number of independently observable components in the mixtures. Assuming that $m<n$, the SVD gives the matrices **U** (*m*x*m*), **S** (*m*x*m*) and **V** (*n*x*m*). These matrices are truncated by removing the right-hand columns and bottom rows to give $\tilde{\mathbf{U}}$ (*m*x*p*), $\tilde{\mathbf{S}}$

(pxp) and $\tilde{\mathbf{V}}$ (nxp), where p is the "pseudorank" of **X**, or the number of independently observable components. In practice, p is usually unknown, but can be estimated by statistical means or cross-validation. In this work, the tilde ("~") will be used to distinguish the truncated matrices and the quantities derived from them. The truncation gives $\tilde{\mathbf{X}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^{\mathrm{T}} = \tilde{\mathbf{T}}\tilde{\mathbf{V}}^{\mathrm{T}}$, where $\tilde{\mathbf{X}}$ is the estimated data matrix and $\tilde{\mathbf{T}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}$ is called the scores matrix for the truncated solution. Alternatively, in a model and parameters framework, we have,

$$\mathbf{X} = \tilde{\mathbf{T}}\tilde{\mathbf{V}}^{\mathrm{T}} + \mathbf{E} \tag{5.2}$$

where **E** is the mxn matrix of residuals. PCA obtains $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{V}}$ by minimizing the sum of the squares of the elements in **E**. This estimation is optimal in a maximum likelihood sense as long as p represents the true pseudorank and the measurement errors for the elements of **X** are *iid* normal.

This reduction in the dimensionality of the problem is the key to PCR, since it improves the reliability of the solution. The actual regression is carried out using orthogonal projections of the spectra onto the subspace determined by PCA, *i.e.* the scores. The regression assumes a model of the form:

$$\mathbf{y} = \tilde{\mathbf{T}}\mathbf{q} + \mathbf{f} \tag{5.3}$$

where **y** is the mx1 vector of analyte concentrations for the analyte set, **q** is a px1 regression vector, and **f** is an mx1 vector of errors. The least squares solution to this problem is:

$$\hat{\mathbf{q}} = \left(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}}\right)^{-1}\tilde{\mathbf{T}}^{\mathrm{T}}\mathbf{y} = \tilde{\mathbf{S}}^{-1}\tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \tag{5.4}$$

In this equation and elsewhere, the caret ("^") is used to indicate an estimated quantity. In the prediction step, the scores for the unknown spectrum are given by:

$$\tilde{\mathbf{t}}_{\mathrm{unk}} = \mathbf{x}_{\mathrm{unk}}\tilde{\mathbf{V}} \tag{5.5}$$

where $\tilde{\mathbf{t}}_{\mathrm{unk}}$ and $\mathbf{x}_{\mathrm{unk}}$ are row vectors of length p and n, respectively. The unknown concentration is then estimated by:

$$\hat{y}_{\mathrm{unk}} = \tilde{\mathbf{t}}_{\mathrm{unk}}\hat{\mathbf{q}} \tag{5.6}$$

More typically, the intermediate step of calculating the scores is incorporated into an nx1 regression vector, $\hat{\mathbf{b}}$, that is multiplied directly by the spectrum to obtain the concentration estimate:

$$\hat{y}_{\mathrm{unk}} = \mathbf{x}_{\mathrm{unk}}\hat{\mathbf{b}} \tag{5.7}$$

$$\hat{\mathbf{b}} = \tilde{\mathbf{V}}\hat{\mathbf{q}} = \tilde{\mathbf{V}}\tilde{\mathbf{S}}^{-1}\tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \tag{5.8}$$

In conventional PCR, the representations in Equations 5.6 and 5.7, are equivalent, but this is not the case for MLPCR, as discussed in the next section.

5.2.2 Maximum Likelihood PCR

When applied in the proper context, conventional PCR is a powerful tool for the quantitative analysis of multicomponent mixtures. However, it suffers from a number of weaknesses. One of these is that it relies on SVD to obtain a reliable estimation of the p-dimensional subspace that contains the component spectra. In essence, the eigenvectors produced by SVD (the columns of $\tilde{\mathbf{V}}$) describe a p-dimensional hyperplane in the n-dimensional wavelength space and should contain all of the pure spectral vectors. As long as the measurement errors in all of the calibration spectra are all *iid* normal, the *p*-dimensional hyperplane determined by SVD will be the optimal model for the data in a maximum likelihood sense (assuming the system is linear and the pseudorank, *p*, has been correctly specified). However, if the measurement errors are not independent with uniform variance, this will no longer be true and the estimation of the subspace will be suboptimal.

There are several potential solutions to this problem. First, it may be possible to scale the data in such a way that all of the measurement standard deviations become equal. It has been shown, however, that in order for this to work in a manner which preserves the structure of the data, the matrix of measurement standard deviations must have a rank of unity [32] (*e.g.* when the uncertainty at each wavelength is independent of signal amplitude). This restriction is frequently violated for experimental data sets, making it impossible to obtain an optimum solution through simple scaling. A second approach to the problem is to perform wavelength selection, removing those channels that significantly violate the assumption of *iid* errors. This assumes, however, that noise is a function only of wavelength and not signal amplitude. Furthermore, although a portion of the spectrum may appear noisy, it may also be the region which is richest in information about the analyte of interest. Wavelength selection has also been performed by using leave-one-out cross-validation. In addition to being very time consuming, this approach only mitigates the problem of finding the optimal subspace and does not address the source of the problem.

What is needed is a modeling method which accounts for spectral measurement errors in the estimation of the spectral subspace. Maximum likelihood principal component analysis, described in Chapters 3 and 4, is such a method and forms the basis of MLPCR. In MLPCA, the eigenvectors are chosen to provide the optimal estimation of the *p*-dimensional hyperplane in a maximum likelihood sense. The optimality of the estimation is, strictly speaking, contingent on the assumption of normally distributed measurement errors with known variances and covariances, but relaxation of these conditions (*i.e.* near normality and/or`estimated variances) still yields significantly improved estimates of the PCA subspace. For uncorrelated measurement errors, MLPCA minimizes a weighted sum of squared residuals:

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(x_{ij} - \hat{x}_{ij}\right)^2}{\sigma_{ij}^2} \qquad (5.9)$$

In this equation, x_{ij} is a measurement (an element of **X**), $\hat{x}_{ij}$ is the maximum likelihood estimate of that measurement, and σ_{ij} is the corresponding measurement error standard deviation. (In practice σ_{ij} is typically replaced by its estimate, s_{ij}.) The MLPCA decomposition can be represented as:

$$\mathbf{X} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{T}}\hat{\mathbf{V}}^{\mathrm{T}} + \mathbf{E} = \hat{\mathbf{X}} + \mathbf{E} \qquad (5.10)$$

where **X**, $\hat{\mathbf{X}}$ and **E** are *mxn*, $\hat{\mathbf{U}}$ and $\hat{\mathbf{T}}$ are *mxp*, $\hat{\mathbf{S}}$ is *pxp*, and $\hat{\mathbf{V}}$ is *nxp* for a *p*-dimensional model. The caret "^" above the matrices **U**, **S**, **V**, and **E** has been used here to distinguish the MLPCA solution from the truncated PCA solution. It must be reiterated that, unlike PCA, MLPCA solutions are not nested; that is, the rank *p* model cannot be obtained simply by truncating higher rank models. Instead, the dimensionality of the model needs to be specified before initiating the decomposition. Although this tends to make MLPCA more cumbersome to use, the superior results often make it worthwhile. Another difference is that in conventional PCA, the estimate for any 1x*n* spectral vector, $\mathbf{x}_i$, is given by an orthogonal projection into the spectral subspace:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i \tilde{\mathbf{V}} \tilde{\mathbf{V}}^T \tag{5.11}$$

In contrast, the maximum likelihood estimate of $\mathbf{x}_i$ is given by a projection which is not generally orthogonal, but rather one which is weighted by the errors in the measurements:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i \Sigma_i^{-1} \hat{\mathbf{V}} \left(\hat{\mathbf{V}}^T \Sigma_i^{-1} \hat{\mathbf{V}} \right)^{-1} \hat{\mathbf{V}}^T \tag{5.12}$$

Here, Σ_i is the *n*x*n* covariance matrix for $\mathbf{x}_i$ (note that any multiple of Σ_i could also be used). For uncorrelated errors, this will be a diagonal matrix whose diagonal elements are the variances for the corresponding spectral measurements. It is clear that Equation 5.12 will result in an orthogonal projection when all of the measurement errors are uncorrelated and have equal variances; *i.e.* the PCA projection is equivalent to a maximum likelihood projection under these conditions. For measurement errors which are correlated only within a spectrum (*i.e.* row correlations but no column correlations), Equation 5.12 is still valid, but the function minimized by MLPCA is modified to:

$$S^2 = \sum_{i=1}^{m} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \Sigma_i^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \tag{5.13}$$

which reduces to Equation 5.9 for uncorrelated errors. If measurement errors are correlated among both rows and columns, a somewhat modified version of MLPCA is needed. This is described in Section 4.2.3 and will not be treated here except to note that MLPCA can handle any measurement error covariance structure provided that the error covariance matrix can be estimated. As noted in Section 5.1, uncorrelated measurement errors have been assumed throughout this work. Although this assumption is not generally valid for experimental data, the error covariance structure is rarely known in practical situations, so it is intended to reflect a realistic implementation of the methods described.

The regression model in MLPCR is developed in a manner analogous to that in PCR, following Equations 5.2 and 5.4:

$$\hat{\mathbf{q}} = \left(\hat{\mathbf{T}}^{\mathrm{T}}\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}^{\mathrm{T}}\mathbf{y} = \hat{\mathbf{S}}^{-1}\hat{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \tag{5.14}$$

However, unlike conventional PCR, a maximum likelihood projection is used to determine the scores for the unknown sample, which are then used to estimate the concentration:

$$\hat{\mathbf{t}}_{\mathrm{unk}} = \mathbf{x}_{\mathrm{unk}}\boldsymbol{\Sigma}_{\mathrm{unk}}^{-1}\hat{\mathbf{V}}\left(\hat{\mathbf{V}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{unk}}^{-1}\hat{\mathbf{V}}\right)^{-1} \tag{5.15}$$

$$\hat{y}_{\mathrm{unk}} = \hat{\mathbf{t}}_{\mathrm{unk}}\hat{\mathbf{q}} \tag{5.16}$$

Note that in MLPCR there is no longer an analog to a universal regression vector, $\hat{\mathbf{b}}$, for all unknown samples, as defined in Equations 5.7 and 5.8 for PCR. This is because the projection matrix depends on the measurement error covariance matrix, which can be different for each unknown sample. This, however, is one of the main advantages of MLPCR, since the projection of the unknown sample onto the spectral subspace will exploit those measurements that have the smallest errors in order to obtain the best estimate of the scores.

To summarize, MLPCR improves the quality of the regression over PCR in two ways. First, it uses MLPCA in conjunction with measurement error information to obtain a more reliable estimate of the subspace containing the pure spectral vectors. Because measurements in the calibration set are appropriately weighted, a maximum likelihood estimate of the PCA model is obtained which is generally superior to that obtained by SVD. This is important because it is the determination of this initial space that ultimately affects the sensitivity of the calibration procedure. The second advantage of MLPCR derives from the projection of the measurements (calibration and unknown) onto the subspace determined by MLPCA. Because the projection is not orthogonal but rather optimized through the use of measurement uncertainties, the maximum information is extracted for the best estimation of the true measurements. These factors tend to lead to a superior calibration model.

5.2.3 Maximum Likelihood Latent Root Regression

Although MLPCR can offer a significant improvement over conventional PCR, it is still not a "pure" maximum likelihood approach to calibration because of the final regression step. For this step to be optimal from a maximum likelihood perspective, the absolute uncertainties in the scores need to be much smaller than those in the concentration values. Since this will not always be true, it would be useful to develop a method which could accommodate an arbitrary error in the final regression step. This can be done by incorporating a variation of latent root regression (LRR).

Unlike PCR, LRR [64-66] is not well known among chemists. With this technique, the original calibration matrix of response variables is augmented by the corresponding concentration vector(s). PCA is then carried out on the augmented matrix. In this way, the reduction of dimensionality and the determination of the calibration model are performed simultaneously. Using the previous example for the estimation of the concentration of a single component, we have:

$$\left[\tilde{\mathbf{X}}\,|\,\tilde{\mathbf{y}}\right]=\tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^{\mathrm{T}} \qquad (5.17)$$

As before, **X** is a matrix of m spectra measured at n wavelengths, **y** is an mx1 vector of concentrations for the component of interest, and the tilde indicates that the SVD results are truncated to pseudorank p. If $\tilde{\mathbf{V}}$ (which now has dimensions (n+1)xp) is partitioned into the upper $\tilde{\mathbf{V}}_1$ (nxp) and lower $\tilde{\mathbf{V}}_2$ (1xp), then the regression vector will be given by:

$$\hat{\mathbf{b}}=\tilde{\mathbf{V}}_1\left(\tilde{\mathbf{V}}_1^{\mathrm{T}}\tilde{\mathbf{V}}_1\right)^{-1}\tilde{\mathbf{V}}_2^{\mathrm{T}} \qquad (5.18)$$

such that the predicted concentration is $\hat{y}_{\mathrm{unk}}=\mathbf{x}_{\mathrm{unk}}\hat{\mathbf{b}}$, where $\mathbf{x}_{\mathrm{unk}}$ is 1xn.

LRR is similar to PCR in its approach to calibration, but for some reason it has been virtually ignored in chemistry. It is possible that it is simply more cumbersome and less intuitive than PCR, and in many cases it does not offer significant advantages. Another difference between the two methods is that the predictive ability of PCR is unaffected by changes in the scale of the y variable. This is because the regression step in PCR implicitly assumes (for the maximum likelihood solution) that all of the error resides in y, so a vertical projection is always used. In contrast, LRR is consistent with a maximum likelihood solution if the absolute uncertainties in all of the measured quantities (x and y) are the same (*iid* normal), leading to results that will change with the scale of y. The situation is exactly analogous to the differences between ordinary least squares and PCA when used for modeling purposes, as discussed in Chapter 2. In reality, neither set of assumptions is likely to be valid. It would therefore be useful to have a single step modeling procedure like LRR which accounts for all of the uncertainties. Such a method is presented here as maximum likelihood latent root regression (MLLRR).

The procedure for MLLRR is similar to that for MLPCR, except that, as in LRR, an augmented matrix is used. In a manner analogous to Equation 5.17, the augmented matrix is decomposed using MLPCA rather than PCA. This requires a companion matrix of measurement variances, also augmented to include the variances in the concentration values. In the absence of other information, measurement uncertainties are usually assumed to be uncorrelated, but correlated errors can be accommodated by MLPCA as well. Once MLPCA has been carried out, prediction is performed using an augmented spectral vector:

$$\left[\hat{\mathbf{x}}_{\text{unk}} \mid \hat{y}_{\text{unk}}\right]=\left[\mathbf{x}_{\text{unk}} \mid 0\right]\Sigma_{\text{unk}}^{-1}\hat{\mathbf{V}}\left(\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_{\text{unk}}^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^{\mathrm{T}} \qquad (5.19)$$

In this case, Σ_{unk} is the error covariance matrix of the augmented row vector for the unknown, and $\hat{\mathbf{V}}$ is the loadings matrix obtained from applying MLPCA to the augmented calibration matrix. The equation is written so that it produces a row vector, since this is the manner in which the spectra appear in the original calibration matrix. Note that Equation 5.19 is simply a maximum likelihood projection of the unknown spectrum into the MLPCA subspace. The key is that, since y_{unk} is the quantity sought, the last entry in the error covariance matrix, Σ_{unk}, is set to be numerically equivalent to infinity, forcing this value to be predicted from the others. Thus the last entry in the first row vector on the right hand side is unimportant and is set to zero in the equation. Extension of Equation 5.19 to additional components is easily accomplished by further augmentation of the calibration and prediction matrices. As with MLPCR, there is no universal regression vector for MLLRR unless the covariance matrices are identical for all of the spectra obtained.

MLLRR is more general in its treatment of measurement errors than MLPCR in that it includes uncertainties in the concentration values. It is an optimal modeling method in the maximum likelihood sense, subject to the usual restrictions (linear model of known pseudorank, normally distributed errors with a known covariance structure).

5.3 EXPERIMENTAL

To examine the two new methods proposed here, three simulated and two experimental data sets were employed. To distinguish the new data sets from those described in preceding and following chapters, they will be designated as data sets 8 through 11 (data set 6, described in Chapter 4, was the remaining data set). The simulated data sets were used to test the methods under carefully controlled conditions to evaluate their potential. Each of these was generated from a model of a three-component mixture. The pure component spectra consisted of Gaussian profiles centered at 480, 500 and 520 nm (for components 1,2 and 3, respectively) with widths (σ) of 20 nm. Spectral data points were generated at 5 nm intervals between 400 and 600 nm. Calibration and prediction data sets consisted of 20 and 100 samples, respectively, whose component concentrations were generated randomly from a uniform distribution between 0 and 1. In all simulations, normally distributed measurement errors were added using a Gaussian random number generator.

Data set 8 was characterized by wavelength dependent noise which was essentially uniform near the center of the spectral range but amplified on either side. To accomplish this, a baseline noise level of σ_0 was first selected. This standard deviation was then multiplied by a wavelength dependent function to give the standard deviation for a particular wavelength. The function used in this work was a "double sigmoidal" mask, with a value close to unity near the center of the spectral region and values of r_{max} at the limits. The profile of this mask is shown along with the individual spectral profiles in Figure 5.1. Using this mask, the standard deviation at wavelength λ is given by:

$$\sigma(\lambda) = \left[1 + \left(r_{max} - 1\right)\left(\frac{1}{1+e^{a(\lambda-\lambda_1)}} + \frac{1}{1+e^{a(\lambda_2-\lambda)}}\right)\right]\sigma_0 \qquad (5.20)$$

In this equation, λ_1 is the inflection point of the sigmoid on the left hand side of the range and λ_2 is that on the right hand side. The parameter a determines the slope of the sigmoidal curves such that:

$$a = \frac{2\,ln9}{\Delta\lambda} = \frac{4.394}{\Delta\lambda} \qquad (5.21)$$

where $\Delta\lambda$ is the 10% to 90% rise range of the sigmoid. In this work, λ_1 = 460 nm, λ_2 = 540 nm, and $\Delta\lambda$ = 40 nm. The standard deviation of the baseline noise level, σ_0, was taken to be 1% of the maximum absorbance in the noise free calibration data. The noise amplification factor, r_{max}, was varied between 1 and 20 for this work. The concentration data used for calibration with data set 8 were assumed to be error free.

Data set 9 was the same as data set 8 except for the noise structure. In this case, both proportional and constant error were added to the signals to give measurement uncertainties that depended on signal amplitude. The formula used to calculate the standard deviation for a given absorbance, A_{ij}, was:

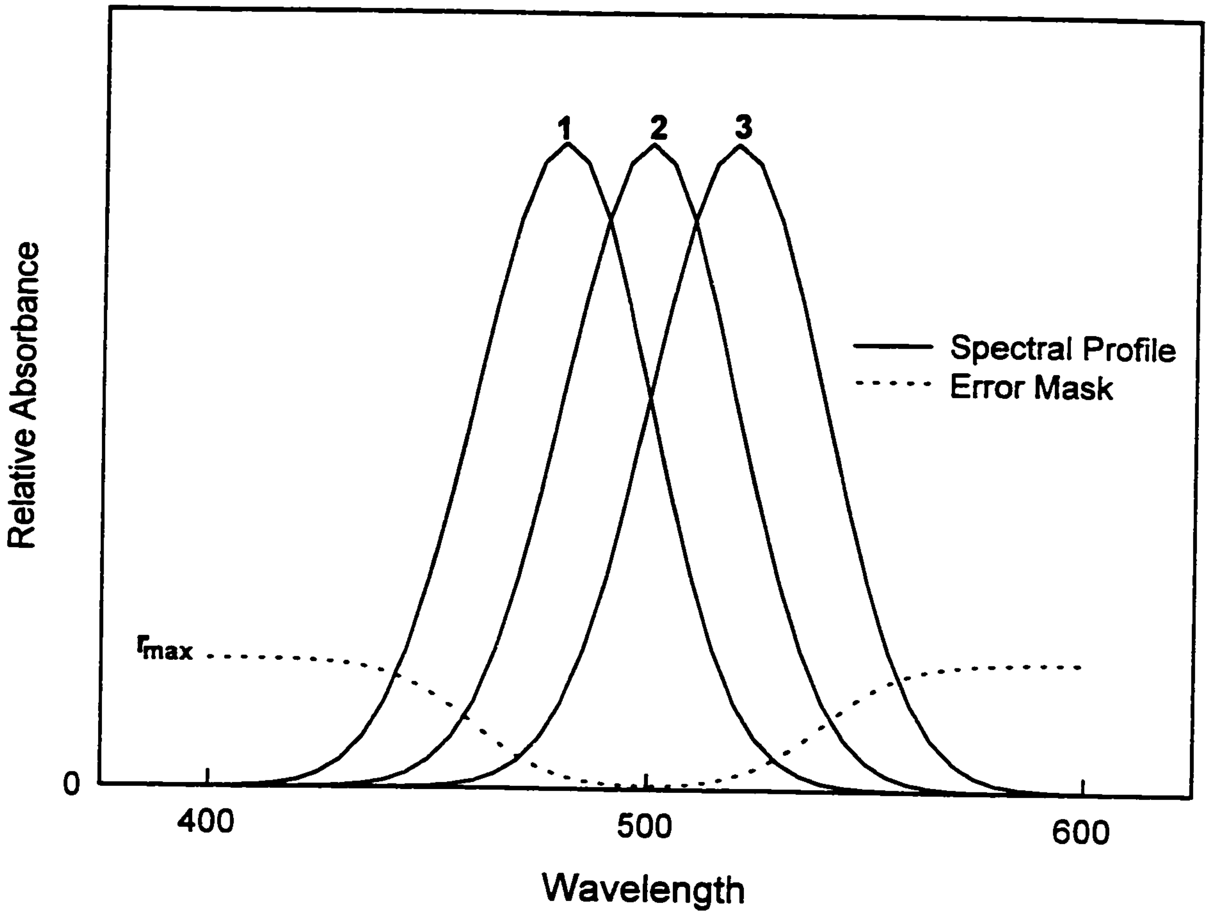


Figure 5.1 Spectral profiles for simulated three component mixtures (data sets 8, 9 and 10) and error mask for data set 8.

$$\sigma_{ij} = \sqrt{\sigma_0^2 + \left(pA_{ij}\right)^2} \tag{5.22}$$

where σ_0 is the level of the constant noise component (in this case, 1% of the maximum signal in the calibration matrix) and p is the level of proportional noise (varied between 0 and 0.20 in this work).

Data set 10, which was intended to exaggerate the differences between MLLRR and MLPCR, included errors in the calibration concentrations in addition to those in the spectral measurements. As for data set 9, the errors in the spectral measurements included both a constant term (1% of the maximum in the calibration set) and a proportional term (in this case fixed at a level of 2% of the pure signal). To simulate non-uniform errors in the calibration concentrations, proportional error was added to the reference concentrations. The proportional error had standard deviations that ranged from 0 to 20% of the true concentration.

To demonstrate the utility of the maximum likelihood calibration methods for practical applications, two experimental data sets were also examined. Data set 11, the first of these, was obtained through a carefully designed experiment involving three-component mixtures of metal ions (Co(II), Cr(III), Ni(II)), a system suggested from the work of Osten and Kowalski [67]. Stock solutions of the nitrates were prepared with concentrations of 0.172, 0.0764, and 0.393 M for Co, Cr and Ni, respectively, in 4% HNO_3. All chemicals used throughout this work were analytical reagent grade or better unless otherwise specified. A three-level, three-factor calibration design was used in which 1, 3, or 5 mL aliquots of the various stock solutions were combined and diluted to 25 mL with 4% HNO_3. Unfortunately, insufficient Ni stock remained for one solution

(3:5:5 Co:Cr:Ni), so the calibration set consisted of 26 rather than 27 solutions. Final concentration ranges were 6.88 to 34.40 mM for Co, 3.06 to 15.29 mM for Cr, and 15.70 to 78.58 mM for Ni. Five replicate spectra were obtained for each sample using randomized blocks (*i.e.* 5 blocks of all 26 solutions, randomly ordered within each block). To minimize the effects of instrument drift, a reference spectrum was run prior to each new sample. Spectra were recorded over the range of 350-650 nm on an HP 8452 diode array spectrophotometer (Hewlett-Packard, Palo Alto, CA) using a standard 1 cm quartz cuvette. Measurements were made at 2 nm intervals with a 1 s integration time. In order to introduce non-uniform noise characteristics, a dichroic bandpass filter (green, no. 67) was placed between the source and the sample to decrease the source intensity at high and low wavelengths for all measurements. The spectra of the individual components and the optical filter are shown in Figure 5.2.

The second experimental data set employed in this work was data set 6 which consisted of near-infrared spectra for three-component mixtures containing toluene, chlorobenzene, and heptane. A further description of this data set can be found in Section 4.3. The purpose of this data set was to demonstrate that MLPCR can utilize all of the available data to obtain superior predictive ability by extracting the optimum amount of information at each wavelength, provided measurement variance information is available. Unfortunately, standard deviation information for this data set was only available from replicate scans of one sample. Not only will this fail to be a precise representation of the standard deviations from the remaining 30 samples, but it also does not reflect all of the sources of measurement error (*e.g.* cell positioning, sample preparation) for the sample

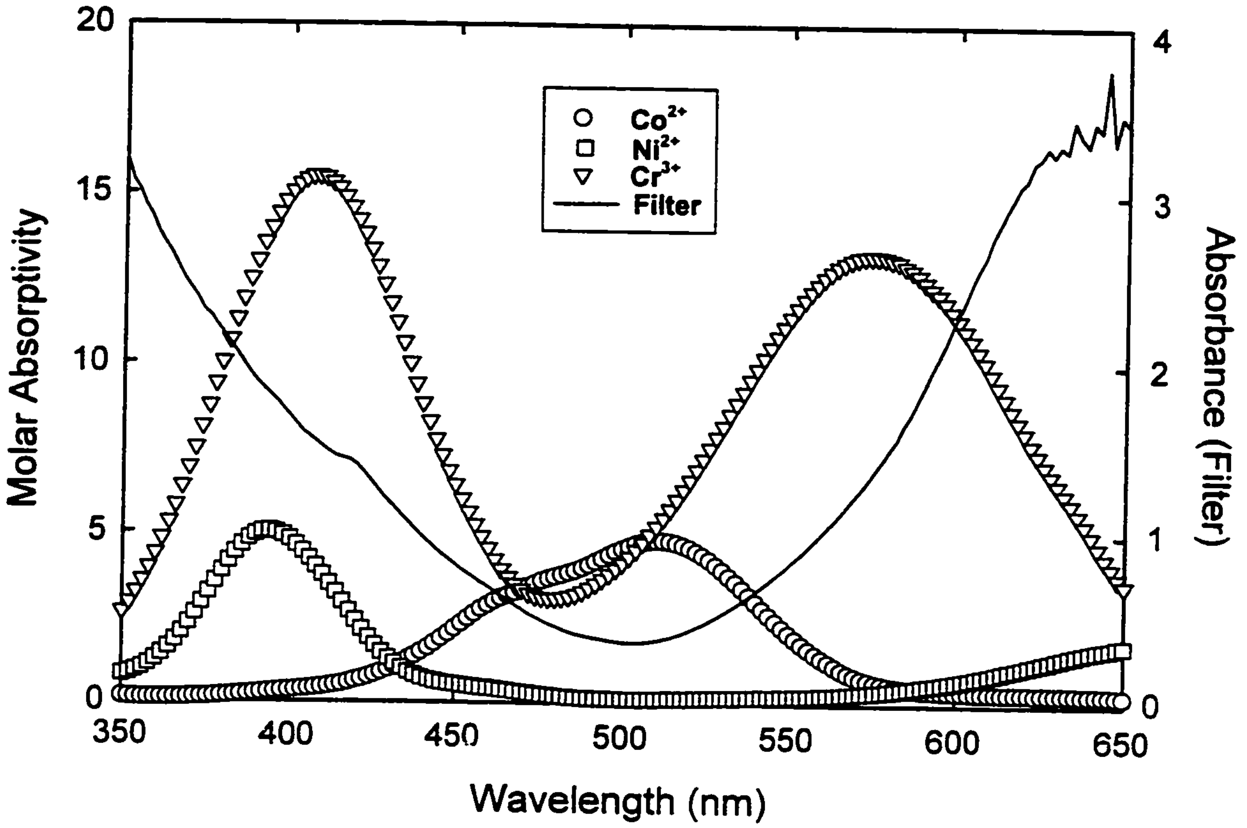


Figure 5.2 Pure component spectra for data set 11 and absorbance profile of bandpass filter applied to source beam for noise amplification.

for which it does apply. Nevertheless, it will be shown that MLPCR can utilize even this approximate information to provide better performance than conventional approaches.

The calculations performed in this work utilized a variety of computational platforms including: (1) 486 and Pentium-based personal computers, (2) a Digital Equipment Corporation 3000/300X workstation with a 175 MHz clock speed and 96 MB of memory, and (3) a Sun Microsystems Sparc Server 1000 with 230 MB of memory and four 50 MHz SuperSPARC CPUs. All calculations were performed in Matlab (The MathWorks, Natick, MA).

5.4 RESULTS

5.4.1 Simulated Data

Initially, simulated spectroscopic data sets were used to assess the new calibration methods. Data set 8 was used to examine the effects of measurement errors whose standard deviation varies as a function of wavelength, but is constant at any given wavelength. Such situations commonly arise when source intensity or detector sensitivity changes with wavelength, or when there is a strongly absorbing (constant) background component. Even when they are not coincident with the regions of the spectrum containing relevant information, noisy measurements can still influence the analysis, since the variance still needs to be accounted for by PCA. Although wavelength selection can often reduce this problem, the task of selection becomes difficult when noisy regions overlap regions of spectral significance, since the selection then relies on choosing the correct balance between the signal and noise retained in a given measurement. Maximum likelihood methods simplify the analysis by extracting the appropriate amount of information from each measurement.

Figure 5.3 shows the results of a comparison among PCR, PLS, MLPCR, and MLLRR for data set 8. Results are presented in terms of a root-mean-square error of prediction for components 1 and 2 in the three component mixture (by symmetry, the results for component 3 will be statistically equivalent to those for component 1). The RMSEP is calculated as:

$$RMSEP = \sqrt{\sum_{i=1}^{N_{pred}} \left(y_i^{pred} - y_i^{true}\right)^2 \Big/ N_{pred}} \qquad (5.23)$$

where y_i^{pred} and y_i^{true} are the predicted and actual concentrations of the analyte in prediction sample i, respectively, and N_{pred} is the number of prediction samples (100 in this case). In carrying out this calculation, the optimum number of latent variables was taken to be three for PCR, MLPCR and MLLRR, since this should be the pseudorank of the calibration matrix by the constraints of the simulation. To permit greater flexibility for PLS, the optimum number of latent variables was selected by cross-validation (below an amplification factor of 8, the optimum number of latent variables was 3; from 10 to 12 it was 2; and above 12 only 1 was needed). For MLPCR and MLLRR, the standard deviation values known from the simulation were used for the measurement error estimates. In actual practice, these standard deviations would likely be determined from experimental replicates and would therefore be known with less accuracy, but the true values were used here to avoid introducing the number of replicates as a variable and also to examine a best-case scenario. For comparison, however, the MLPCR and MLLRR simulations were also run using variances estimated from five replicates. The results

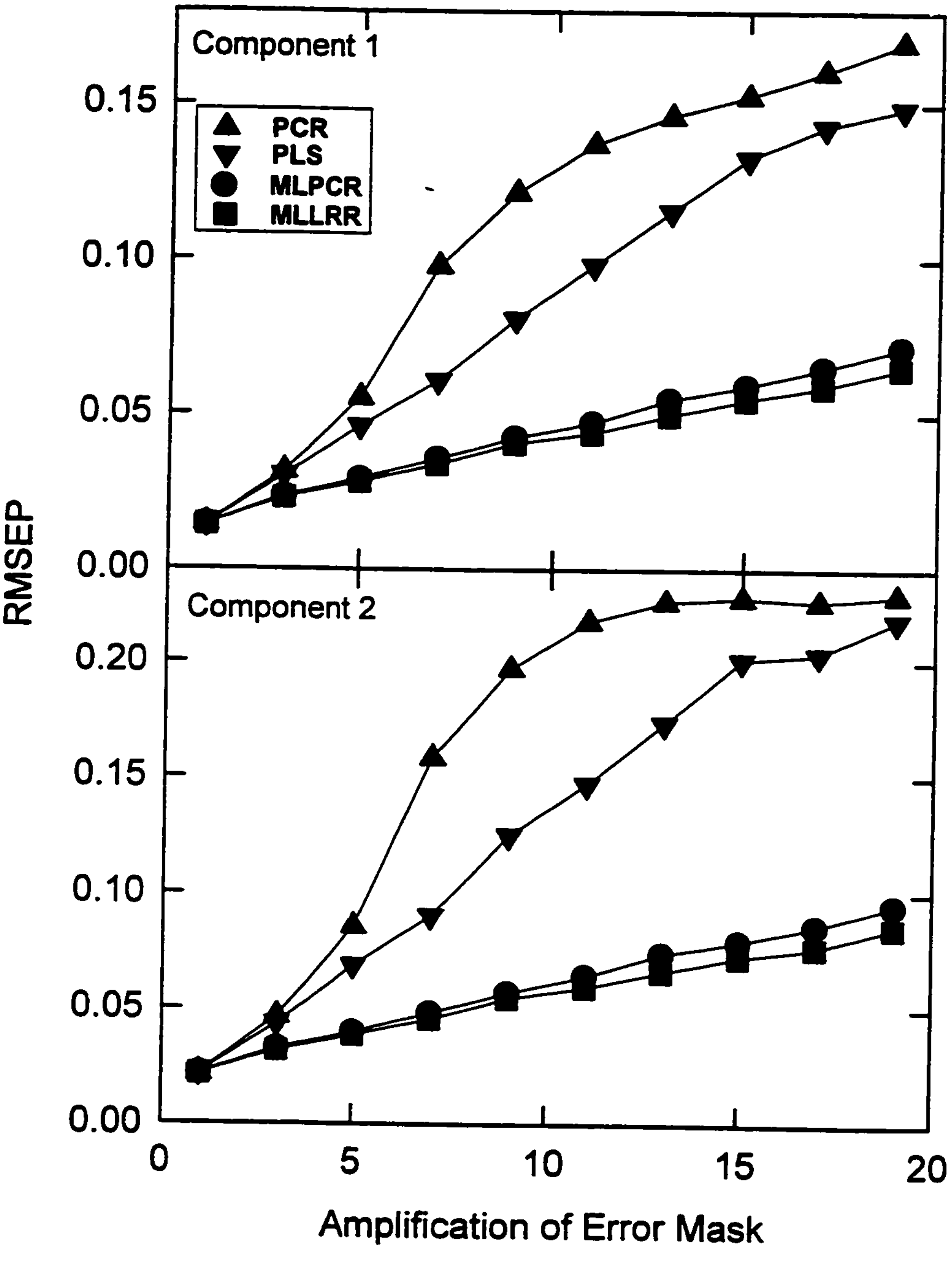


Figure 5.3 Comparison of calibration methods applied to simulated data subjected to a non-uniform error mask (data set 8).

under these conditions were virtually identical, with all of the prediction errors falling within 3% of the values obtained when known variances were used. Thus, at least in this case, the use of estimated variances did not have a large impact. Experimental data presented later illustrates the case where standard deviations are estimated.

From Figure 5.3, it is apparent that when the noise amplification factor, $r_{\max}$, is unity (uniform noise) all of the methods perform equally well. This is expected since the maximum likelihood methods reduce to PCR under these conditions and there is unlikely to be any advantage of PLS over PCR. As the noise amplification factor is increased so that the variance on either side of the spectral range is amplified, the performance of all methods declines (prediction error increases). This is also expected, since increasing the noise obscures the information content of the data and increases the uncertainty. As the noise level is increased, the prediction error for the maximum likelihood methods remains significantly smaller than either PCR or PLS, illustrating the advantages of these techniques. It should be pointed out that even at the limits of this study, the amplified noise on the wings of the spectrum represents only about 20% of the maximum signal in the calibration set, and this is not an unrealistic level. Nevertheless, the prediction errors obtained by the maximum likelihood methods are a factor of 2 to 3 smaller at this point than the conventional methods. Comparison of the conventional methods indicates that PLS performs somewhat better than PCR in this case. This is due in part to the selection of an optimum number of latent variables for PLS, but a more important factor is likely to be the fact that PLS places some significance on correlation with the y variable in extracting latent variables, and so is not entirely based on x-variance.

It will also be noted that MLLRR consistently performs better than MLPCR in this example, although the difference is not substantial. The difference arises from the regression step in MLPCR, which assumes that the errors in y are uniform and much greater than the errors in the scores (for maximum likelihood estimation). In this example, however, the y values were generated with no errors, so the situation is the exact opposite of the second assumption and`MLLRR produces superior results. In a real calibration problem, it is likely that y will be determined by a reference method which has a significant uncertainty, so the assumptions of MLPCR may be more valid. It has been observed throughout this work that MLLRR generally yields results superior to those produced by MLPCR (because the most appropriate weighting of x and y is used), but that the two methods rarely give large differences.

A final point worth noting here is that the magnitudes of the errors are comparable for the two components in this example. In general, one might expect significantly larger prediction errors for component 2, since it is overlapped by two interferences (as opposed to one for components 1 and 3) and therefore should give a smaller net analyte signal [13]. However, the lower sensitivity of the method for component 2 is offset by the lower noise level near the center of the spectral range, so the prediction errors turn out to be comparable. The results presented here represent a limited study and an infinite number of variations (spectral resolution, noise profiles, etc.) are of course possible. However, for all of the cases of this type that were examined, the maximum likelihood methods gave lower prediction errors than the conventional methods.

Although data set 8 clearly shows the advantages of MLPCR and MLLRR, comparable results for this data set could have been obtained simply by scaling each wavelength channel by its corresponding standard deviation prior to performing PCR. In other words, because the standard deviation matrix is rank 1, optimal scaling is possible. For this data set, scaling would be the preferred approach for reasons of computational speed, but optimal scaling is not possible in cases where the noise depends on signal magnitude. For this reason, data set 9, which contains both proportional and constant components of error, was employed for further comparison. A combination of errors was used to make the simulation more realistic, since purely proportional errors are rarely encountered.

Figure 5.4 provides a comparison of the prediction errors for the same four calibration methods applied to data set 9, as well as for two additional methods described in the following paragraph. Results are shown as a function of the level of proportional noise added to the data. Again, component 3 is omitted because of statistical equivalence, and, again, all methods are equivalent in the presence of uniform noise (0% proportional error). For PLS, the optimum number of latent variables was 3 up to 4% proportional noise and 4 thereafter. As before, the maximum likelihood methods show a significant improvement over the conventional methods, with the same order of performance as for data set 8. The improvement for component 1 is more striking than that for component 2 in this case, possibly because the central region of the spectrum remains more uniform in magnitude and therefore more uniform in noise. As with data set 8, the use of estimated variances with MLPCR and MLLRR gave only small differences in results (<7%).

To demonstrate that simple scaling is not sufficient to provide an improvement equivalent to the maximum likelihood methods for cases where the noise depends on

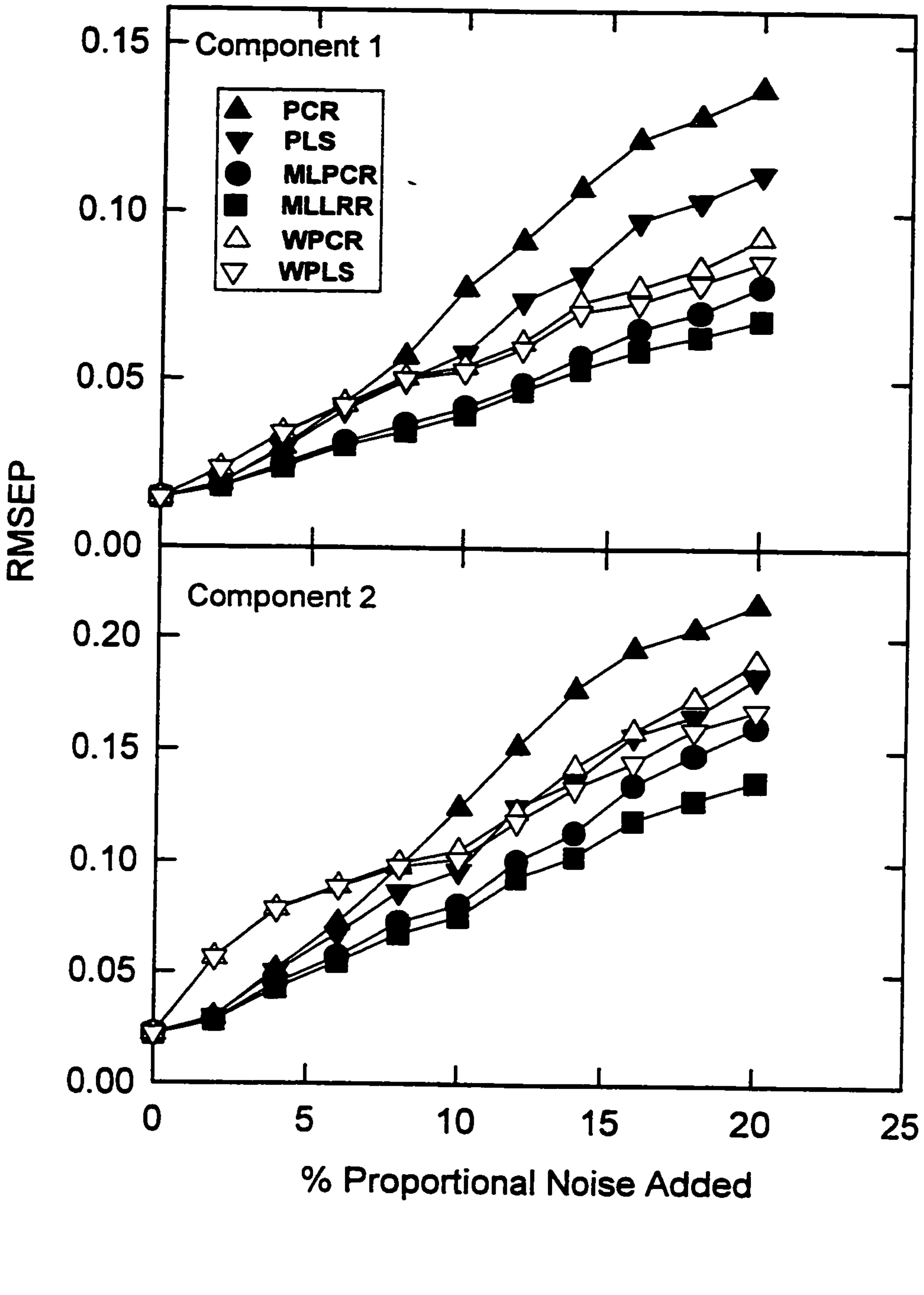


Figure 5.4 Comparison of calibration methods applied to simulated data with constant+proportional errors in spectra (data set 9).

signal amplitude, data set 9 was also examined using "weighted" PCR and PLS, designated as WPCR and WPLS in Figure 5.4. The data sets were scaled by the inverse of a pooled standard deviation at each wavelength. As the figure shows, this often results in smaller prediction errors compared to PCR and PLS (with notable exceptions for small proportional errors), but the extent of improvement is less than what is achieved with the maximum likelihood methods. Although such suboptimal scaling may provide satisfactory results in some cases, MLPCR and MLLRR are preferable because of their optimal performance in the general case, regardless of the error structure.

In the first two data sets, comparable performance was observed for MLPCR and MLLRR. It was speculated that differences would be exaggerated if significant non-uniform errors were added to the concentrations in the calibration set. For this reason, proportional errors were added to the reference concentrations in data set 10. This data set is also more realistic in the sense that multivariate calibration methods often use a reference method to determine concentrations in the calibration mixtures and such measurements are prone to uncertainty. Figure 5.5 shows the prediction errors for components 1 and 2 using PCR, MLPCR and MLLRR as the level of proportional error in the reference concentrations is increased from 0 to 20%. The plot shows the actual prediction errors, *i.e.* the errors from the true concentrations in the prediction set rather than concentrations with errors added. As anticipated, the differences between MLLRR and MLPCR become more pronounced as the errors in the calibration concentrations increase, with MLLRR always providing superior results. In this example, there are only marginal differences between MLPCR and PCR since the level of proportional noise is small enough to make the spectral measurements close to uniform error.

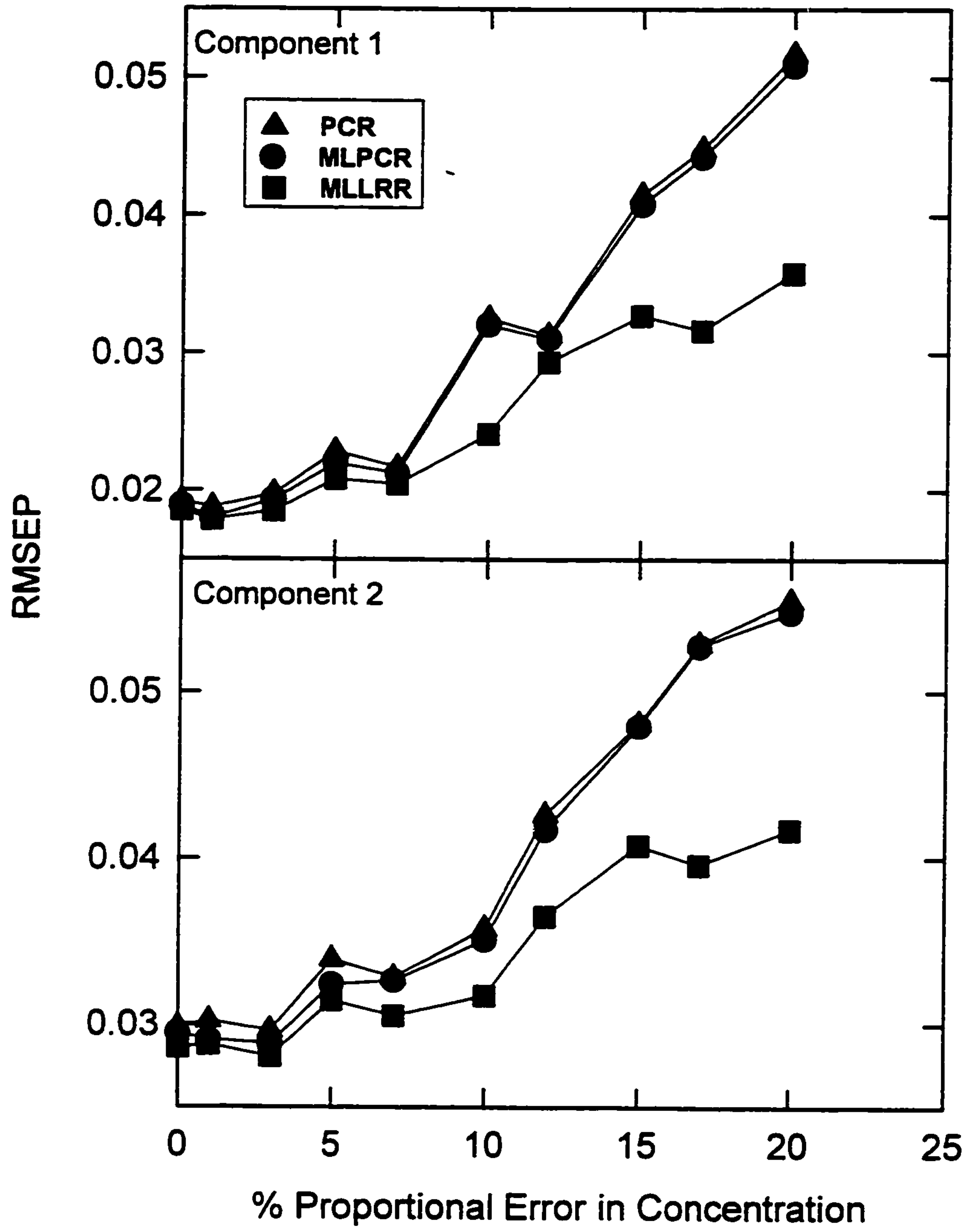


Figure 5.5 Comparison of calibration methods applied to simulated data with errors in both spectra and concentrations (data set 10).

At this point, a comment should be made regarding the augmented error covariance matrix used for MLLRR. Throughout this chapter, it has been assumed that the errors in the calibration concentrations are uncorrelated with the errors in the spectral measurements. Strictly speaking, in a "designed" experiment, this may not be true. A designed experiment is one in which the mixtures are prepared by adding known amounts of the analytes to the calibration mixture. As such, there is no reference measurement used other than the gravimetric or volumetric data in the preparation. An error in these measurements can be considered to be correlated with the true error in the spectral measurements since it will affect these measurements proportionately. However, in practical circumstances, instrumental measurement errors are often much greater than the preparation errors and the correlation can be considered insignificant. More importantly, multivariate calibration procedures more often employ "natural" calibration, where the concentrations in the calibration set are determined by a reference method which should be uncorrelated to errors in the spectral measurements.

5.4.2 Experimental Data

The first of the experimental data sets examined, data set 11, consisted of mixtures of Co, Cr and Ni ions in dilute nitric acid. Spectra for these mixtures are shown in Figure 5.6a, with the corresponding measurement standard deviations for each set of replicates in Figure 5.6b. Increased noise levels are apparent at either end of the spectral range as the result of the optical filter placed in the light path. Two groups of samples were also found to have inordinately high standard deviations near the center of the spectrum, a problem that was traced to two of the samples out of the 130 which appeared to have an offset. The questionable samples were excluded from subsequent analysis, although it was found that their inclusion did not greatly affect the results. In the analysis of this data set, a diagonal error covariance matrix consisting of the variances for each measurement was used (*i.e.* uncorrelated errors were assumed). Although this assumption is known to be invalid, it was made to demonstrate the enhanced performance of maximum likelihood methods even when the error covariance information is unavailable.

To examine the predictive ability of various calibration methods, the technique of leave-one-out cross-validation was employed (alternative methods could be used, but this approach is the most widely used in chemometrics). In this approach, the calibration model is first constructed for a particular analyte using all but one sample. The concentration of the analyte in the excluded sample is then predicted using the model, and the deviation from the expected concentration is measured. This process is repeated so

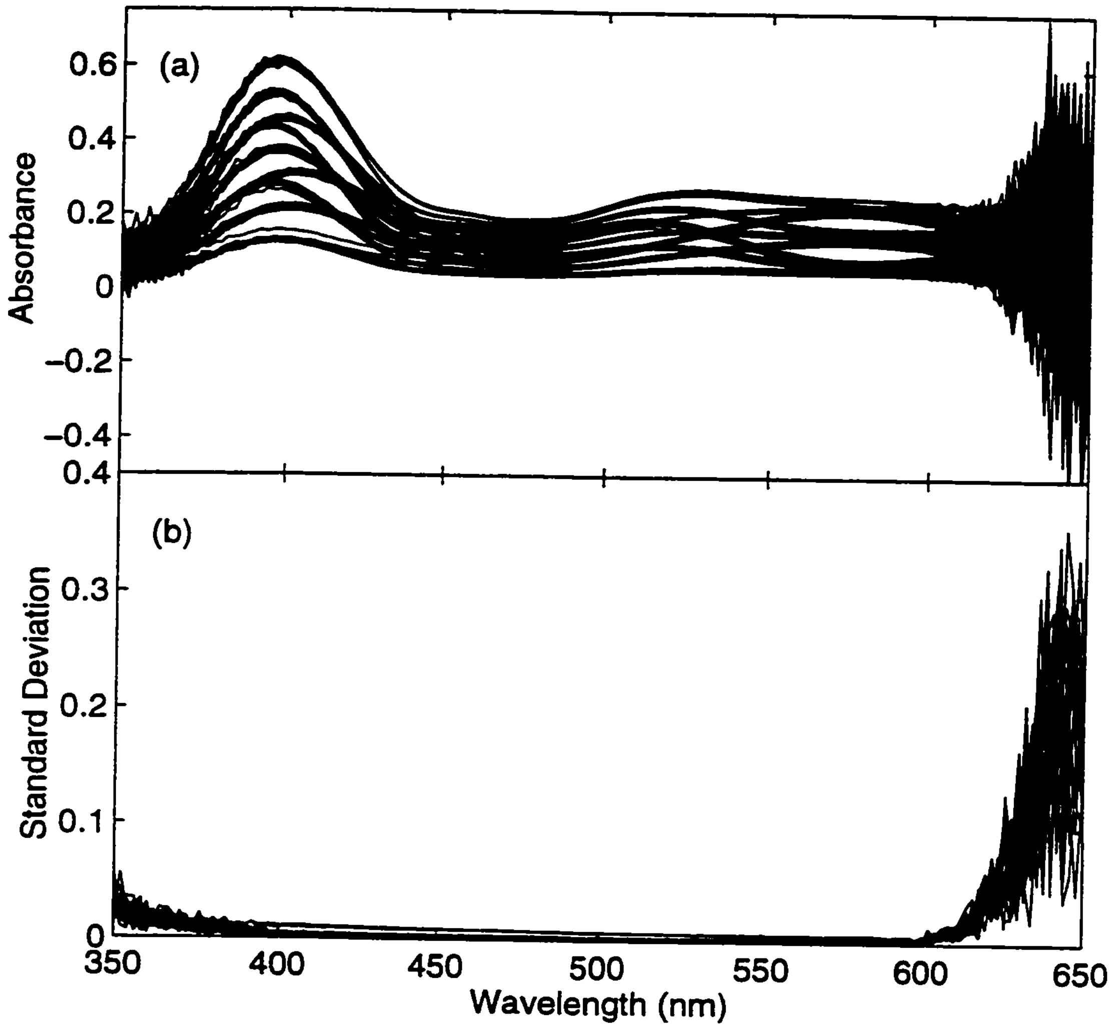


Figure 5.6 (a) Spectra for metal ion mixtures (data set 11) and (b) corresponding standard deviations.

that each of the 128 calibration samples is excluded once and a root-mean-square error of cross-validation (RMSECV) is calculated by:

$$RMSECV = \sqrt{\sum_{i=1}^{N_{cal}} \left(y_i^{pred} - y_i^{ref}\right)^2 \Big/ N_{cal}} \tag{5.24}$$

where y_i^{pred} and y_i^{ref} are the predicted and reference concentrations, respectively, of the analyte in the excluded sample, and N_{cal} is the number of calibration samples. The RMSECV was calculated for each of the three analytes in the mixtures. An overall, or total RMSECV, was also calculated from:

$$RMSECV_{tot} = \sqrt{\left(RMSECV_{Co}^2 + RMSECV_{Cr}^2 + RMSECV_{Ni}^2\right)/3} \tag{5.25}$$

The RMSECV values calculated in this way give an indication of the predictive ability of the model. However, it should be pointed out that, for the PCR methods, two different approaches can be used for cross-validation. In what will be referred to as "leave-one-sample-out" cross-validation, PCA or MLPCA (as appropriate) is carried out on the subset of 127 calibration samples and the results are used for calibration. In "leave-one-score-out" cross-validation, all 128 samples are used for PCA or MLPCA, and these results are retained for all subsequent calibrations, leaving the appropriate score out when building the calibration models by regression. In other words, the basis set is developed using all 128 samples, which are then projected onto the basis to obtain the scores. The regression is carried out on the scores for each combination of 127 samples, leaving one set of sample scores out in each case for cross-validation. This approach is faster, since PCA or (especially) MLPCA is only performed once (or once for each model dimensionality in the case of MLPCA). Although leave-one-score-out cross-validation can be considered a legitimate approach in that it doesn't employ concentration information about the prediction sample in the calibration procedure, purists may argue that it is not as valid as the leave-one-sample-out approach, which is completely blind to the prediction sample. For this reason, both approaches are included in the results presented here. As expected, the differences are very small and the time savings of the leave-one-score-out method is a factor of N_{cal}, an important consideration with MLPCR, which is substantially slower than PCR.

The results for data set 11 are presented in Table 5.2, which shows the RMSECV for each method and analyte as a function of the number of latent variables. The appropriate number of latent variables for this data set should be three, but since experimental realities such as offsets and nonlinearities can affect the optimum number of latent variables, results are given for up to six factors. The results are presented in tabular rather than graphical format because the range of values and number of methods would obscure a conclusive graphical interpretation. In addition to PCR, PLS, MLPCR, and MLLRR, results are also given for weighted PCR and PLS, using pooled standard deviations at each wavelength as weighting factors. For all of the methods examined, the predictive ability is poor when one or two latent variables is used, as expected. The maximum likelihood methods generally reach a performance plateau around three latent variables, where the prediction errors level off, although there is some marginal improvement with additional factors. For PCR and PLS, the plateau is less distinct, with additional factors continuing to bring further improvement. However, even with the addition of more latent variables than are shown in the table, the cross-validation errors for PCR and PLS did not reach the level of those for the maximum likelihood methods

Table 5.2 Comparison of calibration methods for data set 11 (mixtures of Co, Cr and Ni). Values given are the root-mean-squared errors of cross-validation (RMSECV) in mM.

Calibration Method	Species	Number of Latent Variables					
		1	2	3	4	5	6
MLPCR	Co	10.68	6.34	0.32	0.32	0.19	0.17
	Cr	3.07	3.11	0.11	0.11	0.07	0.07
	Ni	24.35	17.98	0.38	0.37	0.33	0.33
	Total	15.45	11.15	0.29	0.29	0.23	0.22
MLPCR*	Co	10.67	6.30	0.32	0.32	0.17	0.16
	Cr	3.07	3.09	0.11	0.11	0.07	0.07
	Ni	24.37	17.93	0.38	0.37	0.33	0.32
	Total	15.46	11.12	0.29	0.29	0.22	0.21
MLLRR	Co	10.89	7.08	0.33	0.17	0.16	0.16
	Cr	3.47	3.43	0.11	0.07	0.07	0.07
	Ni	24.48	16.00	0.38	0.35	0.35	0.36
	Total	15.60	10.30	0.30	0.23	0.23	0.23
PCR	Co	11.53	8.47	8.39	8.94	6.07	2.62
	Cr	3.51	3.11	3.15	3.29	2.44	0.85
	Ni	20.69	11.51	11.73	12.42	8.03	3.44
	Total	13.82	8.44	8.52	9.04	5.98	2.54
PCR*	Co	11.53	8.42	8.27	8.33	5.80	2.46
	Cr	3.50	3.11	3.11	3.09	2.33	0.79
	Ni	20.69	11.48	11.57	11.53	7.63	3.18
	Total	13.82	8.41	8.40	8.41	5.69	2.37
PLS	Co	11.58	9.43	1.72	1.49	0.63	0.60
	Cr	3.55	2.87	0.79	0.58	0.46	0.42
	Ni	20.41	8.83	2.35	1.09	0.97	0.70
	Total	13.70	7.64	1.74	1.12	0.72	0.58
WPCR	Co	10.14	5.40	0.32	0.29	0.20	0.16
	Cr	3.08	3.02	0.11	0.10	0.08	0.07
	Ni	25.36	19.76	0.42	0.36	0.34	0.32
	Total	15.87	11.95	0.31	0.27	0.23	0.21
WPLS	Co	10.14	5.43	0.32	0.29	0.23	0.16
	Cr	3.08	3.03	0.11	0.10	0.08	0.07
	Ni	25.37	19.85	0.42	0.37	0.36	0.32
	Total	15.87	12.01	0.31	0.28	0.25	0.21

* Asterisk indicates leave-one-score-out cross-validation as opposed to leave-one-sample-out cross-validation.

(the minimum total error for both methods was 0.38 mM, attained at 16 latent variables for PCR and 10 for PLS). The differences between the maximum likelihood methods and the conventional techniques is dramatic. Compared to PCR, the cross-validation errors for the maximum likelihood methods are more than an order of magnitude smaller in most cases. PLS fares somewhat better, but the RMSECV values are still substantially higher. Among the maximum likelihood methods, the results for MLPCR and MLLRR are very similar in most cases for this application. It will be also be noted that there is little difference between the leave-one-score-out and leave-one-sample-out cross-validation methods, as expected. In this example, the weighted regression methods (WPCR and WPLS) perform almost identically to the maximum likelihood methods, but this is expected, since the variances are primarily dependent on the wavelength channel in this absorbance range. As demonstrated for data set 9, differences from the weighted methods are more obvious with errors that depend on signal magnitude. For this application, we would expect to see a greater effect at higher absorbance values. In any case, the utility of the maximum likelihood methods is that they guarantee an optimal estimation of the PCA subspace, which is not always assured with scaling.

The remarkably poor performance of PCR in this example motivated further examination of the reasons underlying the differences observed. In conducting this investigation, it was decided to focus on a comparison of PCR and MLPCR, since these two methods are the most complementary. Two of the most important factors influencing the performance of an analytical method are the sensitivity of the technique and the noise in the measurements. It is anticipated that, because of the nature of the geometric projections used, the uncertainty in the scores will be smaller for MLPCR than PCR, but since such differences can be difficult to quantify in the general case, it was decided to focus on the sensitivity aspect. For first-order calibration methods, the sensitivity is related to the net analyte signal (NAS) by:

$$SEN = \|\mathbf{NAS}\| \tag{5.26}$$

where $\|\bullet\|$ indicates the Euclidean norm, or length, of the NAS vector [13,68]. The NAS for a given analyte is that part of the pure analyte spectrum that is orthogonal to the spectra of all other constituents in the mixture. The pure component spectra for a *p*-component mixture can be represented as vectors in an *n*-dimensional absorbance space (n = number of wavelength channels) and will define a *p*-dimensional subspace (hyperplane) within that space. If the vector representing the spectrum of analyte *i* (the analyte of interest) is now excluded, it is possible to identify a vector that is orthogonal to the remaining vectors and lies in the subspace defined by all *p* spectra. This vector is called the *contravariant* vector [65], and it is the projection of the analyte spectral vector onto the unit vector in this direction that defines its NAS. Mathematically, if the pure component spectra of all constituents are known, the NAS is defined as:

$$\mathbf{NAS}_{\mathrm{i}} = \left(\mathbf{I} - \mathbf{R}_{\mathrm{i}}\left(\mathbf{R}_{\mathrm{i}}^{\mathrm{T}}\mathbf{R}_{\mathrm{i}}\right)^{-1}\mathbf{R}_{\mathrm{i}}^{\mathrm{T}}\right)\mathbf{r}_{\mathrm{i}} \tag{5.27}$$

where $\mathbf{R}_{\mathrm{i}}$ is an nx(p-1) matrix whose columns consist of the pure component spectra for all constituents except the analyte, $\mathbf{r}_{\mathrm{i}}$ is an nx1 vector containing the analyte spectrum (normalized to unit concentration), **I** is the *nxn* identity matrix, and $\mathrm{NAS}_{\mathrm{i}}$ is the net analyte signal vector for analyte *i*.

An obvious problem with Equation 5.27 is that the spectra of all constituents must be known. For situations where there are unknown constituents, methods such as PCR are used to estimate the NAS by regression against concentration. To make the problem mathematically tractable, PCA is used as the first step in PCR to identify the subspace of the pure component spectra. Calibration spectra are projected into this space and regressed against concentration. The NAS determined in the subspace, which will be designated as NAS*, can then be transformed back to the original space. The important equations are:

$$\mathbf{q}_i = \tilde{\mathbf{V}}^T \mathbf{r}_i \tag{5.28}$$

$$\mathbf{Q}_i = \tilde{\mathbf{V}}^T \mathbf{R}_i \tag{5.29}$$

$$\mathbf{NAS}_i^* = \left(\mathbf{I} - \mathbf{Q}_i\left(\mathbf{Q}_i^T\mathbf{Q}_i\right)^{-1}\mathbf{Q}_i^T\right)\mathbf{q}_i \tag{5.30}$$

$$\mathbf{NAS}_i^{PCR} = \tilde{\mathbf{V}} \cdot \mathbf{NAS}_i^* \tag{5.31}$$

In these equations, $\mathbf{q}_i$ (px1) and $\mathbf{Q}_i$ (px(p-1)) are analogous to $\mathbf{r}_i$ and $\mathbf{R}_i$ in Equation 5.27 and represent "abstract spectra" in the principal components space. **NAS*** represents the px1 net analyte signal vector in the subspace, and **NAS**^{PCR} is the same vector in the original absorbance space. Note that **NAS**^{PCR} is distinguished from the "true" **NAS** in Equation 5.27 since they will only be identical in the ideal case. If, for example, PCA does not correctly determine the subspace of the component spectra, projection of individual spectra will result in a shorter vector and reduced sensitivity.

In the present study, pure component spectra are available for the three components in the mixture, and therefore it is possible to obtain the NAS directly as well as by PCR and MLPCR. Figure 5.7 shows the results of this calculation for cobalt using three latent variables. Similar results were obtained for chromium and nickel, which are not shown. Note that the NAS obtained from direct calculation and **NAS**^{MLPCR} are very

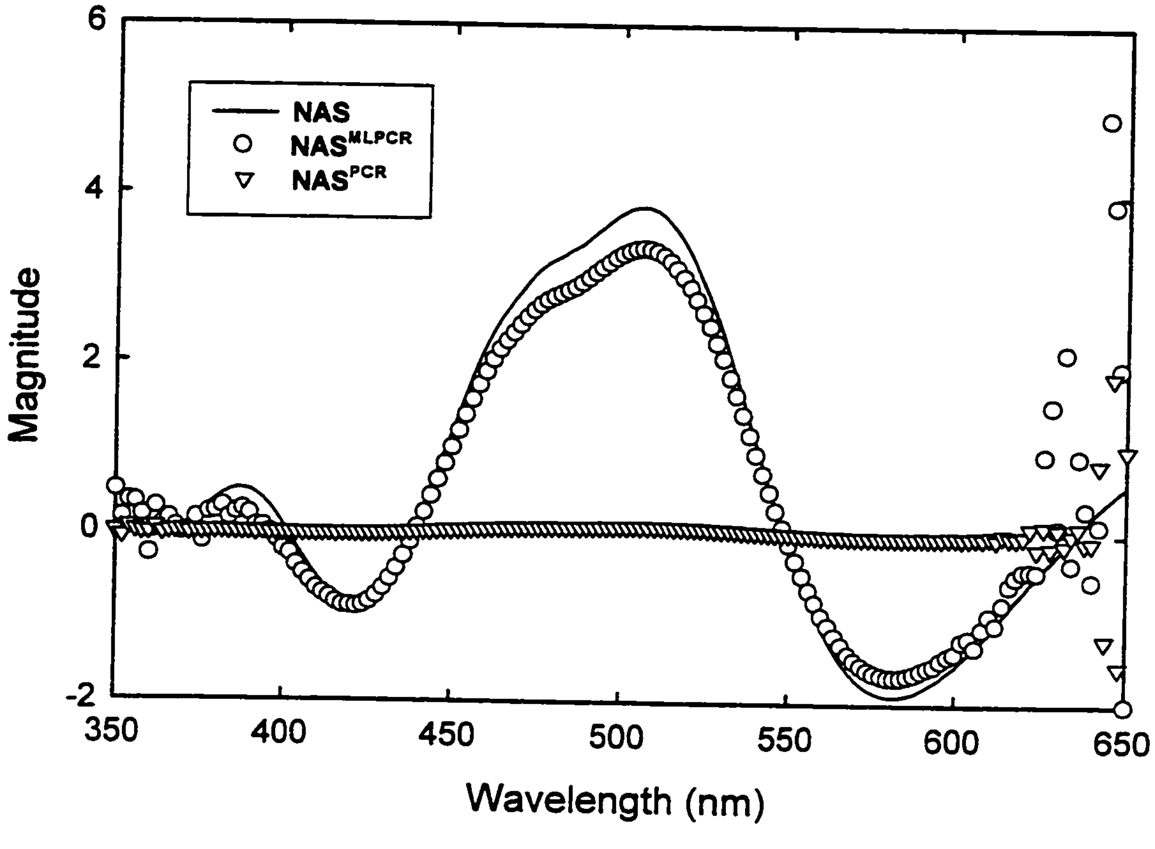


Figure 5.7 Comparison of net analyte signal vectors calculated for Co (data set 11) using different methods.

similar and have the expected shape. However, $\mathbf{NAS}^{PCR}$ is much smaller in magnitude than the other two and it is clear even without resorting to the calculation of Equation 5.27 that the sensitivity of PCR will be much lower. These observations are consistent with results of Table 5.2. Note that the small NAS for PCR does not derive from the regression step, since none is used in this direct calculation method. Instead, it is believed that the spectral space is poorly estimated by PCA as compared to MLPCA, and subsequent projection into this space reduces the sensitivity of PCR. A comparison of eigenvectors produced by PCA and MLPCA is made in Figure 5.8, which shows the loadings (abstract spectra) for each of the first three factors. It will be noted that the first two factors are virtually identical for both methods, but there are radical differences in the third factor. While the third factor for MLPCA shows some meaningful structure in the spectrally active region, the PCA results are essentially flat in this region and show contributions mainly in the region dominated by noise. In other words, at the point at which the third principal component is extracted, the residual variance in the data set is dominated by the noise, and these are the regions modeled by PCA. MLPCA, on the other hand, is able to better account for the systematic variations. This is clearly indicated by the calibration results.

As a final illustration of the power of the maximum likelihood calibration methods developed here, consider data set 6. Based on the typical spectrum shown in Figure 4.3, one would normally choose to carry out PCR on a subset of the full spectral range, *e.g.* in the region of 700 to 1600 nm. If the high noise regions are included, the PCR results are very poor due to the tendency of the PCA decomposition to model the noise variance. On the other hand, selecting a single region excludes other regions that

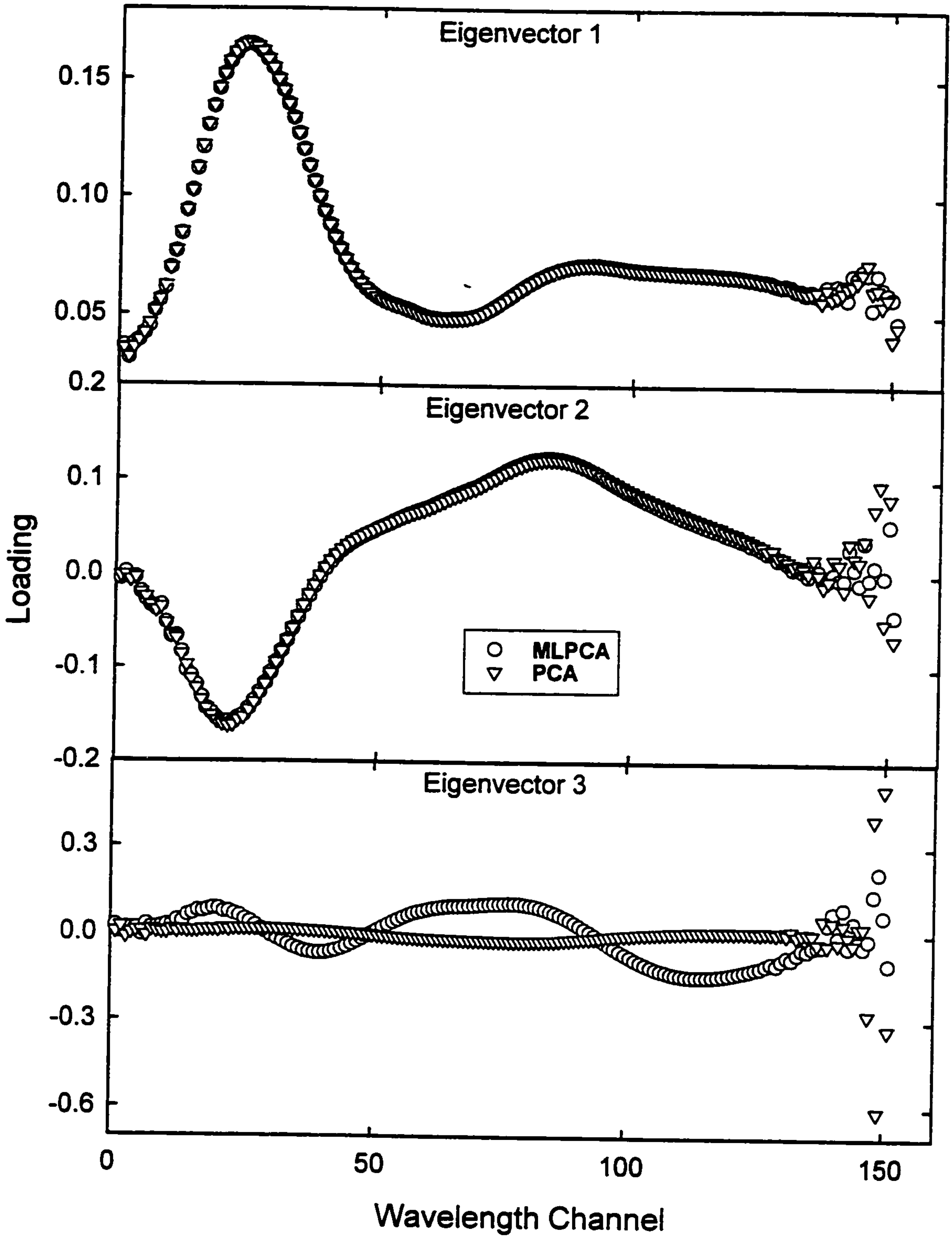


Figure 5.8 Comparison of eigenvector loadings for PCA and MLPCA applied to data set 11.

may be useful for calibration. A more refined variable selection procedure could be used, but this normally relies on cross-validation and is extremely time consuming. A better approach is to apply MLPCR to the entire data set and allow the variance information to determine the importance of each channel.

A comparison of PCR and MLPCR for data set 6 is presented in Table 5.3 in terms of cross-validation errors (leave-one-score-out method). PCR was carried out over the region 700-1600 nm, while MLPCR was applied to the entire data set. For both methods, optimum performance occurs around three latent variables, as expected. It is clear that MLPCR generates models with significantly better predictive ability for all three components. Although it is not necessarily obvious from the spectra, it is apparent from the results that the inclusion of additional wavelength channels in the analysis improves the calibration model through MLPCR. This is because important information exists in the region above 1600 nm on the shoulders of peaks that saturate the detector. Thus, valuable information lost through sub-optimal wavelength selection can be recovered through MLPCR.

It is also important to note that the results for data set 6 did not rely on precisely correct standard deviation estimates since, for all samples, these were based on 400 replicate scans for just one sample. Correlations in the measurement errors, which are known to exist, were also ignored. Nevertheless, this approximation was sufficient to improve the calibration model. This suggests that even approximate information on measurement errors, such as that which might be provided by a skilled spectroscopist, can be used to advantage in multivariate calibration.

Table 5.3 Comparison of PCR and MLPCR for data set 5 (organic mixture). Values given are the root-mean-squared errors of cross-validation (RMSECV) in weight percent.

Calibration Method	Analyte	Number of Latent Variables					
		1	2	3	4	5	6
PCR* (700-1600 nm)	toluene	16.60	7.68	0.61	0.60	0.60	0.46
	chlorobenzene	16.55	10.32	0.57	0.54	0.54	0.42
	heptane	3.13	2.66	0.15	0.14	0.14	0.14
MLPCR* (400-2500 nm)	toluene	20.95	7.96	0.12	0.13	0.13	0.12
	chlorobenzene	13.08	10.32	0.13	0.11	0.11	0.11
	heptane	2.84	2.65	0.09	0.07	0.07	0.06

* Leave-one-score-out cross-validation was used for both methods.

5.5 CONCLUSIONS

It has been the objective of this chapter to describe the theoretical basis of maximum likelihood multivariate calibration methods (MLPCR and MLLRR) that are based on MLPCA, and to present results demonstrating their ability to provide superior calibration models over conventional methods in certain cases. This objective has been accomplished through the use of both computer-generated and experimental data sets which showed that significant improvements over PCR and PLS can be realized by including measurement error information in the calibration procedure. In the majority of cases, MLLRR provided better results than MLPCR, but the improvement was often marginal for the cases examined here.

This study was not intended to be exhaustive in its investigation of the new methods and leaves open many issues concerning, for example: situations under which maximum likelihood methods should offer significant improvements, the relative merits of MLPCR and MLLRR under different measurement conditions, the role of measurement error covariance in the quality of a calibration model, and more extensive comparisons with other methods. Nevertheless, the underlying reasons for the improved results have been described from a fundamental perspective using standard figures of merit for multivariate calibration.

Two of the most common arguments against methods such as MLPCR and MLLRR relate to the requirement for measurement error variance estimates and the extended computation time necessitated by the algorithm. The first argument asserts that methods such as PCR require no variance information and are therefore more universally applicable. This argument is deceptive, since the use of PCR implicitly assumes that the measurement errors are uniform, so variance information is, in fact, required. In the absence of any knowledge of measurement error characteristics whatsoever, an assumption of uniform errors may be reasonable, but practitioners of PCR and similar methods should be aware of the limitations that such assumptions impose. It is the author's contention that some instinct for measurement error characteristics on the part of the analyst is almost always present. Even if measurement error variances are not directly available, reasonable approximations of the error structure can be used effectively with the maximum likelihood techniques, as was demonstrated with data set 6. This should also be true even when the error distribution is only approximately normal, or when an exact covariance structure is not known. Finally, the results presented here support the case for designing instruments which provide measurement error information. Some instruments presently have this capability, but more often the information is unavailable, even when the instrument has the fundamental ability to provide it routinely from replicate scans (*e.g.* FTIR spectrometers).

It is true that the maximum likelihood methods presented here are more computationally intense than traditional PCA-based methods. However, the basic MLPCA algorithm (presented in Table 4.1 and Appendix C) is quite simple to implement and converges reliably without the need for any "fine tuning" like many algorithms. Actual computation times vary with the size of the matrix and error structure and have been described in Chapter 4. In this study, time for calculations ranged from several minutes to several days, with the longest times being observed for leave-one-sample-out cross-validation for MLPCR and MLLRR. As demonstrated here, leave-one-score-out cross-validation is generally equivalent for MLPCR and reduces computation time by a factor equal to the number of samples. This might typically take a few hours. Unfortunately, it is not possible to perform leave-one-score out cross-validation for MLLRR because of the inclusion of concentration information, which is a drawback to this method. In any case, the time spent on calibration is still much less than that typically required to obtain the experimental data, and past history has demonstrated that computational barriers erode quickly with`advancing technology.

Beyond the broad utility that these methods may find in practical situations, there is a more important aspect of their development. Whereas many new techniques are simply modifications of conventional methods designed to improve their utility, MLPCR and MLLRR are generalizations of PCR and LRR, respectively. In other words, PCR and LRR are special cases of the parent techniques that apply under conditions of uniform error variances. The development of general principles and methods for incorporating measurement uncertainties into the calibration process will allow the limitations and strengths of other calibration techniques to be appreciated from a wider perspective, a feature which is inherently valuable.

In the context of the preceding statement, the performance of PLS in the results presented here can be examined. Direct comparisons with PLS have been avoided until now because of basic differences in the fundamental philosophy towards the calibration process. It is generally viewed that, for systems with a well-defined rank, PLS should provide results comparable to PCR when the correct number of latent variables is used (although PLS may provide better results than PCR when fewer latent variables are used). We have found this to be the case when uniform measurement errors prevail, but in cases where measurement errors are significantly non-uniform, PLS consistently performed better than PCR, although worse than the maximum likelihood methods. This is likely because PLS uses correlation with concentration data to help exclude much of the noise variance. Based on this observation, one can speculate that the presence of non-uniform noise in many other applications may be partly responsible for the relative popularity of PLS over PCR in practical environments. This factor may also be important in the relative success of wavelength selection methods for some methods but not for others. Whatever the reasons for these observations, further investigation is warranted and the maximum likelihood calibration methods presented here provide a unifying framework from which to better understand the application of multivariate calibration methods to chemical problems.

The advantages maximum likelihood techniques can offer the area of calibration have been illustrated in this chapter. The next chapter demonstrates the application of these principles to other areas, such as the analysis of incomplete data sets and calibration transfer.

6
APPLICATIONS OF MAXIMUM LIKELIHOOD PRINCIPAL COMPONENT ANALYSIS TO INCOMPLETE DATA SETS AND CALIBRATION TRANSFER

6.1 INTRODUCTION

It has been noted throughout this thesis that of all the different techniques available to analyze multivariate data sets, none is more widely used than principal component analysis (PCA). The strength of PCA stems from its ability to represent multivariate data using a smaller number of variables, called principal components. To do this, the information in many variables is compressed into the first p principal components. If these reproduce the data within experimental error, then p represents the intrinsic rank, or pseudorank, of the data. However, for PCA to be implemented directly, a complete data set (*i.e.* no missing measurements) is needed, and this is not always the case.

Incomplete data sets commonly arise in a number of situations in chemistry. When modeling chemical data or in exploratory data analysis, it is conceivable that some measurements for a particular sample may not have been recorded or are impossible to obtain experimentally. In other instances, insufficient sample may be available for all measurements, or some measurements may be excluded as erroneous. In the analysis of a time series (*i.e.* in multivariate statistical process control), measurements may be missing due to sensor failure for a period of time. The problem of calibration transfer, where one wishes to transfer calibration results from a "master" instrument to a "slave" instrument based on a small subset of samples, can also be regarded as a missing data problem. In this case, whole spectra from the slave instrument are unavailable. Given the pervasiveness of the missing data problem, it would be extremely useful to have a simple and reliable technique for minimizing the influence of missing measurements in PCA without excluding the incomplete samples entirely. Furthermore, such a method should have a sound theoretical basis and be capable of predicting missing measurements according to some recognized criterion of optimality.

A number of methods have been developed over the years to handle missing measurements in the multivariate analysis of chemical data [69-74]. Most of these exploit the fact that the intrinsic rank of the data matrix is substantially smaller than the full rank. However, there has been no consensus on a procedure that allows PCA to address this problem in an optimal manner. In this chapter, a new approach to the missing data problem is introduced through the application of maximum likelihood principal component analysis (MLPCA), a method described earlier in this thesis. MLPCA is a generalized form of PCA which allows estimates of measurement uncertainty to be incorporated in the decomposition step. Thus, missing data can be accommodated simply by assigning very large variances to these measurements prior to implementing MLPCA. In this study, the feasibility and advantages of this approach are demonstrated using a number of experimental data sets.

6.2 THE PROBLEM OF MISSING DATA

As noted in the preceding section, there have been a number of techniques developed to address the problem of missing data. The simplest of these involve preprocessing of the data prior to the decomposition step in PCA. If the number of missing values is small relative to the size of the data matrix, the usual approach is to discard either all of the samples or all of the sensors with missing measurements (*i.e.* delete entire rows and/or columns of the matrix). Unfortunately, potentially useful information from the eliminated sensors (or samples) will be lost. As the proportion of missing values increases, this loss of information can be very significant and, in some cases, the pattern of missing data makes this approach impossible. An alternative would be to "predict" the missing values prior to PCA. The simplest variation of this approach involves the substitution of a zero wherever a missing point is encountered. In certain instances this type of preprocessing may produce acceptable results, but it is not recommended since it can seriously distort the underlying structure of the data. Usually, a better estimate can be obtained via a substitution of the mean of the observed points [2]. Contrary to a popular view, however, substitution of mean values is not neutral from a modeling perspective, and again serious distortion can result.

A more sophisticated approach uses a covariance calculation on the data. Usually, PCA is performed on the original *mxn* data matrix, **X**, using singular value decomposition (SVD) which gives the decomposition,

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{6.1}$$

Alternatively, the covariance matrix of **X** can be decomposed by SVD yielding

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{V}\mathbf{S}^{2}\mathbf{V}^{\mathrm{T}} \tag{6.2}$$

or

$$\mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{U}\mathbf{S}^{2}\mathbf{U}^{\mathrm{T}} \tag{6.3}$$

If the covariance matrix of **X** can be calculated by excluding missing values and reducing the degrees of freedom accordingly, it may be possible to estimate **U** or **V** in this manner. However, the decomposition will weight all of the matrix elements equally and therefore is unlikely to yield an optimum solution. Furthermore, in cases where there are a large number of missing values, this covariance calculation may not be possible.

An extension to the above procedure has been described by Wise in the PLS_Toolbox [75]. In this method, the covariance matrix is calculated and then decomposed by SVD. The intrinsic rank of the data, *p*, is determined before the analysis and estimates of the missing points are obtained through a regression of the loadings. These values are then substituted into the original data matrix and this updated matrix is used to improve the estimates of the SVD model. This procedure is performed iteratively until either the change in the estimated points falls below a predetermined tolerance or a maximum iteration value is reached. Although this technique proves useful in many instances, it is somewhat cumbersome and its statistical basis has not been explored. Also, if the number of missing values is large, the algorithm may become unstable.

This work investigates the application of MLPCA to the missing data problem. This technique performs a PCA-like decomposition of the data but, unlike PCA, uses measurement error variance information to choose the *p*-eigenvectors. The procedure generates a decomposition which is optimal in a maximum likelihood sense for a model of given dimensionality (*p*) provided that the measurement errors are distributed as multivariate normal and their variances/covariances are exactly known. In practice, this restriction is seldom met, but that does not detract from the utility of the method (in the same way that the assumptions implicit in linear regression do not limit its use to cases where those assumptions are valid). Usually, independent, normally distributed measurement errors are assumed and sample variances are used. Although these approximations may not be valid, they permit the application of the MLPCA algorithm, which will incorporate the variance information in the decomposition (in contrast to PCA which assumes equal error variances for all measurements). Similar to conventional regression methods, MLPCA minimizes a weighted residual sum of squares which is given by:

$$S^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(x_{ij} - \hat{x}_{ij}\right)^2}{\sigma_{ij}^2} \tag{6.4}$$

where $\hat{x}_{ij}$ is the maximum likelihood estimate of measurement x_{ij}, and σ_{ij} is the corresponding measurement error standard deviation. The MLPCA decomposition is carried out using a form similar to Equation 6.1:

$$\hat{\mathbf{X}} = \hat{\mathbf{U}}\hat{\mathbf{S}}\hat{\mathbf{V}}^{\mathrm{T}} \tag{6.5}$$

where $\hat{\mathbf{X}}$ is *mxn*, $\hat{\mathbf{U}}$ is *mxp* , $\hat{\mathbf{S}}$ is *pxp* and $\hat{\mathbf{V}}^{\mathrm{T}}$ is *pxn*. It is important to point out here that one of the major differences between PCA and MLPCA is that MLPCA does not have nested solutions (*i.e.* MLPCA must be performed for each change in the rank estimate, *p*). Another difference, and a significant advantage, is that the projection of the original data onto the MLPCA eigenvectors is performed using a maximum likelihood projection, which weights the direction of the projection in proportion to the magnitude of the measurement error variances. That is to say, while PCA projections are orthogonal and take the form:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i \tilde{\mathbf{V}}\tilde{\mathbf{V}}^{\mathrm{T}} \tag{6.6}$$

where $\mathbf{x}_i$ is a row vector of **X**, and $\tilde{\mathbf{V}}$ is the loading matrix truncated to p principal components, MLPCA projections take the form:

$$\hat{\mathbf{x}}_i = \mathbf{x}_i \Sigma_i^{-1} \hat{\mathbf{V}} \left(\hat{\mathbf{V}}^{\mathrm{T}} \Sigma_i^{-1} \hat{\mathbf{V}} \right)^{-1} \hat{\mathbf{V}}^{\mathrm{T}} \tag{6.7}$$

where Σ_i^{-1} is the *nxn* covariance matrix for the row vector $\mathbf{x}_i$ (diagonal matrix of variances for independent error). In a projection of this type, less importance will be given to those measurements with large uncertainties. When the errors in all the measurements are the same, the Equation 6.7 will reduce to Equation 6.6.

In this chapter, the advantages of performing a maximum likelihood decomposition on incomplete data sets will be demonstrated by showing that: (a) even when many of the data are missing, the scores and loadings can retain most of the original information; (b) missing data can be reliably predicted; and (c) MLPCA can be used as an alternative approach to calibration transfer and, under the right conditions, produces results similar to those that would be observed if all the spectra on the slave instrument were measured.

6.3 EXPERIMENTAL

Three experimental data sets (data sets 6, 12 and 13) were employed to demonstrate the advantages mentioned above. To distinguish the two new data sets from those described in earlier chapters, they have been designated as data sets 12 and 13. Data set 12 is a widely used archaeological data set described by Kowalski *et al* [76]. These data were obtained in a study intended to relate the origin of obsidian artifacts collected at three different locations to samples taken from four quarries and consist of X-ray fluorescence data on 10 elements for 75 samples. Data set 13 consists of results compiled by de Ligny *et al* [77] in a study of chromatographic retention characteristics. This set contains transformed retention data for 39 solutes with 2 eluents on 6 adsorbents, but is only about 80% complete. Data set 6 (see also Section 4.3) was part of an Infometrix (Seattle, WA) calibration transfer study [57] and consisted of near infrared (NIR) absorbance spectra acquired for 31 samples (mixtures of toluene, chlorobenzene and heptane) measured on two instruments. Further details on each of these data sets are included in the appropriate section of the RESULTS.

The calculations in this work were performed using Matlab 4.2c.1 (The Mathworks, Natick, MA) on two computer platforms: (1) a Pentium-based personal computer and (2) a Sun Microsystems SparcServer 1000 with 230 Mb of memory and four 50 MHz SuperSPARC CPUs. Under normal circumstances, convergence of the algorithm is quite reliable, although it is considerably slower than for conventional PCA. With missing data, the convergence times are often further extended and may be a problem in extreme cases. Results in this work required times ranging anywhere from under an hour to more than a day. Although this is a drawback to this method, we feel that its sound theoretical basis and versatility justifies its use and are confident that algorithmic improvements (particularly in obtaining initial estimates for the solution) will greatly improve its performance.

6.4 RESULTS

6.4.1 Exploratory Data Analysis

One of the methods commonly employed to visualize the characteristics of multidimensional data sets is to utilize a projection of the original data into the space described by the first two or three principal components. Samples with similar features often appear "clustered" together in space after this dimensionality reduction. As an illustration of this approach, we will consider data set 12. No pretreatment of the data was performed prior to analysis and, in the absence of other reliable information, uniform measurement errors with a variance of unity were assumed. In the case of a complete data set and uniform errors, the actual magnitude of the error variance is unimportant, since MLPCA is equivalent to PCA under these conditions. In the case of incomplete data, it is only important that the magnitude of variance for missing measurements is much greater than that for the measurements present (a factor of 10^{10} was used in this case).

Figure 6.1a shows the MLPCA projection of the uncensored data onto the first two eigenvectors. This projection is equivalent to the PCA projection because the measurement error variances are uniform. The grouping of results according to class (1 to 4 are different quarries; 5 to 7 are different artifact locations) is consistent with

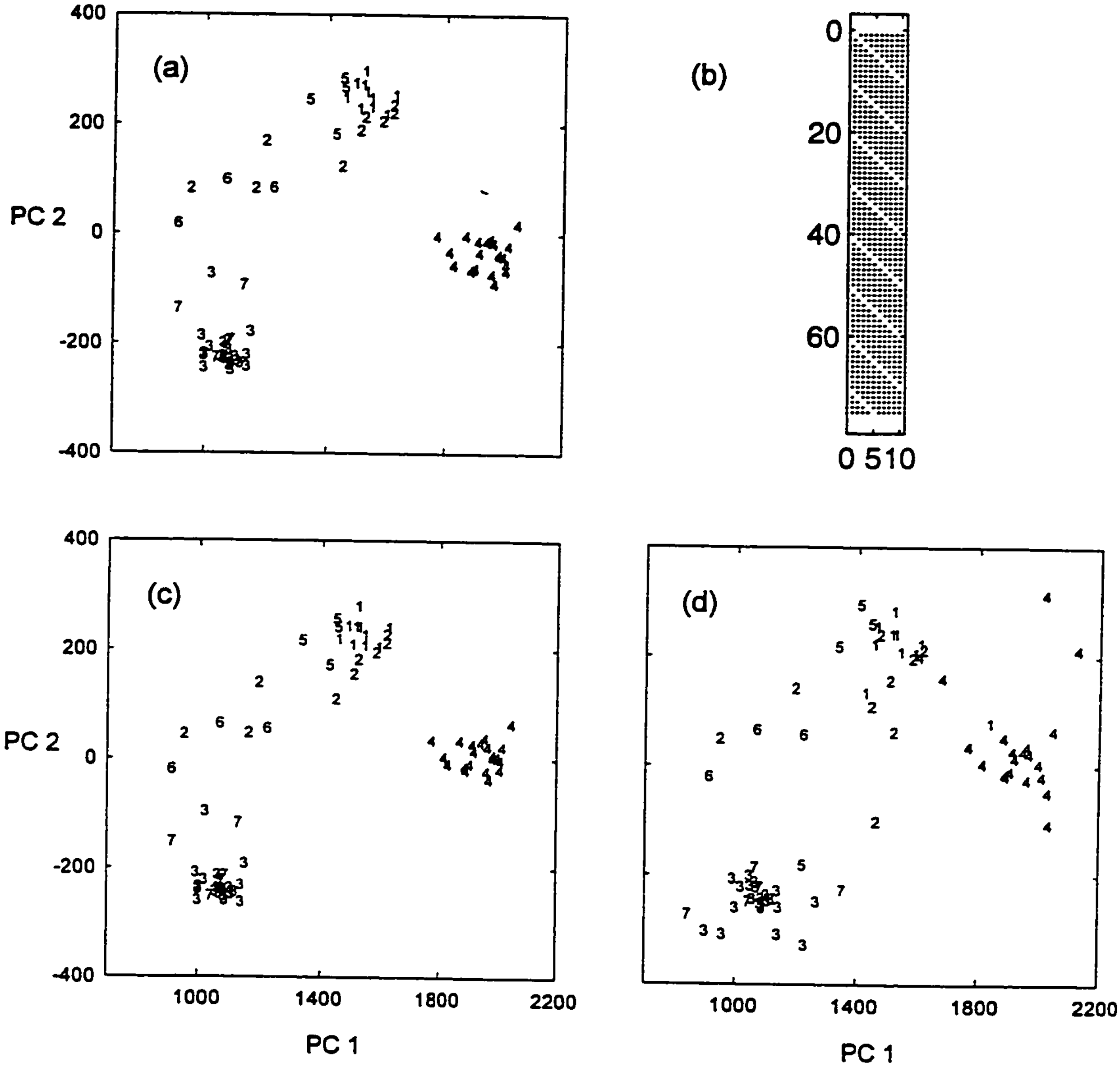


Figure 6.1 (a) Projection of archaeological data onto the first two eigenvectors determined by MLPCA. (b) Pictorial representation of data censoring mask for 10% missing data (the white spaces indicate where the data have been removed). (c) Projection of original data into subspace determined by MLPCA for censored data. (d) Maximum likelihood projection of censored data into MLPCA subspace for censored data.

earlier studies and reveals an association between the quarries and the artifacts. However, the actual interpretation of the data is not of interest here, but rather the behavior of the principal components when data are deleted.

To test the ability of MLPCA to deal with missing measurements, 10% of the measurements were eliminated in a systematic fashion. A "censoring" mask, as shown in Figure 6.1b, was applied to the original`data and was generated by removing the first measurement from row 1 (sample 1), the second measurement from row 2, and so on, cycling back to the beginning when the end of each row was reached. Note that, because each sample has at least one measurement missing, this data set cannot be analyzed by conventional PCA by eliminating rows or columns. However, MLPCA easily handles this situation by employing inflated variances for the missing measurements and the projections are shown in Figures 6.1c and 6.1d. Figure 6.1c shows the projection of the original data (*i.e.* uncensored) onto the first two eigenvectors which were determined by MLPCA performed on the censored data set. This can be regarded as "cheating" because if the data set were indeed incomplete, the measurements that had been removed would not be available for this projection. However, the intent is to demonstrate that even with the missing data, the eigenvectors retain nearly all the information that was contained in the original data. This conclusion is verified by comparison of Figures 6.1a and 6.1c, which are virtually identical. Figure 6.1d is a more accurate portrayal of what one would observe if the measurements were truly absent. In this case, maximum likelihood projections of the censored data onto the MLPCA eigenvectors were used. Although there are small perturbations, it can be seen that the projection very closely resembles the projection for the complete data set in Figure 6.1a, particularly with regard to the clusters observed. Slightly better results were obtained when the data were mean-centered, but this method of preprocessing would not be valid if measurements were truly missing. Of course, the success of such an application will depend on the data set and the correlation of the measurements, but the availability of a statistically optimal procedure for treating missing data is a significant advance.

Data set 12 was also analyzed under more extreme conditions, with as many as 54% of the measurements removed at random. The "censoring" mask used in this case is presented in Figure 6.2b. Under these circumstances, the representation of the two-dimensional subspace remained quite good. This is illustrated in Figure 6.2c which shows the projection of the original data onto the eigenvectors for the censored data. Although there is a reorientation of the eigenvectors, the spatial relationship among the samples is largely unchanged. As might be expected, projection of the censored data was not as successful. In this case, the projection is dominated by a few outlying samples which are missing critical measurements. If these are eliminated, the projection in Figure 6.2d results. Although there is clearly a loss of information in this extreme case, some of the associations are still apparent.

As well as projecting the individual samples into the subspace, a similar treatment may be carried out on the uncertainties associated with the samples. If the *nxn* measurement error covariance matrix for a sample (*i.e.* a row of **X**) is given by Σ_i , then the *pxp* error covariance matrix for the scores of that sample, Ψ_i, will (by propagation of error) be given by Equation 6.8.

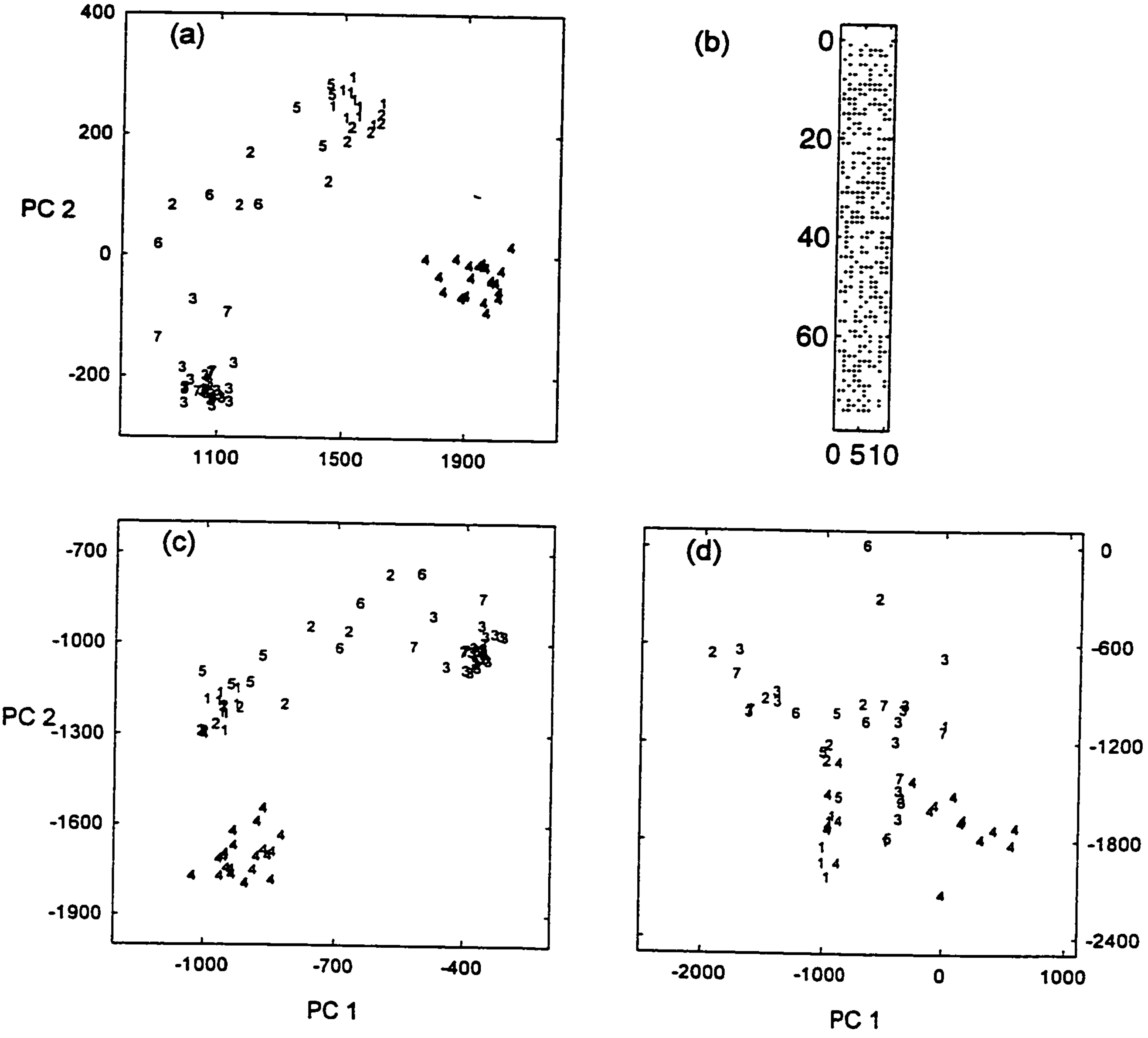


Figure 6.2 (a) Projection of archaeological data onto the first two eigenvectors determined by MLPCA. (b) Pictorial representation of data censoring mask for 54% missing data (the white spaces indicate where the data have been removed). (c) Projection of original data into subspace determined by MLPCA for censored data. (d) Maximum likelihood projection of censored data into MLPCA subspace for censored data.

$$\Psi_i = \left[\left(\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_i^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_i^{-1}\right]\Sigma_i\left[\left(\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_i^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_i^{-1}\right]^{\mathrm{T}} \tag{6.8}$$

$$= \left(\hat{\mathbf{V}}^{\mathrm{T}}\Sigma_i^{-1}\hat{\mathbf{V}}\right)^{-1}$$

While this equation does not take into account the uncertainty in the eigenvectors, and so cannot be regarded as a true measure of variance in the scores, it is useful in identifying outliers which are the result of incomplete data. Normally, MLPCA reduces the influence of uncertain measurements through the projection, but in some cases there is insufficient information to do this, resulting in an outlier (*e.g.* sample 4 in the above example, which only has one measurement). For the archaeological data, a projection of the error covariance matrix for a given sample with uniform, uncorrelated errors will yield a diagonal matrix (2x2) with measurement variances on the diagonal. If the projection of the covariance matrix for some samples with missing data differs from this, the magnitude of the difference will indicate the uncertainty of the scores in the new space. Error bars may be obtained by taking the square roots of the diagonal elements of the error covariance projection and error contours may be constructed utilizing the whole error covariance projection. While caution should be used in interpreting these as true variances of the scores, this approach is useful in the identification of outliers.

6.4.2 Modeling Incomplete Data

The treatment of missing data is expected to be even more successful when strong correlations exist among the measurements. For data set 13, retention volumes (V_N) were measured for 19 solutes (monosubstituted benzenes and polycyclic aromatic hydrocarbons), 2 eluents (n-hexane and 35% v/v methylene chloride in n-hexane) and 6 silica-based adsorbents, and transformed according to:

$$y = \log(V_N / W) \tag{6.9}$$

where W is the weight of the adsorbent. Out of a possible 228 measurements, however, only 183 were available. These data have been analyzed in the past using a physical model proposed by Snyder [78] and three-way factor analysis by de Ligny [77]. The objective in the development of such models is to enable prediction of retention characteristics of solutes with a minimum number of parameters. Estimation of such parameters is made difficult by incomplete data.

Because this data set is third-order, some preprocessing had to be performed prior to implementation of MLPCA. The easiest manipulation would be to analyze each adsorbent or eluent individually and combine the results. Unfortunately, if much or all of the data are missing for a particular solute in one subset, the predictive power of this technique will be greatly reduced. Therefore, it would be beneficial to use any extra information that may be contained in data for the other adsorbents or eluent in the decomposition step. For this work, the third-order data were "unfolded" to give a 19x12 matrix where the first 6 columns correspond to the transformed retention data for eluent 1 and the latter 6 for eluent 2.

MLPCA was applied to this unfolded matrix and the simplest model that produced reliable results had a rank of 3. The maximum likelihood projections were then used to reconstruct the missing values, which are listed in Table 6.1. Also included in the table are the values predicted by de Ligny [77] and Snyder [78]. From the table, it is

Table 6.1 Predictions of missing data for data set 13 (transformed chromatographic data).

Adsorbent	Eluent	Sample	MLPCA	de Ligny	Snyder
1	1	5	2.08±.07	2.22±.18	2.07
2	1	18	0.53±.06	0.54±.13	0.31
3	1	13	0.90±.03	0.91±.12	0.53
4	1	18	1.54±.13	1.53±.21	1.92
4	1	19	2.29±.13	2.25±.23	3.07
5	1	18	1.55±.12	1.53±.20	1.94
5	1	19	2.30±.12	2.26±.23	3.09
6	1	18	1.61±.13	1.64±.22	2.00
6	1	19	2.32±.12	2.41±.25	3.12
1	2	7	−0.91±.12	−0.83±.18	−0.86
1	2	9	−0.86±.08	−0.81±.15	−0.90
1	2	10	−1.05±.12	−0.96±.19	−1.18
2	2	8	−0.60±.05	−0.66±.13	−1.09
2	2	14	−0.54±.04	−0.56±.12	−0.81
2	2	18	−0.41±.04	−0.42±.13	−0.86
3	2	1	−0.56±.07	−0.58±.16	−0.69
3	2	2	−0.75±.07	−0.78±.17	−0.74
3	2	7	−0.70±.11	−0.69±.20	−0.67
3	2	8	−0.71±.08	−0.78±.17	−0.86
3	2	10	−1.06±.11	−1.05±.23	−0.94
3	2	13	−0.45±.04	−0.45±.12	−0.30
4	2	1	−0.53±.08	−0.52±.17	−0.42
4	2	2	−0.64±.08	−0.64±.18	−0.65
4	2	7	−0.52±.11	−0.49±.19	−0.48
4	2	8	−0.48±.09	−0.51±.18	−0.64
4	2	9	−0.42±.07	−0.39±.15	−0.28
4	2	10	−0.99±.11	−0.94±.23	−0.54
4	2	18	0.45±.12	0.46±.20	0.62
4	2	19	1.07±.12	1.08±.23	1.62
5	2	1	−0.45±.07	−0.50±.17	−0.39
5	2	2	−0.58±.07	−0.65±.18	−0.63
5	2	7	−0.50±.10	−0.52±.19	−0.48
5	2	8	−0.46±.08	−0.54±.18	−0.65
5	2	9	−0.38±.06	−0.41±.15	−0.29
5	2	10	−0.90±.10	−0.96±.23	−0.56
5	2	18	0.42±.11	0.44±.20	0.61
5	2	19	1.00±.10	1.07±.23	1.60
6	2	1	−0.67±.10	−0.58±.18	−0.36
6	2	2	−0.75±.10	−0.67±.19	−0.51
6	2	7	−0.58±.13	−0.49±.20	−0.23
6	2	8	−0.54±.11	−0.51±.19	−0.37
6	2	9	−0.48±.08	−0.40±.16	−0.02
6	2	10	−1.18±.13	−1.00±.24	−0.24
6	2	18	0.54±.15	0.50±.22	0.89
6	2	19	1.28±.14	1.63±.25	1.88*

clear that the MLPCA estimates of the missing data are in excellent agreement with those determined by de Ligny [77] and, with the exception of the last entry, fall within the confidence intervals determined in that work. This agreement is somewhat remarkable given that de Ligny *et al* use a trilinear model (75 parameters) whereas this study used an unfolded bilinear model (84 parameters). For a comparison, it would be useful to have an estimate of the uncertainty associated with the data predicted by MLPCA. This may be accomplished using an extension to Equation 6.8 which projects the covariance in the scores back into the original space:

$$\Sigma_i^{pred} = \hat{\mathbf{V}}\left(\hat{\mathbf{V}}^{T}\Sigma_i^{-1}\hat{\mathbf{V}}\right)^{-1}\hat{\mathbf{V}}^{T} \tag{6.10}$$

The diagonal elements of Σ_i^{pred} will contain the variance information for a given sample. Although the magnitude of the uncertainties used in MLPCA will have no bearing on the decomposition as long as it is the same for all of the data present, a valid estimate of measurement uncertainty is required for use in Equation 6.10. De Ligny suggests that an approximate measurement error variance for known values may be obtained using the root-mean-squared error (RMSE) of the fit for the non-missing data. Confidence intervals may then be estimated for data points using Equation 6.11.

$$95\%\,\mathrm{CI} = \hat{x}_{ij} \pm 1.96\left(\hat{\sigma}_{jj}\right)_i \tag{6.11}$$

In this equation, $\hat{x}_{ij}$ is the predicted measurement and $\left(\hat{\sigma}_{jj}\right)_i$ is the square root of the j^{th} diagonal element of Σ_i^{pred}. It can be seen from Table 6.1 that the confidence intervals for the MLPCA prediction are much smaller than their respective counterpart predicted by de

Ligny. This results from a better fit of the known data by the MLPCA approach (RMSE = 0.041) than that obtained by de Ligny (RMSE = 0.10). Although there is a difference in the number of parameters used in the fit, the degree of improvement appears much greater than one would expect.

While the MLPCA and de Ligny results were in good agreement with each other, the estimates by Snyder were often at the limits of the confidence interval of de Ligny and beyond those of the MLPCA model. This suggests that there may be some deficiencies in Snyder's approach. Although MLPCA produced a better fit than de Ligny *et al*, it cannot be said with certainty which is the best approach since the models are substantially different. However, MLPCA does have certain advantages, namely: (1) it is based on a well-established statistical criterion and is easy to apply; and (2) if error estimates are available for the known data, these can be incorporated into the decomposition step.

6.4.3 Calibration Transfer

With the growing use of multivariate instruments in the workplace, a serious issue that has arisen is that of calibration transfer or instrument standardization. The problem is that a calibration model determined for a particular instrument may lead to very poor prediction if used with data collected on another instrument. For example, calibration parameters obtained on a laboratory spectrometer may not yield the same results as an instrument used in an industrial setting, even if they are the same instrument model. A variety of sources may give rise to the failure of a calibration model, including variations in bandwidth, noise or sensitivity, differences in wavelength registration, and changes in instrument characteristics with time or the operating environment. One solution to this problem is to rerun all of the calibration samples on each instrument. Unfortunately, this procedure can be very time-consuming, especially if the number of calibration samples and/or instruments is large. Compounding this, environmental changes may require frequent recalibration of an individual instrument, which may in turn require the maintenance of a large number of standards. Therefore, it would be desirable to obtain calibration data on a "master" instrument and then transfer it to a "slave" instrument using only the data from a few representative samples. The subject of calibration transfer has been addressed in several recent articles which introduce a variety of techniques [79-82]. An alternative approach, based on MLPCA, is presented here to illustrate the potential of maximum likelihood methods. This study is not intended to be comprehensive, and a more complete analysis and comparison with other calibration transfer methods should be a subject for future work.

The approach to this problem is similar to the missing data examples addressed above, except whole samples from the slave instrument are now regarded as missing. For this work, data set 6 is used to illustrate the typical implementation of this technique. The data from the first spectrometer (NIRSystems Model 6500) has dimension 30x201 (A data, shown in Figure 6.3a) and will be regarded as originating from the "master" instrument (range is 1100-1300 nm with 1 nm resolution). The second spectrometer (Guided Wave Model 300P) represents the “slave” and the data from this instrument (B data, also shown in Figure 6.3a) has the same dimensionality (range is 1100-1500 nm with 2 nm resolution). The data matrix used for analysis by MLPCA was an augmentation of the A data with part of the B data and has dimension 30x402. The left half of this matrix corresponds to the full A data matrix, while the right half consists of

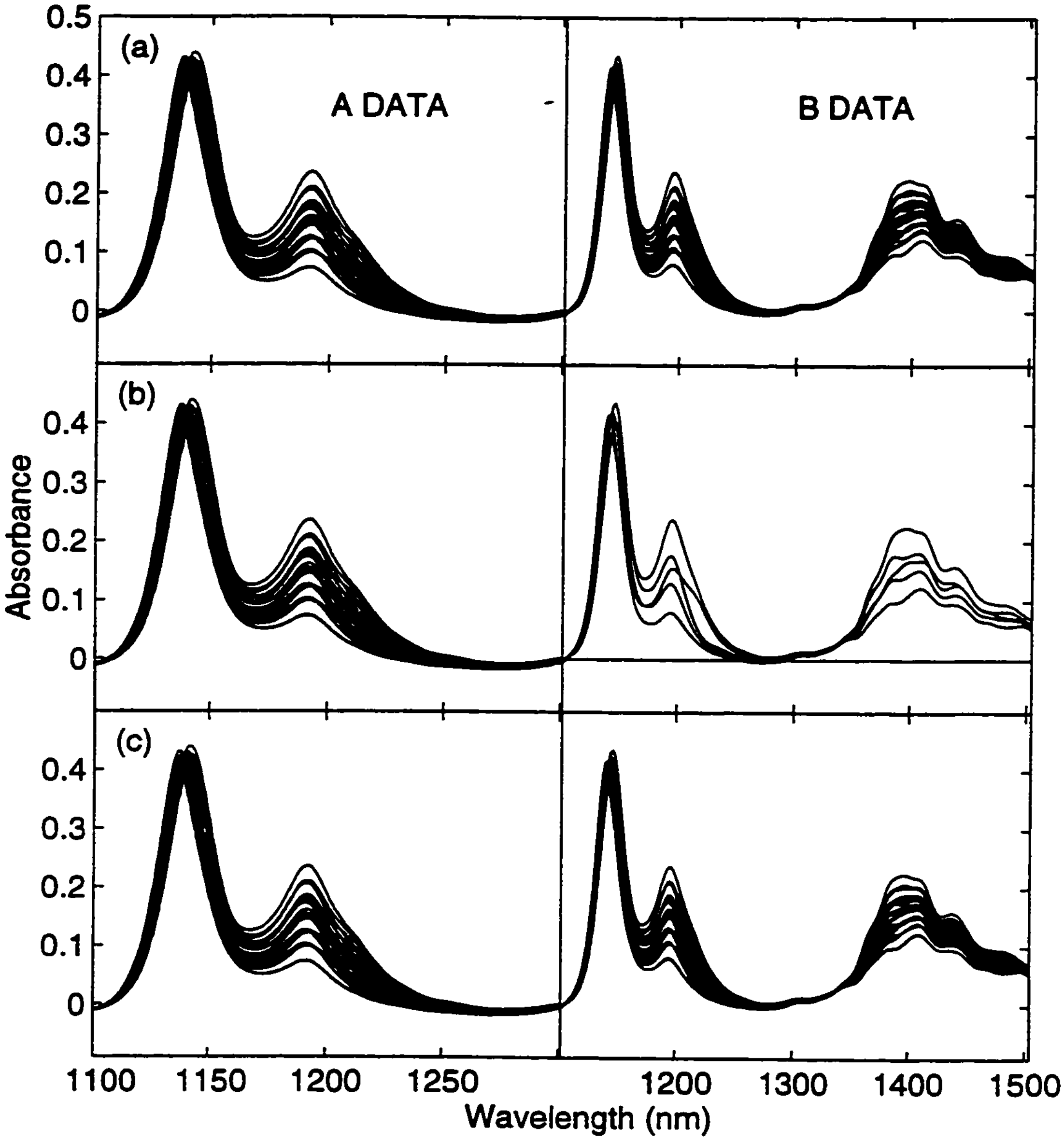


Figure 6.3 (a) Plot of master (A) and slave (B) spectra for data set 6. (b) Graphical representation of augmented data matrix with 5 representative samples from B. (c) Plot of reconstructed data following analysis by MLPCA.

five representative samples (rows) from B. The remaining samples in B are represented by row vectors of zeros (shown in Figure 6.3b). The "known" spectra are assigned uniform variance (the actual value is unimportant, but for this work selected as unity) while the "missing" spectra were assigned a measurement variance equal to 10^{10}. The issue of which subset of samples best represents the B data was addressed using the procedure outlined by Kennard [83].

After application of MLPCA of rank three (determined by cross-validation of A), the augmented data matrix was reconstructed (shown in Figure 6.3c) using the maximum likelihood estimates of the scores and loadings. It can be seen that this reconstruction very closely resembles the original data (Figure 6.3a), and therefore it appears that this is a valid technique for the transfer of spectral data from one instrument to another. Once the maximum likelihood estimates of the spectra on instrument B were obtained, they were used to build a calibration model for that instrument (also three factors). This model was then used to predict concentrations from the actual spectra obtained from instrument B. The RMS errors of prediction are listed in Table 6.2. For comparison, the results from the cross-validation of the original A and B data sets are also included in the table as RMS errors of cross-validation (RMSECV). From the table, it is clear that the prediction errors obtained after calibration transfer are comparable or slightly better than those obtained from the original data for instrument B. Of course there are many issues to be considered in more detail, such as the nature of the transfer subset and the rank of the transfer model. However, this work has shown that, when calibration transfer is regarded as an incomplete data problem, it is possible to use MLPCA to predict calibration spectra

Table 6.2 Comparison of prediction errors from MLPCA calibration transfer with cross-validation errors from master (A) and slave (B) instruments. (Results in units of weight percent; based on five calibration transfer standards and a rank three PCR model).

Component	Instrument A RMSECV	Instrument B RMSECV	Calibration Transfer RMSEP
toluene	0.23	0.30	0.29
chlorobenzene	0.25	0.27	0.24
heptane	0.12	0.12	0.11
total	0.21	0.24	0.23

on the slave instrument and obtain a calibration model similar to that which would be obtained if all of the calibration samples were actually run on that instrument.

6.5 CONCLUSIONS

The problem of incomplete data sets is a pervasive one in the multivariate analysis of chemical measurements. In this chapter, it has been demonstrated that MLPCA is a convenient and reliable approach to solving this problem. While the assumptions for maximum likelihood estimation may not be generally valid for the application of MLPCA to all data sets (*i.e.* known variances, normally distributed errors), MLPCA should be a better alternative for the analysis of multivariate data sets when these assumptions are approximately valid. Furthermore, MLPCA provides a legitimate statistical framework for addressing these problems. The application of MLPCA to a range of missing data problems in exploratory data analysis, modeling, and calibration transfer has shown its versatility and utility.

7
CONCLUSIONS AND FUTURE WORK

7.1 CONCLUSIONS

In order to deal with the increasing complexity of chemical data, the disciplines of analytical chemistry and statistics have combined in order to develop an arsenal of techniques to analyze these data and extract as much useful information as possible. Unfortunately, most of the techniques commonly used today virtually ignore the role of measurement error when modeling data. This work has demonstrated how the use of maximum likelihood methods can accommodate these errors while potentially improving the quality of information that would be obtained using existing methods and providing a unifying framework for common multivariate techniques.

In Chapter 2, the concept of data modeling in two dimensions was introduced. The low dimensionality of the data allowed for a simple graphical and statistical comparison among existing univariate techniques. The concept of maximum likelihood was introduced by considering errors in both axes when building the mathematical model. It was shown that MWR (the maximum likelihood method) was a generalized approach when compared to the other techniques discussed. The maximum likelihood approach taken demonstrated how each method is a specific case of the general model (MWR) and therefore guidelines were proposed regarding which model should be used when a given error structure arises. In addition, for the data sets examined, this technique consistently yielded a smaller mean-squared-error (MSE) and was less biased than any other nonequivalent method, indicating that this approach is the most desirable from a statistical standpoint.

In Chapter 3, the principles which were developed in two-dimensions were extended to higher dimensionality. The resulting model, called maximum likelihood principal component analysis (MLPCA), demonstrated how contentious issues, such as mean-centering and scaling, can be easily addressed. The limitation of developing this model as a parameter-based method, which utilizes a simplex to determine the optimum model, was noted.

In Chapter 4, the MLPCA model was restated using a PCA and singular value decomposition (SVD) framework and its theoretical foundations were explored. It was shown, using simulated and experimental data, that the model estimated by MLPCA is statistically superior to the model produced by PCA for the same data. It was also demonstrated that, for the first time, correlated measurement errors and intercepts may be accommodated in a PCA-like decomposition.

In Chapter 5, MLPCA was applied to the problem of multivariate calibration. Two new methods were introduced, maximum likelihood principal component regression (MLPCR) and maximum likelihood latent root regression (MLLRR), and it was demonstrated, through both simulated and experimental data, that these methods provide superior calibration models over their conventional analogues for the cases considered. It was shown that this improvement arises from better estimation of the subspace used in the calibration. It was also seen, for the cases considered, that there is little difference in model performance when estimates of the standard deviation are used instead of the true standard deviations. Finally, maximum likelihood calibration provides a unifying framework for which to better understand the relationship between calibration model and chemical data.

In Chapter 6, the MLPCA method that was developed in Chapters 3 and 4 was applied to the problems of incomplete data and calibration transfer. It was demonstrated that even when many of the data are missing, the eigenvectors obtained using MLPCA retain most of the original information. This has important implications for exploratory data analysis and modeling. By projecting the errors, it is possible to obtain an estimate of the uncertainty associated with the predicted values corresponding to the missing points. It was also shown that when calibration transfer is regarded as a missing data problem comparable results to those obtained by conventional means may be obtained.

7.2 FUTURE WORK

The work presented here has demonstrated the implications of accommodating measurement errors in the modeling process. While advantages of such maximum likelihood techniques have been shown for a number of situations, more complete studies need to be conducted and many other potential applications can be envisioned. For example, the natural progression of this work would be the development of a maximum likelihood technique for the analysis of third- and higher-order data. In general, any technique which relies on an accurate estimate of a subspace of a data set, such as target testing, should benefit from the application of MLPCA.

It was shown that the maximum likelihood analogues of principal component regression and latent root regression can dramatically improve the predictive ability of these techniques. The role of these new techniques should be examined further, under different measurement conditions, so that guidelines may be drawn regarding their use. Up until now, the issue of correlated errors in spectroscopic data has been acknowledged, but generally ignored. Therefore, the effect of error covariance on calibration should be explored. Additional methods commonly used in multivariate calibration, such as partial least squares, should be considered for potential improvement by a maximum likelihood approach. While it was shown that maximum likelihood methods have the potential to simplify the issue of calibration transfer, this should be studied more rigorously. In particular, the questions regarding the choice of transfer samples, the number of transfer samples and the rank of the transfer model should be considered.

One of the most significant contributions of this work has been the development of a reliable and efficient alternating least squares algorithm for performing the MLPCA decomposition. However, MLPCA is still significantly slower than techniques such as SVD and this will hamper its potential for routine use in the analysis of complex chemical data sets. Therefore, additional work is required to improve the convergence time of the algorithm. Furthermore, practical issues concerning the inversion of large covariance matrices have yet to be addressed. Finally, no efficient algorithm has yet been developed for the inclusion of offsets (intercept terms) in the MLPCA model. When these problems are solved, the maximum likelihood approach to multivariate analysis should instigate much more research in the fields of chemometrics and analytical chemistry.

APPENDIX A

Derivation of maximum likelihood prediction equation

Given a vector $\mathbf{x}$ of length m of observed measurements and the corresponding error covariance matrix, Ψ, the multivariate probability density function at $\mathbf{x}$ is given by:

$$L = \frac{1}{(2\pi)^{m/2}|\Psi|^{1/2}}\, exp\left[-\tfrac{1}{2}(\mathbf{x}-\mathbf{x}^{\circ})^{\mathrm{T}}\Psi^{-1}(\mathbf{x}-\mathbf{x}^{\circ})\right] \qquad \text{(A-1)}$$

where $\mathbf{x}^{\circ}$ represents the (unknown) vector of true values. The vector of maximum likelihood estimates of $\mathbf{x}^{\circ}$ for a given $\hat{\mathbf{A}}$, designated as $\hat{\mathbf{x}}$, is obtained by maximizing the probability density function subject to the parametric model $\hat{\mathbf{x}} = \hat{\mathbf{A}}\hat{\mathbf{x}}_{\mathrm{p}}$. This corresponds to minimizing the function:

$$l = (\mathbf{x}-\hat{\mathbf{x}})^{\mathrm{T}}\Psi^{-1}(\mathbf{x}-\hat{\mathbf{x}}) \qquad \text{(A-2)}$$

with respect to $\hat{\mathbf{x}}$. Substitution of the parametric model gives:

$$\begin{aligned} l &= \left(\mathbf{x}-\hat{\mathbf{A}}\hat{\mathbf{x}}_{\mathrm{p}}\right)^{\mathrm{T}}\Psi^{-1}\left(\mathbf{x}-\hat{\mathbf{A}}\hat{\mathbf{x}}_{\mathrm{p}}\right) \\ &= \mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{x}-\mathbf{x}^{\mathrm{T}}\Psi^{-1}\hat{\mathbf{A}}\hat{\mathbf{x}}_{\mathrm{p}}-\hat{\mathbf{x}}_{\mathrm{p}}^{\mathrm{T}}\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\mathbf{x}+\hat{\mathbf{x}}_{\mathrm{p}}^{\mathrm{T}}\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\hat{\mathbf{A}}\hat{\mathbf{x}}_{\mathrm{p}} \end{aligned} \qquad \text{(A-3)}$$

Using standard relations for derivatives of vectors [84], this gives:

$$\frac{\partial l}{\partial \hat{\mathbf{x}}_{\mathrm{p}}} = 0-\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\mathbf{x}-\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\mathbf{x}+2\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\hat{\mathbf{A}}\hat{\mathbf{x}}_{\mathrm{p}} \qquad \text{(A-4)}$$

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

$$\hat{\mathbf{x}}_{\mathrm{p}} = \left(\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{A}}^{\mathrm{T}}\Psi^{-1}\mathbf{x} \qquad \text{(A-5)}$$

which is the same as Equation 4.7 and leads directly to Equation 4.8.

APPENDIX B
Derivatives of S^2

The calculation of the derivative of S^2 with respect to the rotation angles (in the absence of intercept terms) begins by finding the differential of S^2:

$$\begin{aligned} \mathrm{d}S^2 &= \sum_{j=1}^{n} \mathrm{d}\left(\Delta\mathbf{x}_j^{\mathrm{T}}\Psi_j^{-1}\Delta\mathbf{x}_j\right) \\ &= \sum \left[\mathrm{d}\left(\Delta\mathbf{x}_j^{\mathrm{T}}\right)\Psi_j^{-1}\Delta\mathbf{x}_j + \Delta\mathbf{x}_j^{\mathrm{T}}\Psi_j^{-1}\mathrm{d}\left(\Delta\mathbf{x}_j\right)\right] \\ &= 2\sum \Delta\mathbf{x}_j^{\mathrm{T}}\Psi_j^{-1}\mathrm{d}\left(\Delta\mathbf{x}_j\right) \end{aligned} \qquad \text{(B-1)}$$

It will be assumed that the that the symmetric error covariance matrix, Ψ, is known for each $\mathbf{x}$. For convenience, from this point on the subscript "j" in Equation B-1 will be dropped but will be implied. Also, for simplicity, we will use $\mathbf{U}$ in place of $\hat{\mathbf{U}}$. We can write $\Delta\mathbf{x}$ as:

$$\begin{aligned} \Delta\mathbf{x} &= \left(\mathbf{I} - \mathbf{U}\left(\mathbf{U}^{\mathrm{T}}\Psi^{-1}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}\Psi^{-1}\right)\mathbf{x} \\ &= \left(\mathbf{I} - \mathbf{T}\mathbf{U}_0\left((\mathbf{T}\mathbf{U}_0)^{\mathrm{T}}\Psi^{-1}\mathbf{T}\mathbf{U}_0\right)^{-1}(\mathbf{T}\mathbf{U}_0)^{\mathrm{T}}\Psi^{-1}\right)\mathbf{x} \\ &= \left(\mathbf{I} - \mathbf{T}\mathbf{U}_0\left(\mathbf{U}_0^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\Psi^{-1}\mathbf{T}\mathbf{U}_0\right)^{-1}\mathbf{U}_0^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\Psi^{-1}\right)\mathbf{x} \end{aligned} \qquad \text{(B-2)}$$

This gives:

$$\begin{aligned} \Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathrm{d}(\Delta\mathbf{x}) = -\Big[&\Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}(\mathrm{d}\mathbf{T})\mathbf{U}_0\left(\mathbf{U}^{\mathrm{T}}\Psi^{-1}\mathbf{U}^{\mathrm{T}}\right)^{-1}\mathbf{U}^{\mathrm{T}}\Psi^{-1}\mathbf{x} \\ &+ \Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{U}\left(\mathrm{d}\left(\mathbf{U}_0^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\Psi^{-1}\mathbf{T}\mathbf{U}_0\right)^{-1}\right)\mathbf{U}^{\mathrm{T}}\Psi^{-1}\mathbf{x} \\ &+ \Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{U}\left(\mathbf{U}^{\mathrm{T}}\Psi^{-1}\mathbf{U}^{\mathrm{T}}\right)^{-1}\mathbf{U}_0^{\mathrm{T}}\left(\mathrm{d}\,\mathbf{T}^{\mathrm{T}}\right)\Psi^{-1}\mathbf{x}\Big] \end{aligned} \qquad \text{(B-3)}$$

If we make the substitution $\mathbf{H} = \left(\mathbf{U}^{\mathrm{T}}\Psi^{-1}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}\Psi^{-1}$, eqn B-3 can be further expanded by conventional means [55] to give:

$$\begin{aligned} \Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathrm{d}(\Delta\mathbf{x}) = -\big[&\Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}(\mathrm{d}\mathbf{T})\mathbf{U}_0\mathbf{H}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{U}\mathbf{H}(\mathrm{d}\mathbf{T})\mathbf{U}_0\mathbf{H}\,\Delta\mathbf{x} \\ &- \Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{U}\mathbf{H}(\mathrm{d}\mathbf{T})\mathbf{U}_0\mathbf{H}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\Psi^{-1}(\mathrm{d}\mathbf{T})\mathbf{U}_0\mathbf{H}\,\Delta\mathbf{x}\big] \end{aligned} \qquad \text{(B-4)}$$

The differential of the transformation matrix can be expanded as:

$$dT = (dT_1)T_2T_3\ldots T_{m-1} + T_1(dT_2)T_3\ldots T_{m-1}+\ldots \tag{B-5}$$

where,

$$dT_i = \frac{dT_i}{d\alpha_i}d\alpha_i = J_i d\alpha_i \tag{B-6}$$

Here J_i is the derivative of the rotation matrix corresponding to α_i. For example,

$$J_1 = \begin{bmatrix} -\sin\alpha_1 & -\cos\alpha_1 & 0 & \cdots & 0 \\ \cos\alpha_1 & -\sin\alpha_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad J_2 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & -\sin\alpha_2 & -\cos\alpha_2 & \cdots & 0 \\ 0 & \cos\alpha_2 & -\sin\alpha_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \text{ etc.} \tag{B-7}$$

Furthermore, it is easily shown that:

$$J_i = L_i T_i \tag{B-8}$$

where,

$$L_1 = \begin{bmatrix} 0 & -1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \text{ etc.} \tag{B-9}$$

From this we can write:

$$\begin{aligned} dT &= J_1(T_2T_3\ldots)d\alpha_1 + T_1J_2(T_3T_4\ldots)d\alpha_2+\ldots \\ &= L_1T_1(T_2T_3\ldots)d\alpha_1 + T_1L_2T_2(T_3T_4\ldots)d\alpha_2+\ldots \\ &= L_1Td\alpha_1 + T_1L_2T_1^{-1}T_1T_2(T_3T_4\ldots)d\alpha_2+\ldots \\ &= L_1Td\alpha_1 + \left(T_1L_2T_1^T\right)Td\alpha_2+\ldots \\ &= G_1Td\alpha_1 + G_2Td\alpha_2+\ldots+G_{m-1}Td\alpha_{m-1} \end{aligned} \tag{B-10}$$

where,

$$G_i = (T_1T_2\ldots T_{i-1})L_i\left(T_{i-1}^T T_{i-2}^T\ldots T_1\right) \tag{B-11}$$

Substitution of eqn B-10 into eqn B-4 and recognizing that $\mathbf{U} = \mathbf{T}\mathbf{U}_0$ and $\mathbf{P} = \mathbf{U}\mathbf{H}$ gives:

$$\begin{aligned}\Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathrm{d}(\Delta\mathbf{x}) = -\sum \Big[&\Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{G}_{\mathrm{i}}\mathbf{P}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{P}\mathbf{G}_{\mathrm{i}}\mathbf{P}\Delta\mathbf{x} \\ &- \Delta\mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{P}\mathbf{G}_{\mathrm{i}}\mathbf{P}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\Psi^{-1}\mathbf{G}_{\mathrm{i}}\mathbf{P}\Delta\mathbf{x}\Big]\,\mathrm{d}\alpha_{\mathrm{i}}\end{aligned} \qquad \text{(B-12)}$$

This result leads to Equation 4.18 in the paper. Although this equation is correct and faster than numerical evaluation of the derivatives, it is somewhat cumbersome. Each of the $m-1$ parameters to be optimized requires the calculation and storage of an mxm matrix $\mathbf{G}_i$, which is the product of $2i-1$ matrices that are mxm. Even though these matrices are sparse, the calculations are still time consuming and awkward. Some simplification of eqn B-12 is possible by examining the characteristics of **G**. The matrix $\mathbf{G}_i$ is antisymmetric with zero elements everywhere except for the first i elements of column i+1 and row i+1. If the rotation angles are small, the rotation matrices approach the identity matrix and a good approximation is:

$$\mathbf{G}_{\mathrm{i}} \approx \mathbf{L}_{\mathrm{i}} \qquad \text{(B-13)}$$

We have found that this approximation works well in practice, particularly if $\mathbf{U}_0$ is updated as convergence is approached so that the angles remain small.

APPENDIX C

Listing of MatLab Code for MLPCA (No Intercept, No Covariance)

```
function [U,S,V,SOBJ,ErrFlag] = mlpca(X,stdX,p);
%
%                 MLPCA.M   v. 4.0
%
% This function performs maximum likelihood principal components analysis with missing data.  The
% variables passed to the function are:
%
%   X    is the mxn matrix of observations (measurements).
%   stdX is the mxn matrix of standard deviations associated with the observations in X.  For
%        missing measurements, stdX should be set to zero.
%   p    is the dimensionality of the model (p<n, p<m).
%
% The parameters returned are:
%
%   U,S,V   are pseudo-svd parameters (mxp, pxp, and nxp).  The maximum likelihood estimates are
%           given by:  XML=U*S*V'
%   SOBJ    is the value of the objective function for the best model.
%   ErrFlag indicates the termination conditions of the function;
%             0 = normal termination (convergence)
%             1 = maximum number of iterations exceeded
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Initialization
%
convlim=1e-10;             % convergence limit
maxiter=50000;             % maximum no. of iterations
XX=X;                      % XX is used for calculations
varX=(stdX.^2);            % convert s.d.'s to variances
[i,j] = find(varX==0);     % find zero errors and convert to large
errmax = max(max(varX));   % errors for missing data
for k=1:length(i);
   varX(i(k),j(k)) = 1e+10*errmax;
end
n=length(XX(1,:));         % the number of columns
%
% Generate initial estimates from covariance matrix assuming homoscedastic errors.
%
for i=1:length(X(:,1))
  for j=1:length(X(:,1))
    CV(i,j)=(X(i,:)*X(j,:)')/min([nnz(X(i,:)) nnz(X(j,:))]);
  end
end
[U,S,V]=svd(CV,0);         % decompose adjusted matrix
U0=U(:,1:p);               % truncate solution to rank p
%
% Loop for alternating regression
%
count=0;                   % initialize loop variables
Sold=0;
ErrFlag=-1;
while ErrFlag<0;           % check for termination
   count=count+1;          % increment iteration count
%
% Evaluate objective function
%
   Sobj=0;                    % Initialize objective function
   MLX=zeros(size(XX));       % and maximum likelihood estimates
   for i=1:n                  % loop for each column
      Q=sparse(diag(varX(:,i).^(-1)));    % covariance matrix
      F=inv(U0'*Q*U0);
      MLX(:,i)=U0*(F*(U0'*(Q*XX(:,i))));  % ML projection
      dx=XX(:,i)-MLX(:,i);                % residuals
      Sobj=Sobj+dx'*Q*dx;                 % objective function
   end
%
% Check for convergence or excessive iterations
%
   if rem(count,2)==1             % Check on odd iterations only
      if (abs(Sold-Sobj)/Sobj)<convlim    % Convergence?
         ErrFlag=0;                       % Yes
      elseif count>maxiter                % Excessive iterations?
         ErrFlag=1;                       % Yes
      end
   end
[abs(Sold-Sobj)/Sobj Sobj]                % display info
%
% Now flip matrices for alternating regression
%
   if ErrFlag<0                  % Only do this part if not done
      Sold=Sobj;                 % Save most recent obj. function
      [U,S,V]=svd(MLX,0);        % Decompose ML values
      XX=XX';                    % Flip matrix
      varX=varX';                % and the variances
      n=length(XX(1,:));         % Adjust no. of columns
      U0=V(:,1:p);               % V becomes U for transpose
   end
end
%
% All done.  Clean up and go home.
%
[U,S,V]=svd(MLX,0);
U=U(:,1:p);
S=S(1:p,1:p);
V=V(:,1:p);
SOBJ=Sobj;
```

APPENDIX D

Listing of MatLab Code for MLPCA (No Intercept, Covariance)

```
function [U,S,V,SOBJ,ErrFlag] = mlcov(X,Cov,covtype,p);
%
% This function performs maximum likelihood principal component analysis with covariance in the
% errors (non-diagonal covariance matrices). The variables passed to the function are:
%
% X is the mxn matrix of observations (measurements).
% Cov is the mnxmn error covariance matrix associated with the observations in X.
% covtype indicates the type of covariance matrix passed to the function.
%    1 - means Cov is a stacked (mnxn) matrix consisting of nxn row covariances (no column
%        covariance)
%    2 - means Cov is a stacked (mnxm) matrix consisting of mxm column covariances (no row
%        covariance)
%    3 - means Cov is a full (mnxmn) covariance matrix  for vecX
% p is the rank of the model
%
% The parameters returned are:
%
% U,S,V - are pseudo-svd parameters (mxp, pxp, and nxp). The maximum likelihood estimates are given
% by :  XML=U*S*V'
% SOBJ is the value of the objection function for the best model.
% ErrFlag indicates the termination conditions of the function;
%             0 - normal termination (convergence)
%             1 - maximum number of iterations exceeded
%
% External functions: Uses the function mlsmall.m (MLPCA, No Intercept, No Covariance) to obtain
% initial estimates.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Initialization
%
convlim=1e-10;              % convergence limit
maxiter=200000;             % maximum no. of iterations
XX=X;                       % XX is used for calculations
m=length(XX(:,1));          % the number of rows
n=length(XX(1,:));          % the number of columns
mn=m*n;                     % total no. of elements
%
% Calculate the inverse of the full covariance matrix
%
% For only row or column covariance, calculate the inverse blockwise.
%
if covtype==1
   Q=spalloc(mn,mn,mn);
   for i=1:m
      indx=(i-1)*n;
      Tmp=Cov(indx+1:indx+n,1:n);
      Q(indx+1:indx+n,indx+1:indx+n)=pinv(Tmp);
   end
elseif covtype==2
   Q=spalloc(mn,mn,mn)
   for i=1:n
      indx=(i-1)*m;
      Tmp=Cov(indx+1:indx+m,1:m);
      Q(indx+1:indx+m,indx+1:indx+m)=pinv(Tmp);
   end
else
   Q=pinv(Cov);
end
%
% Now find the commutation matrix for the covariance matrix and apply to Q if row covariances were
% given.
%
```

```
ix=(1:mn)';
iy=reshape((reshape(ix,m,n))',mn,1);
K=sparse(ix,iy,1,mn,mn);
if covtype==1
   Q=K'*Q*K;
end
%
%  Generate initial estimates assuming uncorrelated errors.  First, the matrix of standard
%  deviations on X are required.
%
stdX=zeros(m,n);
if covtype==1
   for i=1:m
      indx=(i-1)*n;
      stdX(i,:)=sqrt((diag(Cov(indx+1:indx+n,1:n)))');
   end
elseif covtype==2
   for i=1:n
      indx=(i-1)*m;
      stdX(:,i)=sqrt(diag(Cov(indx+1:indx+m,1:m)));
   end
else
   stdX=sqrt(reshape(diag(Cov),m,n));
end
%
%  Generate initial guess.
%
[U,S,V,Sobj,ErrFlag]=mlsmall(XX,stdX,p);
%
% Main loop to do alternating regression for MLPCA solution
%
count=0;
Sold=0;
ErrFlag=-1;
while ErrFlag<0;
   count=count+1;
%
%  Vectorize X and create big U.  A Kronecker product is probably prettier for generating Ubig, but
%  probably not as fast.
%
   vecX=reshape(XX,mn,1);
   Ubig=spalloc(mn,n*p,mn*p);
   for i=1:n
      indx1=(i-1)*m;
      indx2=(i-1)*p;
      Ubig(indx1+1:indx1+m,indx2+1:indx2+p)=U;
   end
%
% Evaluate objective function
%
   F=pinv(full(Ubig'*Q*Ubig));
   vecMLX=Ubig*(F*(Ubig'*(Q*vecX)));
   dx=vecX-vecMLX;
   Sobj=dx'*Q*dx;
   MLX=reshape(vecMLX,m,n);
%
% Check for convergence or excessive iterations
%
   if rem(count,2)==1                  % Check on odd iterations only
      if (abs(Sold-Sobj)/abs(Sobj))<convlim  % Convergence criterion
         ErrFlag=0;
      elseif count>maxiter             % Excessive iterations?
         ErrFlag=1;
      end
   end
%
% Now flip matrices for alternating regression
%
   if ErrFlag<0                        % Only do this part if not finished
      Sold=Sobj;                       % Save most recent obj. function
      [U,S,V]=svd(MLX,0);              % Decompose ML values
      XX=XX';                       % Flip matrix
      Q=K*Q*K';
      K=K';
      m=length(XX(:,1));
      n=length(XX(1,:));           % Adjust no. of columns
      U=V(:,1:p);                  % V becomes U in for transpose
   end
end
%
% All done.  Clean up and go home.
%
[U,S,V]=svd(MLX,0);
U=U(:,1:p);
S=S(1:p,1:p);
V=V(:,1:p);
SOBJ=Sobj;
```

APPENDIX E

Listing of MatLab Code for MLPCA (Intercept, No Covariance)

```
function [U,S,V,B,SOBJ,ErrFlag] = mlint(X,stdX,p,intercep);
%
%                     MLINT.M
%
% This function performs maximum likelihood principal components analysis using models with
% intercept terms.  The variables passed to the function are:
%
%  X    is the mxn matrix of observations (measurements).
%  stdX is the mxn matrix of standard deviations associated with the observations in X.  For
%       missing measurements, stdX should be set to zero.
%  p    is the dimensionality of the model (p< n and m).
%  intercep specifies whether or not the model includes intercept terms or not:
%           intercep = 0 - no intercept terms
%           intercep = 1 - include n column intercept terms (analogous to column mean centering)
%           intercep = 2 - include m row intercept terms (analogous to row mean centering)
%           intercep = 3 - include both row and column intercept terms
%
% The parameters returned are:
%
%  U,S,V represent the maximum likelihood decomposition of X-B.  The maximum likelihood estimate of
%          X is given by:
%          U*S*V'+B
%  B     is the matrix of intercept (offset) values.  It has the same size as X.
%  SOBJ  is the value of the objective function for the best model.
%  ErrFlag indicates the termination conditions of the function;
%           0 = normal termination (convergence)
%           1 = maximum number of iterations exceeded
%
% External functions:  Uses the function objfn.m to evaluate the objective function and its
%                      derivates during opimization.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% Initialization
%
convlim=1e-10;          % convergence limit
maxiter=200000;         % maximum no. of iterations
XX=X;                   % XX is used for calculations
varX=(stdX.^2);         % convert s.d.'s to variances
[i,j] = find(varX==0);    % find zero errors and convert to large
errmax = max(max(varX));  % errors for missing data
for k=1:length(i);
   varX(i(k),j(k)) = 1e+10*errmax;
end
m=length(XX(:,1));      % the number of rows
n=length(XX(1,:));      % the number of columns
iflag=intercep;
%
% Now we have the matrix.  Generate initial estimates assuming homoscedastic errors.
%
C=zeros(n,1);
D=zeros(m,1);
if iflag==1
   pp=p+1;
   [U,S,V,Sobj,ErrFlag]=mlsmall(XX,stdX,pp);
   MLX=U*S*V';
   C=mean(MLX)';
elseif iflag==2
   pp=p+1;
   [U,S,V,Sobj,ErrFlag]=mlsmall(XX,stdX,pp);
   MLX=U*S*V';
   D=(mean(MLX'))';
elseif iflag>2
```

```
   pp=p+2;
   [U,S,V,Sobj,ErrFlag]=mlsmall(XX,stdX,pp);
   MLX=U*S*V';
   Xmu=mean(mean(MLX));
   Xcen=MLX-Xmu;
   C=mean(Xcen)';
   D=mean(Xcen')'+Xmu;
end
B=ones(m,1)*C'+D*ones(1,n);
%
Xtmp=XX-B;
Vtmp=varX;
[U,S,V]=svd(Xtmp,0);                % Decompose adjusted matrix
U0=U(:,1:p);                        % Truncate solution to rank p
count=0;
Sold=0;
ErrFlag=-1;
while ErrFlag<0;
   count=count+1;
   [Sobj,MLX]=objfn(Xtmp,Vtmp,U0,0,0);
   convtst=abs(Sold-Sobj)/Sobj;
   Sold=Sobj;
   Ssave(count)=Sobj;
   save mlint;
   if (convtst<sqrt(convlim)) & (iflag~=0)
%
% Implements variable metric minimization [58]
%
      if iflag==1
         Pest=C';
      elseif iflag==2
         Pest=D';
      else
         Pest=[C' D'];
      end
      nvar=length(Pest);
%
% implements Davidson-Fletcher-Powell minimization [58] (dfpmin)
%
      STPMX=10;
      TOLX=4*eps;
      gtol=1e-10;
      stpmax=STPMX*max([norm(Pest) nvar]);
      [fp,MLX,g]=objfn(XX,Vtmp,U0,Pest,iflag);
      hessin=eye(nvar);
      xi=-g;
      LoopFlag=-1;
      while LoopFlag<0
%
% implements lnsrch (line minimization) [58]
%
         ALF=1e-4;
         tolX=1e-10;
         sum1=norm(xi);
         if sum1>stpmax
            xi=xi*stpmax/sum1;
         end
         slope=g*xi';
         test=max(abs(xi)./max([abs(Pest); ones(1,nvar)]));
         alamin=tolX/test;
         alam=1;
         Stopflag=-1;
         count2=0;
         while Stopflag<0
            count2=count2+1;
            Pnew=Pest+alam*xi;
            fret=objfn(XX,Vtmp,U0,Pnew,iflag);
            if alam<alamin
               Pnew=Pest;
               Stopflag=1;
            elseif fret<=(fp+ALF*alam*slope)
```

```
               Stopflag=2;
            else
               if alam==1
                  tmplam=-slope/(2*(fret-fp-slope));
               else
                  rhs1=fret-fp-alam*slope;
                  rhs2=f2-fold2-alam2*slope;
                  a=((rhs1/alam^2)-(rhs2/alam2^2))/(alam-alam2);
                  b=((-alam2*rhs1/alam^2)+(alam*rhs2/alam2^2))/(alam-alam2);
                  if a==0
                     tmplam=-slope/(2*b);
                  else
                     disc=(b*b)-(3*a*slope);
                     tmplam=(-b+sqrt(disc))/(3*a);
                  end
                  tmplam=min([tmplam (0.5*alam)] );
               end
            end
            if Stopflag<0
               alam2=alam;
               f2=fret;
               fold2=fp;
               alam=max([tmplam (0.1*alam)]);
            end
         end
%
% end lnsrch
%
         fp=fret;
         xi=Pnew-Pest;
         Pest=Pnew;
         test=max(abs(xi)./max([abs(Pest); ones(1,nvar)]));
         if (test<TOLX) | (count2==1)          % Second condition is a kluge
            LoopFlag=0;                        % to prevent getting stuck
         else
            dg=g;
            [ftmp,MLX,g]=objfn(XX,Vtmp,U0,Pest,iflag);
            den=max([fret 1]);
            test=max(abs(g).*max([abs(Pest); ones(1,nvar)]))/den;
            if test<gtol
               LoopFlag=0;
            else
               dg=g-dg;
               hdg=(hessin*dg')';
               fac=dg*xi';
               fae=dg*hdg';
               sumdg=dg*dg';
               sumxi=xi*xi';
               if (fac^2)>(eps*sumdg*sumxi)
                  fac=1/fac;
                  fad=1/fae;
                  dg=fac*xi-fad*hdg;
                  hessin=hessin+fac*xi'*xi-fad*hdg'*hdg+fae*dg'*dg;
               end
               xi=xi-(hessin*g')';
            end
         end
      end
%
% end dfpmin
%
      [Sobj,MLX]=objfn(XX,Vtmp,U0,Pest,iflag);
      if iflag==1
         C=Pest';
      elseif iflag==2
         D=Pest';
      elseif iflag>2
         C=Pest(1:n)';
         D=Pest(n+1:n+m)';
      end
      B=ones(m,1)*C'+D*ones(1,n);
      MLX=MLX-B;
      convtst=abs(Sobj-Sold)/Sobj;
   end
   if (rem(count,2)==1) & (convtst<convlim)
      ErrFlag=0;
   else
      [U,S,V]=svd(MLX,0);
      XX=XX';
      B=B';
      Tmp=C;
      C=D;
      D=Tmp;
      Xtmp=XX-B;
      Vtmp=Vtmp';
      m=length(Xtmp(:,1));
      n=length(Xtmp(1,:));
      U0=V(:,1:p);
      if iflag==1
         iflag=2;
      elseif iflag==2
         iflag=1;
      end
   end
end
%
%  All done. Clean up and go home.
%
[U,S,V]=svd(MLX,0);
U=U(:,1:p);
S=S(1:p,1:p);
V=V(:,1:p);
SOBJ=Sobj;
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
function [Sobj,xml,Gobj] = objfn(xobs,varx,U0,P,iflag);
%
%        OBJFN.M
%
%  This function calculates the value of the objective function and its derivates and is used with
%  MLINT.M
%
p=length(U0(1,:));
m=length(xobs(:,1));
n=length(xobs(1,:));
%
%
% This section for no intercepts
%
if iflag==0
%
% Only the objective function and ML estimates needed
%
   Sobj=0;
   for i=1:n
      Q=sparse(diag(varx(:,i).^(-1)));
      F=inv(U0'*Q*U0);
      xml(:,i)=U0*(F*(U0'*(Q*xobs(:,i))));
      dx=xobs(:,i)-xml(:,i);
      Sobj=Sobj+dx'*Q*dx;
   end
   Gobj=0;
%
%
% This section incorporates intercept terms.
%
%
else
%
   P1=zeros(1,n);
   P2=zeros(1,m);
   if iflag==1
      P1=P;
   elseif iflag==2
      P2=P;
   else
      P1=P(1:n);
      P2=P(n+1:n+m);
   end
%
   if nargout<3
%
% Only the objective function and ML estimates needed
%
      Sobj=0;
      for i=1:n
         Q=sparse(diag(varx(:,i).^(-1)));
         A=inv(U0'*Q*U0);
         b=P2'+ones(m,1)*P1(i);
         xml(:,i)=U0*(A*(U0'*(Q*(xobs(:,i)-b))))+b;
         dx=xobs(:,i)-xml(:,i);
         Sobj=Sobj+dx'*Q*dx;
      end
      Gobj=0;
%
   else
%
% Objective function and derivative needed
%
      Sobj=0;
      Gobj1=zeros(1,n);
      Gobj2=zeros(1,m);
      for i=1:n
         Q=sparse(diag(varx(:,i).^(-1)));
         A=inv(U0'*Q*U0);
         b=P2'+ones(m,1)*P1(i);
         xml(:,i)=U0*(A*(U0'*(Q*(xobs(:,i)-b))))+b;
         dx=xobs(:,i)-xml(:,i);
         Sobj=Sobj+dx'*Q*dx;
         C=dx'*Q;
         D=C-C*U0*A*U0'*Q;
         Gobj1(i)=sum(D);
         Gobj2=Gobj2+D;
      end
      if iflag==1
         Gobj=Gobj1;
      elseif iflag==2
         Gobj=Gobj2;
      else
         Gobj(1:n)=Gobj1;
         Gobj(n+1:n+m)=Gobj2;
      end
      Gobj=-2*Gobj;
   end
end
```

REFERENCES

1. Breen, J.J.; Robinson, P.E., Ed. *Environmental Applications of Chemometrics*; ACS Symposium Series 292; American Chemical Society; Washington, 1985.

2. Sharaf, M.A.; Illman, D.L.; Kowalski, B.R. *Chemometrics*; Chemical Analysis Series 82; Wiley-Interscience; New York, 1986.

3. Deming, S.N. *Clin. Chem.* **1986**, *32*, 1702-1706.

4. Massart, D.L.; Vandeginste, B.G.M.; Deming, S.N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier; Amsterdam, 1988.

5. Haswell, S.J., Ed. *Practical Guide to Chemometrics*; Dekker; New York, 1992.

6. Meloun, M.; Militky, J.; Forina, M. *Chemometrics for Analytical Chemistry*; Ellis Horwood; New York, 1992.

7. Brown, S.D.; Bear, R.S.; Blank, T.B. *Anal. Chem.* **1992**, *64*, 22R-49R.

8. Shenk, J.S. In *Near-Infrared Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications*; Ellis Horwood; New York, 1990, p. 235-240.

9. Robinson, M.R.; Eaton, R.P; Haaland, D.H.; Koepp, G.W.; Thomas, E.V.; Stallard, B.R.; Robinson, P.L. *Clin. Chem.* **1992**, *38*, 1618-1622.

10. Nomikos, P.; MacGregor, J.F. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 97-108.

11. Andrade, J.M.; Prada, D.; Muniategui, S.; Lopez, P. *Fres. J. Anal. Chem.* **1996**, *355*, 723-725.

12. Schultz, C.P; Lui, K.Z.; Johnston, J.B; Mantsch, H.H. *Leukemia Research* **1996**, *20*, 649-655.

13. Booksh, K.S.; Kowalski, B.R. *Anal. Chem.* **1994**, *66*, 782A-791A.

14. Thomas, E.V.; Haaland, D.M. *Anal. Chem.* **1990**, *62*, 1091-1099.

15. Haaland, D.M. *Spectroscopy* **1987**, *2*, 56-57.

16. Sanchez, E.; Kowalski, B.R. *J. Chemom.* **1988**, *2*, 265-280.

17. Booksh, K.S.; Lin, Z.; Wang, Z.; Kowalski, B.R. *Anal. Chem.* **1994**, *66*, 2561-2569.

18. Lawley, D.N.; Maxwell, A.E. *Factor Analysis as a Statistical Method, 2nd ed.*; Butterworths, London, 1971.

19. Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37-52.

20. Malinowski, E.R. *Factor Analysis in Chemistry, 2nd Ed.*; Wiley; New York, 1991.

21. Aries, R.E; Lidiard, D.P.; Spragg, R.A. *Chem. in Brit.* **1991**, 821-824.

22. Meglen, R.R *J. Chemom.* **1991**, *5*, 163-179.

23. Miller, J.C.; Miller, J.N. *Statistics for Analytical Chemistry*; Ellis Horwood; Chichester, England, 1984.

24. Riu, J.; Rius, F.X. *J. Chemom.* **1995**, *9*, 343-362.

25. Orear, J. *Am. J. Phys.* **1982**, *50*, 912-916.

26. Lybanon, M. *Am. J. Phys.* **1984**, *52*, 276-278.

27. York, D. *Can. J. Phys.* **1966**, *44*, 1079-1086.

28. Williamson, J.H. *Can. J. Phys.* **1968**, *46*, 1845-1847.

29. MacDonald, J.R.; Thompson, W.J. *Am. J. Phys.* **1992**, *60*, 66-73.

30. Reed, B.C. *Am. J. Phys.* **1992**, *60*, 59-62.

31. Lybanon, M. *Am. J. Phys.* **1984**, *52*, 22-26.

32. Paatero, P.; Tapper, U. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 183-194.

33. Halvorson, H.R. *Biophys. Chem.* **1981**, *14*, 177-184.

34. Simeon, V.; Pavkovic, D. *J. Chemom.* **1992**, *6*, 257-266.

35. Cochran, R.N.; Horne, F.H. *Anal. Chem.* **1977**, *49*, 846-853.

36. Gabriel, K.R.; Zamir, S. *Technomet.* **1979**, *21*, 489-498.

37. Juntto, S.; Paatero, P. *Environmet.* **1994**, *5*, 127-144.

38. Meloun, M.; Militky, J.; Forina, M. *Chemometrics for Analytical Chemistry Volume 1: PC-Aided Statistical Data Analysis*; Ellis Horwood; New York, 1992, p.100.

39. Fuller, W.A. *Measurement Error Models*; Wiley; New York, 1987.

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

41. Kalantar, A.H.; Gelb, R.I.; Alper, J.S. *Talanta* **1995**, *42*, 597-603.

42. Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons; Chichester, 1989.

43. Moens, P.; De Volder, P.; Hoogewijs, R.; Callens, F.; Verbeeck, R. *J. Magn. Reson. Ser. A* **1993**, *101*, 1-15.

44. Persoone, P.; De Gryse, R.; De Volder, P. *J. Electron. Spectrosc. Relat. Phenom.* **1995**, *71*, 225-232.

45. Jöreskog, K.G.; Wold, H. *Systems Under Indirect Observation, Part I*; North-Holland Publishing Co.; Amsterdam, 1982, Ch. 4.

46. Bollen, K.A. *Structural Equations with Latent Variables*; Wiley; New York, 1989.

47. Thomas, E.V. *Technomet.* **1991**, *33*, 405-413.

48. Paatero, P.; Tapper, U. *Environmet.* **1994**, *5*, 111-126.

49. Nelder, J.A.; Mead, R. *Comput. J.* **1965**, *7*, 308-313.

50. Deming, S.N.; Palasota, J.A.; Nocerino, J.M. *J. Chemom.* **1993**, *7*, 393-425.

51. Pell, R.J.; Seasholtz, M.B.; Kowalski, B.R. *J. Chemom.* **1992**, *6*, 57-62.

52. Bartlett, M.S. *Br. J. Psychol.* **1937**, *28*, 97-104.

53. Bartlett, M.S. *Nature* **1938**, *141*, 609-610.

54. Gleser, L.J.; Yang, H. *Anal. Chim. Acta* **1993**, *277*, 405-419.

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

56. Mandel, J. *Technomet.* **1971**, *13*, 1-18.

57. Dean, T.; Kowalski, B.R.; Pell, R. *Appl. Spec.* submitted for publication.

58. Press, W.H.; Teukolsky, S.A; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes in FORTRAN, 2nd ed.*; Cambridge University Press; New York, 1992, p. 215.

59. Thomas, E.V. *Anal. Chem.* **1994,** *66,* 795A-804A.

60. Kowalski, B.R.; Seasholtz, M.B. *J. Chemom.* **1991,** *5,* 129-125.

61. Beebe, K.R.; Kowalski, B.R. *Anal. Chem.* **1987,** *59,* 1007A-1017A.

62. Larsen, R.J.; Marx, M.L. *An Introduction to Mathematical Statistics and Its Applications, 2nd ed.*; Prentice-Hall; Englewood Cliffs, NJ, 1986.

63. Ingle, J.D.; Crouch, S.R. *Spectrochemical Analysis*; Prentice-Hall; Englewood Cliffs, NJ, 1988.

64. Montgomery, D.C.; Peck, E.A. *Introduction to Linear Regression Analysis*; Wiley; New York, 1982, p. 339.

65. Sanchez, E.; Kowalski, B.R. *J. Chemom.* **1988,** *2,* 247-263.

66. Vigneau, E.; Bertrand, D.; Qannari, E.M. *Chemom. Intell. Lab. Syst.* **1996,** *35,* 231-238.

67. Osten, D..W; Kowalski, B.R. *Anal. Chem.* **1985,** *57,* 908-915.

68. Lorber, A.; Kowalski, B.R. *J. Chemom.* **1988,** *2,* 67-80.

69. Brayden, T.H.; Poropatic, P.A.; Watanabe, J.L. *Anal. Chem.* **1988,** *60,* 1154-1158.

70. Phillips, G.R.; Harris, J.M. *Anal. Chem.* **1990,** *62,* 2351-2357.

71. Pytella, O. *Collect. Czech. Chem. Commun.* **1990,** *55,* 42-54.

72. Rannar, S.; Geladi, P.; Lindgren, F.; Wold, S. *J. Chemom.* **1995,** *9,* 459-470.

73. De Ligny, C.L.; Nieuwdorp, G.H.E.; Brederode, W.K.; Hammers, W.E.; van Houwelingen, J.C. *Technomet.* **1981,** *23,* 91-95.

74. Nelson, P.R.C.; Taylor, P.A.; MacGregor, J.F. *Chemom. Intell. Lab. Syst.* **1996,** *35,* 45-65.

75. Wise, B. *PLS_Toolbox for use with MATLAB, Version 1.5.1*; Eigenvector Technologies; Manson, WA., 1996.

76. Kowalski, B.R.; Schatzki, T.F.; Stross, F.H. *Anal. Chem.* **1972**, *44*, 2176-2180.

77. De Ligny, C.L.; Spanjer, M.C.; van Houwelingen, J.C.; Weesie, H.M. *J. Chromatogr.* **1984**, *301*, 311-324.

78. Snyder, L.R. *Principles of Adsorption Chromatography*; Marcel Dekker; New York, 1968.

79. Blank, T.B.; Sum, S.T.; Brown,-S.D.; Monfre, S.L. *Anal. Chem.* **1996**, *68*, 2987-2995.

80. Wang, Y.; Veltkamp, D.J.; Kowalski, B.R. *Anal. Chem.* **1991**, *63*, 2750-2756.

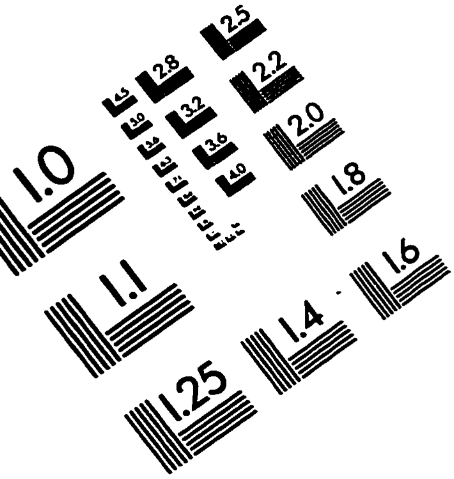
81. Forina, M.; Drava, G.; Armanino, C.; Boggia, R.; Lanteri, S.; Leardi, R.; Corti, P.; Conti, P.; Giangiacomo, R.; Galliena, C.;Bigoni, R.; Quartari, I.; Serra, C.; Ferri, D.; Leoni, O.; Lazzeri, L. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 189-203.

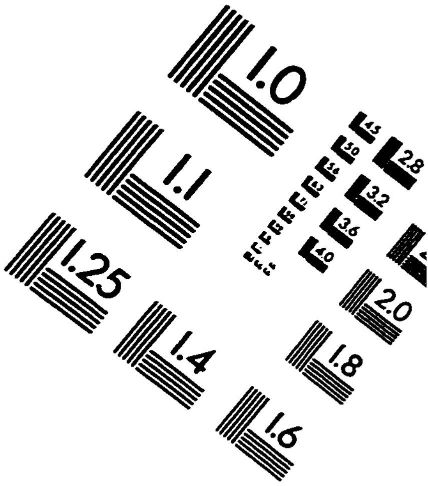
82. Bouveresse, E.; Hartmann, C.; Massart, D.L.; Last, I.R.; Prebble, K.A. *Anal. Chem.* **1996**, *68*, 982-990.

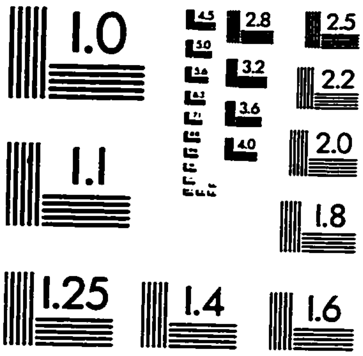
83. Kennard, R.W.; Stone, L.A. *Technomet.* **1969**, *11*, 137-148.

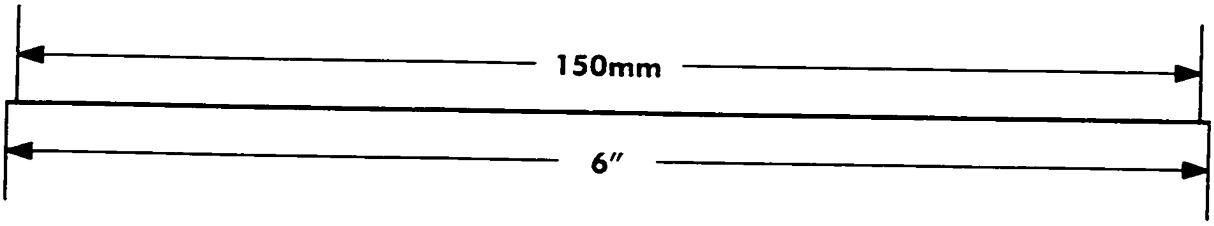
84. Beyer, W.H., ed. *CRC Handbook of Tables for Probability and Statistics*; Chemical Rubber Co.; Cleveland, Ohio, 1966.

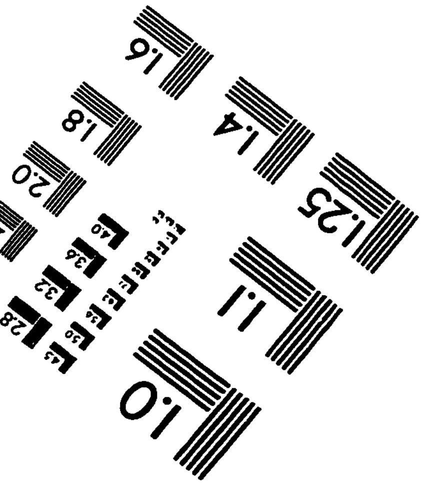
IMAGE EVALUATION TEST TARGET (QA-3)

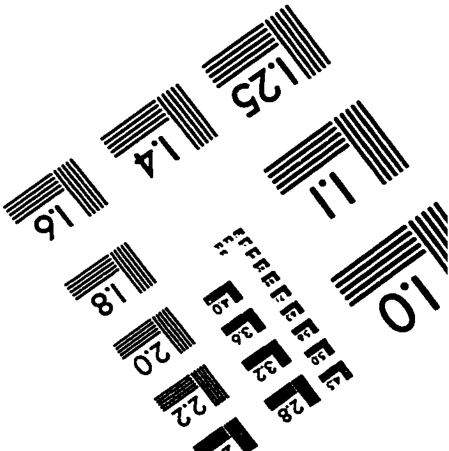












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989