EVALUATION OF EMPIRICAL STRATEGIES TO TREAT CORRELATED AND HETEROSCEDASTIC NOISE IN MULTIVARIATE CHEMICAL MEASUREMENTS

by

Cannon Giglio

Submitted in partial fulfillment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
June 2019

Contents

List of Tables

List of Figures

Abstract

The analysis of multivariate chemical data is often complicated by the presence of errors which are correlated or have non-uniform variance (heteroscedastic). Of the numerous methods used to address these issues, two promising techniques, weighted scatter correction (WSC) methods and principal axis factoring (PAF), are considered in this work.

In near-infrared (NIR) spectroscopy, multiplicative scatter noise occurs due to pathlength changes in samples. This type of noise can obscure chemical information and preprocessing through the use of multiplicative scatter correction (MSC) or standard normal variate (SNV) are routinely applied to mitigate scatter noise. Recently, WSC methods have been proposed as an improvement to MSC and SNV. These methods use regions of a spectrum where the chemical variation is low relative to the scatter variation to estimate the scatter coefficients, ideally resulting in better noise removal.

For many datasets, heteroscedastic noise can be problematic for chemometric tools that model chemical variance, such as principal components analysis (PCA). PAF is an alternative to PCA that has been widely used in the social sciences, but has rarely been applied to the analysis of chemical data. PAF is a decomposition method which is ideally suited for data in which the variables exhibit different measurement uncertainties. PAF tries to simultaneously model the data in a reduced space while also estimating the measurement error variance.

This work critically examines the use of WSC methods and PAF through application to simulated and experimental datasets. It is demonstrated, for multiplicative scatter noise, WSC methods resulted in lower prediction errors than MSC and SNV when the chemical background signal is low and the main chemical analyte signal is large, but that even a modest amount of chemical background variation can be detrimental. In the study of PAF as an alternative to PCA it is shown that, when the measurement errors are heteroscedastic, PAF results in improved subspace estimation and reduced errors, and provides estimates of measurement uncertainties.

List of Abbreviations and Symbols Used

List of Abbreviations

ABV	Alcohol by Volume (%)
Binary SSNV	Selective Standard Normal Variate with Binary Weights
CDR	Chemical Dominant Region
EFA	Exploratory Factor Analysis
FA	Factor Analysis
IDRC	Interference Dominant Region Correction
IE	Imbedded Error
MLFA	Maximum Likelihood Factor Analysis
MSC	Multiplicative Scatter Correction
NIR	Near-Infrared Spectroscopy
PAF	Principal Axis Factoring
PCA	Principal Components Analysis
PLS	Partial Least Squares Regression
RMSECV	Root Mean Squared Error of Cross Validation
RMSEP	Root Mean Squared Error of Prediction
RMSEV	Root Mean Squared Error of Validation
RSD	Relative Standard Deviation
SDR	Scatter Dominant Region
SMSC	Selective Multiplicative Scatter Correction
SNV	Standard Normal Variate
SSNV	Selective Standard Normal Variate
SVD	Singular Value Decomposition
VSN	Variable Selection for Normalization
XE	Extracted Error
XRF	X-ray Fluorescence Spectroscopy

List of Symbols

α	Coefficient of baseline offset (theoretical)
β	Coefficient of multiplicative offset (theoretical)
$\mathbf{b}, \hat{\mathbf{b}}$	$p \times 1$ regression vector
$\mathbf{C}$	Contribution matrix ($n \times r$)
ϵ	Tolerance threshold for VSN
$\mathbf{E}$	Residuals matrix ($n \times p$)
$\mathbf{F}$	PAF scores matrix ($n \times r$)
h	Scale factor for Gaussian peak
i	Row index
j	Column index
k	Number of regions used for IDRC
K	Maximum number of regions used for IDRC
$\mathbf{L}$	Loadings matrix ($p \times r$)
$\mathbf{\Lambda}$	PAF Loadings matrix ($p \times r$)
μ	Mean
n	Number of rows in $\mathbf{X}$ (number of samples)
$N(\mu, \sigma)$	Normal distribution
N_s	Sample pairs for VSN weighting algorithm
N_w	Inner loop iterations for VSN algorithm
$\mathbf{P}$	Profile matrix ($r \times p$)
p	Number of columns in $\mathbf{X}$ (number of variables)
$\mathbf{\Psi}^2$	Diagonal matrix of unique variances ($p \times p$)
$\mathbf{R}$	Covariance matrix of X ($p \times p$)
$\mathcal{R}$	Region of spectrum
r	Number of latent variables
σ	Standard deviation
$\mathbf{S}$	Data correlation matrix ($p \times p$)

List of Symbols (Continued)

$\mathbf{T}$	Scores matrix ($n \times r$)
$U(a, b)$	Uniform distribution in the range of a and b
$\mathbf{W}$	Diagonal weighting matrix (p x p)
$\mathbf{X}$	Data matrix ($n \times p$)
$\mathbf{X}_{MC}$	Mean-centered data matrix
$\mathbf{X}_{SC}$	Autoscaled data matrix
$\mathbf{x}_{ref}$	Reference spectrum for MSC
$\boldsymbol{\xi}$	Wavelength channel indices (1,2,. . . ,p)
$\mathbf{y}$	Property vector ($n \times 1$)

Acknowledgements

Thanks to Peter Wentzell, Steve Driscoll, Mohsen Kompany-Zareh, Jean-Michel Roger, and Maynarhs da Koven for their helpful discussions.

Chapter 1

Introduction

Like other areas of science and technology, chemistry has entered the era of "big data", with multivariate analytical measurements being the norm rather than the exception due to the capabilities of modern instrumentation. Measurement vectors are generated naturally from techniques such as spectroscopy, mass spectrometry, chromatography, electroanalytical methods, X-ray diffraction and many others. Vectors of measurements are also realized through profiling experiments (elements, fatty acids, proteins, nucleic acids, metabolites) and time series studies (kinetics, longitudinal environmental and biological studies, etc.). Often measurement vectors are combined along two dimensions (e.g. wavelength and time) to provide matrices of measurements (two-way data), or further concatenated to provide higher order data structures. Such data pervades all areas of chemical measurement including biology, food science, the environment, industry, forensics, conservation, medicine, pharmaceuticals, diagnostics, and many others.

As a consequence of this evolution, chemometric tools have become increasingly important in chemical analysis, providing access to information that would otherwise be unavailable. For example, multivariate calibration enables the routine estimation of chemical properties (e.g. protein content, octane number, concentrations, drug activity) and are widely used in industry. These tools are used to visualize data and classify data in areas such as proteomics, metabolomics, food science, forensics, archaeology, and process monitoring. Such techniques also provide a better understanding of chemical systems in biology, medicine, the environment, and are widely applied for diagnostics and imaging.

A focus of modern research in chemometrics is continuing to adopt the tools available to better deal with the complex data structures presented by multivariate measurements. One aspect of this is the measurement error structure in the data.

Like all analytical methods, multivariate measurements have errors. Unlike univariate measurements, however, the analysis of multivariate data is complicated by the relationships among the measurement channels. For a variety of reasons, the errors for different measurements may not have the same variance (uncertainty); that is, they are heteroscedastic rather than homoscedastic. Likewise, there may be a statistical relationship among the errors for different variables; that is, they are correlated rather than independent. Chemometric methods are usually optimized assuming that measurement errors are independent and identically distributed and follow a normal distribution (*iid* normal noise). Errors which violate the assumptions of *iid* noise (such as heteroscedastic and correlated errors), may result in suboptimal results when using common chemometric methods. As a consequence a variety of strategies have been developed to handle measurement errors.

For example, a common problem in near-infrared (NIR) spectroscopy is multiplicative offset noise [1, 2]. In NIR spectroscopy measured by diffuse reflectance, multiplicative offset noise is effectively a change in the optical pathlength due to light scattering, and it is difficult to distinguish a change in pathlength from a change in the concentration. This is one manifestation of correlated and heteroscedastic noise in analytical measurements. Various methods have been established for the purpose of correcting multiplicative offset noise, including the standard normal variate (SNV) method [3], and multiplicative scatter correction (MSC) [4], and these two methods have been used for over 30 years. Recent improvements have been proposed for scatter correction, including weighted correction methods which seek to correct the scatter based upon regions where the chemical signals are consistent and show little variation.

Another common problem is heteroscedastic noise, which occurs when the errors are independent, but do not have the same variance. Methods such as principal components analysis (PCA), which is widely used in chemometrics, assume that the errors are *iid* (homoscedastic). When the noise is heteroscedastic, the results of PCA will be sub-optimal. Dealing with heteroscedastic noise typically necessitates scaling of data, which is designed to make the noise in the scaled data behave in an *iid* manner. However, optimal data scaling requires estimates of measurement uncertainties, which are often not available. Factor analytic methods offer an alternative to estimate

measurement uncertainties when information about the measurement error structure is unknown or unavailable.

The purpose of this thesis is to investigate these alternative methods (weighted scatter correction and factor analytic methods) in greater detail. Chapter 1 introduces common multivariate methods and the problems created by non-*iid* measurement errors. Chapter 2 investigates recently proposed improvements to correct for multiplicative offset noise in NIR spectroscopy. Chapter 3 examines factor analysis as an alternative to principal components analysis for datasets that contain heteroscedastic errors. The findings of this work are summarized in the final chapter.

1.1 Notation

Standard notation conventions will be used. Matrices and vectors are designated as boldface symbols, with matrices capitalized and vectors in lowercase, while scalar quantities are italicized. Using standard chemometric conventions, $\mathbf{x}_i$ will denote a row vector of dimensions $1 \times p$ representing the measurements of p variables for sample i, also referred to as the measurement vector or (where appropriate) the spectrum of sample i. A collection of such measurement vectors for n samples will be represented as the $n \times p$ matrix $\mathbf{X}$. A column of $\mathbf{X}$, representing the measurements of n samples for variable j, will be designated by the $n \times 1$ column vector $\mathbf{x}_j$. The transpose of a matrix or vector will be denoted by a superscript "T" (e.g. $\mathbf{X}^{\mathrm{T}}$) and the matrix inverse by a "-1" superscript (e.g. $\mathbf{X}^{-1}$). Where appropriate, a vector of properties (e.g. component concentrations) associated with the n samples will be denoted by the $n \times 1$ vector $\mathbf{y}$.

The sample mean, sample standard deviation, and sample variance of vector $\mathbf{x}$ are denoted by $\mu\{\mathbf{x}\}$ (or $\bar{x}$), $\sigma\{\mathbf{x}\}$, and $var\{\mathbf{x}\}$, respectively. The elementwise multiplication (Hadamard product) of matrices $\mathbf{A}$ and $\mathbf{B}$ will be denoted by $\mathbf{A} \circ \mathbf{B}$, whereas the matrix multiplication of $\mathbf{A}$ and $\mathbf{B}$ is given by $\mathbf{AB}$.

Mean-centering is an operation in which the mean of each variable (column) of $\mathbf{X}$ is subtracted, such that the columns of the mean-centered data $\mathbf{X}_{MC}$ are zero.

$$\mathbf{x}_{MC,j} = \mathbf{x}_j - \mu\{\mathbf{x}_j\} \tag{1.1}$$

$\mathbf{X}_{SC}$ will designate the autoscaled version of $\mathbf{X}$, where the column means are subtracted from $\mathbf{X}$ and then the result is divided by the column standard deviations of $\mathbf{X}$:

$$\mathbf{x}_{SC,j} = \frac{\mathbf{x}_j - \mu\{\mathbf{x}_j\}}{\sigma\{\mathbf{x}_j\}} \tag{1.2}$$

The covariance matrix of $\mathbf{X}$, which has dimensions of $p \times p$ and is denoted as $\mathbf{R}$, is given in Equation 1.3.

$$\mathbf{R} = \frac{\mathbf{X}_{MC}^{\mathrm{T}}\mathbf{X}_{MC}}{n-1} \tag{1.3}$$

The notation $U(a,b)$ will be used to designate a random variable uniformly distributed in the range of a and b and $\mathbf{M} \sim U(a,b)$ indicates that $\mathbf{M}$ is a matrix of random variables drawn from this distribution. Likewise, $N(\mu = 0, \sigma = 1)$ indicates a random variable from a normal distribution with a mean of μ and a standard deviation of σ, and the elements of $\mathbf{M} \sim N(\mu = 0, \sigma = 1)$ are drawn from this distribution.

In Chapters 2 and 3, the term "Gaussian peak" is used to describe a scaled version of the normal probability density function $f\big(h, \mu, \sigma\big)$, which is described using the notation in Equation 1.4, where $\boldsymbol{\xi}$ is a vector of wavelength channel indices (1,2,...,p), μ is the mean, σ is the standard deviation, and h is a scale factor.

$$f\big(h, \mu, \sigma\big) = h\big(\frac{1}{\sigma\sqrt{2\pi}} e^{-(\boldsymbol{\xi}-\mu)^2/(2\sigma^2)}\big) \tag{1.4}$$

1.2 Multivariate Methods in Chemistry

There are many chemometric methods that are employed for the analysis of multivariate chemical data, including classification and clustering methods [5], multi-way methods [6, 7], multivariate calibration and regression [8], and many others. In this chapter, only three general classes most relevant to this work will be considered: multivariate calibration, mixture analysis, and exploratory data analysis for classification. These methods are described in the following subsections, with a a particular emphasis on the implications of error structures..

1.2.1 Multivariate Calibration

The task of multivariate calibration is to use multivariate data (e.g. spectra, chromatograms, mass spectra) to predict some property (e.g. concentration). Multivariate calibration is often based on linear prediction models, which are formulated in

terms of a regression vector. For a matrix of data $\mathbf{X}$ that contains n samples and p measured variables (such as a set of spectra) and a $n \times 1$ property vector $\mathbf{y}$, the regression model is formulated as shown in Equation 1.5, where $\mathbf{b}$ is a $p \times 1$ regression vector, and $\mathbf{e}$ is a $n \times 1$ matrix of residuals.

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1.5}$$

The goal of linear calibration methods is to determine the regression vector $\mathbf{b}$ from a set of training (calibration) data, to predict future samples. Different multivariate calibration methods are characterized by the ways in which they determine the regression vector, $\mathbf{b}$, in Equation 1.5. The four most commonly used regression methods are Classical Least Squares (CLS), Multiple Linear Regression (MLR), Principal Components Regression (PCR), and Partial Least Squares regression (PLS). CLS is only useful when all components of a mixture are known, so it is of limited utility for complex mixtures. The other three techniques are known as Inverse Least Squares (ILS) methods. The most direct solution when the number of training samples, n, is greater than or equal to the number of variables, p, is MLR, which employs the pseudoinverse of $\mathbf{X}$ to minimize the sum of squared residuals (SSR) in $\mathbf{e}$. The solution is given in Equation 1.6, where $\mathbf{X}_{cal}$ and $\mathbf{y}_{cal}$ represent the training data.

$$\hat{\mathbf{b}} = \left(\mathbf{X}_{cal}^{\mathrm{T}}\mathbf{X}_{cal}\right)^{-1}\mathbf{X}_{cal}^{\mathrm{T}}\mathbf{y}_{cal} \tag{1.6}$$

When the number of variables is larger than the number of samples ($p > n$), a common situation in chemistry, the data matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is deemed to be singular or rank-deficient, which means that the matrix inverse shown in Equation 1.6 will not have an exact solution. Even when the condition $n > p$ is met, the solution in Equation 1.6 can be unreliable due to overfitting unless the ratio of samples to variables is high. Various solutions to this problem are commonly employed, including principal components regression (PCR) and partial least squares regression (PLS) [9], and ridge regression (RR) [10]. PCR and PLS are referred to as latent variable methods, since they reduce the number of variables in $\mathbf{X}$ to a smaller number, represented by a scores matrix, $\mathbf{X}$ $(n \times r)$, where $r < n, p$. The scores are linear combinations of the original variables that retain maximal information from $\mathbf{X}$. The regression model now becomes

$$\mathbf{y} = \mathbf{T}\mathbf{d} + \mathbf{e} \tag{1.7}$$

where the least squares solution for the reduced regression vector, $\mathbf{d}$ $(r \times 1)$ is

$$\hat{\mathbf{d}} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \tag{1.8}$$

Through simple manipulation, the reduced regression vector $\hat{\mathbf{d}}$ can be transformed to give a regression vector in the original space, $\hat{\mathbf{b}}$.

The principal difference between PCR and PLS is the manner in which the scores, $\mathbf{T}$, are calculated. PCR scores are calculated using Principal Components Analysis (PCA), through a truncated Singular Value Decomposition (SVD) as described in Chapter 3. The scores in PCA are intended to maximize the representation of variance in $\mathbf{X}$ (i.e. they extract the maximum information from $\mathbf{X}$). In contrast, PLS, which is the most widely used calibration method in chemometrics, extracts the scores in order to simultaneously maximize both the variation in $\mathbf{X}$ and the covariance between $\mathbf{X}$ and $\mathbf{y}$ [11][9]. As a consequence, it is often the case that PLS requires fewer latent variables than PCR.

Both PLS and PCR require selection of the number of latent variables to be used, r. For the calibration (training) data, the estimation (fit) for $\mathbf{y}$ will generally improve with increasing r, but this also increases the risk of overfitting, which means that the regression vector will be good for fitting the calibration data but not for predicting future samples because it is using too many variables. Generally r is determined through a cross-validation procedure in which part of the calibration data is set aside (not used to calculate the regression vector) and then used to evaluate prediction ability for different values of r. Different strategies are used for model validation.

Multivariate calibration is affected by measurement errors in both $\mathbf{X}$ and $\mathbf{y}$. Measurement errors in $\mathbf{y}$ are determined by the reference method used for calibration and set a lower limit on the errors that can be quantified by calibration. Measurement errors in $\mathbf{X}$ have an impact on both the calibration and prediction stages. Measurement errors that are non-iid (heteroscedastic and/or correlated) can lead to calibration models that are suboptimal due to the ineffective extraction of latent variables and/or spurious correlations. Some of these problems will be mitigated by the calibration process itself, which tends to average out variations that are uncorrelated with $\mathbf{y}$. In the prediction step, however, any errors in the measurement vector $\mathbf{x}_i$ for an unknown

sample will be propagated through the regression model,

$$\hat{\mathbf{y}}_i = \mathbf{x}_i \hat{\mathbf{b}} \tag{1.9}$$

Multiplicative offset noise is particularly problematic in this regard because it is large, difficult to distinguish from chemical signals, and is prevalent in NIR spectroscopy, which is widely used. Therefore, this problem is the particular focus of Chapter 2.

1.2.2 Mixture Analysis

Mixture analysis is a common problem in chemical systems, and chemometrics has provided new tools to address this issue. Given a set of measurements from a series of mixtures, the goal is to determine the number of components, their response profiles (e.g. spectra), and concentrations in each mixture. Mixtures can arise in chemical equilibrium or kinetic studies [12], chromatography [13, 14], or in many other situations, such as source apportionment for environmental profiles [15]. The technique known as multivariate curve resolution (MCR) [16] is widely used to address these problems.

MCR uses linear relationships and mathematical constraints (e.g. non-negativity, unimodality, and closure) [17] in an attempt to identify mixture components. In a typical problem involving r components, MCR assumes that a matrix of measurements, $\mathbf{X}$ $(n \times p)$ consists of n mixtures measured at p variables and can be represented as the product of a contribution matrix, $\mathbf{C}$ $(n \times r)$ and a profile matrix, $\mathbf{P}$ $(r \times p)$, along with a matrix of residuals, $\mathbf{E}$ $(n \times p)$.

$$\mathbf{X} = \mathbf{CP} + \mathbf{E} \tag{1.10}$$

This is referred to as a bilinear relationship, where typically $\mathbf{C}$ represents the concentrations of the r components in the n mixtures and $\mathbf{P}$ corresponds to the pure component spectra for the r components. The columns of $\mathbf{X}$ define an r-dimensional subspace within the n-dimensional row space of $\mathbf{X}$, and the rows of p define an r-dimensional subspace within the column space of $\mathbf{X}$.

A common implementation of the MCR algorithm uses a matrix of data reconstructed by principal components analysis (PCA) as an initial starting point, rather than using the raw experimental data matrix [18]. PCA decomposes the original data

matrix into a bilinear product of a scores matrix $\mathbf{T}$ $(n \times r)$ and a loadings matrix $\mathbf{L}$ $(r \times p)$ such that

$$\mathbf{X} = \mathbf{TL} + \mathbf{E} \tag{1.11}$$

While $\mathbf{T}$ and $\mathbf{L}$ are not the same as $\mathbf{C}$ and $\mathbf{P}$, they represent optimal estimates of the subspaces under conditions of *iid* normal measurement errors. PCA is used to project the data from the original space into a subspace of reduced dimensionality. The performance of the MCR algorithm is dependent on how closely the PCA subspace is to the "true" subspace of the data that contains the chemical information about the mixture components. Inaccurate subspace estimation will decrease the reliability of the component estimates, which will harm the performance of the MCR algorithm.

The presence of heteroscedastic noise can have a negative impact on the PCA subspace estimation [19]. When the noise is *iid* normal, then PCA will provide optimal estimates of the subspace, but the PCA subspace estimation will be suboptimal if the noise is not *iid* normal. This problem has long been recognized in MCR, and a variety of solutions have been proposed to optimize subspace estimation in the presence of heteroscedastic errors. These include positive matrix factorization (PMF) [20], multivariate curve resolution with weighted alternating least squares (MCR-WALS) [21], and maximum likelihood principal components analysis (MLPCA) [22]. However, each of these methods requires reliable estimates of the measurement uncertainty to be effective, and often such information is unknown.

In Chapter 3, principal axis factoring (PAF) is explored as an alternative to PCA to simultaneously estimate the subspace of the data and the magnitude of the measurement error variance. Although direct applications of MCR methods are not included in the study, the ability of PCA and PAF to accurately estimate the subspace of the relevant data is explored.

1.2.3 Exploratory Data Analysis

The term exploratory data analysis encompasses any chemometric methodology which is used to examine the characteristics of multivariate data through various reduction strategies, but in this work, the emphasis is on data visualization through linear projection. In particular, PCA has become a data visualization tool that is used extensively for multivariate data, especially in fields such as metabolomics. Typically,

the $n \times p$ data matrix $\mathbf{X}$ is decomposed into scores ($\mathbf{T}$) and loadings ($\mathbf{L}$), and the scores for the first two or three components are plotted in a space of corresponding dimensionality to visualize the relationships among the n objects (samples). Often the goal is to identify clusters (groups of objects belonging to the same class) through their spatial positions in the scores plot.

As with MCR, the results obtained from a PCA projection can be strongly influenced by the presence of heteroscedastic errors. The spatial projection of objects is influenced by measurement noise in two ways: (1) estimation of the subspace, and (2) projection of objects into the subspace. Visualization can often be improved with optimal scaling but, as for MCR, this requires prior knowledge of measurement error variance, which may be unavailable or difficult to measure. In Chapter 3, it is shown that PAF can improve the estimation of the subspace while also providing an estimate of measurement uncertainty. Moreover, the use of a maximum likelihood projection (as opposed to the orthogonal projection used for PCA) improves the projection of objects into the subspace.

1.3 Error Structures in Multivariate Analysis

Multivariate analytical data can exhibit many error structures that deviate from *iid* normal [23]. Measurement errors can exhibit heteroscedasticity, meaning that different measurements show different (non-uniform) error variance. Errors can also be correlated, meaning that the error from one variable (measurement channel) has a statistical correlation with the errors from other variables. If $\mathbf{x}$ is a measurement vector and $\mathbf{x}^0$ represents the error-free measurements, then the measurement errors $\mathbf{e}$ are characterized by the error covariance structure, $\mathbf{\Sigma}_e$, of dimensions $p \times p$, as shown in Equation 1.12.

$$\mathbf{\Sigma}_e = E\Big(\big(\mathbf{x} - \mathbf{x}^0\big)^{\mathrm{T}}\big(\mathbf{x} - \mathbf{x}^0\big)\Big) = E\big(\mathbf{e}^{\mathrm{T}}\mathbf{e}\big) \tag{1.12}$$

In Equation 1.12, the operator $E\big(\big)$ represents the expectation value. The structure of the error covariance matrix is often related to the nature of the analytical measurement, and its characteristics are ultimately related to the analytical methodology, sample characteristics, and instrumentation. In the following sections, two commonly encountered cases, multiplicative offset noise and heteroscedastic errors, are examined in more detail, along with the methods commonly used to treat them.

1.4 Correlated Noise in Near-Infrared Spectroscopy

Near-infrared (NIR)spectroscopy is one of the most widely used analytical techniques for multivariate calibration and other applications. It is commonly used for quantitative analysis across a broad range of application areas to displace more tedious and time consuming reference methods (e.g. Kjeldahl titrations). One advantage is that it can be used to estimate not only concentration (e.g. alcohol content) but also other variables related to chemical composition (e.g. octane number). It is widely used in the food industry to estimate product quality parameters (e.g. protein, moisture and fat content [24]), in the pharmaceutical industry to quickly assess active ingredients and other parameters [25], in the petroleum industry to assess octane number, cetane number and other empirical parameters, and in other industries worldwide for quality control and process monitoring.

NIR spectroscopy measures the absorption or diffuse reflectance of a sample irradiated by light in the NIR region. The NIR region is located between 780-2500 nm, with NIR absorbance arising from broad overtone and combination bands of molecular vibrations [26]. The bonds which are most active in the NIR include O-H, N-H, C-H, and C=O. The near infrared region is not very useful for chemical interpretation, since it consists of broad combination and overtone bands and is quite complex. However, it has several advantages that make it well-suited for quantitative analysis, including simple and inexpensive instrumentation. It is often used in diffuse reflectance mode, which allows it to be employed in a passive, non-invasive way for solid and liquid samples that are non-transparent.

NIR also has the advantage that it is non-specific in its response, making it applicable to a wide variety of sample matrices, but because of the large number of overlapping bands, quantitation using NIR is only possible using multivariate calibration. In a typical application, the NIR spectrum is recorded for a set of calibration samples for which a property of interest (*e.g.* concentration, protein content, octane number) is measured by a reference method (*e.g.* Kjeldahl titration for protein content). A calibration model is then built using the spectra ($\mathbf{X}$) and reference data $\mathbf{y}$. Nominally, NIR spectra are characterized by a high signal-to-noise ratio (S/N) when one considers only the independent component of the measurement errors, but correlated errors represent a serious limitation to the precision of calibration models.

1.4.1 Multiplicative Offset Noise

The particular types of correlated noise observed in NIR spectroscopy are called multiplicative offset noise and baseline offset noise. Multiplicative offset noise is sometimes referred to as multiplicative noise, but this does not make a clear distinction from independent proportional noise, which is generally not a problem in NIR spectroscopy. Multiplicative offset noise can also occur for other types of measurements (e.g. IR spectroscopy), but is a dominant source of variation in NIR spectroscopy. NIR reflectance of solid samples is dependent upon the particle shape, size, and chemical composition [1]. Unwanted variation due to light scattering can occur, as a result of scattering at the surface of particles, and due to changes in spectral pathlength through the sample [3]. The variation in the scattering can often occur due to physical effects, such as variation in sample thickness or particle size [2]. Because, by Beer's law, a change in pathlength is indistinguishable from a change in concentration, this can lead to errors in prediction of the property of interest.

A model for multiplicative offset and baseline offset noise is given in Equation 1.13:

$$\mathbf{x}_i = \alpha_i \mathbf{1} + \beta_i \mathbf{x}_{i,chem} + \mathbf{x}_{i,chem} + \epsilon_i \tag{1.13}$$

where $\mathbf{x}_i$ is the spectrum of the i^{th} sample (i=$\{1, 2, ..., n\}$), α_i is the coefficient of the baseline offset of the i^{th} sample, β_i is the coefficient of the multiplicative offset of the i^{th} sample, $\mathbf{x}_{i,chem}$ is the pure chemical response excluding scatter effects, and $\boldsymbol{\epsilon}_i$ is a vector of *iid* normal errors. In Figure 1.1, examples of baseline offset noise and multiplicative offset noise added to a Gaussian peak are shown.

Multiplicative offset errors are a major complication to analysis and quantification by NIR using multivariate tools. Because of this, a variety of solutions have been proposed, with the most widely used (by far) being standard normal variate (SNV) [3] and multiplicative scatter correction (MSC) [4], which are described in the following sections.

1.4.2 Standard Normal Variate (SNV)

The fundamental underlying assumption of the SNV method is that the variation among samples due to chemical sources is quite small relative to the multiplicative

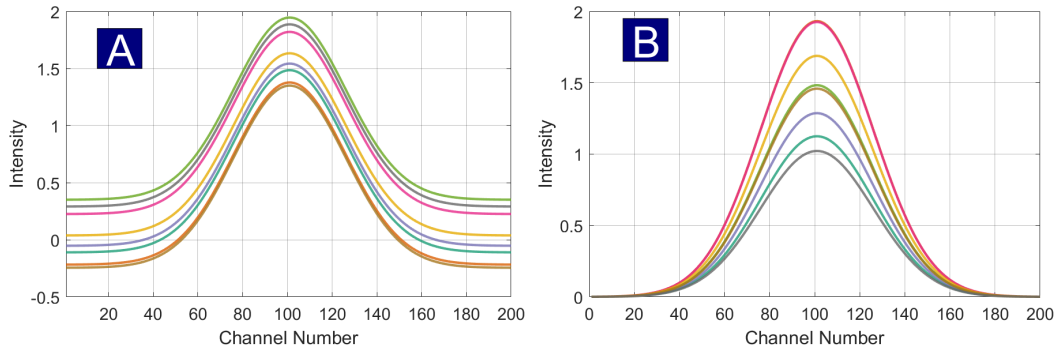

Figure 1.1: Example of baseline offset noise and multiplicative offset noise, added to a Gaussian peak. (A) Gaussian peak with baseline offset noise added; (B) Gaussian peak with multiplicative offset noise added;

offset effect. Taken to the extreme, if there were no chemical differences among spectra and the noise conforms to the assumptions, each spectrum could be scaled to make the SNV-corrected spectra identical, but of course for calibration no predictions could be obtained if all of the spectra contain the same chemical information. SNV assumes that, in the absence of multiplicative and baseline effects, each spectrum will have the same "average magnitude", i.e. the variations among spectra due to scatter effects are much larger than those arising from chemical differences. On this basis, the standard normal variate (SNV) correction subtracts the mean of each spectrum, and divides the result by the standard deviation of each spectrum, as shown in Equation 1.14:

$$\mathbf{x}_{i,SNV} = \frac{\mathbf{x}_i - \mu\{\mathbf{x}_i\}}{\sigma\{\mathbf{x}_i\}} \tag{1.14}$$

where $\mathbf{x}_{i,SNV}$ is the correction for the spectrum of sample i ($\mathbf{x}_i$). When scaled, each SNV-corrected spectrum will have a mean of zero and a variance of unity.

The effects of SNV can be examined by considering what would happen for a set of spectra that have identical chemical signals, no independent errors, and differing amounts of baseline offset and multiplicative offset noise. For this hypothetical case, the mean and standard deviation of each spectrum will be directly related to the scatter coefficients. In such a case, the SNV-treated spectra will be identical. Alternatively, for a hypothetical dataset in which there is no scattering, and the chemical response differs between samples, then the SNV transformation will distort the chemical information in the spectra.

For real chemical datasets, the chemical signals will differ among samples. SNV

is most appropriate when the variation due to scattering effects is significantly larger than the variation due to chemical effects. While SNV is widely used and can be an effective tool to correct for multiplicative offset noise, it is based on the assumption that the underlying, error-free spectra contain the same variance. Deviations from this assumption will introduce bias in the results. However, the assumption is often approximately valid in NIR applications that involve very complex mixtures where the overall composition and chemical responses are not highly variable.

1.4.3 Multiplicative Scatter Correction (MSC)

Multiplicative scatter correction (MSC) is perhaps the most widely used correction method for NIR spectra. MSC was proposed by Geladi *et al* [4]. Like SNV, MSC assumes that the multiplicative offset effects dominate the variation among spectra. Instead of normalizing each spectrum using its own variance, however, MSC normalizes each spectrum to a reference spectrum (typically chosen to be the mean spectrum) by assuming the linear relationship, as shown in Equation 1.15.

$$\mathbf{x}_{i,MSC} = \frac{\mathbf{x}_i - a_{i,MSC}}{b_{i,MSC}} \tag{1.15}$$

where the MSC coefficients $a_{i,MSC}$ and $b_{i,MSC}$ are obtained from the slope and intercept respectively of a simple linear regression of spectrum $\mathbf{x}_i$ against a reference spectrum $\mathbf{x}_{ref}$. The reference spectrum is typically the average spectrum for the training/calibration set. The model for determining the MSC coefficients is given in Equation 1.16:

$$\mathbf{x}_i = a_{i,MSC} + b_{i,MSC}\mathbf{x}_{ref} + \mathbf{e}_{i,MSC} \tag{1.16}$$

where $\mathbf{x}_{ref}$ is the reference spectrum, and $a_{i,MSC}$ and $b_{i,MSC}$ are the offset and multiplicative correction parameters specific to the sample. These parameters are determined by a regression of the elements of $\mathbf{x}_i$ (as a "$\mathbf{y}$") against the elements of $\mathbf{x}_{ref}$, which minimizes the sum of squares of the residual vector $\mathbf{e}_{i,MSC}$. The MSC-corrected values are then obtained by Equation 1.15.

By comparing equation 1.15 with equation 1.13, it is readily apparent that MSC assumes that the reference spectrum is a proxy for the chemical component of the

signal $\mathbf{x}_{i,chem}$. For this assumption to be true, the majority of the difference between spectrum $\mathbf{x}_i$ and the reference spectrum $\mathbf{x}_{ref}$ must be due to effects of the scatter, and the chemical component of the response and independent errors must be similar between $\mathbf{x}_i$ and $\mathbf{x}_{ref}$. If there is a large amount of chemical variation in $\mathbf{x}_i$, then some of that information will be removed when the MSC correction is performed.

The effect of the first step of the MSC correction (subtraction of the offset term) is to correct the spectra for a variable baseline offset. The second step of the correction consists of division by the slope of the regression of $\mathbf{x}_i$ onto $\mathbf{x}_{ref}$. The problem with MSC, as discussed by Fearn [27], is that a spectrum that is nearly orthogonal to the reference spectrum will have a slope that is close to zero. The MSC treatment for samples that are nearly orthogonal to the reference will introduce distortion, and will result in the spectra appearing as outliers.

The application of the regression in MSC assumes that the only difference between the signal and the mean are the scale and offset and the residuals around the line will be randomly distributed. As with SNV, deviations from this assumption are essential for calibration to work, and it is assumed by the method that the errors introduced through this correction are more than compensated for by a significant reduction in the multiplicative offset noise.

1.4.4 Alternative Approaches

Numerous alternatives to SNV and MSC have been developed over the years, and some are reviewed in the introduction section of Chapter 2. In particular, the present work focuses on recent methods that select particular wavelength channels to determine corrections to be applied, or which use a weight vector which places greater emphasis on certain variables when determining the scatter correction parameters. These weighted correction methods are based on the premise that some regions of the spectrum may be dominated by multiplicative offset with little chemical variation, and therefore are more reliable in making corrections. These methods are especially needed for data which contain variables where significant chemical signal is present. The utility and limitations of weighted correction methods are investigated in greater detail in Chapter 2.

1.5 Heteroscedastic Noise

For certain types of multivariate measurements, errors may be independent (uncorrelated) but complicated by non-uniform variance in the measurement errors, which is termed "heteroscedasticity". Even in cases where correlated errors are present, heteroscedastic noise may be a dominant factor in the data analysis. Heteroscedastic noise is a particular problem when the range of error variance is quite large. This is often the case when the variables in a measurement vector represent different types of measurements and/or have different units. For example, a measurement vector for an environmental sample may contain the concentrations of dozens of elements with a wide dynamic range. Alternatively, a measurement vector used to predict the activity of a drug may contain variables which represent various chemical and physical properties with different units or ranges. In other instances, such as mass spectrometry, measurements may have a wide dynamic range and be subject to proportional noise (e.g. due to ion source variation), meaning that the relative standard deviation (RSD) remains relatively constant, but the absolute uncertainty changes. Another example of heteroscedastic errors is in "data fusion", where measurement vectors are constructed by concatenating variables from multiple techniques, such as infrared and Raman spectra, where the measurements have different ranges and uncertainties. Still another case is where some instrument measurement channels may have excessive noise due to the nature of the measurement. For example, absorbance measurements will become increasingly noisy at wavelengths where the source intensity becomes low.

Heteroscedastic noise is often a problem for multivariate analysis methods that seek to simultaneously reduce the dimensionality of the data and exclude the noise. In PCA for example, data compression does not distinguish between chemical variance and noise variance, so it may over-emphasize variables with high noise over more informative variables if the absolute signal is small. As a consequence, the subspace of the measurements is more poorly estimated, resulting in less information and more noise propagating into the compressed data. This has consequences for all multivariate data analysis methods, including multivariate calibration, mixture analysis, classification, and visualization.

1.5.1 Heteroscedastic Noise Structure

If we consider a data matrix $\mathbf{X}$ $(n \times p)$, there is only one way for the errors to be homoscedastic (uniform variance in the errors), whereas there are infinite ways for the measurement errors to be heteroscedastic. In general, however, several broad categories of heteroscedastic noise can be distinguished, as described below.

1. Column heteroscedastic. In this case, the measurements can be considered to have a uniform error variance within a column of data (same variable or measurement channel), but the error variance is different among columns. This could be the case, for example, if one column represented pH and another conductivity. This situation, or at least an approximation of it, is fairly common.

2. Row heteroscedastic. Similar to case 1, this is a situation where the error variance is constant within a row (sample), but varies among the rows. This circumstance is fairly rare, but can occur if measurements with homoscedastic errors are individually normalized for each sample to compare profiles (e.g. chromatograms or mass spectra) resulting in different noise amplification.

3. Systematic heteroscedasticity. This is perhaps the most common case, where each measurement has its own error variance, but these uncertainties have a structure related in a predictable way to the measurement channel, sample, and/or measurement. For example, source flicker noise in spectroscopy can yield uncertainties that are proportional to the measurement, and shot noise (Poisson noise) has a standard deviation proportional to the square root of the signal. Often this type of noise is a composite function of several sources that can also include a homoscedastic *iid* component If the range of signals for a given channel is limited, this type of noise may be approximated by case 1.

4. General heteroscedastic. Although less common, certain measurements may be heteroscedastic with no apparent structure. This is the case, for example, with DNA microarray measurements [28], where the measurement is expressed as a ratio of fluorescence signals, each of which has an uncertainty dependent on the quantity of the particular sensor element and the amount of hybridized DNA.

Several simple examples of different types of noise are shown in Figure 1.2. In the case of *iid* normal noise (Fig. 1.2A), the standard deviation of the noise in each column is the same. In Fig. 1.2B, the noise was generated such that the noise in each column was normally distributed, but the standard deviation of the noise in each column was chosen at random, and as a result the noise is column heteroscedastic and non-systematic (case 1). In Fig. 1.2C, the standard deviation of the noise increases with increasing channel number, and the noise is therefore classified as systematic and column heteroscedastic (case 3).

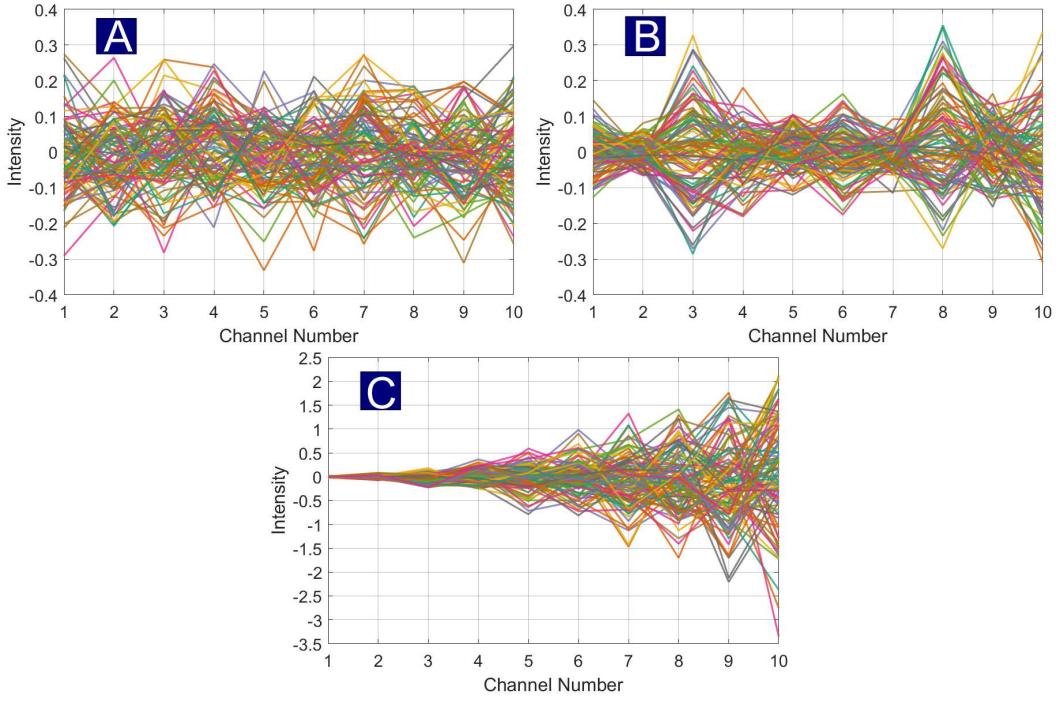

Figure 1.2: Examples of different types of independent noise. (A) *iid* normal noise; (B) Column heteroscedastic, non-systematic; (C) Systematic column heteroscedastic.

A variety of strategies have been developed to deal with heteroscedastic measurement errors, the most common being transformations of the data. While non-linear transformations (e.g. log-scaling [29]) are sometimes used, these can alter the linear structure of the data, so linear transformations (i.e. scaling) are much more common, as discussed in the following section.

1.5.2 Data Scaling

In principle, when heteroscedastic errors are present, optimal subspace modeling can be accomplished using a technique such as maximum likelihood principal components analysis (MLPCA) [22], which incorporates measurement error information to better separate noise variance from chemical variance in a manner analogous to weighted least squares. In practice, however, this requires accurate information about the measurement error structure (error covariance matrices), which is usually unavailable. In the absence of measurement error information, most approaches resort to scaling of the columns of $\mathbf{X}$ as a way to deal with heteroscedastic errors. This means that each column of data is divided (or multiplied) by a specific normalization factor to render the measurement error homoscedastic, or nearly so.

In scaling of the data, there are several assumptions that are made. The first is that the measurement errors follow case (1) in the previous section (column homoscedastic). Error variances can only be made homoscedastic across the matrix by column scaling if they are homoscedastic within the column. While each measurement could, in principle, be scaled independently to give homoscedastic errors, this would destroy the linear structure of the data [30]. Column scaling retains the structure of the data. Even if the data are not strictly column heteroscedastic, this is often a reasonable approximation to remove the effects of gross heteroscedasticity among columns.

A second assumption in scaling is that we know measurement uncertainty in each column, or at least the values relative to each other. Assuming case (1) noise, optimal scaling from a maximum-likelihood perspective (i.e. MLPCA) would involve dividing each column by its measurement error standard deviation [22]. As already noted, however, this information is generally not available. In the absence of measurement error information, an implicit (and reasonable) assumption that is often made is that the relative uncertainty in each column of a measurement is the same. Since the errors also have to be homoscedastic within a column, however, this relative uncertainty has to implicitly reference some measure of the magnitude of a column of data. The value of the relative uncertainty is not required, however, since it is assumed to be the same across all columns. Depending on the metric used for the magnitude, several techniques can be employed:

1. range scaling, where each column is divided by the range (maximum minus minimum) of the measurements in the column.

2. Mean or median scaling, where the divisor is the mean or median of the column.

3. Variance scaling, where the measurements are divided by the standard deviation around the mean (when used in conjunction with mean centering, this is called autoscaling).

Of these methods, variance scaling is by far the most widely used. The reason for this is unclear, since mean scaling would seem more practical for analytical measurements with proportional errors, but it is likely that this practice was adopted historically from other fields. In practice, the type of scaling used is likely to be of far less consequence than the assumption of uniform relative errors. If one column of measurements has a higher relative uncertainty than others (e.g. the baseline region of a spectrum), scaling can actually amplify the effect of noisy variables and degrade the quality of analysis. Based on the details discussed above, scaling to correct for heteroscedasticity without a knowledge of measurement uncertainty can lead to unpredictable results and is one of the reasons that data preprocessing is such a critical step in most chemometric methods.

A simple dataset of periodic trends is used here to illustrate the effects of scaling, and why it is so important. The dataset consists of 14 elements from groups 1, 15, and 17 (Li, Na, K, Rb, Cs, N, P, As, Sb, Bi, F, Cl, Br, I), and 9 properties for each element (bonding radius, atomic radius, ionization potential, electronegativity, melting point, boiling point, heat of vaporization, heat of fusion, specific heat). The raw data are plotted in Figure 1.3, and the elements are colored by group (alkali metals, pnictogens, halogens). The units for melting point and boiling point are significantly larger than the units of the other variables, so some form of scaling is obviously necessary. The periodic trends data were scaled using range scaling, mean scaling, variance scaling, and autoscaling, and the results of each scaling are shown in Figure 1.4. As can be seen from Figure 1.4, the different types of scaling result in different variables being emphasized in different ways. For example, ionization potential (IP) and electronegativity (EN) show large variation relative to the other variables when range scaling is used, whereas, when mean scaling is used, the variation is not quite as

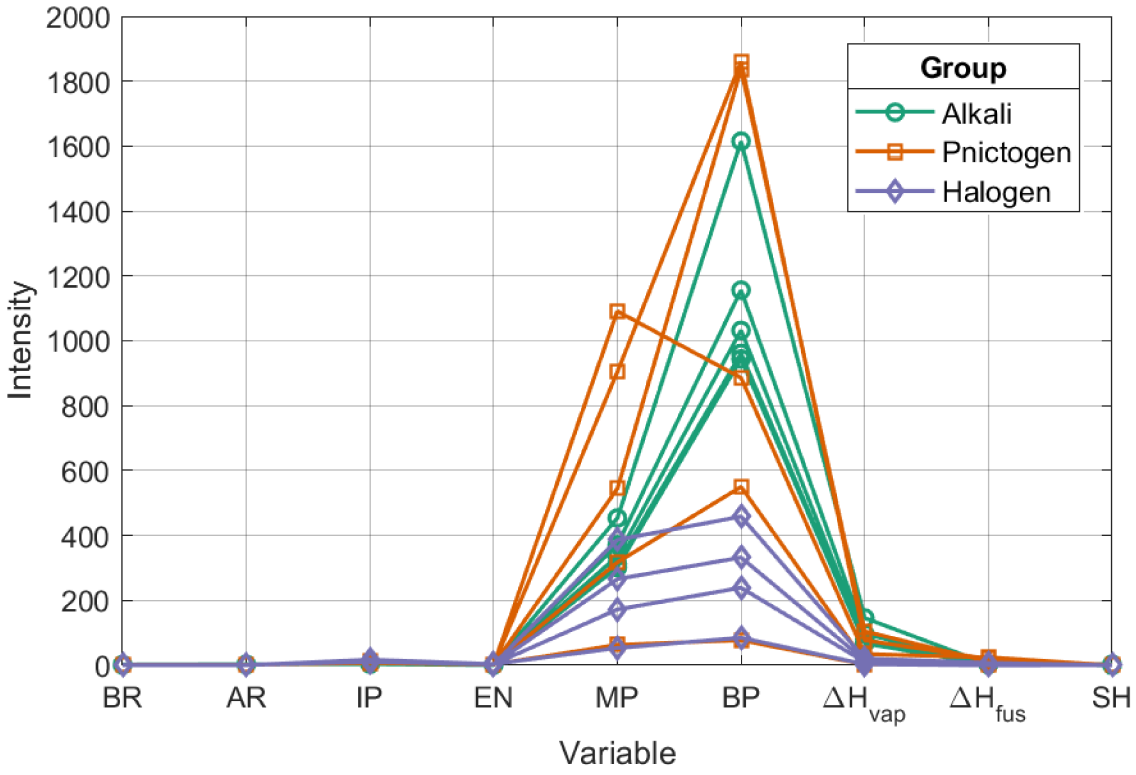

Figure 1.3: Periodic table data for 14 elements. Variable names: bonding radius (BR), atomic radius(AR), ionization potential (IP), electronegativity (EN), melting point (MP), boiling point (BP), heat of vaporization (ΔH_{vap}), heat of fusion (ΔH_{fus}), specific heat (SH).

large relative to the other variables. Without knowing the measurement uncertainties associated with each variable, it is difficult to assess which scaling method is the most appropriate for this data, and the choice of which scaling method to use can have a significant impact on the results.

1.5.3 Factor Analysis

Principal components analysis is one variant of a larger class of methods known as factor analysis (FA). The subspace estimation using PCA is based on an assumption of iid measurement errors and, as noted in the previous section, scaling is often invoked to ensure this condition. The main difficulty in using scaling, however, is a lack of measurement uncertainty estimates. A possible solution to this problem is to employ alternative FA methods that allow the simultaneous estimation of the data subspace and the measurement uncertainties. These methods have rarely been

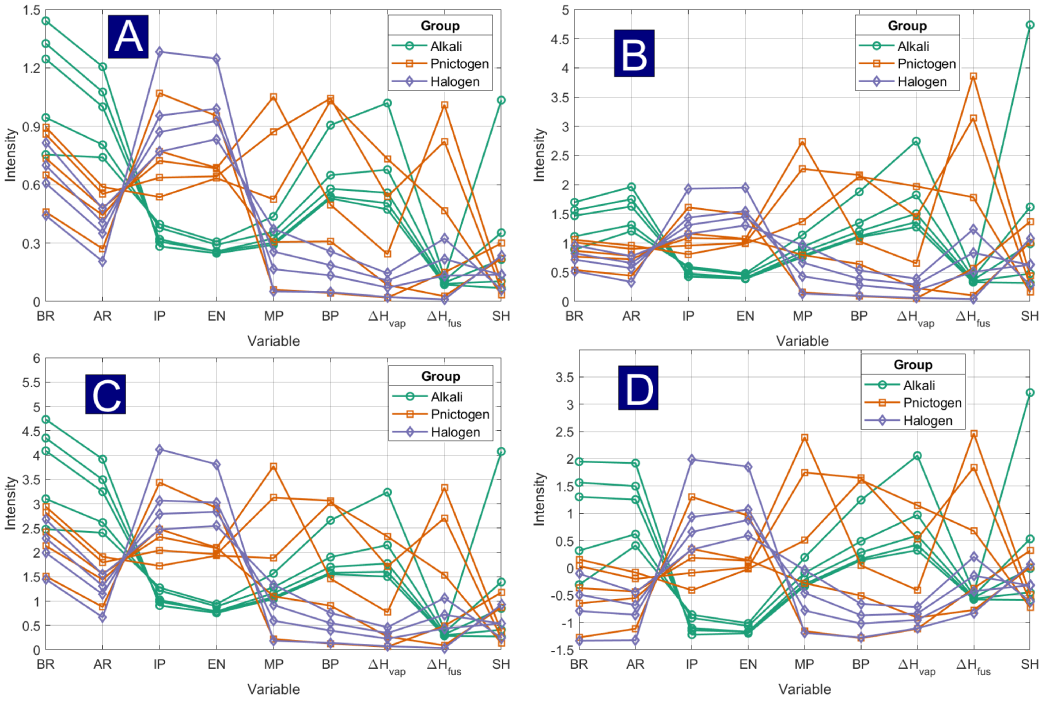

Figure 1.4: Periodic table data for 14 elements, scaled by different methods. (A)Range scaling; (B) Mean scaling; (C) Variance scaling; (D) Autoscaling. Variable names: bonding radius (BR), atomic radius(AR), ionization potential (IP), electronegativity (EN), melting point (MP), boiling point (BP), heat of vaporization ΔH_{vap}, heat of fusion (ΔH_{fus}), specific heat (SH).

used for chemical measurements, but have the potential to simplify preprocessing requirements. This possibility is explored in Chapter 3.

1.6 Summary

Multivariate analytical measurements can be characterized by a wide variety of measurement error structures that represent different levels of heteroscedasticity and/or correlation. The nature of the errors is a complex function of samples, procedures, and instrumentation and transformations applied to the data, but in certain cases a general characterization can be made. It is widely known that chemometric techniques such as calibration, curve resolution, and exploratory data analysis can be significantly impacted by the error structure. Optimal application of these methods generally involves having reliable information about the measurement error structure,

typically expressed in terms of an error covariance matrix. However, this information is often unavailable or unobtainable due to resource limitations or experimental constraints restricting the acquisition of replicate data.

In certain specific applications, it may be possible to estimate the error characteristics of the data (e.g. scattering coefficients in NIR, error variance for heteroscedastic measurements) from the data itself. This thesis explores two such cases. One of these, the use of weighted scatter correction methods, is based on a long standing methodology recently reported in the literature. The other involves the introduction of principal axis factorization, a technique widely used in other fields, as an alternative to PCA in chemometrics.

Chapter 2

Comparison of Weighted Scatter Correction Methods For Multivariate Calibration

As described in Chapter 1, in areas such as near-infrared (NIR) spectroscopy, light scattering effects can introduce multiplicative and/or additive effects on the signal, which can harm the predictive performance of resulting models. These effects are commonly corrected using methods such as the standard normal variate (SNV) and multiplicative scatter correction (MSC) transformations. However, SNV and MSC assume that the majority of the variance in the observed spectra is due to scattering effects, and that the amount of chemical variation is consistent across all variables. Weighted scatter correction methods use a weight function to account for the fact that scattering effects dominate some variables that are relatively unaffected by chemical variation. In this chapter, the circumstances under which weighted correction methods are most useful are evaluated, and three simulated datasets and one experimental dataset are used to investigate the characteristics of two weighted scatter correction methods under different measurement conditions. These methods are compared to each other, as well as conventional correction methods.

This chapter begins with a general review of scatter correction methods for NIR that have been proposed in the literature, with an emphasis on modifications to the standard SNV and MSC approaches. A more detailed theoretical discussion of two recently proposed weighted correction methods, Variable Selection for Normalization (VSN) [31] and Interferent Dominant Region Correction (IDRC) [32], are then presented. The data and methods used in this study are then described, followed by a comprehensive analysis of the results obtained.

2.1 Scatter Correction Methods in NIR Spectroscopy

The two main families of preprocessing techniques for dealing with scattering effects in NIR spectroscopy are derivative-based methods, and scatter-correction methods. A

review of these preprocessing methods has been given by Rinnan *et al* [33]. Derivative-based methods include a first-difference or second-difference of the measured spectra, along with smoothing. The Norris-Williams method of derivatives [34] uses a smoothing function such that the intensity at a given point is the weighted average of the neighboring points. The Norris-Williams derivative uses a "gap derivative", which calculates the first or second difference based upon the smoothed values for points that are separated by a gap distance from a given channel. The other method for derivatives is the Savitzky-Golay [35] approach, which calculates the derivative of a point (channel) i by using a polynomial function to fit a symmetric window of neighboring points. The first derivative will remove all constant sources of variation, and can therefore remove pure baseline offsets, and the second derivative can also remove linear slope terms as well as baseline offsets. However, derivative-based methods cannot remove multiplicative noise completely. completely.

The most common scatter correction methods for NIR spectra are the multiplicative scatter correction (MSC) [4] and standard normal variate (SNV) [3]. MSC uses a linear regression of each sample versus a reference spectrum (usually the average spectrum), and the slope and intercept are obtained. The MSC correction subtracts the intercept from the sample, and then divides by the slope. SNV subtracts the mean of each spectrum and then divides by the standard deviation of the result. The relationship between MSC and SNV has been explored from a theoretical perspective, and it has been found that, while SNV and MSC are generally quite similar, the SNV transformation can introduce curved structures in scores plots, while MSC can sometimes introduce outliers [27].

Scatter correction methods have been developed which extend beyond the framework of MSC and SNV. The extended MSC (EMSC) [36] [37] method can account for other effects beyond simple additive and multiplicative offset terms, such as linear and quadratic wavelength-dependent terms, and the pure component spectra of the main components (if available). The inverted scatter correction (ISC) method [38] uses the estimate of the coefficients based upon a regression in which the reference spectrum is projected onto the spectrum of a given sample (the "inverse" of the MSC, where the sample spectrum is projected onto the reference). An ISC model which also includes wavelength-dependent terms or pure components spectra is termed extended

inverted scatter correction (EISC) [39]. In the least squares fitting, the ISC and EISC methods both assume that the errors in the sample spectrum are smaller than the errors in the reference spectrum, which is a questionable assumption. Loopy MSC [40] is another variant which involves applying the MSC correction, then using the MSC-corrected spectra to re-apply the MSC correction. The Robust normal variate (RNV) [41] method calculates the mean and standard deviation using a percentile of the values in a spectrum. All of these modifications of MSC and SNV have been shown by the authors to produce improved or marginally improved results for the specific applications presented in the corresponding papers. A commonality of these approaches is that they utilize the full spectrum, with the assumption that scattering effects dominate at all wavelength channels. Several alternative approaches have been based on the premise that some regions of the spectrum may be better than others for estimating scattering effects, and these are discussed below.

Piecewise MSC (PMSC) [42] is one method that does not assume that scattering effects are uniform across all wavelengths. PMSC calculates the MSC scatter coefficients for a moving window of neighboring wavelength channels, and it is based upon the assumption that the scattering coefficients can change in different wavelength regions. A somewhat related method, localized SNV (LSNV) [43], entails dividing the spectrum into several regions of equal width, and performing separate SNV corrections on each region. In a 2018 paper by Grisanti *et al* [44], three new scatter correction methods were proposed, termed Dynamic Localized SNV, Peak SNV (PSNV), and Partial Peak SNV (PPSNV). DLSNV is based upon localized SNV, with a modification that it varies the starting points of the SNV windows. PSNV and PPSNV are based upon picking peaks which have a high correlation with $\mathbf{y}$, combining picked points that are near each other, and performing SNV across each region of interest. The main difference between PSNV and PPSNV is that for PSNV, the spectrum is subdivided based upon the points of interest, whereas for PPSNV, a fixed window of neighboring points around each picked point are used. Overall, PSNV and PPSNV are more closely related to piecewise MSC than to weighted correction methods, as the corrections are performed at a local level rather than global level. All of these methods are likely to be less effective if there are regions where chemical variance dominates over the variance due to scatter. Their underlying assumptions differ from

those of weighted scatter correction methods which assume uniform weighted scatter coefficients, but also that those coefficients are best estimated using selected wavelength regions or channels.

The SNV and MSC methods both depend upon the assumption that the dominant source of variation among a set of spectra is due to scattering effects. If significant chemical variation occurs, then SNV and MSC will not be able to completely distinguish the effects of the physical scattering from the chemical signals [37] [32]. To address the issue of chemical variation confounding the scatter correction, methods of weighted scatter correction, which seek to account for the presence of chemical variation in datasets affected by scatter through the use of weighted normalization, have been proposed. A weighted scatter correction involves the calculation of vector of weights for the spectral channels, then the spectra are multiplied by the weights, the scatter correction parameters are calculated using the weighted spectra, and finally the scatter correction is performed by applying the correction parameters to the original spectra. These methods are discussed in the section that follows.

2.2 Weighted Scatter Correction Methods

Several methods of weighted scatter correction have been proposed in the literature, and these methods differ based upon how the weights are calculated, and the underlying model used to perform the correction (SNV, MSC, EMSC, etc.). The original EMSC paper proposed by Martens *et al* in 1991 [36] suggested that a weighted least squares estimation of the EMSC solution could be used, although it did not suggest a procedure for calculating the weights. In a 2005 paper by Gallagher *et al* [45], a weighted EISC procedure was proposed where, for a given weight matrix, a weighted EISC solution is calculated, spectral channels with high residuals are de-weighted, and the procedure is iterated using the new weights until convergence. In a 2019 paper, Wu *et al* proposed a method called Weighted Multiplicative Scatter Correction using Variable Selection (WMSCVS) [46]. WMSCVS assumes an EMSC model and uses a weighted bootstrap sampling method to perform variable selection and it is optimized by trying to find the minimum prediction error for a PLS model using the corrected spectra. WMSCVS uses an orthogonal projection to remove baseline offsets and wavelength-dependent terms, and then solves for the coefficient of the multiplicative

term. When wavelength-dependent terms are not present, the orthogonal projection step is mathematically equivalent to subtracting the mean from each row of the spectral matrix. The orthogonal projection does not use any weighting for the estimation of the baseline and wavelength-dependent terms, so as formulated, the WMSCVS is not a complete weighted scatter correction method. EISC, EMSC, and WMSCVS all assume an "extended" scatter correction model in which wavelength-dependent scatter parameters are assumed. In this work, a conventional scatter correction model is assumed, and so the extended methods are not explored further.

In 2014, Bi *et al* proposed the Interference Dominant Region Correction (IDRC) method [32]. In IDRC, the spectra are divided into evenly spaced regions. One of the regions is selected, and a weighted SNV correction is performed using the mean and standard deviation of the variables in the selected region. As with WMSCVS, a PLS regression model is then built using the corrected spectra to predict a target variable and the prediction errors are calculated. The process is repeated using different regions of varying size. After all regions have been tested, the region which gives the lowest prediction errors is selected, and a weighted SNV is performed, with the weights equal to 1 in the selected region, and with weights of 0 for the variables outside the selected region.

In a 2019 paper by Roger *et al*, a method termed Variable Selection for Normalization (VSN) [31] was proposed. The VSN procedure calculates weights by using the Random Sample Consensus (RANSAC) algorithm [47]. For the regression of a pair of signals on one another, The RANSAC algorithm seeks to fit a line that maximizes the number of points that lie within a tolerance from the line. The VSN approach uses the RANSAC algorithm to fit random pairings of spectral samples, and determine how frequently each wavelength channel lies within the tolerance of the line. The assumption behind VSN is that variables which have little chemical information will follow the same relationship, and will therefore be more likely to lie within the tolerance of the line. The VSN weights can be applied to an SNV or MSC model. The VSN approach can also be used to solve an EMSC model, although the weights must be calculated in a slightly different manner.

In this chapter, simulated datasets will be used to test the effectiveness of the weighting used in the VSN and IDRC methods under a variety of conditions. The

details of these methods are described in the following sections.

2.2.1 Theory of Weighted Scatter Correction

A weighted scatter correction method will typically use a diagonal weighting matrix $\mathbf{W}$ of dimensions $p \times p$, where p is the number of variables, such that the off-diagonal elements (w_{jk}) are 0, and $0 \leq w_{jj} \leq 1$. For a variable selection-based correction, the diagonal elements w_{jj} are equal to either 0 or 1. For weighted SNV, the mean and standard deviation of $\mathbf{x}_i\mathbf{W}$ are used instead of the mean and standard deviation of sample $\mathbf{x}_i$. The weights are structured such that the larger the weight, the more a variable is assumed to be dominated by scatter.

There are two aspects of weighted scatter correction: the weighting procedure used, and the correction itself. Building off of the principles of MSC and SNV, it is possible to understand the circumstances in which weighted scatter correction will be effective, and the circumstances in which it will be ineffective. Weighted scatter correction methods are best suited for datasets which have regions with low chemical signal, and are dominated by scatter, and also have regions where there is a significant amount of chemical variation. The regions dominated by scatter are needed in order for it to be possible to obtain an accurate estimate of the scatter coefficients. The presence of regions that are dominated by chemical signal will cause major issues for conventional MSC and SNV.

2.2.2 Variable Selection for Normalization (VSN)

The VSN algorithm uses the RANSAC (random sample consensus) [47] algorithm as part of the procedure for estimating the weights. Given two spectra that are plotted against one another, the RANSAC algorithm tries to find a line which maximizes the number of points that lie within a certain distance (tolerance) from the line. The points that lie within the tolerance of the line are termed "inliers", and the points that are outside the tolerance are the outliers. In the context of scatter correction, the idea of using RANSAC is that, for variables which are dominated by scattering effects, the same linear relationship will be obeyed and so those variables will be inliers, whereas variables which contain significant chemical signals will be less likely to be inliers.

input : Data Matrix of Spectra ($\mathbf{X}$) ($n \times p$); Tolerance value, ϵ ; Number of sample pairs, N_s ; Number of inner loop iterations, N_w;
output: A vector of weights, $\mathbf{w}$

Initialize the weight vector $\mathbf{w}$, a vector of zeros of dimension $p \times 1$;
for $n_{outer} \leftarrow 1$ **to** N_s **do**
  Draw a pair of samples $\mathbf{x}_1$,$\mathbf{x}_2$;
  $bestn_{inliers}$=0;
  for $n_{inner} \leftarrow 1$ **to** N_w **do**
    Draw a random pair of variables k and ℓ, where $x_{1k} \neq x_{1\ell}$;
    Calculate the coefficients a and b such that $\mathbf{x}_1 = a\mathbf{x}_2 + b$ for variables k,ℓ: ;
    a=$\frac{x_{2k}-x_{2\ell}}{x_{1k}-x_{1\ell}}$;
    b=$x_{2\ell} - ax_1$;
    $\boldsymbol{\delta} = \left|\mathbf{x}_2 - a\mathbf{x}_1 - b\right|$;
    inliers=**which**($\boldsymbol{\delta} < \epsilon$);
    $n\_inliers$=size(**inliers**) ;
    if $n\_inliers > best\_n\_inliers$ **then**
      $best\_n\_inliers = n\_inliers$;
      best_inliers = **inliers**;
    end
  end
  for $j \leftarrow 1$ **to** p **do**
    if $j \in$ ***best_inliers*** **then**
      $\mathbf{w}_j = \mathbf{w}_j + 1$;
    end
  end
end
$\mathbf{w} = \mathbf{w}/N_s$;
$\mathbf{W} = diag(\mathbf{w})$

Algorithm 1: Algorithm for the VSN weighting method based on the RANSAC method

The VSN algorithm has a version for calculating a weighted SNV correction, and a version for calculation of a weighted EMSC correction. The weighting procedure for the VSN algorithm proceeds as follows in Algorithm 1. The VSN weighting algorithm is initialized by choosing a value of the tolerance threshold for distinguishing inliers and outliers, ϵ. The algorithm consists of an outer loop, and an inner loop. The outer loop is based upon the number of sample pairs (N_s). The authors recommend using all possible pairings of the samples in the training set. For n training set samples, the number of unique pairings is equal to $\Big(n\big(n-1\big)\Big)/2$. At each iteration of the outer loop, a pair of samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are used. In the inner loop, the purpose is to optimize the parameters a and b to find the largest set of inliers for fitting the equation $\mathbf{x}_1 = a\mathbf{x}_2 + b$ to within the tolerance of ϵ. For each iteration in the inner loop, a random pair of variables k and ℓ are drawn to calculate trial values for a and b. The number of iterations in the inner loop is determined by the number of variable pairs to test (N_w), which must be set by the user. The recommended N_w is 500-1000. Using the pair of variables k and ℓ, the slope (a) and intercept (b) of a line to estimate $\mathbf{x}_1$ from $\mathbf{x}_2$ is calculated. Using the coefficients a and b, a vector of absolute differences ($\boldsymbol{\delta}$) is used to store the absolute error in the fit of $\mathbf{x}_1$ onto $\mathbf{x}_2$ for each variable (wavelength channel). Next, the difference vector $\boldsymbol{\delta}$ is tested to see which elements are less than the tolerance ϵ, and the indices of the variables are stored in the vector **inliers**. The number of inliers ($n\_inliers$) is calculated, and $n\_inliers$ is compared with the current largest number of inliers ($best\_n\_inliers$). If $n\_inliers$ is larger than $best\_n\_inliers$, then $best\_n\_inliers$ and **best_inliers** are updated. Then the inner loop continues to iterate until N_w iterations are reached. If a variable j is a member of **best_inliers**, the weight $\mathbf{w}_j$ is increased by 1. Then the next iteration of the outer loop is run. After all N_s iterations of the outer loop are completed, the weight vector is normalized by dividing by N_s, so that the weights are on a scale from 0-1. Finally, the weight vector $\mathbf{w}$ is converted to a diagonal weight matrix $\mathbf{W}$.

For VSN, the value of the tolerance ϵ should be varied across several orders of magnitude (for example, from 10^{-5} to 10^{0}). After calculating the weights for several values of the tolerance, the authors of the VSN algorithm recommend choosing the threshold value which results in the largest standard deviation of the weight vector. The reasoning is that, for a weight vector where the weights are all nearly equal to

one another, the standard deviation will be very small, whereas if the weights are all either close to zero or close to unity, the standard deviation of the weights will be large.

There are two different types of corrections that can be applied using VSN. One version is to use a weighted SNV correction, and this version is termed "Selective SNV", or SSNV. A special case of SSNV, in which the weights are either unity or zero, is termed here as "Binary SSNV". SSNV is carried out in a three-step process:

$$\mu_{C,i} = \mu\{\mathbf{x}_i\mathbf{W}\} \tag{2.1}$$

$$\mathbf{x}_{C,i} = \mathbf{x}_i - \mu_{C,i} \tag{2.2}$$

$$\sigma_{C,i} = \sigma\{\mathbf{x}_{C,i}\mathbf{W}\} \tag{2.3}$$

$$\mathbf{x}_{SSNV,i} = \frac{\mathbf{x}_{C,i}}{\sigma_{C,i}} \tag{2.4}$$

In these equations, $\mathbf{x}_{C,i}$ is equal to the spectrum of sample i ($\mathbf{x}_i$) corrected by subtracting the weighted mean ($\mu_{C,i}$), $\sigma_{C,i}$ is the weighted standard deviation of $\mathbf{x}_i$ (calculated using $\mathbf{x}_{C,i}$), and $\mathbf{x}_{SSNV,i}$ is the SSNV-corrected spectrum.

The other variation of VSN employs a weighted projection, which can be used to solve either an MSC or an EMSC model. The VSN solution for the MSC model is hereby termed Selective MSC, or SMSC for short. For the SMSC solution, a projection matrix $\mathbf{M}$ of size $2 \times p$, which consists of a row vector of ones of size $1 \times p$, and a reference spectrum, $\mathbf{x}_{ref}$ is used. Using the projection matrix $\mathbf{M}$, and the weights matrix $\mathbf{W}$, the SMSC coefficients $a_{i,SMSC}$ and $b_{i,SMSC}$ for the spectrum of sample $\mathbf{x}_i$ are solved using the following equation:

$$\begin{bmatrix} a_{i,SMSC} \\ b_{i,SMSC} \end{bmatrix} = \left(\mathbf{MWM}^{\mathrm{T}}\right)^{-1}\mathbf{MW}\mathbf{x}_i^{\mathrm{T}} \tag{2.5}$$

Equation 2.5 is simply a weighted form of the regression used in ordinary MSC, emphasizing the wavelength channels with high weights in performing the fit. If the weights for the variables are all equal, the SMSC will result in exactly the same coefficients as a conventional MSC.

In the paper proposing the VSN method, the MSC-based solution was not explored, and instead the primary focus was on the EMSC-based correction. In the present study, the SMSC approach is investigated, to compare it with SSNV and

Binary SSNV. For the sake of simplicity, and ease of comparison among methods, the simulated datasets did not incorporate wavelength-dependent scattering, which is a fundamental component of the EMSC model.

The authors who proposed VSN claimed that weighted SNV can improve the model interpretation, and use the PLS regression vectors to illustrate this point [31]. However, direct interpretation of regression vectors is a hazardous undertaking for even the most experienced chemometrician, as the shape of a regression vector is a complex function of the pure-component spectra, the experimental design, and the measurement error structure [48].

2.2.3 Interferent Dominant Region Correction (IDRC)

The IDRC method was proposed by Bi *et al* [32]. In the terminology of IDRC, an interferent dominant region (IDR) is used to describe a region which has a low noise level and low level of chemical variation, and which is therefore dominated by scattering effects. An IDR is equivalent to what is termed a scatter dominant region (SDR) in the work presented here. The objective of IDRC is to try to find a SDR by testing corrections using the parameters from local spectral regions. The IDRC algorithm begins by dividing the spectra into a series of k equally spaced regions, where $1 \leq k \leq K$, and K is the maximum number of regions (for example, when K=10, the spectrum can be divided into at most 10 different sub-regions). Let a given spectral region be denoted as $\mathcal{R}$, and let $\mathbf{X}^{\mathcal{R}}$ be the part of the spectra $\mathbf{X}$ in region $\mathcal{R}$. The mean and standard deviation of sample i in region $\mathcal{R}$ are calculated, and are used to perform an SNV-like correction to the spectra:

$$\mathbf{x}_{i,IDRC} = \frac{\mathbf{x}_i - \mu\{\mathbf{x}_i^{\mathcal{R}}\}}{\sigma\{\mathbf{x}_i^{\mathcal{R}}\}} \tag{2.6}$$

This is equivalent to a weighted SNV correction, where the weights are 1 for all variables in $\mathcal{R}$, and the weights are 0 for all variables that are not in region $\mathcal{R}$. The next step in the IDRC algorithm is to calculate an index of differences among samples:

$$t = \sum_{j=1}^{p} var\{\mathbf{x}_{j,IDRC}\} \tag{2.7}$$

where p is the number of wavelength channels, $\mathbf{x}_{j,IDRC}$ is the j^{th} column of the IDRC-corrected spectra ($\mathbf{X}_{IDRC}$). The index of differences among samples is also calculated

using a conventional SNV correction, and is termed t_{SNV}. The index of differences is used to screen the regions, such that regions in which the t is larger than t_{SNV} are excluded from further consideration. If the t of a region is smaller than t_{SNV}, then a PLS regression will be built using the corrected spectra $\mathbf{X}_{IDRC}$ and a target variable $\mathbf{y}$. From the PLS regression, the root mean squared error of cross validation (RMSECV) is calculated. The above process is repeated for all k spectral regions of a given size, then the region size is changed to split the data into $(k+1)$ regions, until $k = K$. After testing out all regions from each possible region size, the region which results in the lowest RMSECV is chosen as the optimal region, and is used for the final IDRC correction. If there are multiple target variables, then a separate IDRC analysis must be performed for each analyte of interest.

2.2.4 Objectives

In the work presented here, it is hypothesized that weighted scatter correction methods (VSN and IDRC) will be most appropriate when used to correct spectra that exhibit two types of regions: a region in which the majority of the variance is due to scatter (multiplicative offset and baseline offset), which is hereby referred to as the scatter dominant region (SDR), and a region in which there is strong variation from chemical signals, which will be called the chemical dominant region (CDR).

This theory is tested using three simulated datasets and one experimental dataset. In the simulated datasets, only baseline offset and multiplicative offset noise are assumed to be present (with a small amount of *iid* noise added), and it is assumed that no wavelength-dependent scattering effects were present. The spectra were designed to have a small amount of background chemical signal in a region that was dominated by scatter while, in another region, the signal was dominated by chemical signals from the analytes of interest. In each dataset, the pure spectra and noise characteristics were the same, but the amounts of background signal and main chemical signals were varied. The experimental dataset consisted of NIR spectra of wine musts. The data contained significant amounts of scatter in some regions of the spectra, while in other regions had significant amounts of chemical variation. For each dataset, the effects of preprocessing using SNV, MSC, VSN, and IDRC were tested, and the PLS regression was used to assess the effect of the different preprocessing methods on

the prediction performance. Each of the preprocessing methods were also assessed qualitatively based upon how well the methods preserved the chemical signal. It was found that the results for the simulated datasets support the theory of weighted scatter correction presented above, and recommendations regarding the usage of SNV and IDRC are proposed on the basis of the results.

2.3 Data

Three simulated datasets and one experimental dataset of NIR spectra of wine musts were used in the present study. The simulated datasets were designed to include features that are present in experimental spectra, but in a somewhat simplified form, and these datasets were also designed to test the assumptions of weighted scatter correction methods by varying the levels of main chemical signal and the chemical background level while holding all other parameters the same. Simulated Dataset 1 was designed to have the optimal setup for weighted scatter correction methods, as it had a large main analyte signal level, and a low background level. Simulated Dataset 2 represents data for which the assumptions of SNV are not violated, as it had a low main chemical signal level, and a low background level (the same level as for Dataset 1). Simulated Dataset 3 was designed to show that when the background level is too high, the scatter correction parameters cannot be accurately estimated, resulting in errors in the corrections, and for this dataset, the main chemical signal was the same as for simulated Dataset 1, while the chemical background signal was significantly larger.

2.3.1 Simulated Data

The three simulated datasets were generated as follows. Each dataset contained $n = 2100$ total samples, and $p = 200$ variables, and was composed of a set of background spectra, chemical analyte spectra, and noise. 100 of the samples were used for optimization of the scatter correction parameters (for VSN and IDRC), while the remaining samples were used for regression (1000 samples for calibration, and 1000 samples for the test set). The large number of samples used for calibration and prediction in these simulation studies far exceeds the number that would typically be available in a real experimental calibration, but were intended to minimize the errors

from these sources. The final prediction errors can be viewed as being a combination of errors in building the calibration model, errors in the prediction step, and errors arising from preprocessing. As the purpose of this study was to examine contributions from preprocessing (scatter correction methods), the large number of calibration and prediction samples minimizes those contributions. The number of samples used to determine the weighting parameters (100) is more typical of what might be available for experimental data.

The noise-free spectra $\mathbf{X}_{chem}$ were composed of the component spectra of the main chemical analytes ($\mathbf{X}_{main}$), and of background components spectra ($\mathbf{X}_{bg}$), and of a base peak shape ($\mathbf{X}_{base}$), such that $\mathbf{X}_{chem} = \mathbf{X}_{main} + \mathbf{X}_{bg} + \mathbf{X}_{base}$. The base spectrum, shown in Figure 2.1A, is intended to represent a common spectral profile shared by all samples of a given experiment and represent an average of the broad spectral features typical of NIR spectra. It can be viewed as a mixture of all of the combination and overtone bands presented by the complex mixture. Ordinarily, since the mean is removed prior to calibration, the base spectrum would have no effect, but in the present simulations, it will effect the magnitude of the multiplicative offset noise. The background spectra $\mathbf{X}_{bg}$, exemplified in Figure 2.1C, represent the multitude of chemical variations between samples not associated with the analyte of interest. These are assumed to be random and uncorrelated with the analyte of interest. Finally, $\mathbf{X}_{main}$ will represent the spectra from the analyte of interest, for which the measured property (e.g. concentration) has been determined. For the purpose of this study, $\mathbf{X}_{main}$ is composed of two chemical components, the analyte of interest and an interfering compound with similar spectral properties. This was done to more accurately represent a typical calibration scenario. Typical spectra of the main analytes are shown in Figure 2.1D, with the noise-free spectra shown in Figure 2.1E.

All of the spectra simulated in this work were simulated using Gaussian functions or combinations of Gaussians to represent the broad spectral features. In general, the Gaussians can be described by a height scale factor (h), the position of the mean μ, and a standard deviation σ, as given in Equation 2.8

$$f_{Gauss}\big(h, \mu, \sigma, \boldsymbol{\xi}\big) = h\big(\frac{1}{\sigma\sqrt{2\pi}} e^{-(\boldsymbol{\xi}-\mu)^2/(2\sigma^2)}\big) \tag{2.8}$$

In this equation, $\boldsymbol{\xi}$ represents the wavelength channel indices, and has the same units as μ and σ. The maximum peak height will occur at μ, and the baseline width of the peak will be approximately 4σ. Spectra are represented as row vectors for $\boldsymbol{\xi} = \big[1, 2, ..., 200\big]$ (or $\big[1, 2, ..., 300\big]$ for the background spectra) and may be represented by the addition of two or more Gaussians.

The base peak shape matrix $\mathbf{X}_{base}$ shown in Figure 2.1A was the same for all samples and resulted from the addition of three Gaussian functions, all with $h = 20$ and $\sigma = 25$. The mean positions of the three Gaussians were $\mu_1 = 30$, $\mu_2 = 120$, and $\mu_3 = 180$. The resulting 1×200 vector was copied n times to give the base profile matrix $\mathbf{X}_{base}$ for subsequent calculations.

The objective for the chemical background spectra was to have a mixture of peaks, with similar mean intensities. These background spectra were designed to model variation due to the presence of minor chemical components. The chemical background spectra were generated as follows. First, the pure component spectra for 31 background components were generated over the range $\boldsymbol{\xi} = \big[1, 2, ..., 300\big]$. Each spectrum consisted of a Gaussian function with $h = 200$ and $\sigma = 200$, with the mean position varying from $\mu = 0$ to $\mu = 300$ in steps of 10. These 31 spectra consisted of the pure spectral profile matrix, $\mathbf{S}_{bg}$ (31×300). To introduce chemical variation, it is necessary to incorporate concentration variations in the background components for each of the 2100 mixture spectra generated. To do this, a random concentration matrix, $\mathbf{C}_{bg}$ (2100×31) was generated by drawing random values from a normal distribution with a mean of μ_{bg} and a standard deviation of σ_{bg}. For all Datasets, μ_{bg} was set to 0.005, but σ_{bg} was varied to improve different conditions for the simulations. For Datasets 1 and 2, σ_{bg} was set to 0.00002, while for Dataset 3 σ_{bg} was 0.001. The background concentrations and background spectral profiles were multiplied to give the background spectra, such that $(\mathbf{X}_{bg} = \mathbf{C}_{bg}\mathbf{S}_{bg}^{\mathrm{T}})$. The last step was that the background spectra were truncated to only include channels 51-250 to remove edge effects. The background spectra $\mathbf{X}_{bg}$ for simulated Dataset 1 are shown in Figure 2.1B-C, which show the spectra before truncation and after truncation, respectively. From examining the plot, the motivation for using extra channels and then trimming becomes apparent, as the full background spectra (before trimming) experienced a drop-off in the intensity at each end of the spectrum, whereas the trimmed background spectra all have an

average intensity of close to 0.1. The signal standard deviation of individual background components was about 0.00016 (for Datasets 1 and 2) and 0.008 (for Dataset 3), but the additive effect led to the values of about 0.0021 and 0.011, respectively, in the final spectra.

The main chemical spectra consisted of two components representing the analyte of interest and an interferent. Although the background spectra can also be considered to represent interferences, the intent here was to include at least one interferent that was constrained to be one of similar magnitude to the analyte to challenge the calibration. The spectra of the analyte and interferent were designed to be similar, each with peaks close to channels 120 and 180. These were on the right hand side of the spectrum, with the potential for the channels on the left-hand side to be used for scatter correction. The analyte spectrum was created by combining two Gaussian functions with h=200, $\sigma = 10$ and $\mu_1 = 115, \mu_2 = 175$. The interferent (component 2) was created in the same way, except with $\mu_1 = 125$ and $\mu_2 = 185$. This ensured some overlap between the components. The combination of these two spectra provided the 2×200 spectral matrix, $\mathbf{S}_{main}$. As with the background matrix, the concentration matrix for the main components, $\mathbf{C}_{main}$ (2100×2), was generated by sampling from a random normal distribution with parameters $N(\mu = 0.10, \sigma_{main})$, where σ_{main} is the standard deviation of the background concentrations. For Datasets 1 and 3, σ_{main} was set to 0.03, while for Dataset 2 σ_{main} was 0.001. The main analyte concentrations and spectral profiles were multiplied to give the main analyte spectra, such that $(\mathbf{X}_{main} = \mathbf{C}_{main}\mathbf{S}_{main}^{\mathrm{T}})$. The main chemical analyte spectra $\mathbf{X}_{main}$ for Dataset 1 are shown in Figure 2.1D. The combination of the base profile $\mathbf{X}_{base}$, background spectra $(\mathbf{X}_{bg})$ and main component spectra $\mathbf{X}_{main}$ results in the noise-free spectra $\mathbf{X}_{chem}$, typified in Figure 2.1E.

The model for the overall spectra, including the individual components of the noise and the noise-free spectra, is shown in Equation 2.9:

$$\mathbf{x}_i = \left(\alpha_i \mathbf{1} + \beta_i \mathbf{x}_{i,chem} + \epsilon_i\right) + \mathbf{x}_{i,chem} \tag{2.9}$$

where $\mathbf{x}_i$ is the spectrum of the i^{th} sample, $\alpha_i \sim N(\mu = 0, \sigma = 0.1)$ is the coefficient of the baseline offset, $\beta_i \sim N(\mu = 0, \sigma = 0.1)$ is the coefficient of the multiplicative offset of the i^{th} sample, $\mathbf{x}_{i,chem}$ is the pure chemical response excluding scatter effects, and $\boldsymbol{\epsilon}_i \sim N(\mu = 0, \sigma = 0.0001)$ is a vector of *iid* normal errors. The result of adding

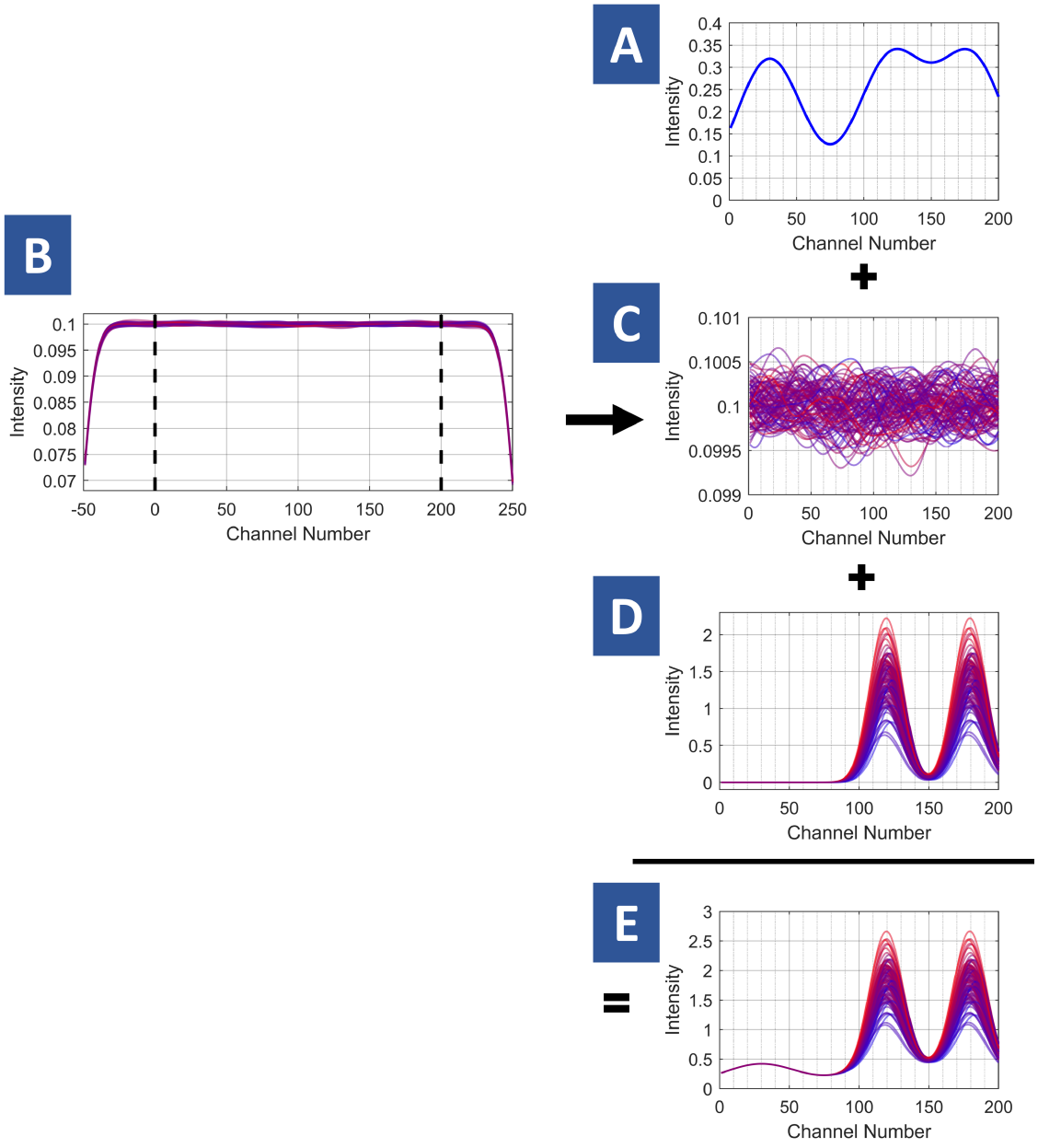

Figure 2.1: Simulated spectra for Dataset 1. (A)Base peak shape $\mathbf{X}_{base}$; (B) Background components spectra $\mathbf{X}_{bg}$ (before trimming wavelengths); (C) Background components spectra $\mathbf{X}_{bg}$ (after trimming wavelengths); (D)Spectra of the main chemical analytes ($\mathbf{X}_{main}$); (E)Noise-free spectra $\mathbf{X}_{chem}$.

noise is shown in Figure 2.2B.

To simulate a small error in the reference method, the main analyte concentration matrix $\mathbf{C}_{main}$ was transformed to obtain the "observed" concentrations $\mathbf{Y}$, using the equation $\mathbf{Y} = \frac{1}{\sigma_{main}}\mathbf{C}_{main} + \mathbf{E}_Y$, where $\mathbf{E}_Y \sim N(\mu = 0, \sigma = 0.10)$ is a matrix of

randomly normally distributed (*iid*) numbers. This transformation gave a range of $\mathbf{Y}$ of about 2%, a reasonable value for most reference methods. This transformation also had the benefit of scaling $\mathbf{Y}$ to the same range across each dataset (making for easier comparisons of RMSEP across datasets), and set a floor of approximately 0.1 for the RMSEP. The benefit of setting a non-zero floor to the RMSEP is that any RMSEP values below 0.1 would be indicative that overfitting had occurred.

The three different datasets simulated represent different levels of signals and noise, but because there are different sources of "noise", it is useful to put these into perspective. The simulations are summarized in Table 2.1 to provide this perspective in terms of variations introduced by different sources. The first three rows give the standard deviations of the three noise sources (baseline offset, multiplicative offset, and *iid* noise. These were fixed for all three simulations. The baseline offset and *iid* noise were in units of the measured signal and it is clear that the offset noise is much larger than the *iid* noise by several orders of magnitude. This is typical for NIR spectra and the *iid* noise has only been included here to provide a limit. The multiplicative offset noise is expressed as a relative standard deviation (RSD), but with the signal range given, this will vary from about 0.025 to 0.25, which is on the same order as the offset noise. The fourth row gives the individual background component variation in concentration units, but for greater utility, row five gives the average background variability in signal units. The chemical background variability is increased by a factor of fifty for Dataset 3, but is still about an order of magnitude below the scatter effects. The analyte variability changes by about a factor of 30 over the simulations in terms of concentration, but this is more challenging to quantify in terms of the signal measurement domain because of the presence of the interferent. The relevant quantity to express is the variability in the net analyte signal (NAS) which is derived from multivariate calibration theory [49] using only the two main components, where the selectivity for analyte is 0.62. The magnitude of the NAS for each simulation is given in row seven. From this, we can calculate an operational value for the ratio of the analyte signal to the chemical background signal, given in row eight. The table shows more clearly how the simulations are intended to examine the effectiveness under different conditions of background chemical noise and signal to background ratio.

Table 2.1: Parameters for Simulated Datasets

Parameter	Dataset 1	Dataset 2	Dataset 3
σ_{base} (signal units)	0.1	0.1	0.1
σ_{mult} (RSD)	0.1	0.1	0.1
σ_{iid} (Signal Units)	0.0001	0.0001	0.0001
σ_{bg} (Concentration Units)	0.00002	0.00002	0.001
σ_{bg} (Signal Units)	0.00214	0.00214	0.0106
$\sigma_{analyte}$ (Concentration Units)	0.03	0.001	0.03
$\sigma_{analyte}$ (Net Analyte Signal)	0.89	0.030	0.89
Signal to Background Ratio	4150	140	84

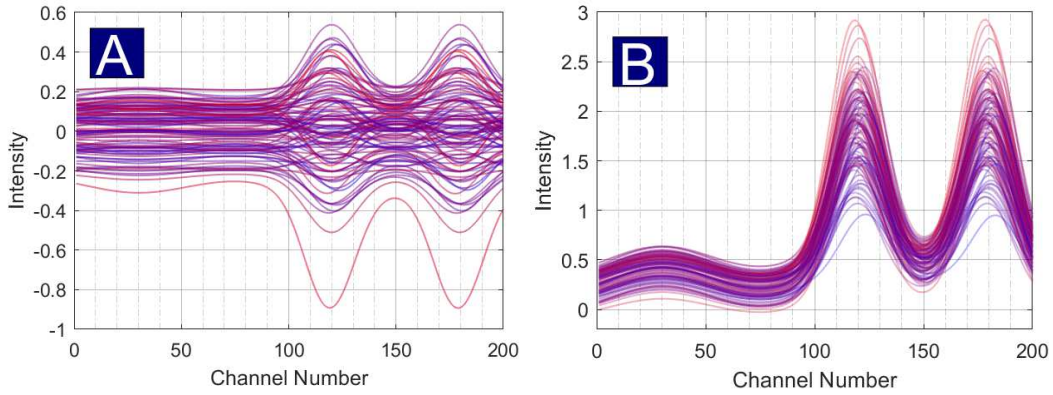

Figure 2.2: Simulated noise and simulated spectra Dataset 1. (A) Noise matrix (including multiplicative offset, baseline offset, and *iid* noise); (B) Simulated spectra matrix, **X**.

The noise matrix for simulated Dataset 1 is depicted in Figure 2.2A, the noise-free spectra are shown in Figure 2.1E, and the spectra for Dataset 1 are shown in Figure 2.2B. These spectra were designed to cause problems for conventional SNV and MSC, because in the region from channels 80-200, the majority of the variation is due to chemical variation, whereas MSC and SNV assume that the majority of the variation in **X** is due to scattering.

The simulated spectra for Dataset 2 and the noise-free spectra are shown in Figure 2.3A and 2.3B. The only feature of Dataset 2 that was different from Dataset 1 was that in Dataset 2 the standard deviations of the pure concentrations were lower. The amount of chemical variation was extremely low for Dataset 2, and as a result the assumptions of SNV and MSC were not strongly violated.

The spectra for simulated Dataset 3 are depicted in Figure 2.4A, and the noise-free

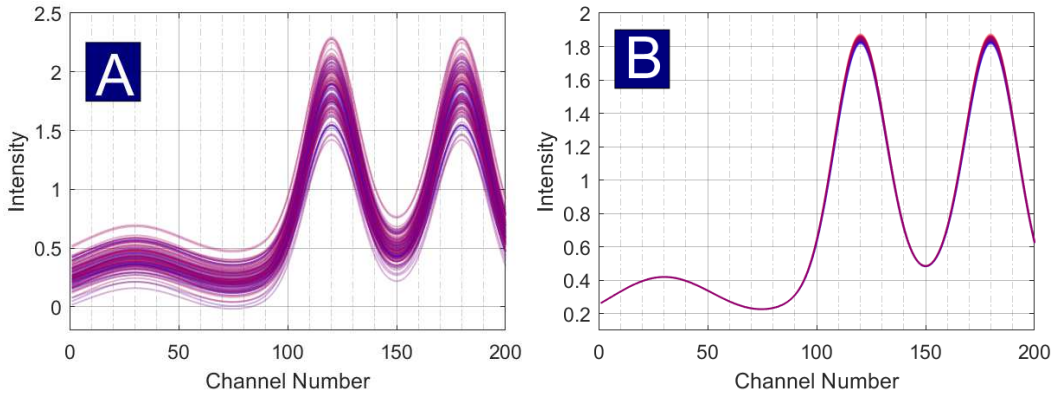

Figure 2.3: Simulated Spectra for Dataset 2. (A) Simulated spectra $\mathbf{X}$; (B) Noise-free spectra $\mathbf{X}_{chem}$.

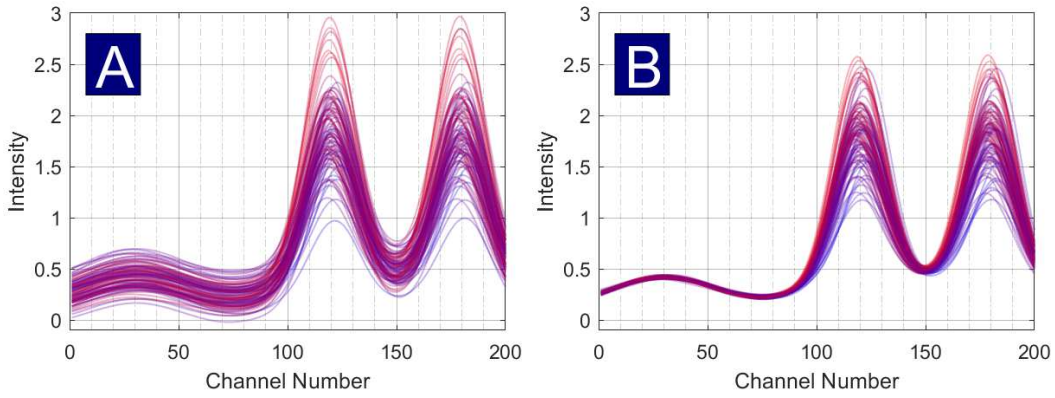

Figure 2.4: Simulated Spectra for Dataset 3. (A) Simulated spectra $\mathbf{X}$; (B) Noise-free spectra $\mathbf{X}_{chem}$.

spectra for Dataset 3 are shown in Figure 2.4B. The only differences between Dataset 3 and Dataset 1 were that for Dataset 3 there was more chemical background variation, as can be seen from the variation in channels 1-80 for the noise-free spectra of Dataset 3 in Figure 2.4B. The increased level of chemical background signal was used to assess how much the background signal would introduce error in the corrections for both weighted and unweighted scatter correction methods.

2.3.2 NIR Wine Must Data

The dataset of NIR wine must data was obtained courtesy of Jean-Michel Roger (IRSTEA, France), and was used in the VSN paper [31]. NIR spectra of 621 samples of wine musts were measured using a double beam JASCO V560 NIR spectrometer. One beam passed through a 1 mm cell filled with water (as a reference), and the other

beam passed through a 1 mm cell filled with must. The absorbance was recorded on 750 wavelengths, located between 800 nm and 2298 nm in 2 nm intervals. The property of interest (the **y** variable) was the alcohol by volume (ABV) content of the musts. The wine must samples were measured during the entire wine making process, including during the beginning of fermentation (when the must is grape juice, and therefore has an alcohol by volume content of 0). As a result, there were 158 samples in the dataset with **y** values of 0.

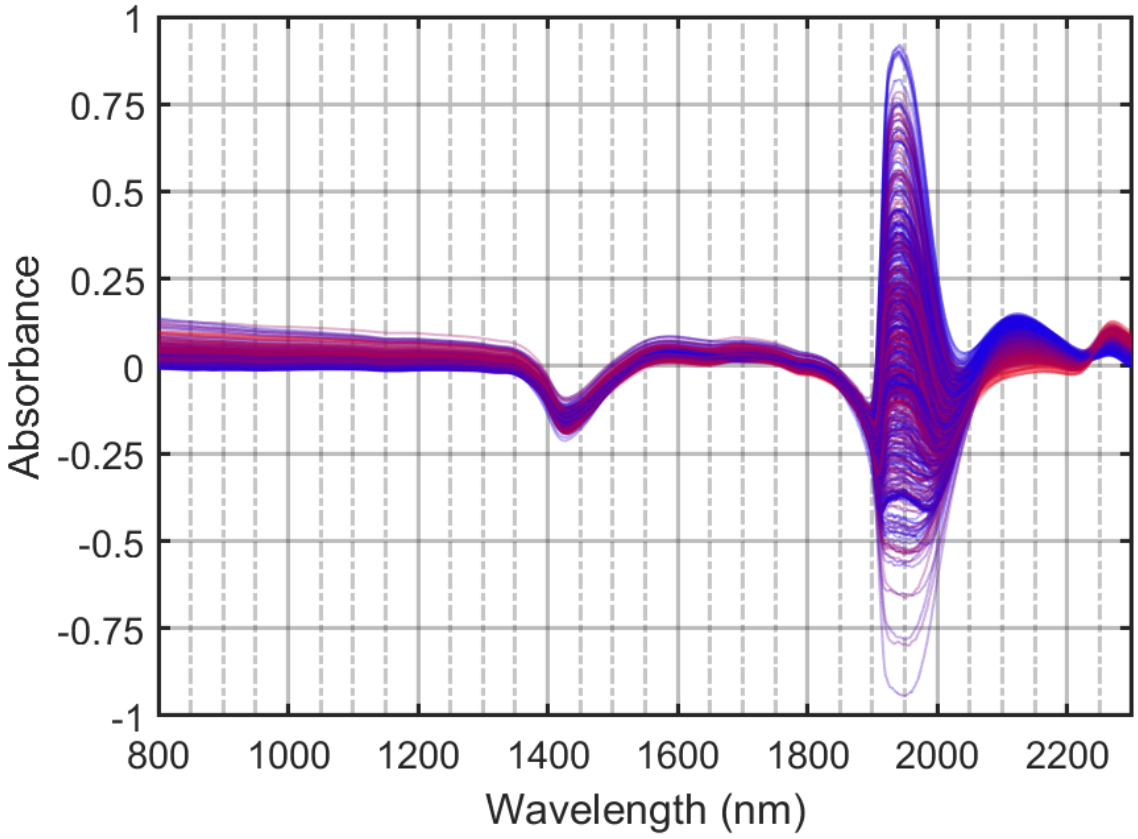

Figure 2.5: NIR spectra of wine musts. Samples are color-coded according to the alcohol by volume (ABV) content such that samples with low ABV values are blue and samples that have high ABV content are red.

The NIR spectra of the wine musts are depicted in Figure 2.5. The spectra of the musts exhibit a baseline due to the turbidity of the musts [31]. Between 1900 nm and 2050 nm, where the major absorption band of water is located, there is a significant amount of chemical variation, In the region between 2050 nm and 2298 nm, there is variation due to ethanol absorption, which can clearly be observed when the samples

are color-coded based upon the alcohol by volume (ABV) content.

2.4 Methods

2.4.1 Simulated Data

All simulations were carried out using MATLAB version R2017b (Mathworks, Natick, MA, USA). The code for the variable selection for normalization (VSN) was obtained courtesy of Jean-Michel Roger (IRSTEA, Montpellier, France). Small modifications to the code were made, including a change that allows the user to use a random fraction of all spectral couples, whereas the original code uses all spectral couples. The code for IDRC was obtained courtesy of Yifan Wu (Chinese Academy of Sciences, Beijing, China).

For VSN, 25 values of the threshold parameter were calculated. The values of the thresholds ranged from 10^{-5} to 10^{0}, and used a sequence which sampled values in the "middle" of the logarithmic range (10^{-3} to 10^{-2}) more heavily than values closer to the extremes, (see Figure 2.6 for an example of the threshold sequence). The threshold selection criterion employed the threshold with the largest standard deviation of the weights. Typical weight values for each of the three simulated datasets are shown in Figure 2.7 Using the optimal VSN weights, both the SSNV and SMSC corrections were applied to the spectra. Additionally, as a reference, a Binary SSNV correction was calculated, with weights of 1 for channels 1-75 (where there is no absorbance from the main chemical sources), and weights of 0 for channels 76-200 (where there is variance from the main chemical sources). The Binary SSNV weight vector is also shown in Figure 2.7. The Binary SSNV represents somewhat of a best-case scenario for weighted SNV, since the scatter-dominant regions and chemical-dominant regions are "known".

For IDRC, the spectra were evaluated for $k = 1$ to $k = 20$ regions. The total number of regions evaluated was 210. PLS models with 25 components were calculated, and five-fold cross validation with random splits was used to calculate the root mean squared error of cross validation (RMSECV). To determine the optimal number of PLS components to select, an F-test [50] based upon the ratio of $RMSECV^2/min(RMSECV^2)$, with significance level $\alpha = 0.25$ was performed. The

computational parameters (five-fold cross validation and F-test) were suggested by the authors of the IDRC method [32].

In addition to the VSN-based methods and IDRC, models were also created using standard SNV and MSC, and using the original spectra (with no preprocessing) for comparison.

The objective of the analysis was to assess how each preprocessing method affected the prediction errors from multivariate calibration models. The challenge for such a study is that the prediction errors are affected not just by the preprocessing technique, as the prediction errors will also vary based upon the realization of the samples/noise (due to statistical sampling error). The simulations were designed with the intent of minimizing the prediction errors due to sampling effects, such that a realistic estimate of the prediction errors based upon the preprocessing technique could be obtained.

Each dataset was split into 3 different sets: a training set with 100 samples, a calibration set with 1000 samples, and a test (prediction) set with 1000 samples. The parameters for the variable selection and weighting algorithms were determined using the samples from the training set. Using the parameters that were found for the training set, the corrections were applied to the calibration and test sets. To choose the optimal number of PLS components, a 50/50 split of the calibration set was used, and a 25-component PLS model was calculated using the first 500 samples of the calibration set, and was used to predict the concentrations of samples 501-1000 of the calibration set. In each PLS regression, the spectra were mean-centered, relative to the mean of samples 1-500 of the calibration set. The root mean squared error of validation ($RMSEV$) was calculated using the prediction errors from samples 501-1000 of the calibration set. The number of components which resulted in the minimum $RMSEV$ were selected. A PLS model was built using the selected number of components and all 1000 calibration set samples, and predictions were made using the test set samples, to obtain the root mean squared error in prediction (RMSEP_{Test}).

To assess the stability and robustness of both the preprocessing methods and the prediction errors, and to obtain an estimate of the uncertainty in the prediction errors, a Monte Carlo method was used. One hundred different realizations of the data matrix $\mathbf{X}$ and the concentrations matrix $\mathbf{Y}$ were generated, with the same parameters used to create each realization. For each realization, the preprocessing parameters

were calculated, and the prediction errors were calculated.

2.4.2 Wine Must Data

For the analysis of the Wine Must Dataset, the dataset was split into a training (calibration) set of 414 samples (2/3), and a test (prediction) set of 207 samples (1/3) using the Duplex algorithm [51]. For VSN, 25 values of the threshold parameter were calculated, using the same methodology as described above for the simulated data. For IDRC, the spectra were evaluated for $k = 1$ to $k = 50$ regions for PLS models with 25 components, using five-fold cross validation and an F-test with significance level $\alpha = 0.25$, to find the region which resulted in the lowest RMSECV.

The number of components for the PLS regressions for each variable selection method were chosen using cross validation (CV) with 10 random 50/50 splits of the training set. For each of the 50/50 splits, the root mean squared error of cross validation (RMSECV) was calculated. The number of components was selected based upon an F-test based upon the RMSECV, with significance level $\alpha = 0.25$ was performed, as was done with IDRC. For both IDRC and for the Wine Must data, using the F-test will result in fewer components selected than selecting the number of components that results in the minimum RMSECV. Cross validation can result in overfitting if too many components are used.

2.5 Results and Discussion

For the figures depicting the simulated spectra, the color of each spectrum corresponds to the concentration of component 1 of ($\mathbf{y}_1$). The values of $\mathbf{y}_1$ were scaled so that $min(\mathbf{y}_{1,scaled}) = 0.1$ and $max(\mathbf{y}_{1,scaled}) = 0.95$, and then the RGB (red, green, and blue) values (on a 0-1 scale) were set equal to $\left[\mathbf{y}_{1,scaled}, 0, (1-\mathbf{y}_{1,scaled})\right]$. Spectra with larger values of $\mathbf{y}_1$ had higher red coloration, and the spectra with the lowest values of $\mathbf{y}_1$ had more blue coloration. This color-coding scheme enables the viewer to more readily discern the chemical information in the spectra.

The same color scheme described above was also applied to the Wine Must Dataset, such that the samples were color coded based upon the $\mathbf{y}$ values (percent alcohol by volume, ABV), such that red colors corresponded with high ABV content, and blue corresponded with low ABV content.

Figure 2.6 shows the standard deviation of the VSN weights versus the threshold value for typical realizations of the three simulated datasets. Comparing these results for the three datasets, the standard deviation of weights was larger for Dataset 1 than for Datasets 2 and 3, and the optimal threshold (the threshold which resulted in the largest standard deviation of the weights) for Dataset 3 was the largest (0.0155), while the optimal threshold for Dataset 2 was the smallest (4.78×10^{-4}). The threshold value which resulted in the largest standard deviation of weights for Dataset 1 was 7.79×10^{-4}. The similarity of the optimal thresholds for Datasets 1 and 2 is not surprising since they have the same levels of background variation. In all three cases, the threshold values are reasonably close to the background signal standard deviation given in Table 2.1 (0.00021 for Datasets 1 and 2, 0.0105 for Dataset 3).

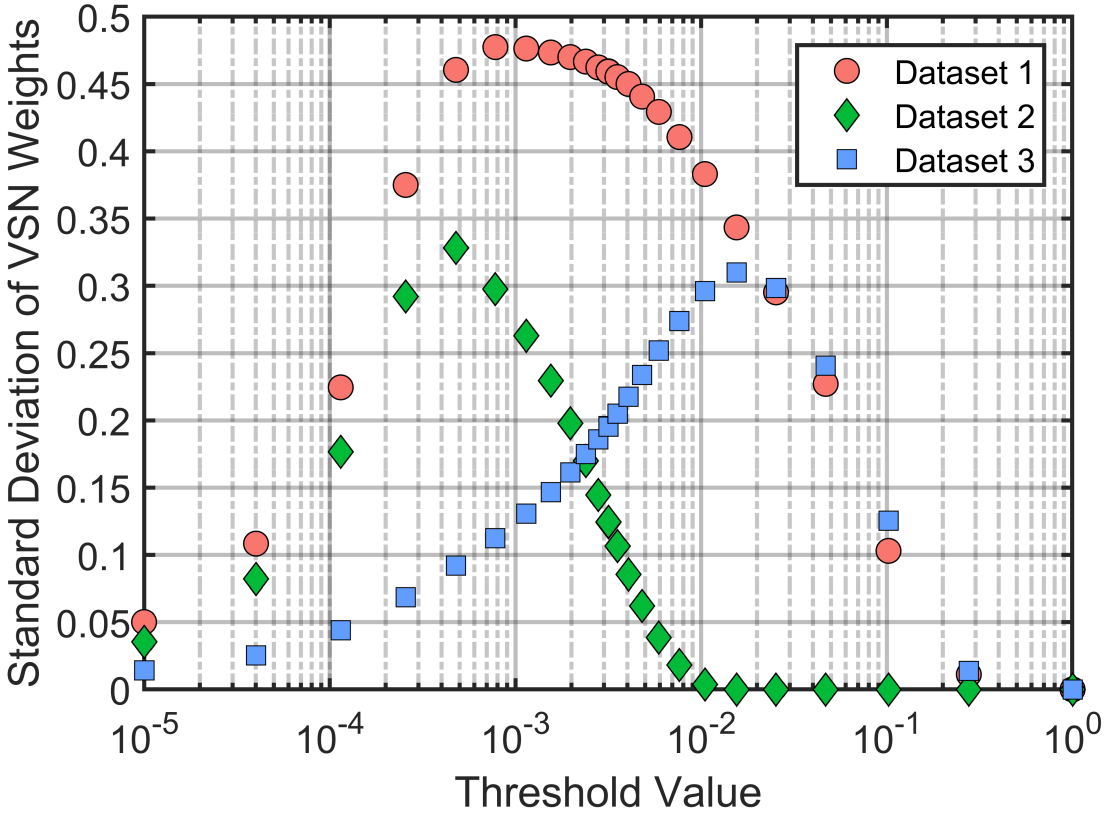

Figure 2.6: Results of the weighting procedure for VSN for each of the three simulated datasets, showing standard deviation of the VSN weights for each threshold value.

Typical weight vectors for VSN which correspond to the optimal thresholds for each simulated dataset, and the binary weight vector that was used, are plotted in

Figure 2.7. For Dataset 1, in the scatter-dominant region (channels 1-80), the weights for VSN were close to unity, while in the chemical-dominant region (channels 100-200), the weights were close to zero. The weights for Dataset 1 were similar to the binary weights. The weight vectors for Datasets 2 and 3 were similar to one another, despite the threshold differences, and are characterized by lower weights (circa 0.9) in the background region and higher weights in the analyte region. In particular, there is an increase in weights around channel 150, which corresponds to the valley between peaks in the analyte and interferent spectra, where the signal will be closer to background. Compared to Dataset 1, this region is more likely to be detected as background for Dataset 2 because of the lower analyte variability and for Dataset 3 because of a higher background signal. These weight vectors are somewhat less than ideal compared to Dataset 1, as anticipated.

It should be noted that in the course of many Monte Carlo simulations using different realizations of the parameters, the optimal weight vectors will change somewhat, but Figures 2.6 and 2.7 are typical representations of the characteristics observed.

2.5.1 Simulated Dataset 1

Dataset 1 was designed to have low background signals in the scatter dominant region and high analyte signals. Therefore, it represents the ideal case for weighted scatter correction methods and the case in which whole spectrum methods are most likely to fail.

The corrected spectra obtained by using SNV and MSC are shown in the top row of Figure 2.8. The spectra that were processed by SNV and MSC both exhibit severe distortion, as the variation in the peaks centered at channels 120 and 180 was much smaller than the variation in that region for the noise-free spectra (see Fig. 2.1E), and in the region from channels 1-90, artificial correlation with $\mathbf{y}$ were introduced as a result of the corrections. The reason why SNV and MSC resulted in such problematic corrections was that there was a significant contribution from chemical signals in the calculation of the scattering coefficients.

Using the optimal VSN weights for Dataset 1, the SSNV (Fig. 2.8C) and SMSC

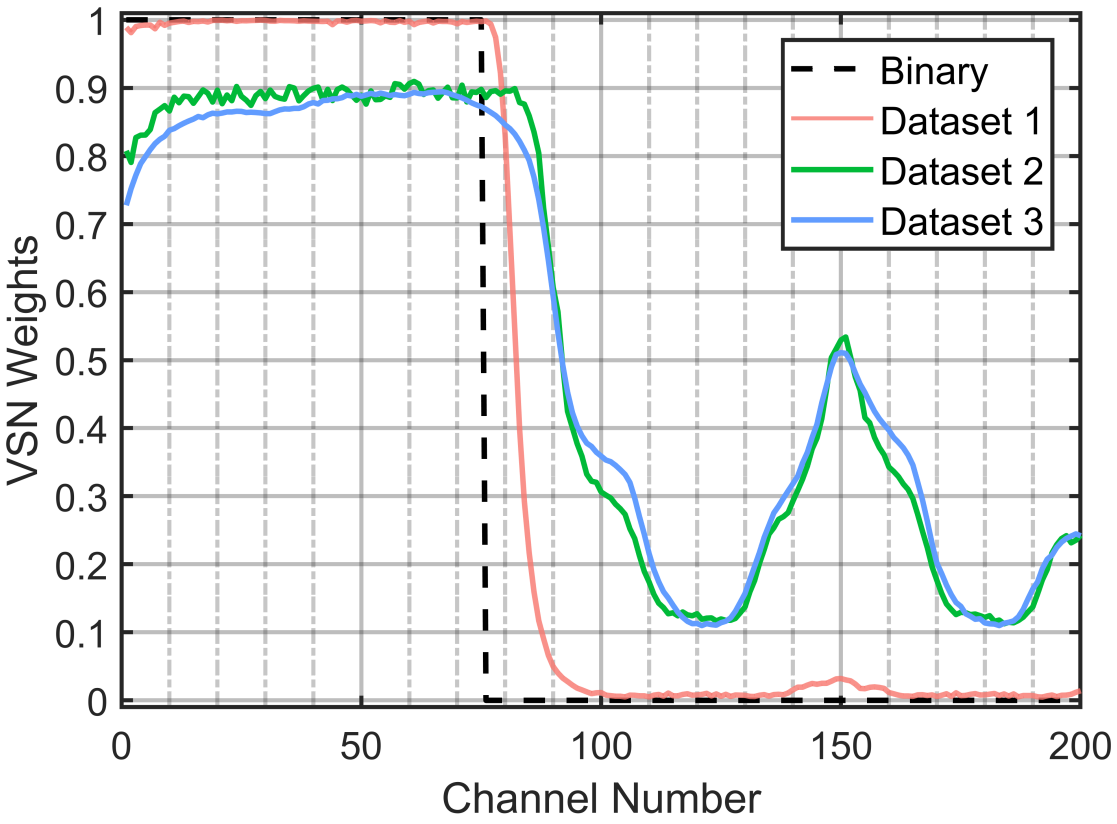

Figure 2.7: Binary VSN weight vector (black), as well as optimal VSN weight vectors for each simulated dataset.

(Fig. 2.8D) methods were used to correct the spectra. The spectra that were processed by SSNV appeared to be similar to the Binary SSNV corrected spectra, although in Fig. 2.8C it can be observed that there is still some variation present in the baseline region, which indicates that the SSNV correction was not quite as accurate at reproducing the behavior of the noise-free spectra. In contrast to SSNV, the spectra processed by SMSC looked very similar to the spectra obtained by using SNV and MSC, exhibiting a reduced amount of variation in the main chemical peak regions when compared with the noise-free spectra. Thus, it appears that SMSC is not as effective at correcting for scatter as SSNV, even though the same weights were used.

The IDRC method resulted in the selection of channels 34-44 for this realization of the data. For the optimal IDRC results, the RMSECV was 0.0914, and 2 latent variables were used in the training phase. For this dataset, the spectra obtained

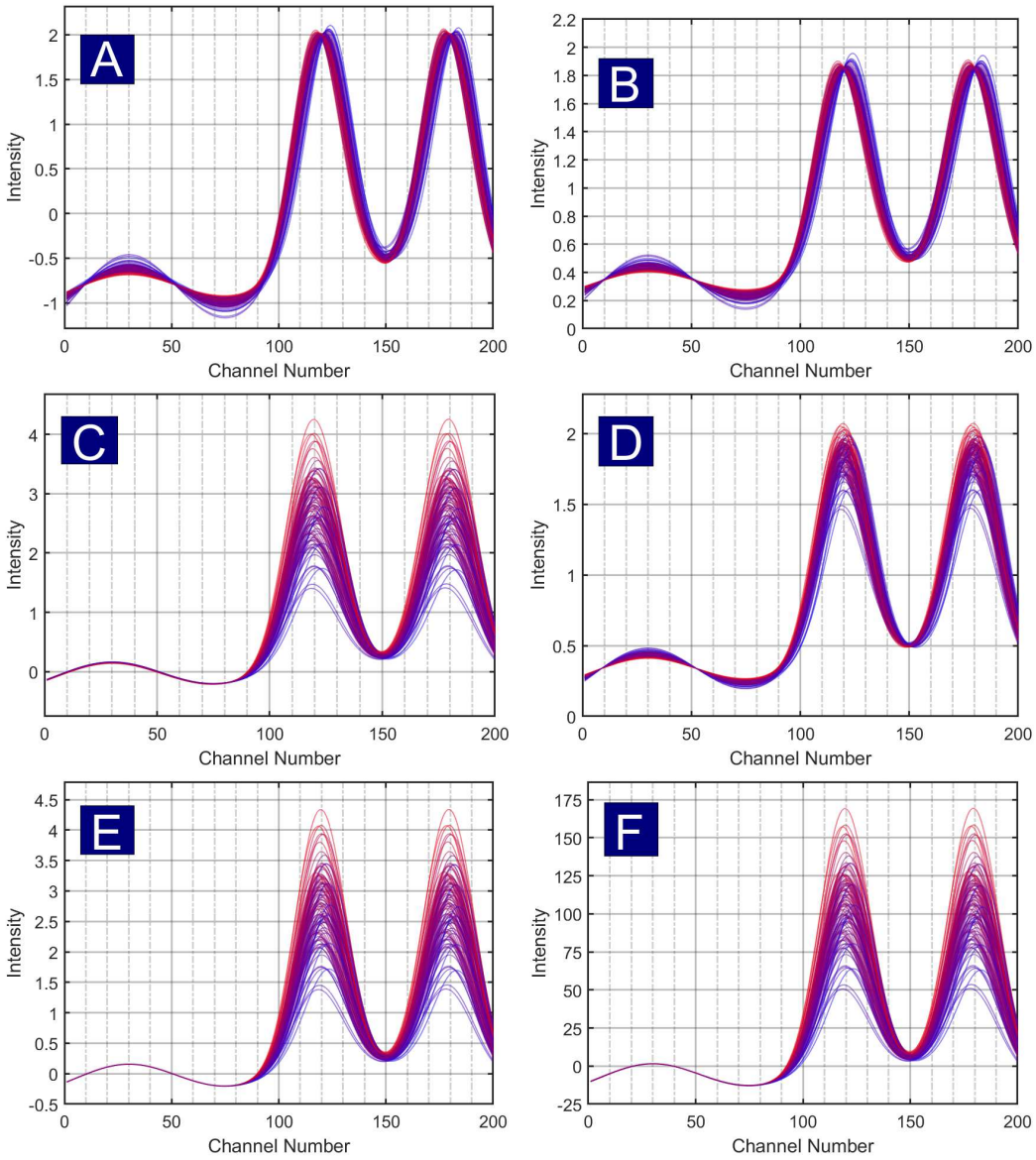

Figure 2.8: Results for the preprocessed spectra for Dataset 1. (A) SNV; (B) MSC; (C) SSNV; (D) SMSC; (E) Binary SSNV; (F) IDRC.

with IDRC were very similar to the spectra that were obtained by using the ideal Binary SSNV. Both the Binary SSNV and IDRC were able to adequately correct for the scattering effects, as can be seen by comparison of the Binary SSNV and IDRC-processed spectra (Fig. 2.8E and 2.8F, respectively) with the noise-free spectra in Fig 2.1E.

In the spectra that were processed using SMSC, information from the CDR leaked into the other regions of the spectrum, as can be seen by the fact that there are clear

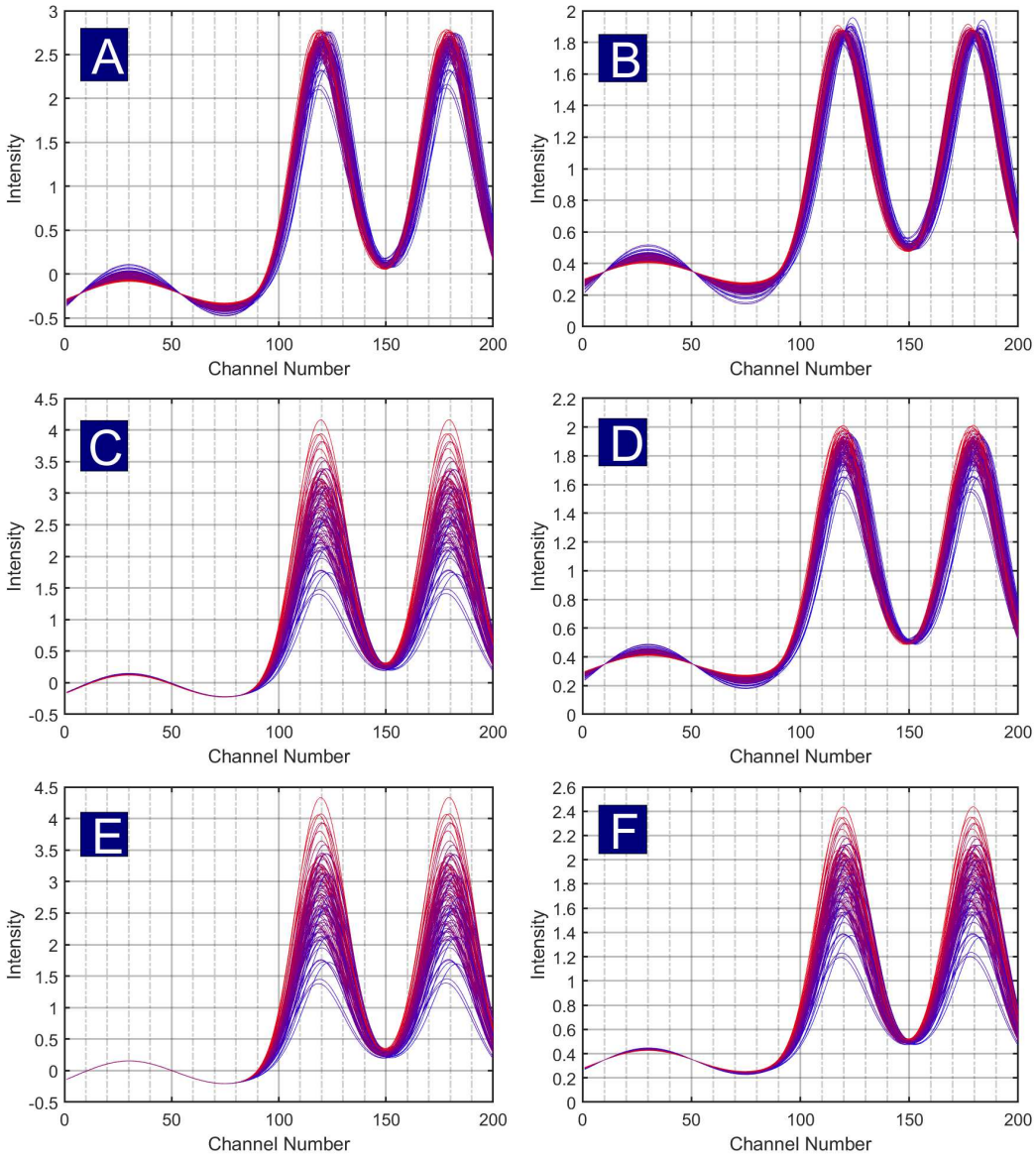

Figure 2.9: Effects of VSN weights in simulated Dataset 1. Binary weights, $[w_1, w_2]$ were applied to channels 1-75 (w_1) and channels 76-200 (w_2). (A) SSNV with weights=[1,0.10]; (B) SMSC with weights=[1,0.10]; (C) SSNV with weights=[1,0.01]; (D) SMSC with weights=[1,0.01]; (E) SSNV with weights=[1,0.001]; (F) SMSC with weights=[1,0.001].

patterns of red and blue in the region from channels 1-80, which should not have any chemical information. Since the information leaked from the chemical dominant region to the scatter-dominant region, it can be concluded that even the small weights in the CDR were large enough to have a significant effect on the scatter correction.

To test the effect of the magnitude of weights in the chemical-dominant region

on the SSNV and SMSC processed spectra, a series of tests were performed using a scheme similar to that used for binary weighted SNV. Using weight vectors that consisted of weights of 1 for channels 1-75 (the SDR) and weights of either 0.100, 0.010, or 0.001 for channels 76-200 (the CDR), SSNV and SMSC corrections were calculated for each possible case. The results of the SSNV and SMSC corrections for each level of weighting are shown in Figure 2.9. When the weights in the CDR were 0.10, both SSNV (Fig. 2.9A) and SMSC (Fig. 2.9B) experienced significant distortion, and appeared somewhat similar to the spectra obtained by using conventional SNV and MSC. When the weights in the CDR were 0.01, using SSNV (Fig. 2.9C) resulted in spectra which adequately corrected for the scattering effects, although a small amount of baseline variation is evident. The shape of the spectra that resulted from using SMSC (Fig. 2.9D) with weights of 0.01 in the CDR were somewhat improved from when the weights were 0.10, with reduced signal variation in the SDR, and less information loss in the chemical region, but the spectra were still not quite ideal in their behavior. When the weights in the chemical-dominant region were 0.001, the SSNV-processed spectra (Fig. 2.9E) were nearly perfect at correcting the scatter, and to the naked eye appear to be the same as the Binary SSNV spectra in Fig 2.8C. For the SMSC-processed (Fig. 2.9F) spectra which used weights of 0.001 in the CDR, it can be observed that the scatter correction was still imperfect in the scatter-dominant region although much improved. At each weight level, the SSNV correction was better able to reproduce the behavior of the noise-free signal than SMSC.

Overall, the results of the tests that are displayed in Figure 2.9 indicate that both SSNV and SMSC can be significantly affected by small weights in regions where significant chemical absorbance occurs, with SMSC being more sensitive to this effect than SSNV. The impact of the small weights on each method can be understood by consideration of the equations used to calculate SSNV and SMSC. For SSNV, the contribution of channel j to the weighted mean is equal to $w_j\mathbf{x}_j$, and for the weighted standard deviation the contribution is equal to $w_j\left(\mathbf{x}_j - \bar{\mathbf{x}}_w\right)^2$. As a result, the contribution of the impact of spectral channel j to the model is approximately linearly proportional to the value of the weight w_j. As far as the contribution to the weighted mean is concerned, a channel with a measured absorbance of 1 and a weight of 0.1 is equivalent to a channel with a measured absorbance of 0.1 with a weight of

Table 2.2: PLS regression results for the various scatter correction methods applied to simulated Dataset 1.

Method	RMSEV[a]	RMSEP^{b}_{Test}	NLV[c]
None	0.1413 (0.0062)	0.1417 (0.0044)	4.03 (0.17)
SNV	0.287 (0.028)	0.288 (0.030)	2.76 (0.82)
MSC	0.293 (0.031)	0.294 (0.035)	2.16 (0.66)
Binary SSNV	0.1006 (0.0035)	0.1006 (0.0019)	2.08 (0.27)
SSNV	0.1011 (0.0036)	0.1012 (0.0020)	2.60 (0.70)
SMSC	0.200 (0.016)	0.200 (0.013)	1.98 (0.43)
IDRC	0.1008 (0.0037)	0.1009 (0.0020)	2.92 (0.58)

Average values from 100 realizations of Monte Carlo Procedure (standard deviations in parentheses)
[a] RMSEV=Root mean squared error from calibration
[b] RMSEP_{Test}=Root mean squared error of prediction for test set
[c] NLV=number of latent variables

1. For SMSC, the picture is somewhat more complex, as SMSC uses a weighted least squares projection to solve for the scatter parameters.

The PLS regression results from the Monte Carlo procedure for Dataset 1 are summarized in Table 2.2. For each method, the difference between the average $RMSEV$ and the average $RMSEP_{Test}$ was negligible. The smallest prediction errors were obtained by the models which used Binary SSNV, IDRC, and SSNV, which resulted in nearly the same average $RMSEP_{Test}$ (to within their respective uncertainties) of 0.1006 (Binary SSNV) to 0.1012 (SSNV). These prediction errors were nearly as low as could be obtained, as the the error in $\mathbf{y}$ set a floor of 0.10 for the RMSEP. The data that were not preprocessed resulted in an average $RMSEP_{Test}$ of 0.1417. SMSC led to a larger average $RMSEP_{Test}$ than the non-preprocessed data, with an average $RMSEP_{Test}$ of 0.200. SNV and MSC had the largest prediction errors, with average $RMSEP_{Test}$ values of 0.288 and 0.294, respectively. For each of the methods tested, the number of latent variables used was very small, averaging between 2 and 4 components. Due to the small level of the chemical background variation relative to both the variation from the main chemical components, and also to the *iid* noise.

In Figure 2.10, the Monte Carlo procedure results for the variable selection and weighting for IDRC and VSN are shown. For IDRC, for each run of the Monte Carlo procedure, the optimal selected variables were stored. The indices of the selected variables for IDRC for each of the 100 trials are shown in Fig. 2.10A. The IDRC

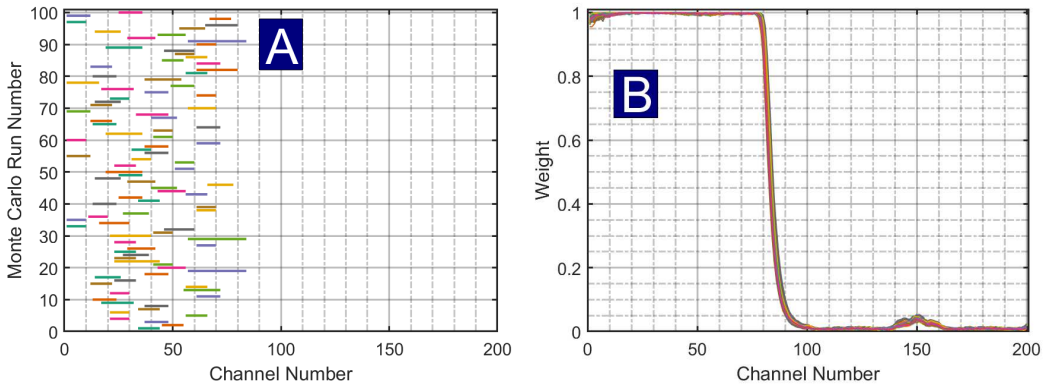

Figure 2.10: Monte Carlo results for IDRC and VSN for Dataset 1. (A) Indices of the variables selected by IDRC for each Monte Carlo run are shown using horizontal lines; (B) Optimal VSN weight vectors for each run.

method did not consistently select the same regions each time, but for each trial the selected variables were always in the scatter dominant region (channels 1-83). For VSN, the optimal weight vectors for each of the 100 trials are shown in Fig. 2.10B. The optimal VSN weights did not differ much from one trial to another. While both methods led to accurate predictions, the VSN results were more consistently reproducible for the conditions present for simulated Dataset 1.

The results for Dataset 1 were largely as anticipated, with the weighted methods performing significantly better than the unweighted methods, with the exception of SMSC. The standard correction methods (SNV and MSC) showed poorer prediction than no preprocessing at all, which highlights the need for proper preprocessing.

2.5.2 Simulated Dataset 2

In contrast to Dataset 1, Dataset 2 was designed to represent a case where the analyte signal approached that of the background and was lower than the effects of scattering. Under these conditions, it was expected that traditional scatter correction methods should be effective.

The spectra for Dataset 2 that were corrected by each of the preprocessing methods are shown in Figure 2.11. Overall, the preprocessed spectra for each method were very similar, as nearly all of the scatter was successfully removed.

The PLS regression results from the Monte Carlo procedure are summarized in Table 2.3. The scatter correction methods (SNV, MSC, Binary SSNV, SMSC, and

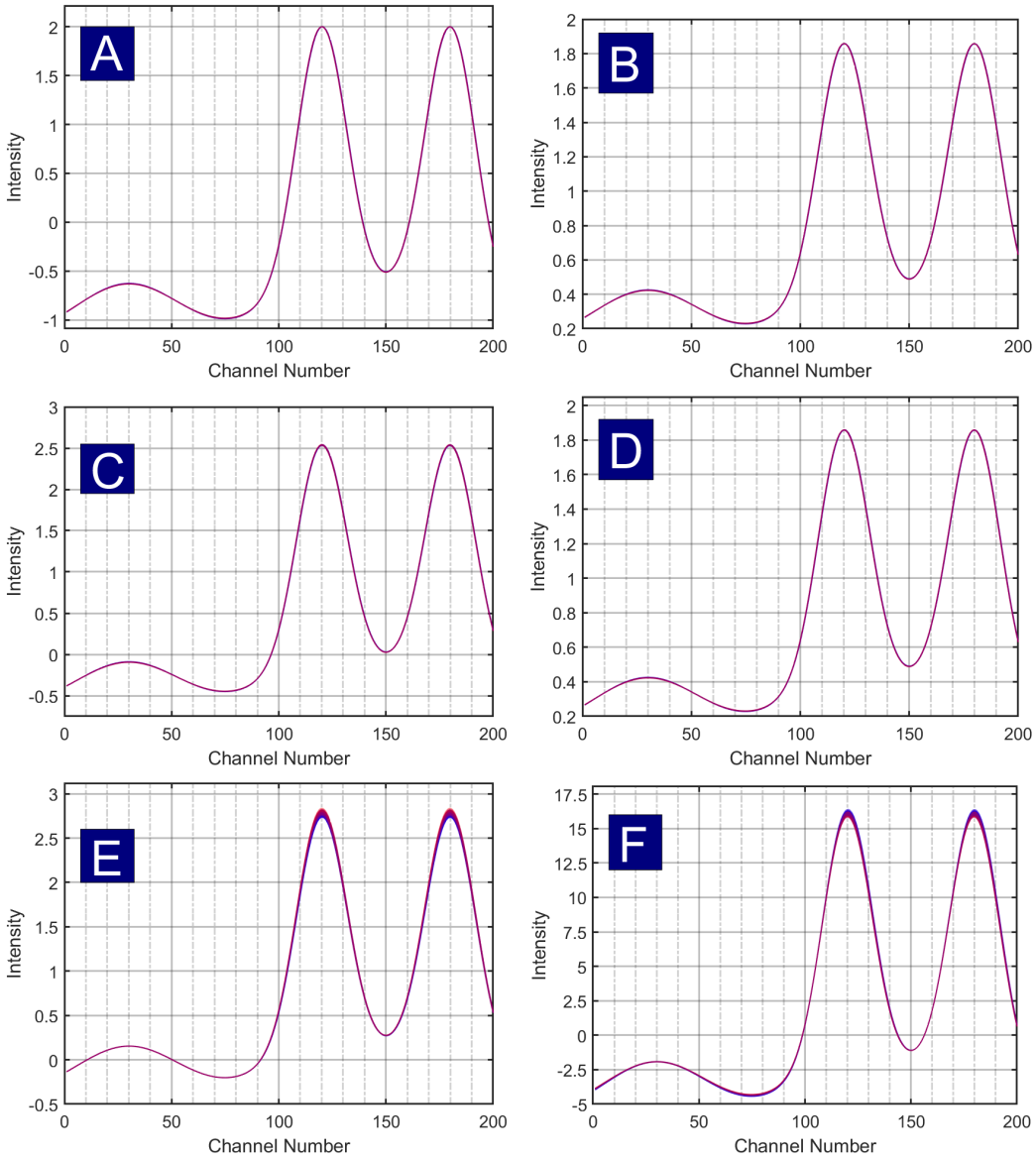

Figure 2.11: Results for the preprocessed spectra for Dataset 2. (A) SNV; (B) MSC; (C) SSNV; (D) SMSC; (E) Binary SSNV; (F) IDRC.

IDRC) all resulted in nearly identical average prediction errors, both for the calibration set ($RMSEV$) and the test set ($RMSEP_{Test}$), with average $RMSEP_{Test}$ values ranging between 0.1611 and 0.1618. As a result, it can be concluded that for Dataset 2, the weighted scatter correction methods did not result in any significant improvements in the prediction performance over conventional SNV and MSC. In contrast, the results with no preprocessing error had a prediction uncertainty of about 0.19. It will also be noted that the number of latent variables used for the PLS models

Table 2.3: PLS regression results for the various scatter correction methods applied to simulated Dataset 2.

Method	RMSEV[a,b]	RMSEP$_{Test}$[a,c]	NLV[a,d]
None	0.1908 (0.0062)	0.1904 (0.0046)	7.1 (1.1)
SNV	0.1617 (0.0049)	0.1615 (0.0041)	5.6 (1.1)
MSC	0.1617 (0.0049)	0.1615 (0.0041)	5.6 (1.1)
Binary SSNV	0.1614 (0.0051)	0.1611 (0.0040)	6.4 (1.1)
SSNV	0.1616 (0.0050)	0.1615 (0.0041)	5.7 (1.1)
SMSC	0.1616 (0.0049)	0.1616 (0.0042)	5.4 (1.1)
IDRC	0.1617 (0.0049)	0.1618 (0.0042)	6.4 (1.2)

[a] Averages from Monte Carlo procedure
(standard deviations in parentheses)
[b] NLV=number of latent variables

was higher in all cases for Dataset 2 compared with the PLS models for Dataset 1. This is expected since the chemical background poses a larger interference with the smaller analyte signal and requires more components to account for the effects of the background.

For the VSN results for Dataset 2, the effectiveness of some of the assumptions of the VSN algorithm were also tested. The optimal VSN weights were chosen based upon the standard deviation of the weights. In Figure 2.12A, the standard deviation of the VSN weights are plotted for the various values of the threshold parameter, and the points are plotted using a rainbow-color palette that goes from red at the smallest threshold tested, to blue at the largest threshold tested. In Figure 2.12B, the VSN weight vectors for various threshold values are shown, with the colors of the lines corresponding to the color that was plotted for each threshold value in Fig 2.12A. For the weight vectors that were calculated using the smallest threshold values (red and orange lines), the weights remained small throughout the spectral range, whereas for the larger values of the weights (cyan and blue lines), the weights were larger than 0.5 for all variables. At intermediate threshold values (yellow and green lines), the weight values were consistently large (above 0.9) across channels 1-90, while from channels 91-200 the weights varied significantly depending on the exact value of the threshold that was used. From channels 91-200, the smallest weights were typically at channels 120 and 180 (corresponding to analyte and interferent regions), while around channel 150 the weights were larger (in the valley of the analyte/interferent spectra).

Overall, the weight vectors which resulted in the largest standard deviations of the weights were those which resulted in a contrast in the weights (large weights in some regions, small weights in other regions), whereas the weight vectors which resulted in smaller standard deviations were the ones in which the weights were similar for all the variables. This suggests that using the standard deviation of the weights as a criterion to choose which VSN threshold to use is a reasonable choice.

To test the effects of how the number of sample pairs used (N_s) impact the VSN weights, the VSN weights for the optimal threshold value were calculated using all possible sample couples, and the weights were also calculated by using a random sample of 10%, and 1% of all sample pairs. The results of this test of the VSN weighting procedure are depicted in Figure 2.13. Since there were 100 samples in the training set, the number of combinations of sample pairs was $\big(100(100-1)\big)/2 = 4950$, so 10% of possible sample pairs was 495 pairings, and 1% was obtained from using 49 pairings. Overall, it appears that varying the fraction of sample pairs used does not have a significant impact on the overall trend of the VSN weights, although using a larger fraction of sample pairings does result in a somewhat smoother weight vector. Additionally, the computational times were calculated ten times for each of the three fractions of sample pairings (100%, 10%, 1%), and the average computation times were 17.5 seconds for 100% of sample pairs, 2.43 s for 10% of sample pairs, and 0.259 s for 1% of sample pairs. It can reasonably be concluded that using a larger fraction of sample pairs decreases the uncertainty in the weights due to the effects of statistical sampling. However, using a smaller fraction of weights could result in a more efficient algorithm, especially when the VSN weights are calculated for multiple values of the threshold parameter.

In Figure 2.14, the Monte Carlo procedure results for the variable selection and weighting for IDRC and VSN for Dataset 2 are shown. For IDRC, for each run of the Monte Carlo procedure, the optimal selected variables were stored. The indices of the selected variables for IDRC for each of the 100 trials are shown in Fig. 2.14A. The IDRC method was highly inconsistent in the variables which where selected. The three most common regions of the spectrum that were selected were channels 1-20, channels 80-120, and channels 180-200. The large variation in the selected regions may indicate that the choice of variables for IDRC did not significantly impact the

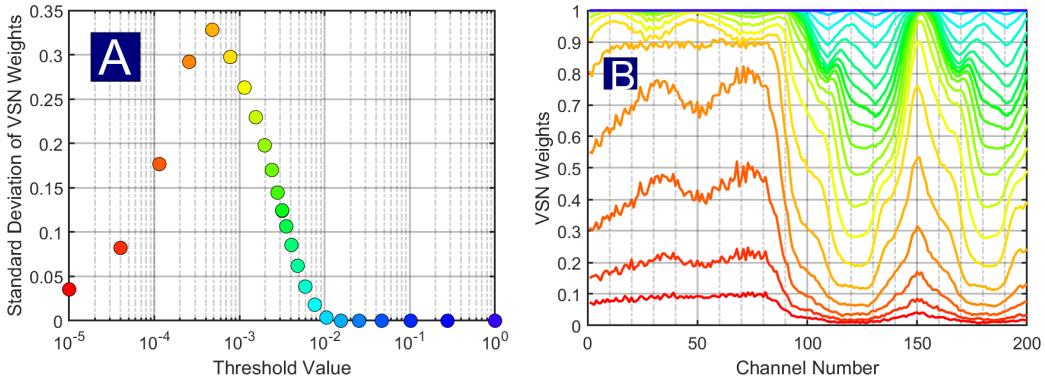

Figure 2.12: Results of the weighting procedure for VSN for simulated Dataset 2. (A) Standard deviations of VSN weights vs. threshold; (B) VSN weights for various threshold values.

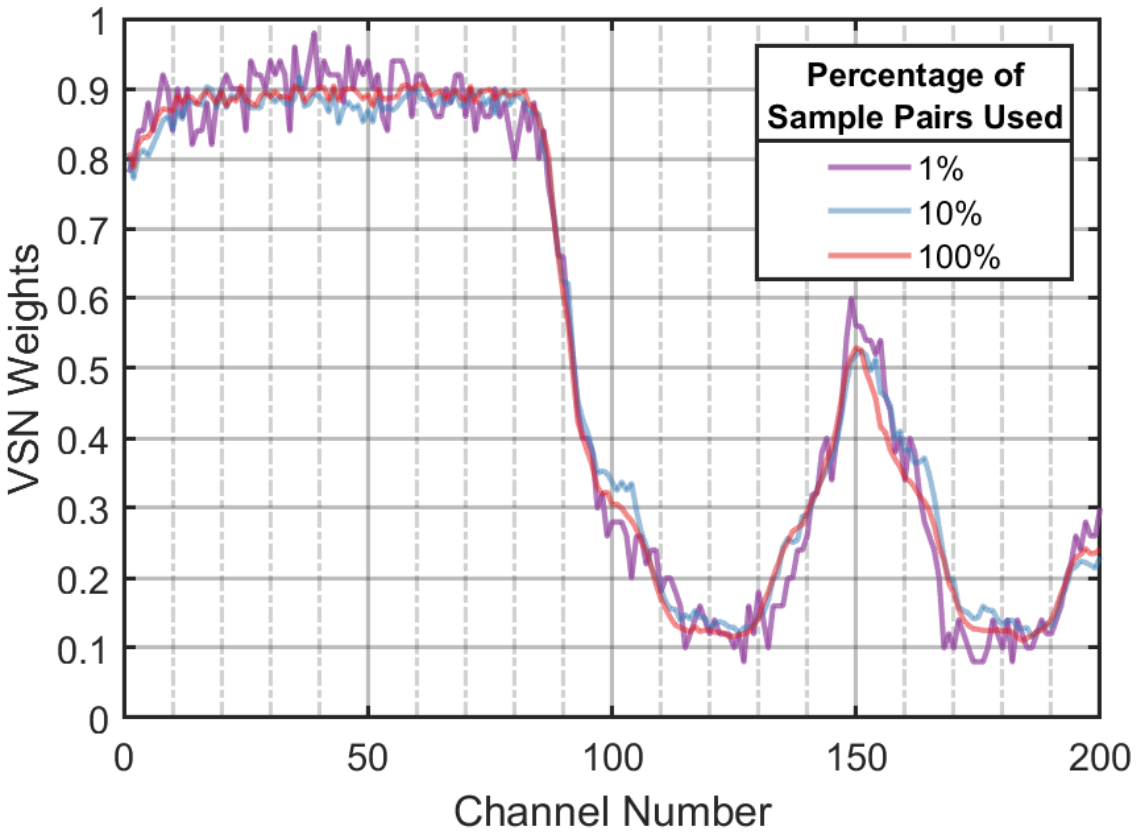

Figure 2.13: VSN weights for Dataset 2, calculated using different percentages of the number of sample couples (N_s).

prediction errors at the optimization phase, and as a result the variable selections were not very meaningful. For VSN, the optimal weight vectors for each of the 100 trials

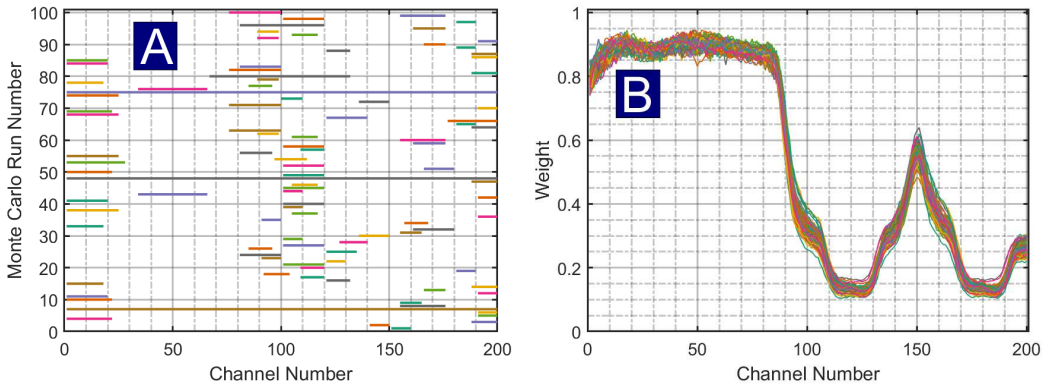

Figure 2.14: Monte Carlo results for IDRC and VSN for Dataset 2. (A)Variables selected by IDRC for each run; (B)Optimal VSN weight vectors for each run.

are shown in Fig. 2.14B. The optimal VSN weights did not differ much from one trial to another. For most of the wavelength channels, the range of VSN weights (largest weight minus smallest weight) was 0.1 or less. As was the case with Dataset 1, both VSN and IDRC methods led to accurate predictions, but the VSN weight vectors were consistently reproducible for the conditions present for simulated Dataset 2 while the IDRC variable selections were highly inconsistent.

Comparing the VSN weights for Dataset 2 with the VSN weights for Dataset 1, some interesting observations can be made. For Dataset 2, the weights were larger throughout the chemical-dominant region, while in the scatter-dominant region the weights were not as large in that region as compared with the VSN weights for Dataset 1. Based upon the results for Dataset 1, in which it was observed that the value of the weights in the chemical-dominant region can have a significant impact on the signal shape for both SSNV and SMSC, in might be expected that the weights obtained for Dataset 2 would result in a large amount of signal distortion. The fact that the weights for Dataset 2 resulted in an accurate correction suggest that the interaction of the signal characteristics and weight size matters. In Dataset 1, the chemical component of the signal was very large, which meant that the weights for variables with significant chemical signals need to be very small to counteract the influence of such variables. For Dataset 2, the correction was still effective because there was less of a chemical component of the signal.

The main conclusion that can be drawn from Dataset 2 is that, when traditional

scatter correction methods (SNV, MSC) are appropriate, there are no particular disadvantages to the application of weighted methods.

2.5.3 Simulated Dataset 3

Dataset 3 was similar to Dataset 1, except that the level of chemical background interference has been increased by a factor of 50. This makes it more difficult to distinguish variation from scatter from chemical variation, so scatter correction will be less reliable.

Figure 2.6, presented earlier, shows the standard deviation of the VSN weights versus the threshold value for Dataset 3. The threshold value which resulted in the largest standard deviation of weights was 2.53×10^{-2}, comparable to the background signal variation (0.011). The weight vector for the VSN function with optimal weighting is plotted in Figure 2.7. The smallest VSN weights were approximately 0.20, and the largest VSN weights were roughly 0.90.

The spectra that were processed by using SNV and MSC are shown in Fig. 2.15A and 2.15B, while the spectra that were processed by using SSNV and SMSC are shown in Fig. 2.15C and 2.15D. These four methods resulted in highly similar looking spectra. In each instance, the corrected spectra showed severe distortion of the original signal shape.

The spectra that were processed by using Binary SSNV are shown in Figure 2.15E. In the chemical-dominant region, the Binary SSNV appears at first glance to have preserved the general shape of the noise-free spectra. Upon closer examination and comparison with Fig. 2.4B, it can be observed that the samples which had the highest intensity for the Binary SSNV spectra do not agree very closely with the relative peak intensities of the noise-free spectra. Further, the patterns of signal shapes in the scatter-dominant region differ slightly between the Binary SSNV processed spectra and the noise-free spectra. Because the Binary SSNV represents the "ideal" case, this suggests that the problem arises with the background variability, and not with the implementation of the weighted scatter correction.

The spectra that were processed using IDRC used spectral channels 41-80 to perform the correction. For the optimal IDRC results, the RMSECV was 0.274, and 4 latent variables were used in the training phase. The IDRC-processed spectra,

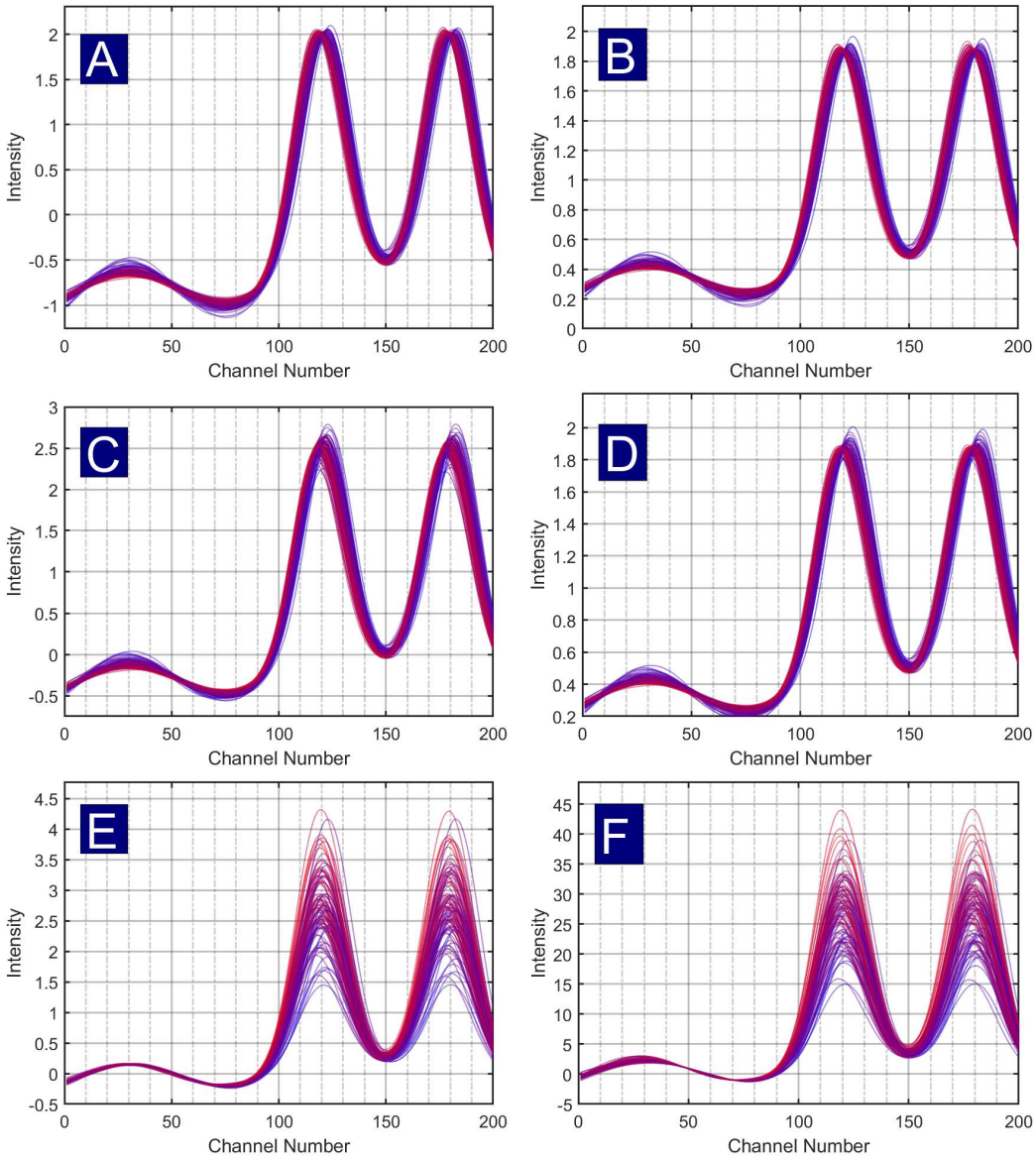

Figure 2.15: Results for the preprocessed spectra for Dataset 3. (A) SNV; (B) MSC; (C) SSNV; (D) SMSC; (E) Binary SSNV; (F) IDRC.

which are shown in Fig. 2.15F, were significantly better at preserving the signal shape of the noise-free data as compared with the results for SSNV and SMSC, but the IDRC-processed spectra did not perform quite as well as the spectra which were processed using Binary SSNV. Compared with the noise-free spectra, the IDRC-corrected spectra should have had relatively uniform levels of variation in channels 40-80, but instead the IDRC correction resulted in very small variance in channels 40-80, and larger variance in channels 1-40. This is no doubt a consequence of the

Table 2.4: PLS regression results for the various scatter correction methods applied to simulated Dataset 3.

Method	RMSEV[a]	RMSEP$^{a}_{Test}$	NLV[a,b]
None	0.209 (0.008)	0.209 (0.005)	19.9 (2.0)
SNV	0.326 (0.029)	0.329 (0.022)	15.3 (4.1)
MSC	0.330 (0.031)	0.335 (0.025)	14.3 (4.2)
Binary SSNV	0.211 (0.008)	0.211 (0.006)	19.2 (1.6)
SSNV	0.282 (0.014)	0.284 (0.010)	17.1 (3.1)
SMSC	0.322 (0.028)	0.326 (0.022)	14.8 (3.9)
IDRC	0.224 (0.015)	0.223 (0.012)	18.5 (1.4)

[a] Averages from Monte Carlo procedure
(standard deviations in parentheses)
[b] NLV=number of latent variables

region selected, since the variance in the selected region is minimized.

The PLS regression results from the Monte Carlo procedure for Dataset 3 are summarized in Table 2.4. For each method, the difference between the average $RMSEV$ and the average $RMSEP_{Test}$ was in the third decimal place. The lowest average $RMSEP_{Test}$ values were for the non-preprocessed spectra, and for the Binary SSNV-preprocessed spectra, for which the averages were 0.209 and 0.211, respectively. Because the analyte variation is relatively large, the results for no preprocessing are better than all of the other methods, suggesting that the distortion in the data by the other methods outweighs the errors introduced by scatter. Only the "ideal" Binary SSNV result approaches the raw data, indicating the deleterious effects that chemical background variation can have on scatter correction. The next lowest prediction errors were obtained by IDRC, which had an average $RMSEP_{Test}$ of 0.223, which is actually quite good under the circumstances. SMSC, SNV, and MSC produced the worst average prediction errors, with average $RMSEP_{Test}$ values of 0.326, 0.329 and 0.335. SSNV resulted in intermediate prediction performance, with an average $RMSEP_{Test}$ of 0.284. On average, the PLS models used 14-20 components. With the variation due to the various chemical background components, more components were needed to fully model the variation that was present.

In Figure 2.16, the Monte Carlo procedure results for the variable selection and weighting for IDRC and VSN are shown. The indices of the selected variables for

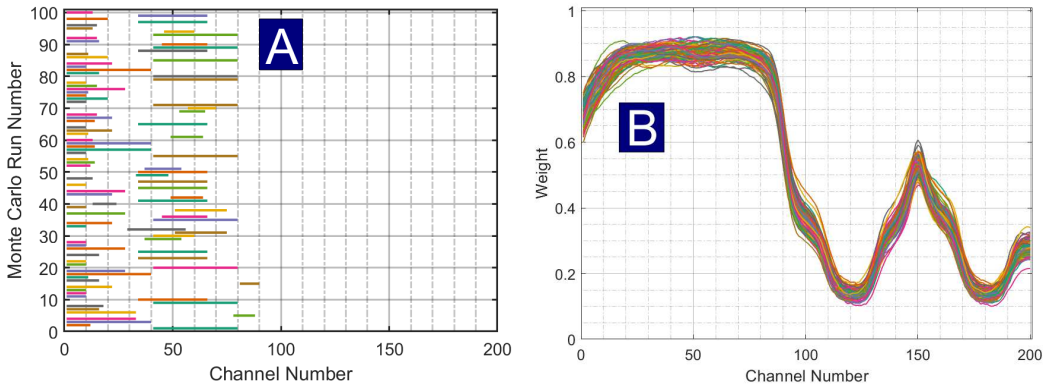

Figure 2.16: Monte Carlo results for IDRC and VSN for Dataset 3. (A)Variables selected by IDRC for each run; (B)Optimal VSN weight vectors for each run.

IDRC for each of the 100 trials are shown in Fig. 2.16A. The IDRC method consistently selected variables in one of two regions: between channels 1-20, and channels 40-70. The peak of the base spectrum was located at channel 30, and the peak was somewhat symmetric, so it appears that only the channels on one side of the peak or the other were needed. For VSN, the optimal weight vectors for each of the 100 trials are shown in Fig. 2.16B. Within each channel, the range of VSN weights was between 0.06 and 0.17, and there was greater variance in the weights than for Datasets 1 and 2.

Based upon the results for simulated Dataset 3, it can be concluded that when the chemical background variation exceeds a certain amount, then a weighted scatter correction approach becomes less effective. Weighted scatter correction is dependent upon a low background signal level, to accurately be able to estimate the scatter coefficients. As the level of background signal and/or noise from independent sources of error increase, the error in the correction due to the inaccuracy of the estimated scatter coefficients will increase as well. When the background signal variation exceeds a certain amount, any weighted scatter correction approach may be worse than doing no correction at all.

The weighted scatter correction methods did not lead to improvements in RMSEP compared with using the original spectra for Dataset 3 but they did lead to different levels of improvements over unweighted methods, with significant improvements for IDRC and only marginal improvements for VSN. This could have been predicted from merely looking at the VSN weight vector. Dataset 3 contained a large amount of

background signal, and as a result the variables in channels 1-90 were not completely dominated by scattering effects. This would explain why the optimal threshold for Dataset 3 was larger than the thresholds for Datasets 1 and 2, and why the VSN weights had a very low standard deviation. VSN could still be somewhat helpful as a diagnostic method, to check whether or not a set of spectra have consistent variables.

2.5.4 Wine Must Data

The Wine Must Dataset was included as an experimental example in this work since it was employed in the original development of the VSN method [31] and exhibited some of the characteristics that weighted scatter correction methods are intended to address, namely a high analyte signal variability and (presumed) scatter dominant region. Since all of the factors affecting real complex samples are difficult to include in simulations, the examination of experimental signals can provide some context. It also afforded the opportunity to evaluate IDRC and other modifications that were not included in the original paper.

The spectra for the Wine Must Data were presented earlier in Figure 2.5. Unlike the simulated data, the SDR, if it exists, was unknown, but it was presumed it could be located below 1900 nm, since longer wavelengths show clear chemical variation attributable to water and ethanol.

The spectra that were processed using SNV and MSC are shown in Fig. 2.17A and 2.17B. Based on visual inspection alone, it is clear that SNV failed to remove the scatter from the spectra, and instead the baseline variation was larger than it was in the raw data, which suggests that SNV failed to adequately correct the scatter in the data. The large peak located between 1900 nm and 2000 nm likely had a sizable influence in the calculation of the mean and standard deviation of each sample, which in turn likely corrupted the SNV correction parameters. The large baseline present throughout the MSC-corrected spectra suggests that MSC was also ineffective at correcting the scatter, likely for similar reasons.

Figure 2.18A shows the standard deviation of the VSN weights versus the threshold value. The threshold value which resulted in the largest standard deviation of weights was 1.97×10^{-3}. The weight vector for the VSN function with optimal weighting, and a "stepped" VSN weight vector, are plotted in Figure 2.18B. For the VSN

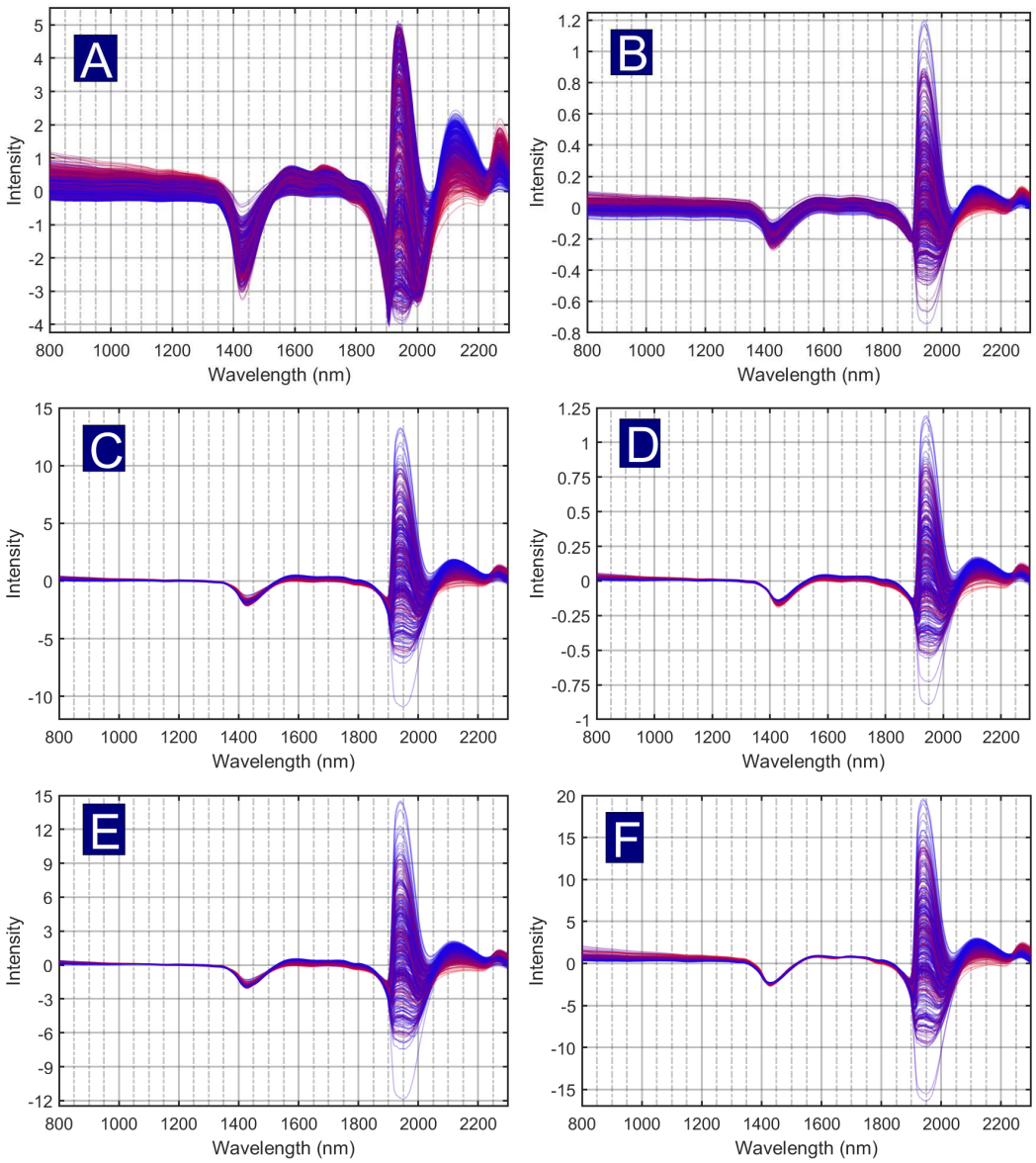

Figure 2.17: Results for the preprocessed spectra for Wine Must Data. In each subfigure, the samples were color-coded based upon the alcohol by volume (ABV) content (red=high ABV content; blue=low ABV content). (A) SNV; (B)MSC; (C) SSNV; (D) SMSC; (E) Binary SSNV; (F) IDRC.

weight vector, the variables in the region from approximately 800 nm to 1300 nm received the largest weights (greater than 0.7), while the variables in the region between approximately 1400 nm and 2298 nm received weights of less than 0.30. The stepped VSN weight vector was used to calculate a stepped SSNV correction analagous to the Binary SSNV for the simulated data, but with three levels of weights to reflect the characteristics of the weight vector determined. The stepped VSN weights shown in

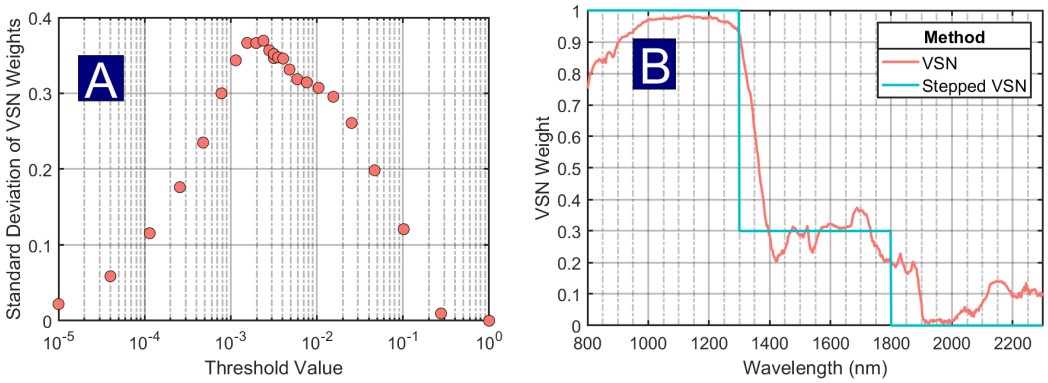

Figure 2.18: Results of the weighting procedure for VSN for wine must data. (A) Std. Dev. of VSN Weights vs. Threshold; (B) VSN weights for optimal threshold value.

Figure 2.17B were used to test how a simpler version of the optimal VSN weights would perform compared with SSNV using the optimal VSN weights. The stepped weights were equal to 1 for channels 1-250 (between 800 nm and 1348 nm), in channels 251-500 (between 1350 nm 1798 nm) the weights were 0.3, and between channels 501-750 (from 1800 nm to 2298 nm) the weights were 0.

The spectra that were processed by using SSNV and SMSC are shown in Fig. 2.17C and 2.17D, while the spectra processed using stepped SSNV are shown in Figure 2.17E. These spectra resulted in a decrease in the baseline variation between 800 nm and 1900 nm, but there may have been some information lost in the region between 2050 nm and 2298 nm.

The spectra that were processed using IDRC used wavelength channels 251-500 (1300nm to 1798 nm) to perform the correction. The IDRC-processed spectra, are shown in Fig. 2.17F. The IDRC-corrected showed some similarities with the spectra corrected by SSNV and SMSC, with the biggest difference being that the variation in the 1300-1900 nm region was smaller for IDRC than for the SSNV and SMSC, while in the 800 nm-1300 nm region, the IDRC-corrected spectra exhibited greater variation than the spectra corrected by SSNV and SMSC. This is not surprising given that the VSN weight vector emphasized regions below 1300 nm, while the IDRC selected channels between 1300 and 1798 nm.

The results for the prediction of ABV by each method are displayed in Table 2.5. SNV and MSC both resulted in very large prediction errors. The lowest value of

Table 2.5: PLS regression results for the various scatter correction methods applied to wine must data.

Method	RMSECV[a]	RMSEP_{Test}	NLV[b]
None	0.240 (0.011)	0.239	8
SNV	0.903 (0.046)	0.962	5
MSC	0.68 (0.16)	0.63	5
Stepped SSNV	0.404 (0.025)	0.382	15
SSNV	0.428 (0.038)	0.407	10
SMSC	0.383 (0.022)	0.399	10
IDRC	0.322 (0.016)	0.325	13

[a] Mean RMSECV, (standard deviation in parentheses)
[b] NLV=Number of Latent Variables

RMSEP_{Test} was 0.239, and was obtained by the spectra which were not corrected. The next-lowest prediction errors were obtained by IDRC, which had a RMSEP_{Test} of 0.325. Binary SSNV, SSNV, and SMSC resulted in RMSEP values of between 0.38 and 0.41. The largest prediction errors were obtained using MSC, with an RMSEP_{Test} of 0.63, and SNV, which had a RMSEP_{Test} of 0.962. MSC had a large standard deviation in the RMSECV results, which was likely due to variation in the reference (average) spectrum. As was the case with simulated Dataset 3, the prediction errors for the weighted scatter corrections were lower than the prediction errors for conventional MSC and SNV, but higher than the prediction errors for the non-preprocessed spectra. This suggests that the selected scatter dominant regions for the experimental data still contain a large amount of chemical background.

In the VSN paper, prediction results were reported for only SNV and SSNV [31]. It was reported that for SNV, the RMSECV was 0.890 and the RMSEP was 0.963, while for SSNV the RMSECV was 0.653 and the RMSEP was 0.701, and both methods used five latent variables. The results for SNV reported in this chapter were in close agreement with the results in the VSN paper, but the results for SSNV were significantly different, as the prediction errors reported in this chapter were significantly smaller (0.407 for the $RMSEP_{Test}$). Through email correspondence with Jean-Michel Roger, it was confirmed that the data splitting procedure was the same, and that the test set prediction errors for SNV were the same. A PLS model was calculated using an SSNV correction with the optimal parameters reported by Roger (threshold value of 3.16×10^{-3}, five components), and an $RMSEP_{Test}$ of 0.462 was obtained. However,

it may be the case that there were differences in the VSN parameters used, such as the number of inner loop iterations N_w.

In the region between 800 nm and 1300 nm, the spectra do exhibit a pattern which is consistent with multiplicative scatter. However, the behavior of the spectra in the regions from 1400 nm to 2298 nm do not follow the same patterns as in the 800 nm - 1300 nm region. If the difference in spectral behavior was due to the presence of significant chemical variation in addition to the scatter, one would expect that such regions would have a greater variation than the scatter-dominant regions, but this is not the case for the wine must spectra. In the region between 2050 nm and 2298 nm, where a strong correlation with alcohol by volume is present, the variables do not seem to be effected by scattering at all. Further, there is an isobestic-like point near 2230 nm where the range of values becomes very small, which would be impossible if an additive baseline offset was present throughout the whole spectrum. Consequently, any correction method that is applied to the wine must data will likely degrade the strength of relationship between the spectra in the region between 2050 nm and 2298 nm and the ABV content, and such a degradation would likely harm the predictive performance of the model.

2.6 Conclusions

In this chapter, the principles of weighted scatter correction methods were described, and their characteristics were examined using three simulated datasets and one experimental dataset. The central premise of this study was that weighted scatter correction methods will be maximally effective for spectral data featuring scattering from baseline and multiplicative effects, with one or more regions exhibiting very low levels of chemical signals and/or independent errors, and region(s) in which the variation due to chemical signal equals or exceeds the variation due to scattering. This was supported by the findings from the three simulated datasets.

In simulated Dataset 1, SSNV, Binary SSNV, and IDRC were all effective at correcting the scatter while preserving the chemical signals. When using VSN with such a dataset, the exact values of the weights in the chemical-dominant region(s) may be of great importance, and it may be beneficial to set the weights to zero for variables that are obviously dominated by chemical variation. It was found for Dataset 1 that

the SSNV method is less sensitive to the effects of variables with small weights than the SMSC correction. In Dataset 2, it was shown that if the amount of chemical variation is not sufficiently large, conventional SNV and MSC will work reasonably well, and so weighted scatter correction can result in only marginal improvements, even under otherwise ideal circumstances. In Dataset 3, it was shown that if the chemical background variation is too large, then weighted scatter correction methods will not be able to accurately estimate the scattering parameters.

The VSN method resulted in several interesting findings. In Dataset 1, the SMSC correction performed poorly, whereas the SSNV correction resulted in low prediction errors. To examine why this was the case, SSNV and SMSC corrections were performed using different weight vectors. It was found that the SMSC correction is more sensitive to the effects of small weights than SSNV. In practice, it may be beneficial for any practitioners of VSN to employ a cutoff value for the weights, such that any weights less than the cutoff (e.g. 0.10) will become zero. Variables with very small weights can reasonably be considered to be uninformative for calculating the scatter parameters, so it makes sense to zero the weights for such variables. The Binary SSNV method appeared to result in improved corrections compared with SSNV for each of the simulated datasets. However, due to the design of the simulated datasets, it is unclear as to whether Binary SSNV is always better than SSNV, as in a hypothetical dataset in which many variables contain weights between 0.3 and 0.7, using completely binary weights might not make as much sense. In Dataset 2, the VSN algorithm was tested to see the effect of the number of sample pairs used had on the calculated weight vector, and it was found that using a random fraction of sample pairs did not have a significant impact on the shape of the weight vector. For large numbers of samples, a random fraction of sample pairs is recommended when trying to optimize the threshold as it significantly reduces the computational time without changing the weights very much. After choosing the threshold which results in the largest standard deviation of weights, the weights can be re-calculated using all sample pairs. The last noteworthy point for VSN is that the weight vectors appear to provide very interesting information about the structure of the spectra, especially when different values of the threshold parameter are used. Examination of the VSN weights may be useful by itself as a simple exploratory technique.

The IDRC correction method had mixed results when compared with VSN-based methods and traditional SNV and MSC. In Datasets 1 and 3, IDRC was only bested by Binary SSNV as far as the shape of the corrected spectra, and prediction errors were concerned, while for Dataset 2 IDRC performed slightly worse than all the other scatter correction methods tested for both prediction errors and preservation of signal shape. The findings from Dataset 2 indicated that the RMSECV is not guaranteed to be a reliable index for choosing which variables to use to perform the weighted correction. Further, the IDRC methodology as it currently exists cannot be used for curve resolution or for exploratory analysis since validation is required. Also, since only one region is selected in IDRC, there is no easy way for IDRC to be used for datasets that have multiple scatter-dominant regions that are non-contiguous.

It is not presently clear to what extent weighted scatter corrections are relevant to experimental NIR data. For these techniques to be useful the conditions for Dataset 1 (low chemical background variation, high analyte signal variation) need to be met. Under conditions of Dataset 2, where the analyte signal is relatively low, traditional SNV and MSC appear to be adequate. Higher chemical background variation, even with high analyte signals, as in the case of Dataset 3, are problematic because errors in the estimation of the scatter coefficients have a multiplicative effect. In other words, higher analyte signals do not give better results because the errors are also bigger. The Wine Must Dataset was the most prototypical dataset that could be accessed for this study, but it appears to fall under the category of Dataset 3 rather than Dataset 1. It is therefore unclear how many experimental datasets meet the criteria to justify weighted scatter correction methods.

The primary intent of this study was to illustrate fundamental concepts pertinent to weighted scatter correction, and several points may require further study. The following parameters, which were not accounted for, may impact the performance of weighted scatter correction methods. The size of the training set, and the composition of the samples in the training set, could have an impact on the accuracy of the weights. The simulations used in the present study, as well as the simulations from the VSN paper, both featured a scatter-dominant region that comprised about half of the spectral channels. Given the mechanics of the RANSAC algorithm that is used by VSN, it could be the case that if a scatter-dominant region is very small,

then the algorithm may not weight the variables appropriately (since the RANSAC seeks to find the largest subset of inliers). The pure spectral profiles of the two major components were highly similar for the simulated data in this study, whereas real chemical components may have highly different pure spectra.

The simulations here did not incorporate wavelength-dependent terms in the scattering, and so the performance of EMSC-based weighted corrections was not able to be assessed. Future work could investigate the factors which are most important for using SEMSC and WMSCVS. Additionally, since second derivatives can remove constant baseline offsets and linear slope components, a method could be developed which would apply the second derivative correction, and then use a weighted correction to remove the multiplicative component of the scattering.

Chapter 3

Comparison of Principal Components Analysis (PCA) and Princpal Axis Factoring (PAF) in the Presence of Heteroscedastic Noise

In Chapter 2, methods that correct for the effects of correlated errors, specifically multiplicative offset noise and baseline offset noise, were investigated. Another issue that occurs in the analysis of multivariate data is that of heteroscedastic noise. As described in Chapter 1, PCA is commonly used to perform data compression for purposes of curve resolution, classification, calibration, and exploratory analysis. However, because PCA assumes *iid* normal errors, subspace estimation may be suboptimal when the data contain heteroscedastic noise. The issue of heteroscedastic noise is typically addressed through the use of data scaling, where optimal scaling is based on measurement error standard deviations. The main difficulty in scaling is the unavailability of measurement error information. An alternative to data scaling methods is to use exploratory factor analysis methods which seek to simultaneously estimate the subspace and the measurement error variance.

This chapter seeks to evaluate the use of principal axis factoring (PAF), an exploratory factor analysis method, and PCA, under both *iid* normal and column heteroscedastic conditions, and examines the subspace estimation and estimation of the measurement error uncertainties for each method. Three simulated datasets and two experimental datasets were used to investigate the relative performance of PAF and PCA. The chapter begins with a general review of PAF and PCA. A more detailed theoretical discussion of PCA and PAF is then presented. The data and methods used in this study are then described, followed by a thorough analysis of the results obtained.

3.1 Background

The method of Exploratory Factor Analysis (EFA) is nearly a century old, and traces back to Spearman's work on the topic of "general intelligence" in psychology [52, 53]. In broad terms, the goal was to relate the performances of a variety of measures of intelligence (e.g. tests/exams) to a small number of underlying determinants of overall intelligence (e.g. mathematical and language ability). A number of approaches were proposed, but it was not until Lawley and Maxwell's development of the theory of Maximum Likelihood Factor Analysis (MLFA) that the statistical foundations of factor analysis were firmly established [54]. Principal Components Analysis (PCA) originated with the work of Pearson (1901) [55] and Hotelling (1933) [56]. Its foundations were somewhat different from other factor analysis methods, but, because of a commonality of purpose, it was captured under the umbrella of factor analysis methods. In the days before the widespread availability of computers, calculations using any factor analytical method were difficult, but PCA was more tractable than MLFA. Principal Axis Factoring (PAF) evolved as a more computationally accessible alternative to MLFA. These three methods (PCA, PAF, and MLFA) are the three most widely used factor analysis methods in the social sciences and other fields, but in chemistry, PCA is used almost exclusively.

All of these factor analysis techniques seek to model latent variables ("hidden" variables) that describe a larger set of variables, but they do so slightly differently. The objective is to find latent variables that describe the variation in "common" factors, and separate the variation due to "unique" factors. If we consider an $n \times p$ matrix of measurements $\mathbf{X}$, where the columns represent the variables, and the rows correspond with different samples, the model for all factor analysis methods can be represented in terms of the sample covariance matrix, $\mathbf{R}$, as shown in Equation 3.1.

$$\mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Psi}^2 + \mathbf{E} \tag{3.1}$$

In this equation, $\mathbf{\Lambda}$ is a $p \times r$ matrix (where $r < p, n$) whose columns give the linear combinations of original variables making up the r latent variables. In factor analysis, these are referred to as the common factors, while in PCA they are called loadings (when normalized, they are also referred to as eigenvectors). The diagonal matrix $\mathbf{\Psi}^2$ ($p \times p$) is the matrix of unique factors or unique variances, and $\mathbf{E}$ is a $p \times p$ matrix

of residuals. For chemical measurements, $\mathbf{\Psi}^2$ can be viewed as representing the error variance associated with each variable. Practically speaking, PCA does not make a distinction between the unique variances and the residuals, and simply tries to find the set of common factors that explain the greatest amount of variation in $\mathbf{R}$, whereas PAF and MLFA try to solve for the common factors and unique factors to minimize the residuals. MLFA makes the assumption that the measurements in $\mathbf{X}$ follow a multivariate normal distribution [57, 58], an assumption often violated for chemical data.

In the context of this chapter, the term "factor analysis" is used to refer to describe mathematical algorithms which perform a bilinear decomposition of data. This means that the original data matrix $\mathbf{X}$ $(n \times p)$ can be expressed as the product of two matrices of smaller dimensions, generally referred to as the scores, $\mathbf{T}$ $(n \times r)$, and loadings $\mathbf{L}$ $(p \times r)$. Here r is the number of latent variables and $\mathbf{L}$ is consistent with the common factor loadings $\mathbf{\Lambda}$ in Equation 3.1. The decomposition, including residuals $\mathbf{E}$ $(n \times p)$ is represented in Equation 3.2.

$$\mathbf{X} = \mathbf{T}\mathbf{L}^{\mathrm{T}} + \mathbf{E} \tag{3.2}$$

In the chemometrics literature, the term "factor analysis" often refers to the use of PCA with rotated loadings. An early source of confusion may lie in the use of the term factor analysis as a synonym for PCA [59]. To further add to the potential confusion, there is the technique of evolving factor analysis, which is a method that is commonly used in curve resolution [60], and parallel factor analysis (PAFAFAC) [61, 62] is used for multiway data. Despite the names, both of these are more similar to PCA than other factor analysis methods.

Although widely used in other fields, FA methods other than PCA have not been widely applied in analytical chemistry and only a few examples were identified in the course of this research. In 1995, Campisi *et al* [63] used PAF for the comparison two different types of mandarin essential oils for a dataset composed of 17 volatile compounds whose peak areas were obtained by gas chromatography. In 2002, Pletnev *et al* [64] used PAF for a dataset which consisted of stability constants for 25 cations, with approximately 4000 different ligands, to assess the similarity of different classes of metal ions. The authors found that PAF "did not appear informative" for clustering the cations. Factor analysis has been used in geochemistry for the

purposes of identifying and mapping patterns of geochemistry across a broad survey area [65]. Maximum-likelihood factor analysis (MLFA) was used in several papers published between 1991 and 1995 by P. De Volder, which were in the research area of Auger Electron Spectroscopy [66, 67, 68, 69] and electron paramagnetic resonance spectroscopy [70]. The authors claimed that MLFA was superior to PCA, although no definitive results were presented in this regard. Likewise, no comprehensive comparisons of FA methods were made in the other studies.

The principal motivation of this work is to investigate whether there are situations in which other factor analysis methods, specifically PAF, have an advantage over PCA and, if so, to what extent. Unlike PCA, PAF does not assume homoscedastic (*iid* normal) measurement errors across all variables, although it does assume homoscedasticity within variables (columns). Therefore, PAF may provide more reliable results when heteroscedasticity is present. Moreover, PAF provides estimates of the variable measurement uncertainty, which can be valuable information in the characterization of a dataset.

3.1.1 Principal Components Analysis (PCA)

Principal components analysis (PCA) has been called "one of the most powerful tools in chemometrics" [71]. Depending on the nature of the data, PCA can be applied to the original data matrix, $\mathbf{X}$, the column mean-centered data, $\mathbf{X}_{MC}$, or the autoscaled data, $\mathbf{X}_{SC}$ (autoscaling refers to centering columns around their mean followed by division by the column standard deviation). PCA is generally implemented using singular value decomposition (SVD) which represents $\mathbf{X}$ as the product of a $n \times n$ matrix of left eigenvectors ($\mathbf{U}$), a $n \times n$ diagonal matrix of singular values ($\mathbf{D}$), and a $p \times n$ matrix of right eigenvectors ($\mathbf{V}$)

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} \tag{3.3}$$

The PCA solution typically consists of truncating the SVD result to include only the first r columns of $\mathbf{U}$ and $\mathbf{V}$, and using only the first r singular values in $\mathbf{D}$. From the truncated left eigenvectors matrix and the singular values, a scores matrix ($\mathbf{T} = \mathbf{U}_r\mathbf{D}_r$) of dimension $n \times r$ is calculated and the truncated eigenvectors matrix ($\mathbf{V}_r$) is the same as the loadings matrix $\mathbf{L}$ ($p \times r$) in Equation 3.2. The truncation

of the scores and loadings means that some of the variance in $\mathbf{X}$ is not accounted for in the scores and loadings, and as a result the unexplained parts are found in the residuals matrix $\mathbf{E}_{PCA}$.

If the data have been mean-centered, which is often the case since the mean vector does not contain information about the differences in the measurements, reconstruction of the original data involves the addition of the vector of column means, $\bar{\mathbf{x}}$ $(1 \times p)$. The PCA reconstruction of the original data matrix is termed $\hat{\mathbf{X}}_{PCA}$, as shown in Equation 3.5.

$$\mathbf{X}_{MC} = \mathbf{T}\mathbf{L}^{\mathrm{T}} + \mathbf{E}_{PCA} \tag{3.4}$$

$$\hat{\mathbf{X}}_{PCA} = \mathbf{T}\mathbf{L}^{\mathrm{T}} + \mathbf{1}_n\bar{\mathbf{x}} \tag{3.5}$$

Here, $\mathbf{1}_n$ indicates an $n \times 1$ vector of ones. Likewise, autoscaling is often used when it is expected that the relative standard deviation of the measurement errors in each column of $\mathbf{X}$ are approximately the same. Reconstruction of the original data in this case involves multiplication by the vector of column standard deviations, $\mathbf{s}_x$ $(1 \times p)$.

$$\mathbf{X}_{SC} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}_{PCA} \tag{3.6}$$

$$\hat{\mathbf{X}}_{PCA} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \circ \left(\mathbf{1}_n\mathbf{s}_x\right) + \mathbf{1}_n\bar{\mathbf{x}} \tag{3.7}$$

Here $\circ$ indicates the Hadamard (element-wise) product.

When measurement errors are *iid* normal and the assigned rank of the data r is correct, PCA provides the maximum likelihood (unbiased, minimum variance) estimate of the error-free data. The root-mean-squared error of the residuals provides an estimate of the measurement error standard deviation.

$$\hat{\sigma}_{meas} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2}{(n-r)(p-r)}} \tag{3.8}$$

This equation is valid when the data have not been scaled. In cases where autoscaling is used, or where heteroscedasticity is expected, the measurement uncertainty for each column, $\hat{\sigma}_{PCA,j}$, can be calculated from individual column standard deviations of the residuals.

$$\hat{\sigma}_{PCA,j} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}{(n-1)}} \tag{3.9}$$

It may be convenient to represent this as a relative standard deviation (RSD) by dividing by the column means. The metrics mentioned above are used in the interpretation of results in this work.

3.1.2 Principal Axis Factoring (PAF)

Exploratory factor analysis (EFA) is a family of methods which includes both PAF and maximum-likelihood factor analysis (MLFA). The EFA model seeks to describe the correlation matrix **S**. Because these methods do not assume *iid* errors across **X** and estimate the measurement error variance associated with each column of data, scaling is less important for PAF than for PCA. However, for mathematical convenience, EFA methods are generally applied to autoscaled data. This means that the covariance matrix, **R**, shown in Equation 3.1 becomes the correlation matrix, **S**. The elements of the correlation matrix are simply the covariance matrix elements divided by the product of the corresponding column standard deviations, σ_i and σ_j.

$$s_{ij} = \frac{r_{ij}}{\sigma_i \sigma_j} \tag{3.10}$$

Equation 3.1 can now be modified to

$$\mathbf{S} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Psi}^2 + \mathbf{E} = \mathbf{\Sigma} + \mathbf{E} \tag{3.11}$$

A convenient property of the correlation matrix is that the diagonal elements are unity, which simplifies calculations. Once a solution is obtained, it can be transformed back to the original space of the data.

The PAF algorithm is solved in an iterative manner as shown in Algorithm 2 [72]. In step 2, the $p \times p$ eigenvectors matrix **W** and the $p \times p$ diagonal matrix of singular values **K** are found by calculating the singular value decomposition of the reduced correlation matrix ($\mathbf{S} - \mathbf{\Psi}^2$). Initially (step 1), the $\mathbf{\Psi}^2$ is set to equal a matrix of zeros. In step 3, the common factor loadings $\mathbf{\Lambda}$ are estimated by multiplying each of the first r columns of **W** by the square root of the corresponding singular values in **K**. In step 4, the unique factors $\mathbf{\Psi}^2$ are estimated using the approximated loadings that were calculated in step 3. The equations in steps 2-4 are iterated until either a stable solution is reached, or a maximum number of iteration have been performed.

The $p \times r$ matrix of loadings $\mathbf{\Lambda}$, produced by this algorithm are analogous to the loadings that would be produced from PCA on autoscaled data, **L**, in that they are

input : Correlation Matrix ($\mathbf{S}$) of dimension $p \times p$, model rank, r

output: Factor loadings ($\mathbf{\Lambda}$), unique variances $\mathbf{\Psi}^2$

1 Set $\mathbf{\Psi}^2$ to equal a $p \times p$ matrix of zeros;

2 Decompose $(\mathbf{S} - \mathbf{\Psi}^2)$ by SVD to give
$(\mathbf{S} - \mathbf{\Psi}^2) = \mathbf{W}\mathbf{K}\mathbf{W}^{\mathrm{T}}$;

3 Compute a rank r estimate of $\mathbf{\Lambda}$ using
$\mathbf{\Lambda} = \mathbf{W}_r\mathbf{K}_r^{1/2}$;

4 Estimate the diagonal elements of $\mathbf{\Psi}^2$ as
$\psi_{jj}^2 = 1 - \boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^{\mathrm{T}}$,
for $j = \{1, 2, ..., j, ..., p\}$, where $\boldsymbol{\lambda}_i$ is the j^{th} row of $\mathbf{\Lambda}$;

5 Repeat from Step 2 until convergence, or until maximum number of iterations

Algorithm 2: PAF algorithm

an orthogonal representation of the space containing the data. However, the two sets of loadings will not be coincident because of the algorithmic differences.

Factor Scores

To estimate the measurements in the scaled or original spaces, it is necessary to multiply the loadings matrix by the factor scores. The scores represent the coordinates of the estimated measurements in the subspace defined by the loadings. For PCA, the scores are obtained directly from SVD. Alternatively, the scores vector $\mathbf{t}$ $(1 \times r)$ for a given sample can be estimated by orthogonal projection of a measurement vector, $\mathbf{x}$ $(1 \times p)$, onto the space defined by the loadings, $\mathbf{L}$ $(p \times r)$.

$$\mathbf{t} = \mathbf{x}\mathbf{L} \tag{3.12}$$

In PCA, the orthogonal projection provides a maximum likelihood estimate consistent with an assumption of *iid* measurement errors. For PAF, the scores do not result directly from Algorithm 2, but must be computed. Rather than using the orthogonal projection as is used in PCA, PAF employs a maximum likelihood (ML) projection based on the estimated measurement error variances for $\mathbf{X}_{SC}$.

$$\mathbf{F} = \mathbf{X}_{SC}\mathbf{\Psi}^{-2}\mathbf{\Lambda}\left(\mathbf{\Lambda}^{\mathrm{T}}\mathbf{\Psi}^{-2}\mathbf{\Lambda}\right)^{-1} \tag{3.13}$$

In Equation 3.13, $\mathbf{F}$ is analogous to the PCA scores ($\mathbf{T}$), and $\mathbf{\Lambda}$ is analogous to the PCA loadings $\mathbf{L}$, but notational conventions consistent with the PAF literature have been adopted here. The ML projection used by PAF should, in principle, lead to more reliable estimates of the scores and, hence, the original data. However, it depends on reliable estimates of the measurement errors, which can show a high variability, especially for small datasets.

Once the scores have been obtained, the data can be estimated in the usual way as discussed in Section 3.1.1. For the autoscaled data, the estimates are calculated using Equation 3.14.

$$\hat{\mathbf{X}}_{SC} = \mathbf{F}\mathbf{\Lambda}^{\mathrm{T}} \tag{3.14}$$

To transform the data back to the original space, a method similar to Equation 3.15 can be used.

$$\hat{\mathbf{X}} = \mathbf{F}\mathbf{\Lambda}^{\mathrm{T}} \circ \left(\mathbf{1}_n \mathbf{s}_x\right) + \mathbf{1}_n \bar{\mathbf{x}} \tag{3.15}$$

Heywood Cases

A complication that can sometimes occur with EFA methods (but not PCA) is that one or more diagonal elements of $\mathbf{\Psi}^2$ can become negative. This is known as a "Heywood case", and is problematic because, of course, measurement error variances cannot be negative. Such cases are also a problem in the reconstruction of the data because the computed scores cannot be considered to be reliable. Heywood cases are a consequence of the algorithms used and the stochastic nature of the data, i.e, they represent a statistical anomaly. They are most often observed in cases where the number of samples is small, the number of components is inaccurately estimated, and/or the error variance is small.

When Heywood cases occur, they can be dealt with by several methods. One approach is thresholding, in which a minimum allowable threshold for the diagonal elements of $\mathbf{\Psi}^2$ is set. Thresholding is simple to implement, but the choice of threshold value is likely to be somewhat arbitrary. Another possible solution is to use estimates of the measurement error variances derived from PCA when any Heywood cases occur. Resampling, which can entail using random subsets of samples, and/or random subsets of variables, and calculating the PAF solution multiple times by using multiple repetitions, could be used, although it is not guaranteed to work. In this work,

Heywood cases were observed in one of the experimental datasets, but the presence of these Heywood cases did not substantially effect the results, so no further actions were needed.

3.1.3 Comparison Metrics

The purpose of this work is to compare the results of PCA and PAF. Several metrics were used for comparison purposes and are described in the following subsections.

Subspace Angles

Many chemical datasets are bilinear in nature. For example, spectroscopic data of chemical mixtures ($\mathbf{X}_{pure}$, $n \times p$) can be decomposed into a matrix of pure component concentrations ($\mathbf{C}_{pure}$, $n \times r$) and a matrix of pure component spectra ($\mathbf{P}_{pure}$, $p \times r$).

$$\mathbf{X}_{pure} = \mathbf{C}_{pure}\mathbf{P}_{pure}^{\mathrm{T}} \tag{3.16}$$

PCA is often used to estimate the subspaces of $\mathbf{C}_{pure}$ and $\mathbf{P}_{pure}$ through the scores $\mathbf{T}$ and loadings $\mathbf{P}$. The mathematical values of the subspaces in the scores and loadings is not important, but instead it is the angle between the estimated subspace and the subspace of the noise-free data that is important. One method of comparing how close the subspaces for PCA and PAF are to the original subspaces, is to compare the subspace angles. The angle between the subspaces of $\mathbf{C}_{pure}$ and $\mathbf{T}$ is called the scores subspace angle, while the angle between the subspaces of $\mathbf{P}_{pure}$ and $\mathbf{L}$ is called the loadings subspace angle.

Imbedded Errors

Factor analysis methods will remove some of the errors in the data through the modeling process. The difference between the experimental data $\mathbf{X}$ and the reconstructed data $\hat{\mathbf{X}}$ is termed the "extracted error" [73]. However, some of the measurement errors will remain imbedded in the data, which are termed "imbedded errors". The imbedded error (IE^2) can be used as a performance metric, and is defined as the sum of squared differences of the reconstructed data and the noise-free data, over the total

number of elements in the data (np).

$$IE^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{\left(\hat{x}_{ij} - x_{pure,ij}\right)^2}{np} \tag{3.17}$$

Measurement Error Estimates

The measurement error estimates derived from PAF and PCA can also be used as a performance metric. For PAF, estimates of the measurement error variances are found in the diagonal matrix $\mathbf{\Psi}^2$. The PAF estimate of the measurement error standard deviation in the original measurement (unscaled) space of the j^{th} column of $\mathbf{X}$ ($\hat{\sigma}_{PAF,j}$) is found by taking the square root of the j^{th} diagonal element of $\mathbf{\Psi}^2$ and multiplying by the standard deviation of the j^{th} column of $\mathbf{X}$. In this conversion, the square root is used because the standard deviation is the square root of the variance, and multiplication by $s_{\mathbf{x}_i}$ is used to convert the estimate from the autoscaled space to the original data space.

$$\hat{\sigma}_{PAF,j} = \left(\sqrt{\psi_{jj}^2}\right)s_{\mathbf{x}_j} \tag{3.18}$$

For PCA, several metrics for the measurement uncertainty can be employed. Since PCA is based on the assumption of *iid* errors, there is technically only one estimate of the measurement uncertainty, which can be calculated from the data reconstructed in the space of the decomposition,

$$\hat{\sigma}_{PCA} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2}{(n-r)(p-r)}} \tag{3.19}$$

If the PCA is carried out on scaled data, the individual variable standard deviations in the original space can be estimated from the above result by multiplication by the scaling factor

$$\hat{\sigma}_{PCA,j} = \hat{\sigma}_{PCA(Scaled)} \cdot s_{\mathbf{x}_j} \tag{3.20}$$

However, these results are predicated on the assumption that the errors (in the original or scaled space) are *iid* normal. To better explore the heteroscedasticity of the errors, an alternative approach is to estimate the measurement uncertainty in each variable from the standard deviation of the residuals in each column for the reconstructed data.

$$\hat{\sigma}_{PCA,j} = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}{(n-1)}} \tag{3.21}$$

This calculation may be done in the original or scaled space, depending on the context of the comparison. This was the equation used for the PCA uncertainties in this work.

For simulated data, the standard deviations of the added noise is known, while for experimental data with replicates, the measurement error standard deviations in each column can be estimated by pooling the standard deviation of the replicates in each column. In other cases, empirical estimates of the measurement uncertainties may be available from a knowledge of the nature of the measurement. The estimates of the measurement error standard deviations in each column derived from using PCA ($\hat{\sigma}_{PCA,j}$) and PAF ($\hat{\sigma}_{PAF,j}$) can be compared to each other, and to benchmark estimates of the measurement errors derived from the added noise (for simulated data) or from replicates or estimates (experimental data).

Scores and Loadings

The scores and loadings resulting from PAF and PCA can be compared in several ways. The most direct method of comparison is to plot the scores and loadings using scores plots and loadings plots. A "scores plot" is a scatter plot of the scores for the first 2-3 components (factors) in two or three dimensions. For data where clustering is expected, the clusters formed in the PAF scores plots can be observed with the clusters for the PCA scores plots. Where Monte Carlo studies are done on simulated data, uncertainty ellipses can also be plotted for each score value. For loadings, the two methods can be compared by plotting the loadings for each factor/component, like spectra, where the x-axis is the wavelength channel number (for spectral data), the y-axis is the loading value, and the loadings for each component are plotted as a separate line.

3.2 Data

To compare the performance of PAF and PCA, three simulated datasets and two experimental datasets were used. All calculations were carried out using MATLAB version R2017b (Mathworks, Natick, MA, USA).

Datasets 1 and 2 consisted of simulated spectra of chemical mixtures with three components. The purpose of these datasets was to assess how PAF and PCA are effected by the measurement error structure. To investigate this, the noise-free data

were the same for both Dataset 1 and 2, while the noise structure was varied. In Dataset 1, the noise added was heteroscedastic, while in Dataset 2 the noise was *iid* normal. The third simulated dataset (the "Discrete Dataset") was designed to model data with discrete variables, such as elemental analysis data, where each column corresponds to the concentration of an element or ion. Such datasets may include major, minor and trace components measured in different units. With such large range in the size of variables, scaling is often employed. The errors between variables are likely to be heteroscedastic, even after scaling. PAF could be useful in such a situation as an alternative to conventional data scaling methods.

The first of the two experimental datasets, referred to as the Metals Data, consisted of visible spectra of mixtures of metal ions. The spectra were measured using an optical filter which introduced heteroscedastic noise, which made the dataset suitable for testing the performance of PAF and PCA. The other experimental dataset, referred to as the Obsidian Data, consisted of obsidian samples for which the concentrations of 10 elements were determined by X-ray fluorescence (XRF) spectroscopy. This dataset is often employed as a benchmark for exploratory analysis.

3.2.1 Simulated Datasets 1 and 2

The simulated spectra for Dataset 1 consisted of $n = 100$ samples, $p = 20$ variables, and $r = 3$ components. The noise-free data ($\mathbf{X}_{pure}$) were generated by multiplying a 100×3 matrix of pure concentrations ($\mathbf{C}_{pure}$) by a 20×3 matrix of pure spectral profiles ($\mathbf{S}_{pure}$). The choice of $n = 100$ samples was motivated by trying to balance the competing interests of having enough samples to define the covariance/correlation matrix, while having a number of samples that is realistic for datasets in the chemometric literature (where the number of samples is often less than 100). The pure concentration matrix $\mathbf{C}_{pure}$ was generated by sampling the normal distribution with parameters $N(\mu = 0.20, \sigma = 0.040)$. The pure spectral profiles $\mathbf{S}_{pure}$ consisted of three Gaussian functions, all with a height constant $h = 20$, and standard deviation $\sigma = 2.5$, for the wavelength channel indices $\boldsymbol{\xi} = \big[1, 2, ..., 20\big]$. The pure spectral profiles, shown in Figure 3.1, were designed to be partially overlapping. If the profiles were too close together, the number of components could have been more ambiguous, as peaks that are too close together cannot be distinguished from one another. The

noise-free spectra for the mixture are shown in Figure 3.2A.

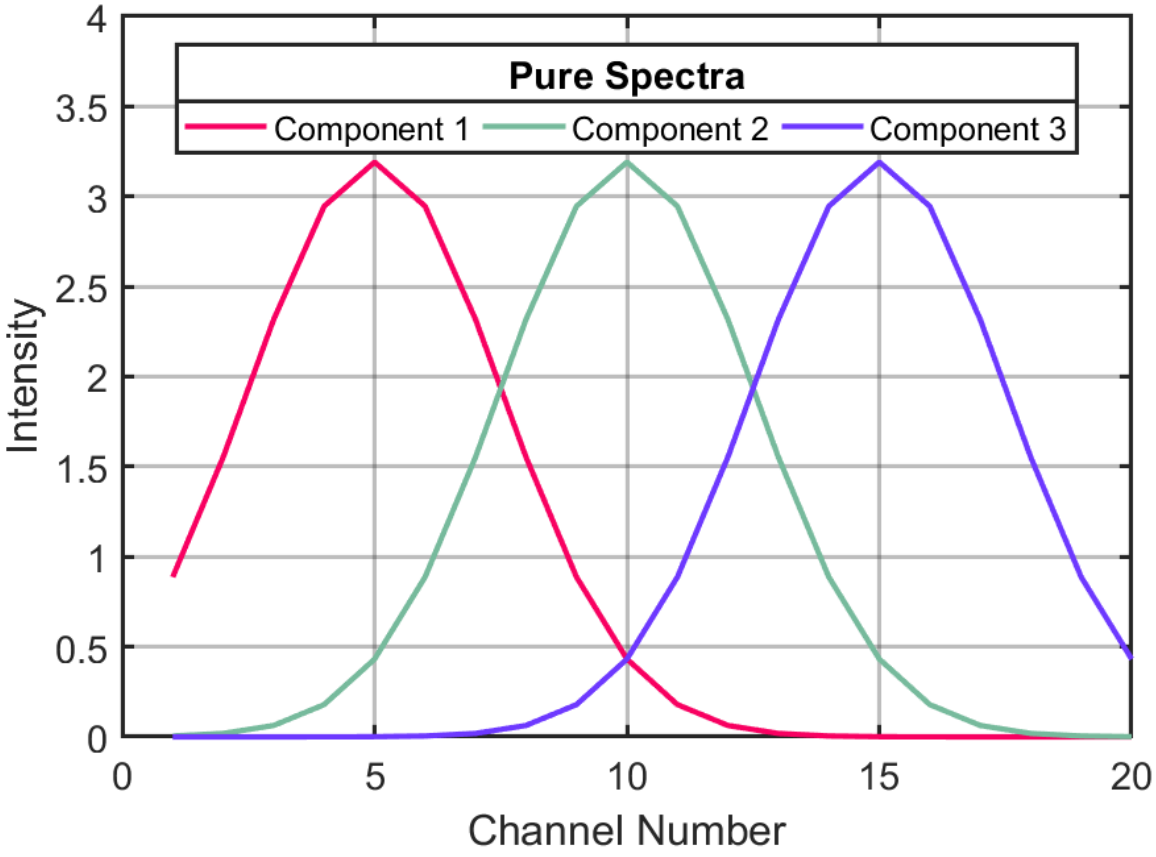

Figure 3.1: Pure component spectra (**S**), used for Datasets 1 and 2.

The noise added for Datasets 1 and 2 consisted of a mixture of column heteroscedastic noise and *iid* noise. This noise composition reflects many physical measurements where a proportional noise structure can dominate for large signals, but is limited by a baseline noise as signals become smaller. The proportional (heteroscedastic) component was represented by σ_{het}, which gives the RSD of the noise with respect to the mean of the noise-free data. The *iid* component was characterized by σ_{iid}, an absolute standard deviation. The j^{th} column of the error matrix, **E** $(n \times p)$, is defined by Equation 3.22

$$\mathbf{e}_j = \sigma_{het} \cdot \bar{x}_{pure,j} \cdot \epsilon_{het,0,1} + \sigma_{iid} \cdot \epsilon_{iid,0,1} \tag{3.22}$$

Here, $\bar{x}_{pure,j}$ is the mean of the j^{th} column of the noise-free data matrix, and $\epsilon_{het,0,1}$ and $\epsilon_{iid,0,1}$ both represent $n \times 1$ vectors of random variables drawn from a normal distribution with a mean of zero and a standard deviation of unity $(N(0,1))$. For

both datasets, σ_{het} was set to 0.02 (2% RSD). In Dataset 1, σ_{iid} was 0.0002, making proportional errors the dominant noise component, while σ_{iid} was set to 0.02 in Dataset 2, making the noise approximately homoscedastic. The theoretical and observed standard deviation for the added errors in Datasets 1 and 2 are shown in Figure 3.2B.

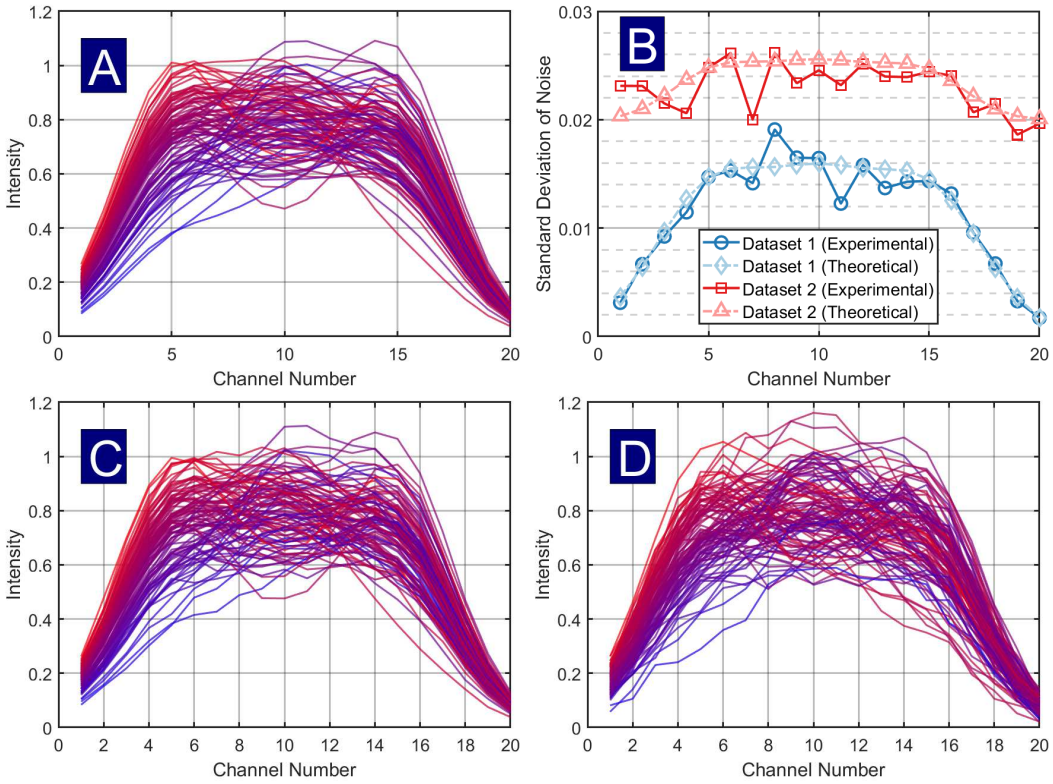

Figure 3.2: Figures for simulated Datasets 1 and 2. (A) Noise free spectra $\mathbf{X}_{chem}$ used for Datasets 1 and 2; (B) Column standard deviations of noise matrix $\mathbf{E}$ that was used for Dataset 1 and Dataset 2; (C) Matrix of spectra $\mathbf{X}$ for Dataset 1; (D) Matrix of spectra $\mathbf{X}$ for Dataset 2.

The noise for Dataset 1 follows a column heteroscedastic structure, as the standard deviation of the noise varies based upon the signal intensity. The noise for Dataset 2, however, can be characterized as approximately homoscedastic noise, since the standard deviation of the noise is mostly consistent across the different channels, with the standard deviations of the noise ranging from about 0.018 to 0.026. For both Dataset 1 and Dataset 2, the simulated spectral matrix ($\mathbf{X}$) was calculated by simply adding the noise-free data ($\mathbf{X}_{pure}$) and the noise ($\mathbf{E}$). The spectra for Datasets 1 and 2 are shown in Figure 3.2C-D, respectively.

3.2.2 Simulated Discrete Data

There were several objectives that influenced the decisions in making the Discrete Dataset. One objective was to recreate the differences in range of variables that often occur in compositional data (e.g. elemental analysis, metabolomics, fatty acid profiles), such that the range of the variables differed by several orders of magnitude. For the measurement errors, the goal was to add noise using realistic assumptions. A third objective was to create the data in such a way that the variation in the scores due to different realizations of the noise matrix could be calculated and visualized.

The data consisted of $n = 50$ samples, $p = 20$ variables, and $r = 2$ components. The choice of using only 2 components was motivated by the objective of being able to monitor and visualize the variation in the scores. With 2 components, a 2-dimensional scatter plot of the scores, with the first column of the scores on the x-axis and the second column of the scores on the y-axis, could be used.

In Simulated Datasets 1 and 2, the noise-free data ($\mathbf{X}_{pure}$) were generated by multiplying a pure concentration matrix ($\mathbf{C}_{pure}$) times a pure spectral profile matrix ($\mathbf{S}_{pure}$). The Discrete Data were generated in a similar manner, although for the Discrete Dataset the matrix $\mathbf{C}_{pure}$ is called the "compositional profiles", while $\mathbf{S}_{pure}$ is termed the "variable profiles", which are more accurate descriptors of what these matrices represent. The compositional profiles matrix ($\mathbf{C}_{pure}$) was of dimension 50×2, and was generated by sampling the normal distribution with parameters $N(\mu = 1.0, \sigma = 0.30)$. The variable profiles were generated in a two-step process. In the first step, a 2×20 matrix of "base" variable profiles ($\mathbf{S}_{pure,0}$) was created by sampling from a uniform distribution ($U(0, 1)$). In the second step, the variable profiles $\mathbf{S}_{pure}$ were calculated by multiplying the base profiles $\mathbf{S}_{pure,0}$ by a 1×20 vector of values that were evenly spaced along the logarithmic scale from 10^1 to 10^4. The first step was designed to ensure that the two components were independent of one another, while the second step ensured that the variable profiles would differ by several orders of magnitude, which was a main objective. A log-scale plot showing the pure spectral profiles $\mathbf{S}$ is shown in Figure 3.3A.

For the noise generation, the goal was to add noise using realistic assumptions. For real elemental data, the noise would likely be proportional to concentration, but

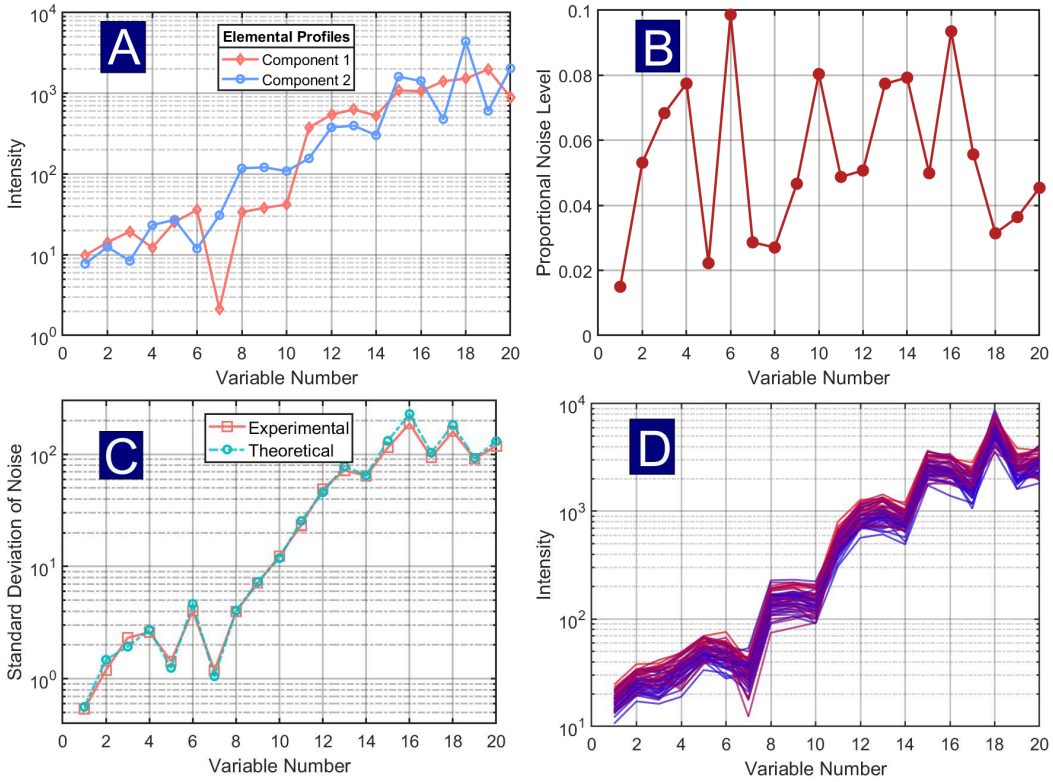

Figure 3.3: Simulated Discrete Data. (A)log-scale plot of the pure spectral profiles; (B) Proportional noise levels for each channel; (C) Log-scale plot of the standard deviation of the noise for each element; (D) log-scale plot of the data.

depending on the measurement techniques used, some elements would have larger relative errors than other elements. This error assumption was implemented by making proportional heteroscedastic noise, with randomness introduced so that the standard deviation of the proportional noise was different for each element, and a small amount of *iid* noise was added. For the heteroscedastic component of the noise, each column of data (element) was assigned an individual noise RSD, designated as $\sigma_{het,j}$ in the range of 0.01 to 0.10 (1% to 10% RSD) by sampling random numbers from a uniform distribution, $U(0.01, 0.10)$. The columns of the $n \times p$ noise matrix $\mathbf{E}$ were generated according to Equation 3.23.

$$\mathbf{e}_j = \sigma_{het,j} \cdot \bar{x}_{pure,j} \cdot \epsilon_{het,0,1} + \sigma_{iid} \cdot \epsilon_{iid,0,1} \tag{3.23}$$

As before, $\bar{x}_{pure,j}$ is the mean of column j of the noise-free data and $\epsilon_{het,0,1}$ and $\epsilon_{iid,0,1}$ both represent $n \times 1$ vectors of random variables drawn from $N(0, 1)$. The values for the RSD for each variable ($\sigma_{het,j}$) are shown in Figure 3.3B. The value for σ_{iid} was

set to 0.5 for all columns. For column j of $\mathbf{X}$, this gives the theoretical (population) error standard deviation of

$$\sigma_{err,j} = \sqrt{\left(\sigma_{het,j} \cdot \bar{x}_{pure,j}\right)^2 + \sigma_{iid}^2} \tag{3.24}$$

These values are shown on a log scale in Figure 3.3C, along with the calculated standard deviations for the data generated. The matrix of errors, $\mathbf{E}$, was added to the matrix of pure data, $\mathbf{X}_{pure}$, to generate the simulated data, $\mathbf{X}$, shown in Figure 3.3D.

For the Discrete Data, PCA was performed using the autoscaled data, as is the norm in datasets in which the ranges of the variables are significantly different. For the calculation of the measurement error comparison metrics, some modifications to the procedure for calculating the metrics was necessary. Due to the data range differences, using the unscaled raw data space to calculate the subspace angles and other metrics would result in the variables with large values having significantly more impact than the variables with small values. To facilitate comparison of different methods (PCA, PAF) with the noise-free results, it is necessary to evaluate metrics in a space of equivalent scaling and also one that accurately reflects the information retained by each method. Although the scaled space would be better in this regard than the original space, it is subject to variations in the scaling estimates and the heteroscedastic RSDs in the errors. Instead the procedure used was to (1) calculate the reconstructed data, (2) scale the reconstructed data to the original space ($\hat{\mathbf{X}}$), and (3) rescale each column of $\hat{\mathbf{X}}$ by its theoretical measurement error standard deviation, $\sigma_{err,j}$, given in Equation 3.24. The columns of this rescaled matrix, $\hat{\mathbf{X}}_{ESC}$, are given by

$$\hat{\mathbf{X}}_{ESC,j} = \frac{\hat{\mathbf{X}}_j}{\sigma_{err,j}} \tag{3.25}$$

Note that this scaling method was applied for each set of results (PCA, PAF, error-free data). By scaling the original data by the error standard deviation, optimal scaling is applied so that each column has an equivalent error variance. This was regarded as the most appropriate space to compare metrics.

A Monte Carlo approach was used to account for the variation in the relevant comparison metrics due to the effects of noise realizations. In the Monte Carlo approach, 1000 different realizations of the noise matrix $\mathbf{E}$ were calculated using the

same statistical parameters for the *iid* and heteroscedastic noise, and the noise matrix **E** was added to the noise-free data matrix $\mathbf{X}_{pure}$ (which was the same for each realization of the noise). For each realization of the noise, the comparison metrics were calculated for both the PAF-reconstructed scaled data and PCA-reconstructed scaled data, so that their statistical characteristics could be examined.

3.2.3 Metals Data

The Metals Dataset was originally described by Wentzell *et al* [74]. The data consisted of mixtures of the metal ions Co(II), Cr(III), and Ni(II) in solution, measured using visible spectrophotometry. The solutions were made using a 3-level, 3-factor design, such that 1, 3, or 5 mL aliquots of the stock solutions were added and diluted to 25 mL with 4% HNO_3. The stock solutions of the metal nitrates were prepared in 4% HNO_3 with concentrations of 0.172 M Co(II), 0.0764 M Cr(III), and 0.393 M Ni(II). The amount of Ni stock remaining was insufficient to prepare the solution with a 3:5:5 ratio of Co:Cr:Ni aliquots, so the dataset consisted of 26 solutions (rather than 27). Five replicate spectra were obtained for each solution using five randomized blocks, and a reference spectrum was measured prior to each new sample. The spectra were measured using an HP 8452 diode array spectrophotometer (Hewlett-Packard, Palo Alto, CA) using a 1 cm quartz cuvette. The measurements were made over the wavelength range 300-650 nm in 2 nm intervals with a 1 s integration time. To introduce nonuniform noise characteristics, a dichroic band-pass filter (green, no. 67) was placed between the source and the sample to decrease the source intensity at high and low wavelengths. Two spectra were removed as outliers.

The Metals Data was analyzed for two versions of the dataset. In the "Full" version of the dataset, all 176 wavelength channels (300-650 nm) were used. In the "Cut" version, only the 136 wavelength channels were used, in the wavelength range 350-620 nm. The two versions of the Metals Data are shown in Figure 3.4. Due to the use of the band-pass filter, the noise in the channels at each end of the spectrum (300-350 nm and 620-650 nm) contain a significant amount of heteroscedastic noise. In fact, the heteroscedastic noise is a major source of variation, which poses a problem for PCA. In the results section, it will be shown that PCA performs poorly for the Full spectra, whereas PAF does not have major issues. For this dataset, the problems

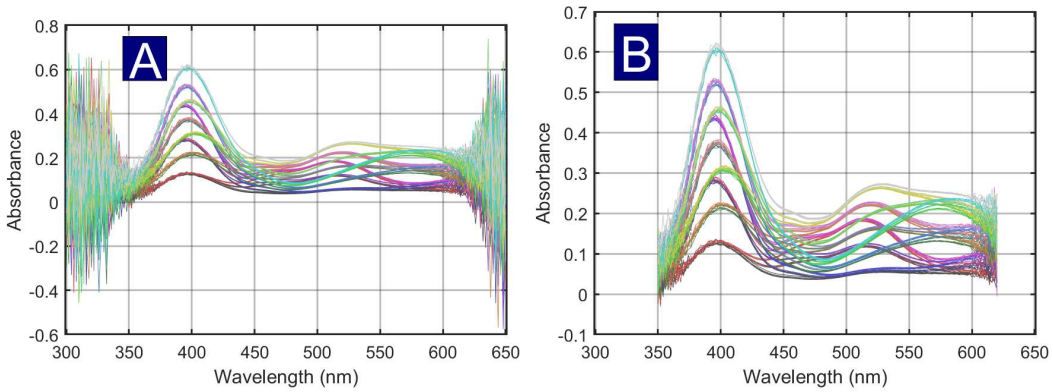

Figure 3.4: Plot showing the spectra for the two different versions of the Metals Data. (A) Metals spectra with all wavelengths used (Full); (B) Metals spectra with truncated wavelengths (Cut).

that arise from the noisy variables in PCA can be easily avoided by simply removing the noisy variables, as can be seen with the Cut spectra. However, the choice of which variables to remove (loss of information versus contamination of results) may not always be readily apparent.

For analysis of the measurement errors, the replicate spectra were used to estimate the measurement error standard deviations. For a given set of replicates ($\mathbf{X}_{rep}$), the error was calculated by subtracting the average replicate spectrum ($\bar{\mathbf{x}}_{rep}$) from each sample in $\mathbf{X}_{rep}$. For both the Cut and Full versions of the Metals data, 3-component PAF and PCA models were calculated, and the scores, loadings, and measurement error estimates were determined.

3.2.4 Obsidian Data

The analysis of the Obsidian Dataset was originally described in a paper by Kowalski *et al* [75]. The data originates from a study of archaeological sites in Northern California, where 75 samples of obsidian were obtained. The obsidian samples were analyzed using X-ray fluorescence (XRF) spectroscopy. The instrument used was a General Electric XRD-6, with both a tungsten and chromium target, and a LiF crystal. The concentrations of 10 elements were measured: Fe, Ti, Ba, Ca, K, Mn, Rb, Sr, Y, and Zr. A more detailed description of the data collection can be found in a paper by Stevenson *et al* [76]. The objective of the original study was to classify unknown

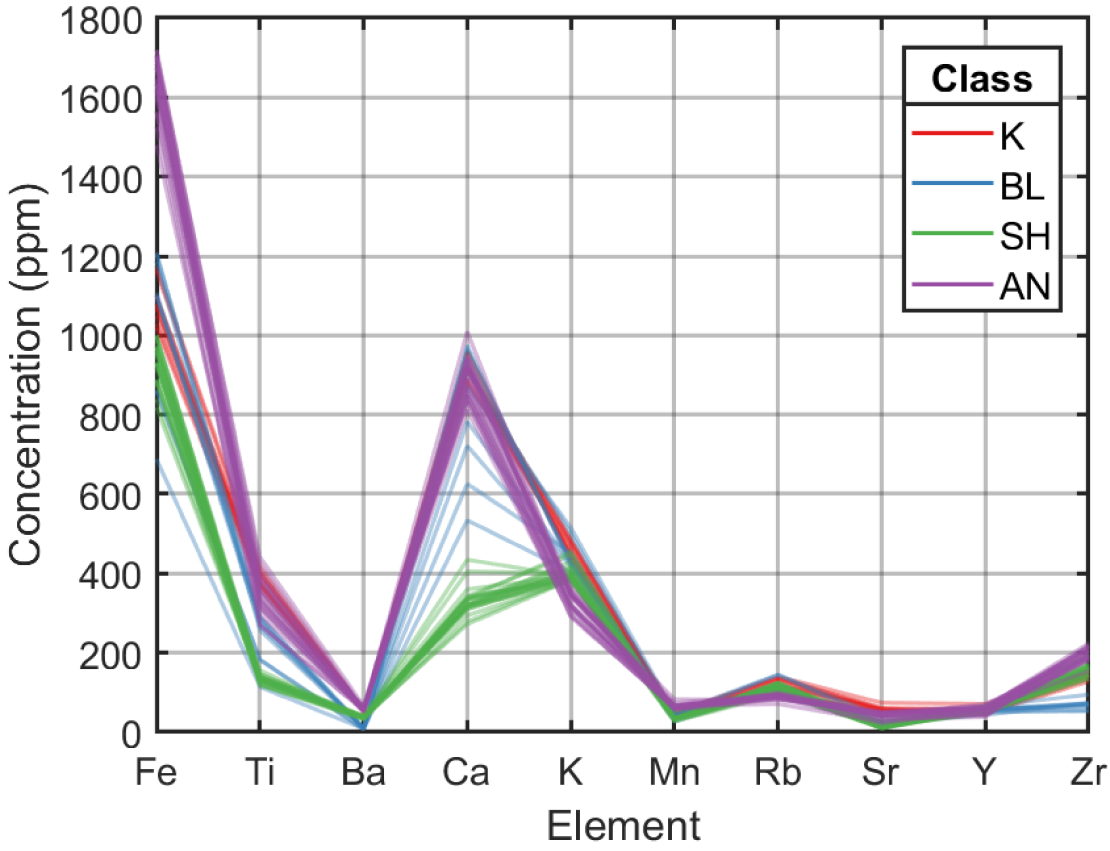

Figure 3.5: Concentrations (in ppm) of each element for the Obsidian Data (training set samples only).

samples on the basis of the geographical origin of the obsidian source. The sources of the obsidian were the following sites: Mount Konocti (K) and Borax Lake (BL) in Lake County; Glass Mountain (St. Helena) (SH) in Napa County; Anadel (AN) in Sonoma County. The obsidian artifacts were obtained in sites in Contra Costa, Napa, Lake, and Mendocino Counties in Northern California.

Sixty-three of the obsidian samples came from the four sites, and twelve unknown samples were obtained from other sites but were expected to have originated from one of the four quarries[77]. The samples from the training set a are plotted in Figure 3.5, and the samples were color-coded according to the classes of the samples. Several patterns can be observed from the plot of the raw data. The samples from the Anadel site (purple) had significantly larger concentrations of iron than the other classes. The Glass Mountain (SH) samples were noteworthy for having lower concentrations of calcium and titanium than the other classes. The samples from the Borax Lake

(BL) site had levels of strontium and barium that were near the limit of detection [76], and the strontium concentrations were below the limit of detection for the Glass Mountain samples as well. Observations which were below the limit of detection were reported using the measured value.

3.3 Results and Discussion

For each of the simulated datasets, the calculated subspace angles and imbedded errors vary somewhat based on the exact realization of the noise. To account for the variation in the relevant comparison metrics due to noise realization, a Monte Carlo approach in which 1000 different realizations of the noise matrix $\mathbf{E}$ were calculated while keeping the noise-free data matrix $\mathbf{X}_{pure}$ the same for each noise realization. Imbedded errors, and the scores and loadings subspace angles were calculated from PAF and PCA models for each realization of the noise. The measurement error estimates were calculated for a single realization of the data.

3.3.1 Simulated Dataset 1

In simulated Dataset 1, the data consisted of a three-component mixture of Gaussian peaks, with heteroscedastic noise added. PCA and PAF models with three components were calculated, where the PCA models used the mean-centered data. The amount of noise added was relatively small (the relative standard deviation was only about 1%), and the noise was not dramatically heteroscedastic. It was expected that there would be some overlap in the performance of PAF and PCA, but that PAF would be slightly better due to the assumption of column heteroscedasticity.

The subspace angles and imbedded errors for PAF and PCA were calculated for 1000 realizations of the noise matrix. In Figure 3.6A, a histogram is used to visualize the distribution of loading subspace angles for PCA and PAF relative to the true subspace (of the error-free data) for 1000 realizations. The histogram shows the distribution of outcomes for both PAF and PCA, as well as the overlap of their respective distributions (purple regions are the result of overlap). The loadings subspace angles for PAF and PCA were almost completely overlapped for this dataset, and from this it can be concluded that the PCA and PAF subspace angles were not significantly different from each other. The overlap of the loading subspace angles is reasonable

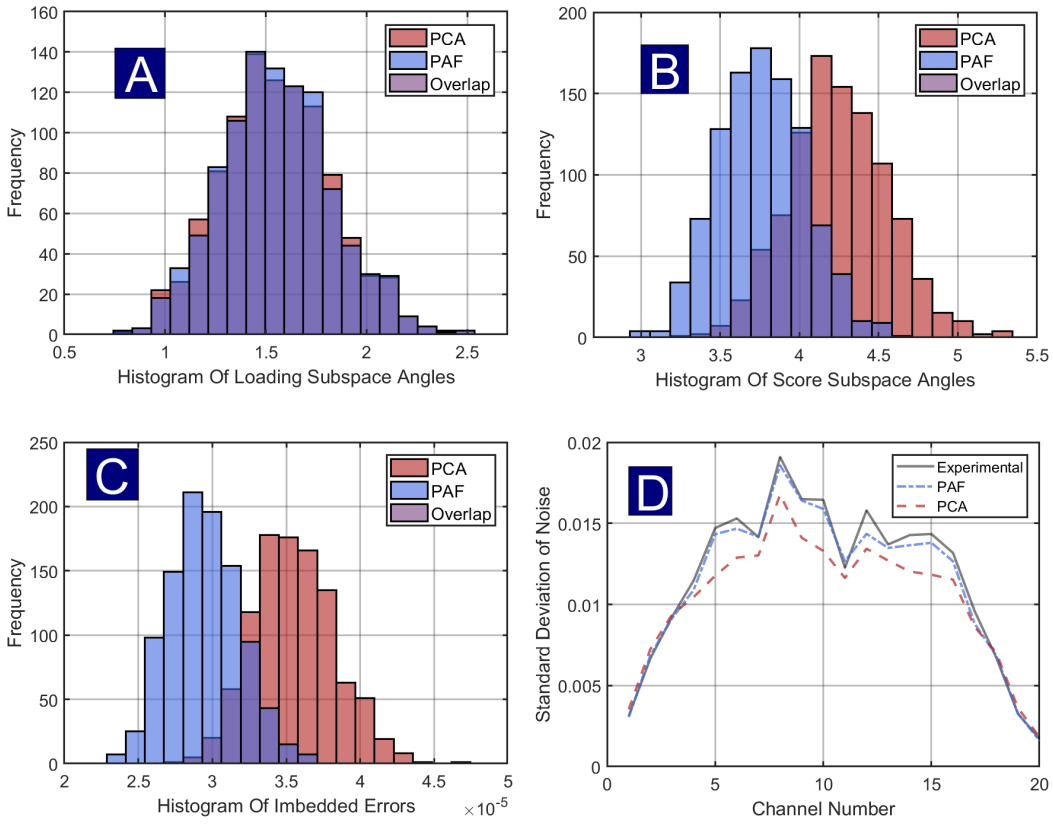

Figure 3.6: Summary of results for simulated Dataset 1. (A) Histogram of loading subspace angles (in degrees); (B) Histogram of score subspace angles (in degrees); (C) Histogram of imbedded errors; (D) Plot of the experimental noise standard deviations for each variable, and the estimated noise standard deviations from PCA and PAF. In subfigures A-C, the blue region of the histogram corresponds with PAF, the red region corresponds with PCA, and the purple region is the result of overlap of PCA and PAF.

because the measurement errors were not very large for Dataset 1, and the difference between PCA and PAF loadings is expected to be small when the measurement errors are small [72]. In contrast, histograms of the score subspace angles and imbedded errors are shown in Figures 3.6B and 3.6C, respectively. On average, PAF resulted in smaller score subspace angles and smaller imbedded errors than PCA, which makes sense when the principles of PAF and PCA are considered. PAF and PCA calculate the loadings in a rather similar manner, but for the calculation of the scores, PCA uses an unweighted (orthogonal) projection (which implicitly assumes that the errors are iid), whereas PAF uses the unique variances $\boldsymbol{\Psi}^2$ to calculate the scores via a

maximum-likelihood (oblique) projection, such that noisier variables receive less emphasis in the calculation. The distributions of the score subspace angles for PAF and PCA followed a highly similar pattern to the distributions of the imbedded errors. In Figure 3.6D, the estimates of the standard deviation of each column of the noise for PAF and PCA are plotted, along with the experimental standard deviation of the noise. These noise standard deviations are equivalent to relative standard deviations of approximately $1-2\%$. The PCA model underestimated the standard deviation of the noise in channels 4-17, although the estimates were not far off from the experimental values. The lower values for PCA are not surprising since the goal of PCA is to obtain the best fit of the data using a strictly least squares criterion. Consequently, it adapts to capture the higher noise variance in the middle part of the spectrum. Since the fit is better, the residual variance (and hence the estimated noise standard deviation) is lower.

The differences in the histograms for the scores and loadings may appear at first to be incongruous, but it is consistent with the fact that different spaces are being modeled. One can think of the estimation of the loadings as fitting each column of data. Because the errors within each column are homoscedastic, both methods produce similar results. Likewise, the scores calculation is analagous to fitting the rows. Since the errors in the rows are heteroscedastic, a weighted fit (analagous to PAF) will be more reliable than an unweighted fit (analagous to PCA). These differences are reflected in the histograms.

3.3.2 Simulated Dataset 2

The data in Simulated Dataset 2 used the same parameters as Dataset 1, except for the noise structure. In both datasets, the data consisted of three-component spectral mixtures, with noise added. In Dataset 1, the noise was a mixture of heteroscedastic noise and *iid* noise, but the level of *iid* noise was small enough that the overall errors followed a heteroscedastic structure. In Dataset 2, the level of heteroscedastic noise was kept the same as for Dataset 1, while the standard deviation of the *iid* noise was significantly increased. Overall, the standard deviation of the noise in each of the columns was greater for Dataset 2 than for Dataset 1, but because most of the noise added to Dataset 2 was *iid* noise, the overall noise structure was classified as

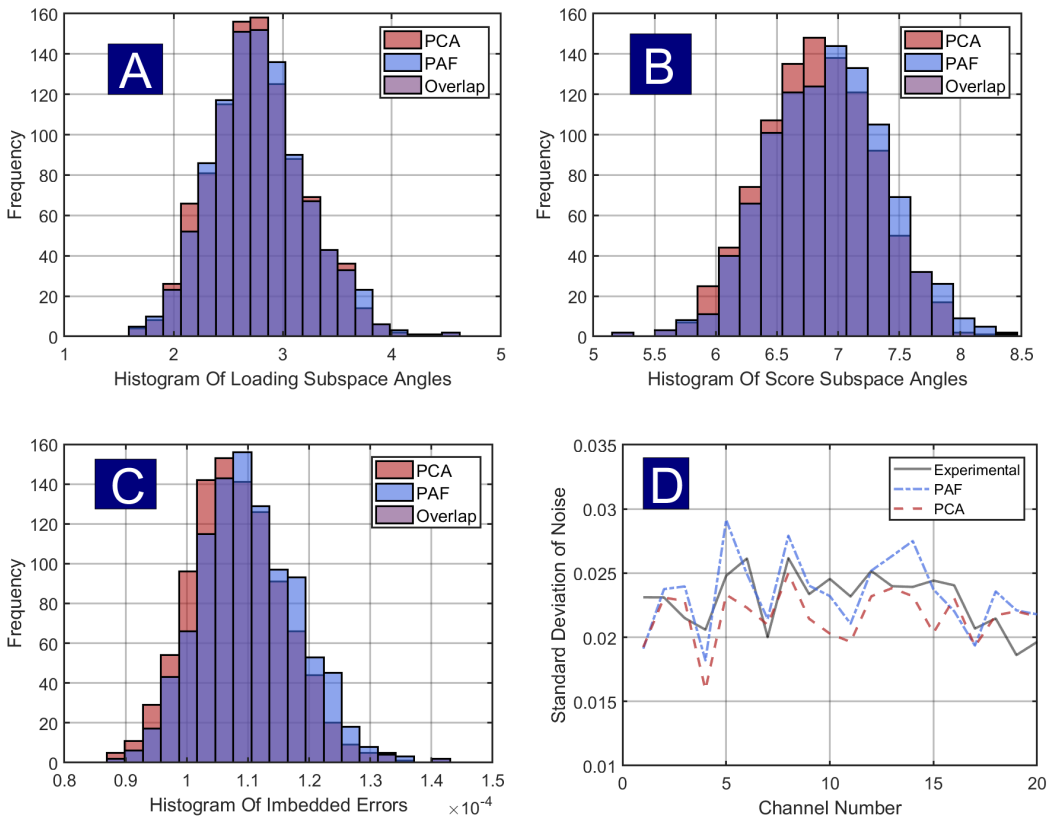

Figure 3.7: Summary of results for simulated Dataset 2. (A) Histogram of loading subspace angles (in degrees); (B) Histogram of score subspace angles (in degrees); (C) Histogram of imbedded errors. (D) Plot of the experimental noise standard deviations for each variable, and the estimated noise standard deviations from PCA and PAF. In subfigures A-C, the blue region of the histogram corresponds with PAF, the red region corresponds with PCA, and the purple region is the result of overlap of PCA and PAF.

approximately homoscedastic. It was expected that, for Dataset 2, the subspace angles and imbedded errors would be larger for both PAF and PCA compared with the results for Dataset 1 due to the increased noise levels, but that the subspace angles and imbedded errors for PAF and PCA would be similar due to the *iid* structure of the noise.

The results for simulated Dataset 2 are summarized in Figure 3.7. In Figure 3.7A, a histogram of the loading subspace angles for PCA and PAF for 1000 realizations of the noise matrix. There was not a significant difference in the loadings subspace angles of PAF and PCA but, as anticipated, the mean loading subspace angle was

larger for PAF. In Figure 3.7B-C, histograms of the scores subspace angles and the imbedded errors are shown, respectively. For both the scores subspace angles and the imbedded errors, PCA actually resulted in a slightly smaller subspace angle than PAF on average as well as slightly smaller imbedded errors, although both histograms were highly overlapped. The slightly better results for PCA can be anticipated by recognizing that it is constrained to assume that the measurement uncertainties are uniform (which is approximately true) whereas PAF is also fitting the uncertainties to improve the model. When the errors are significantly heteroscedastic, PAF has the advantage, but when the errors are close to homoscedastic, PAF will show a higher variance, which is reflected in the subspace angles.

In Dataset 1, the average score and loading subspace angles were approximately 1.5^o and 4^o respectively, whereas for Dataset 2 the average score subspace angle was about about 2.9^o, and the average loading subspace angle was about 7^o. The overall noise level was greater for Dataset 2, so it makes sense that the subspace angles were larger. The estimates of the standard deviation of each column of the noise for PAF and PCA, along with the experimental standard deviation of the noise, are plotted in Figure 3.6D. The estimated noise standard deviations were mostly between 0.02 and 0.025, although within that range the PAF and PCA estimates differed somewhat from one another, especially in channels 5 and 14.

The purpose of Simulated Datasets 1 and 2 was to investigate how the measurement error structure effects the performance of PAF and PCA as determined by subspace estimation, imbedded errors, and estimation of the noise. In simulated Dataset 1, when the noise was column heteroscedastic, the PAF models resulted in smaller score subspace angles, smaller imbedded errors, and more accurate measurement error estimates, when compared with PCA. In simulated Dataset 2, where the noise was distributed in a homoscedastic manner, PCA was (on average) slightly better than PAF as far as the subspace angles and imbedded errors were concerned, an the two methods resulted in estimates of the measurement errors that were similar to the experimental values.

3.3.3 Simulated Discrete Data

In the Discrete Dataset, the measured variables differed by several orders of magnitude in intensity, and the errors were proportional to mean intensity, but the proportionality constants were different for each variable. The reason for creating this dataset was to investigate whether PAF could be used as an alternative to scaling for this type of data, since proper scaling is difficult when there is both a difference in the range and/or units of the variables and heteroscedasticity. The PCA and PAF models of the Discrete Data used two components, and the PCA models were constructed using the autoscaled data. As was the case with Datasets 1 and 2, 1000 realizations of the noise were used to assess the variability in the subspace angles, imbedded errors, and the scores of PAF and PCA. It was expected that, due to the heteroscedastic errors, there would be a difference that would result in smaller score subspace angles and imbedded errors for PAF, but it was unclear as to whether or not the loading subspace angles and measurement error estimates would be significantly different.

The results for the simulated Discrete Data are summarized in Figure 3.8. Figure 3.8A shows a histogram of the loading subspace angles for the autoscaled PCA and PAF for 1000 realizations of the noise matrix. On average, PAF resulted in slightly smaller loading subspace angles than PCA, but the difference was not significant. In Figure 3.8B a histogram of the scores subspace angles is shown, and in Figure 3.8B depicts a histogram of the imbedded errors. For both the scores subspace angles and the imbedded errors, PAF resulted in significantly lower values such that the distributions for PAF and PCA were only partially overlapped. In Figure 3.8D, the estimates of the %RSD for PAF and PCA are plotted, along with the "true" %RSD. The PAF estimates were extremely close to the true values, whereas the estimates obtained from scaled PCA were slightly less close to the true values, although the scaled PCA estimates were still reasonable. Since PCA assumes that the errors are homoscedastic, when autoscaled data (where the variables have unit variance and zero mean) are used to calculate a PCA model, the assumption is that the relative standard deviation of the noise is equal for all variables. For the Discrete Data, the PCA assumption of equal relative standard deviations of the noise was not severely violated. If several of the variables had large relative standard deviations of the noise, a much larger difference between PCA and PAF might have been observed.

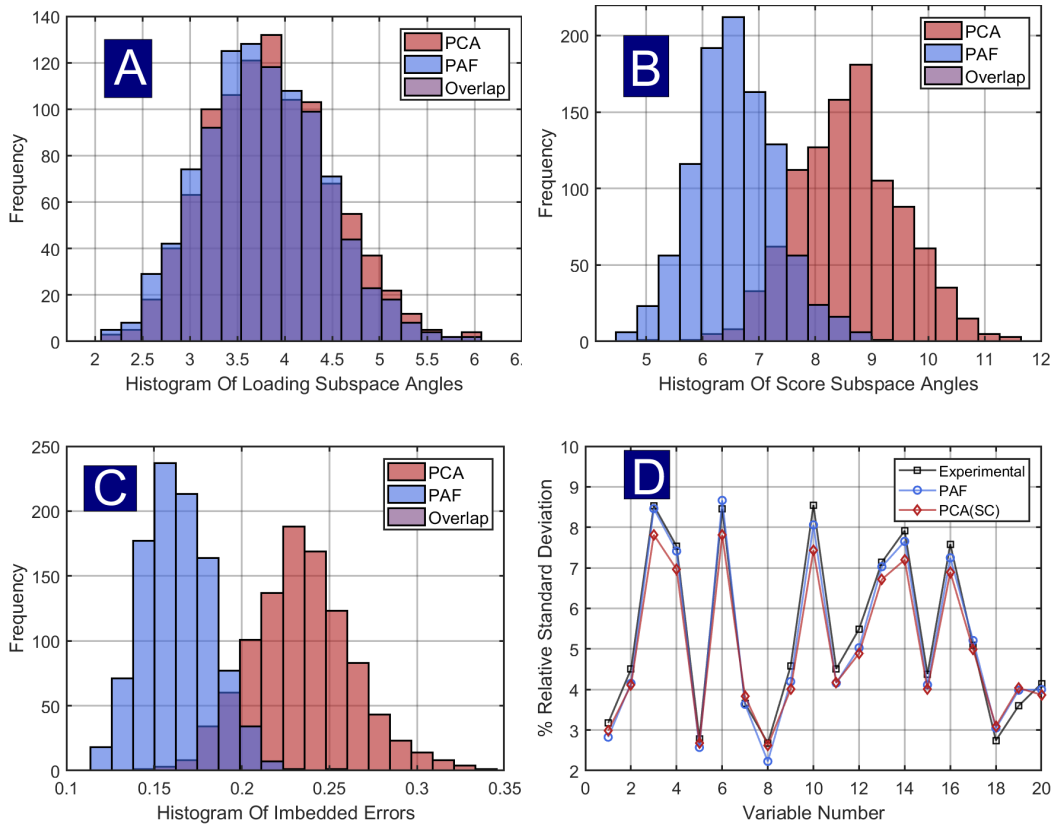

Figure 3.8: Summary of results for Discrete Data. Using the same set of noise-free data, 1000 different realizations of the noise were calculated, and PAF and autoscaled PCA were calculated in each instance. (A) Histogram of loading subspace angles (in degrees). (B) Histogram of score subspace angles (in degrees); (C) Histogram of imbedded errors; (D)Plot of the percent RSD for each variable, and the estimated %RSD values from PCA and PAF. In subfigures A-C, the blue region of the histogram corresponds with PAF, the red region corresponds with PCA, and the purple region is the result of overlap of PCA and PAF.

Another way to visualize the effect of the noise on the PAF and PCA scores is to use a scores plot. In Figure 3.9, a scores plot depicting scores 1 and 2 for samples 1-20 of the Discrete Data is shown. In the scores plot, the average scores are marked using the "+" symbol, and error ellipses of the scores are drawn, with PAF in blue and PCA in red. The average scores and error ellipses were calculated from generating PAF and PCA models for 1000 realizations of the noise, with the error ellipses representing the variation for 95% of the noise realizations. The locations of the average scores for PAF and PCA were nearly identical for each of the samples plotted. However, the error ellipses of the PCA scores were larger than the confidence intervals of the PAF scores,

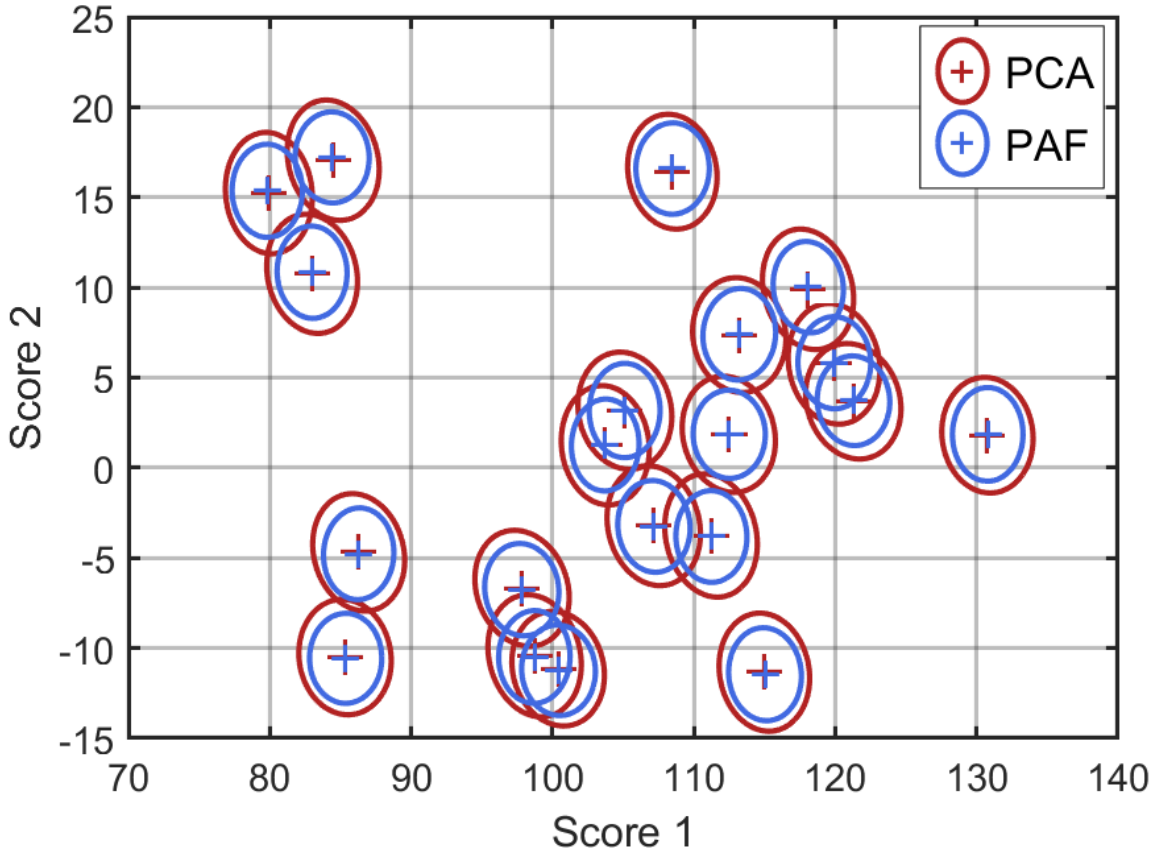

Figure 3.9: Scores plot for samples 1-20 of the data reconstructed by both PAF and PCA, over the course of 1000 different realizations of the noise. The average scores are denoted using the + symbol, and the variation in the each score is shown with the ellipsoids, which correspond to the 95% confidence interval. The PCA scores are plotted in red, and the PAF scores are plotted in blue.

which means that the PCA scores were more sensitive to the effects of the noise. This finding also indicates that the noise effected the variance in each of the scores, but it did not introduce a significant systematic difference in the scores. If the data had more noise, or if the models had inaccurately estimated the errors, the error ellipses would have been larger. The variation in the scores is of interest in classification problems, which often entails assigning boundaries to PCA scores for the purpose of separating the samples into distinct groups. Since PCA appears to introduce more variation in the scores than PAF when the errors are heteroscedastic, this suggests that samples near the class boundaries would be more likely to be misclassified when PCA scores are used.

The main findings for this dataset were that the measurement error estimates and

loading subspace angles were not significantly different between scaled PCA and PAF, whereas PAF resulted in improvements in the score subspace angles and imbedded errors, and PAF also led to decreased score variability compared with PCA.

3.3.4 Metals Data

The Metals Dataset was included in this work since it contained significant amounts of heteroscedastic noise, and had a measurement error structure that could be estimated using replicate measurements. The wavelengths on the ends of the spectra (300-350 nm and 620-650 nm) were extremely noisy, while between 400 nm and 600 nm the noise levels were small. This dataset presents a test for PCA, since PCA calculates the components such that they have maximal explained variance, and for these spectra the variation of the noise was comparable to the variation of the main chemical analytes. It was expected that when PCA was calculated using the Full spectra, that the model would be corrupted by the noise, whereas PAF would not result in significant issues. PCA and PAF models were also calculated for the Cut Metals Data, where the noisiest regions were removed. It was expected that for the Cut metals spectra, the differences between PCA and PAF would be small, since the noise structure for the Cut spectra was less severely heteroscedastic. All models were calculated using three components, and the PCA models were calculated using the mean-centered spectra. Autoscaling was not used for these data, because it generally has detrimental effects when used on spectra, as it tends to decrease the importance of variables with large chemical variation due to the presence of baseline regions.

For the experimental data, it is not possible to calculate subspace angles and imbedded errors because the true measurement space is unknown. However, the measurement error estimates can be compared to those from replicate data, and the scores and loadings from the two methods can be compared. The measurement error estimates for the Full Metals Data are shown in Figure 3.10A, while the measurement error estimates for the Cut Metals Data are shown in Figure 3.10B. For the Full Metals spectra, the noise standard deviation estimates for PAF were similar to the estimates of the noise that were calculated using the replicates, although the PAF estimates of the errors were slightly larger than the replicate-estimates in the regions from 300-350nm and 620-650 nm. The estimates of the noise standard deviations from PCA

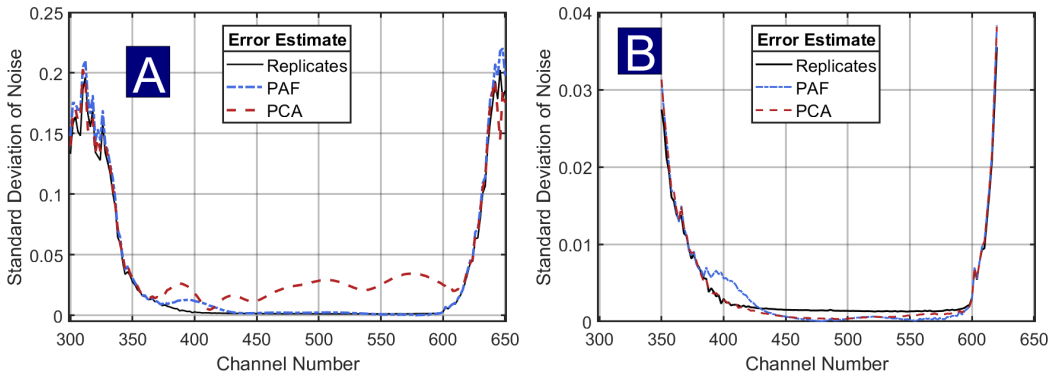

Figure 3.10: Results for estimated standard deviations of noise for Metals Data. (A) Full Metals Data; and (B) Cut Metals Data.

were slightly closer to the replicate-based estimates in the noisiest regions of 300-350 nm and 620-650 nm, but in the main chemical region of the spectra the PCA residual variance was very large relative to the true pattern. Further, in the PCA estimates of the errors, clear peak-like shapes in the estimates can be observed between 350 nm and 600 nm, which indicates that there was chemical variation present in the residuals. This is anticipated, since the principal components will attempt to model some of the noise variance at the expense of the chemical variance, leading to larger residuals from the latter. The PAF uncertainty estimates track the replicate values better in the low noise region, except for a small peak around 400 nm where the uncertainty estimates appear anomalously higher than the replicate estimates. Since this is coincident with a peak maximum, it may be a consequence of nonlinear behavior, but this is only speculation.

For the Cut Metals Data, the noise estimates of PAF, PCA, and the estimates from the replicates, were nearly identical to one another. There were two Heywood cases for the PAF model of the Full Metals spectra, such that the values of $\boldsymbol{\Psi}^2$ at wavelength the channels at 572 nm and 578 nm were slightly negative. For the Cut Metals spectra, the PAF model resulted in six Heywood cases, at the wavelength channels 478 nm, 480 nm, 482 nm, 552 nm, 564 nm, and 572 nm. For spectroscopic data, the Heywood cases are somewhat less of an obstacle to deal with, because the measurement error variance can be considered to be approximately homoscedastic over small intervals of wavelength channels. As a result, for spectroscopic data, Heywood cases can be

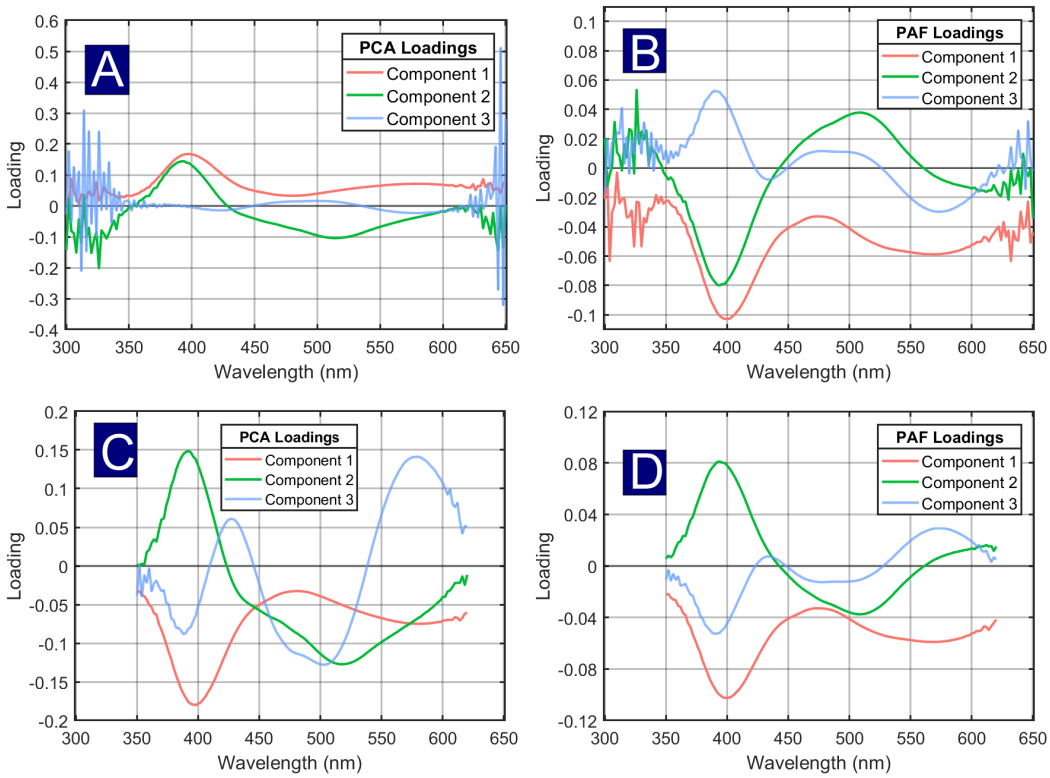

Figure 3.11: Loadings plots for Metals Data. (A) PCA (Full); (B) PAF (Full) (C) PCA (Cut) (D) PAF (Cut).

dealt with using weighted averaging or smoothing of the measurement error estimates. Both the PCA and PAF error estimates track the replicate estimates reasonably well, but both are lower than the replicate values in the middle region and the anomalous peak at 400 nm is still present for PAF. It is suspected that the higher value for the replicates may be due to the presence of baseline offset noise, which is a source of correlated noise that would be partly removed by PCA and PAF.

In Figure 3.11, loadings plots for PAF and PCA for both the Full and Cut Metals Data are plotted. The PAF loadings were calculated in the autoscaled space, and were scaled to the original space by multiplying by the column standard deviations of $\mathbf{X}$. In a loadings plot, the first component corresponds to the largest source of variance, and each successive component accounts for the remaining variance. Generally speaking, for a given component, the importance of the variables is related to the magnitude (positive or negative) of the loadings, such that variables with loadings close to zero are less important. The PCA loadings for the Full Metals Data are shown in Figure

3.11A. For component 1, the loadings were all positive, and the largest loading was for the 400 nm wavelength channel. For component 2, the loadings in the noisy region between 300 and 320 nm were of similar magnitude to the largest loadings in the main chemical regions of the spectra. The PCA loadings for component 3 were significantly impacted by the noise, as the loadings in the main chemical region were all close to zero, while the loadings were quite large in the noisy regions. The PAF loadings for the Full Metals Data are shown in Figure 3.11B. The loadings in the noisy regions were not equal to zero for the PAF loadings, but they were generally smaller relative to the loadings in the regions where the metals absorbed most intensely. In Figure 3.11C, the PCA loadings for the Cut Metals Data are plotted. The sign of the loadings for component 1 were reversed for this figure, to show the similarities with the PAF loadings for the Cut spectra shown in Figure 3.11D. For the Cut Metals Data, any remaining noisy variables do not have a significant impact on either the the PAF or the PCA loadings. The PAF and PCA loadings visually look somewhat different from one another, especially for component 3, but the spaces modeled by the two methods are likely the same (or highly similar).

In Figure 3.12, scores plots for PAF and PCA for the Full and Cut versions of the Metals Data are shown. In each of the scores plots, the samples are color-coded based upon the concentrations of each of the three metals. Scores plots can be useful for visualizing how the samples in a dataset are related to one another, with the expectation that samples with similar chemical composition will have similar scores. For the Metals data, the samples were created using a three-factor, three-level design, with five replicate samples per mixture. Therefore, it is to be expected that the scores of the samples will be related to their concentrations such that the design matrix becomes apparent, and it is also expected that the scores of the replicate samples should be highly similar. The PCA scores for the Full Metals Data are shown in Figure 3.12A. The PCA scores appeared to be distributed in a random manner, and the scores of replicates were not clustered at all. In Figure 3.12B, the factor scores $\mathbf{F}$ from the Full Metals Data clearly form a pattern resembling a data cube that reflects the experimental design, and one can even see where the "missing" mixture component would have been (bottom level, middle column, back row of the cube formed by the scores pattern). Further, the scores of replicate samples were highly

similar to one another. For the PCA scores from the Cut Metals Data in Figure 3.12C, and the PAF scores from the Cut Metals Data in Figure 3.12D, the mixture design is clearly evident as well. For the PAF scores, the design is somewhat more difficult to see from a single perspective, but from rotating the perspective of the scores plot it can be seen that the scores do form a 3D cube.

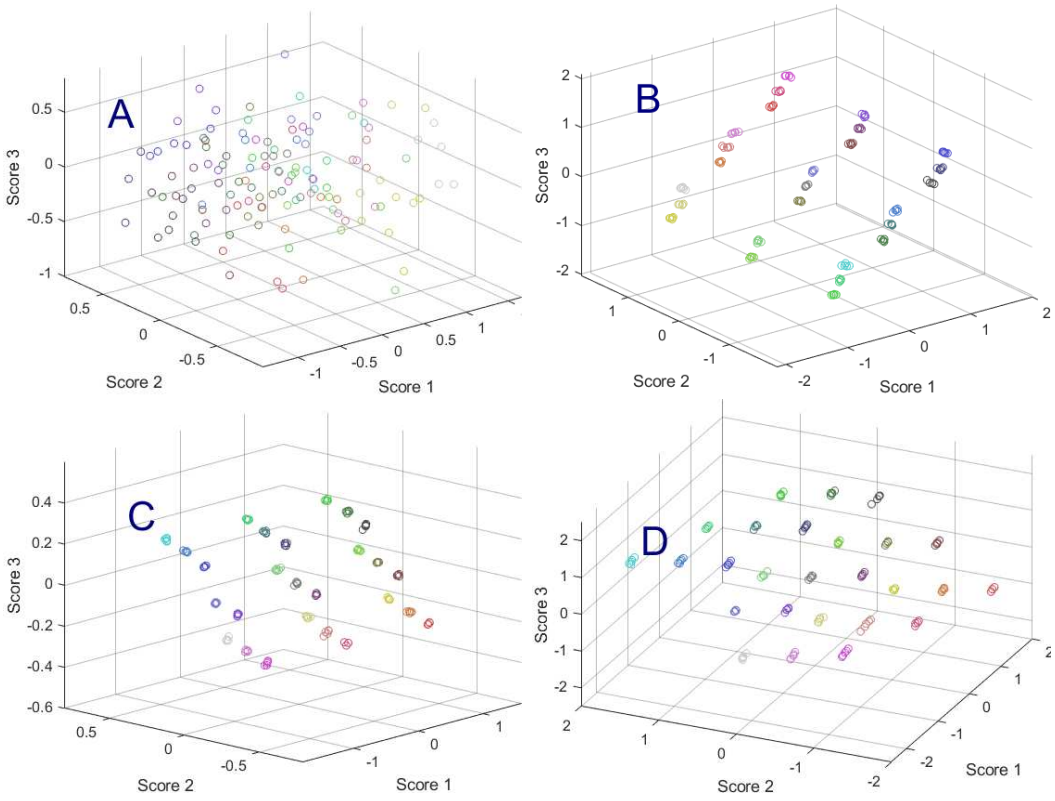

Figure 3.12: 3D scores plots for Metals Data. (A) PCA (Full); (B) PAF (Full) (C) PCA (Cut) (D) PAF (Cut).

Overall, for the Full version of the Metals Data, PCA resulted in a significant propagation of noise into both the scores and loadings, such that the mixture design could not be identified in the scores plot. PAF, on the other hand, was able to give an accurate estimate of the measurement errors, and the experimental design of the data was clearly evident in the scores plot. For the Cut version of the Metals Data, PAF and PCA performed similarly as far as the estimation of the measurement errors, loadings, and scores were concerned. The results demonstrate that, on the one hand, PAF can lead to a strong model performance even in the presence of extremely noisy variables, while the PCA model will be corrupted by the noise. On the other hand,

for this type of spectroscopic data, it is rather easy to identify and remove variables with high noise levels, and after removing the noisiest variables, PCA and PAF will behave similarly to one another.

3.3.5 Obsidian Data

The Obsidian Data consisted of measurements of trace elements for obsidian samples from four quarry locations in California and in this respect can be representative of the Discrete Data case studied in the simulations. This dataset was included in several early chemometrics software packages, and as a result the dataset is well-known in the field. PAF and PCA models with four components were calculated using the samples from the training set, and scores of the test set samples were calculated by projection. PCA models were calculated using both mean-centering and autoscaling. The parameters of interest for this dataset were the scores, and the measurement error estimates. For the scores, the objective was to see whether samples from the same quarry had similar scores or not, and to see whether the scores of the test set samples would cluster with the samples from the known quarries. The measurement error estimates for PCA and PAF were compared with values for the data that were reported by Duewer *et al* [78]. Given the wide range of the elemental concentrations, it was expected that mean-centered PCA would be the worst of the methods. It was also anticipated that PAF would fare slightly better than autoscaled PCA, since the errors in the elements were likely to be somewhat heteroscedastic even after scaling.

The results for the Obsidian Data are summarized in Figure 3.13, where scores plots of the first two components for each of the methods are shown, as well as the estimates of the relative standard deviation of the noise. In Figure 3.13A, the scores of mean-centered PCA are shown. The scores for the samples from the Anadel quarry (purple squares) are clustered in the bottom right quadrant of the plot, while the samples from the Glass Mountain quarry (green diamonds) cluster in the bottom left, although one of the Glass Mountain samples was somewhat different than the rest. The separation of the Anadel samples in the unscaled space is not surprising given the high and differentiating concentrations of iron (see Figure 3.5). The samples from the Konocti (K, red circles) quarry were somewhat closely clustered together, although the scores were overlapped with the samples from the Borax Lake (BL, blue

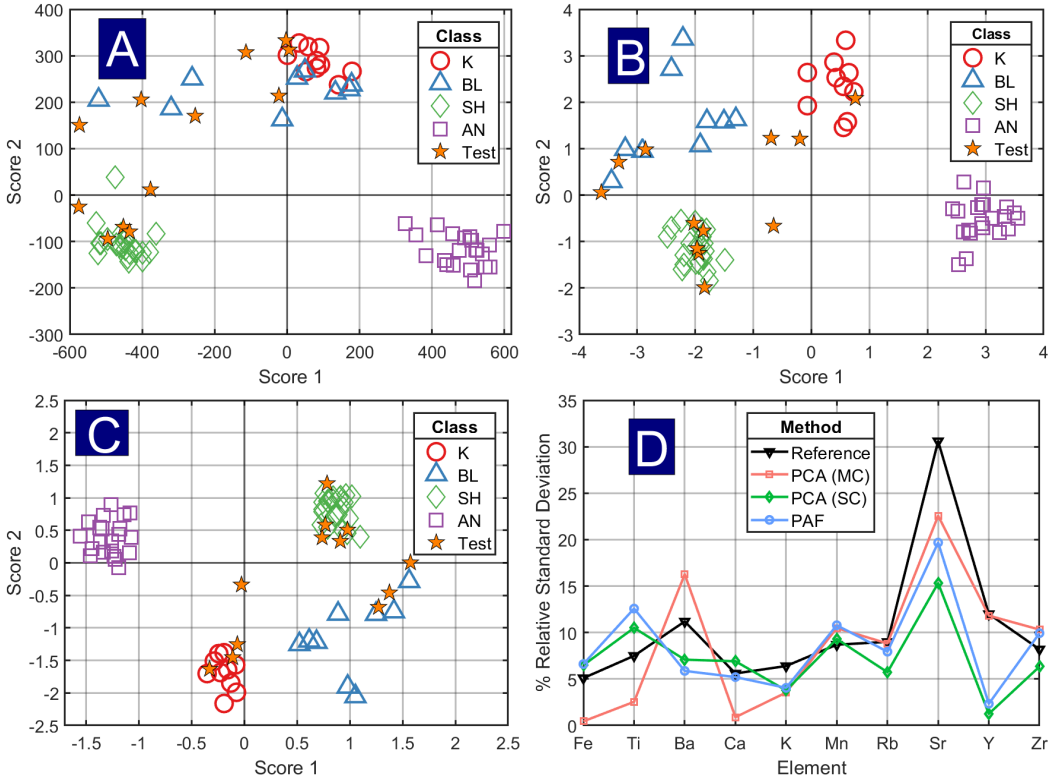

Figure 3.13: Results for the Obsidian Data. (A) Scores plot for mean-centered PCA; (B) Scores plot for autoscaled PCA; (C) Scores plot for PAF (D) Plot of estimates of the relative standard deviation of the noise for each method.

triangles), which had very weak clustering. Due to the overlap between the Borax Lake and Konocti samples, it is somewhat difficult to assign the classes for several of the test set samples. In Figure 3.13B, the scores plot from the autoscaled PCA are shown. The scores of the samples from the four quarries all clustered in different locations, such that the classes can be visually distinguished from each other. Of the unknown samples, three samples clustered with class BL, six samples clustered with class SH, one sample clearly clustered with Konocti, two samples were somewhat in-between the centers of the K and BL classes but perhaps slightly closer to Konocti, and one sample appears near the middle of the figure, and is considered to be an "outlier" that does not belong to any of the other classes [71]. The scores for PAF are shown in Figure 3.13C. For the PAF scores, the scores of the samples in each class were much closer together, with the exception of the Borax Lake quarry samples. For the test set samples, five clustered with the Glass Mountain samples, three were clustered with the Borax Lake samples, three were clustered with the Konocti samples, and

the "outlier" sample appeared near the middle of the plot. For this dataset, the PAF model was better for clustering the samples than either scaled PCA or mean-centered PCA.

In Figure 3.13D, the estimates of the percent relative standard deviation ($\%RSD$) of the noise are shown. The reference values of the percent relative standard deviation ($\%RSD$) of the noise %RSD were reported by Duewer *et al* [78], although the authors did not describe how the measurement uncertainties were calculated. Further details about the measurement errors were also found in a paper by Stevenson *et al* [76], where the authors stated that for these data, the relative standard deviations were, generally speaking, "on the order of 10%". The mean-centered PCA estimates of the measurement errors (shown in red) were very low for iron, titanium, and calcium, which had some of the largest variances, whereas the estimated errors in barium (which had low concentrations, and low variation) was larger than the estimates from the other methods. These results are not surprising, since the unscaled PCA will model the largest variances first so that, when proportional errors are present, it will model the noise for the large variables before the chemical variance in the smaller variables. Consequently, the (relative) residuals for larger variables are attenuated and those for smaller variables are amplified. With the notable exceptions of strontium and yttrium, the estimated measurement errors for PAF (blue) and scaled PCA (green) were between about 4% RSD and 15% RSD, which means that the errors generally were in the same range as both the estimates reported by Duewer, and the "order of 10%" reported by Stevenson. Both sets of error estimates also track the general behavior anticipated by Duewer, except for yttrium, which appears to be much lower than expected. For the estimation of the uncertainty in the strontium concentration, Duewer reported a relative standard deviation of 30.6%. Based upon the data reported by Stevenson, that RSD seems accurate for the samples from the Borax Lake and Glass Mountain sites, where the strontium concentration was near the limit of detection, but the samples from Konocti and Anadel had uncertainties of around 15%, so the overall uncertainty in strontium concentration is likely somewhere around 20%, which is what was reported by PAF. For the estimates of yttrium concentration, it seems unclear as to why PAF and PCA both resulted in such low estimates of the RSD.

The main findings for the Obsidian Data were that PAF led to tighter clusters and all the classes of the test set samples could be clearly identified (except for the outlier sample), whereas with the mean-centered PCA and scaled PCA the clusters were less well-defined and there were multiple test set samples whose class identity was unclear based upon the projected scores. The measurement error estimates for both PAF and scaled PCA were similar for most of the elements, and with the exception of yttrium, the relative standard deviations were similar to values reported in the literature.

3.4 Conclusions

In simulated Datasets 1 and 2, the effects of the measurement error structure on the relative performance of PAF and PCA were observed. When the measurement errors were column heteroscedastic, PAF resulted in improved subspace estimation and error estimation, whereas when the added noise was approximately homoscedastic, PCA was actually marginally better than PAF at subspace estimation. These simulation results were reinforced by the Metals Dataset, an experimental spectroscopic dataset with extreme heteroscedasticity at the edges of the spectrum. It was found that PCA performed significantly worse than PAF when all wavelength channels were used, but when the noisy variables were removed, PCA and PAF performed similarly as far as the estimation of the measurement errors, scores and loadings.

In the simulated Discrete Data, data with heteroscedastic errors and exponential differences in the ranges of the measured variables were modeled. It was found that PAF resulted in improvements in the score subspace angles, more accurate error estimates, and decreased variation in the scores when compared with PCA on the autoscaled data. These results were reflected in the experimental Obsidian Dataset, which consisted of trace element concentrations of samples from different locations. It was found that PAF resulted in improved clustering of samples when compared with PCA.

The findings of this study suggest that using PAF instead of PCA could be beneficial when the measurement errors are significantly heteroscedastic. While PAF generally resulted in more accurate measurement errors than PCA, the PCA estimates of the measurement errors from the residuals were not wildly inaccurate. Further, with the exception of the Full Metals Data, which were extremely heteroscedastic,

the space of the loadings for PAF and PCA were similar. These findings suggest that the primary advantage of PAF in the case of heteroscedastic data is in the use of the estimates of the measurement errors to calculate a maximum-likelihood projection of the scores. As a result, it is possible that the PCA residuals could be used to estimate the measurement errors, and maximum-likelihood scores for PCA could be calculated. Such a method would likely perform similarly to PAF in all but the most extremely heteroscedastic cases.

Chapter 4

Conclusions

Broadly speaking, in the analysis of multivariate data, there are three families of methods for non-*iid* measurement errors.

1. Presumptive methods, which assume that the measurement error follows a given structure, and adopt the data to fit the structure.

2. Determinate methods, which use information about the measurement error structure such as the error covariance matrix.

3. Empirical estimation methods, which try to both model the data and estimate the measurement errors.

Most of the conventional approaches for handling non-*iid* errors, such as autoscaling, fall into the first category. Autoscaling assumes that the relative standard deviation of the measurement errors is the same across the variables, such that when the data are scaled they become homoscedastic. methods may perform adequately when the assumed error structure is valid (or approximately so), but real data are often messy and have complex error structures, such that empirical methods may be inadequate and/or sub-optimal. The second category, of methods include techniques such as maximum likelihood principal components analysis (MLPCA) and multivariate curve resolution with weighted alternating least squares (MCR-WALS). When accurate information about the measurement error characteristics is available, such techniques can be extremely powerful, but measurement error information is often unavailable due to the added costs and time of collecting additional replicate measurements. The methods investigated in this thesis, weighted scatter correction methods (IDRC and VSN-based methods) and PAF, can be seen as belonging to the third category of simultaneous estimation methods.

The effects of weighted scatter correction methods were tested using three simulated datasets, and one experimental dataset of NIR spectra. It was found that when

the level of chemical background signal variation was low, and the level of the main chemical analyte variation was high, the corrections made by SSNV and IDRC resulted in significant reductions in prediction errors when compared with conventional SNV and MSC corrections. However, it was found that when the level of the chemical background signal was high, corrections using SSNV and IDRC resulted in larger prediction errors than using no correction at all. It remains to be seen how useful weighted correction methods will be for real data. The assumptions of SNV and MSC (low chemical background variation relative to scatter variation) are frequently satisfied for measurements of samples that are mostly homogeneous in composition. For weighted scatter correction methods, the assumptions are that a dataset contains a region with significant chemical variation while also having low background chemical variation, which are potentially contradictory. If spectra exhibit significant chemical variation in at least one region, there must be variation in the chemical composition of the samples, then there must be a certain amount of inhomogeneity in the samples. When the chemical (or physical) variation in the composition varies significantly, the effects on the spectra (especially those of NIR spectra) are often complex and nuanced, whereas the weighted correction methods still assume that a simple model of scatter is still valid.

PAF was shown to have advantages over PCA for data with heteroscedastic errors. PAF could potentially be extremely useful for a variety of problems in chemometrics. There are multiple areas where the measured variables are discrete and heteroscedastic, such as elemental analysis, metabolomics, and data fusion. One potential issue for using PAF in real chemical datasets is that of rank ambiguity. A real complex mixture, such as a gasoline sample or corn meal, may contain a small number of major analytes but may also have dozens (or possibly hundreds) of minor or trace analytes. As a result, there are often components with levels of variance at or around the noise level, such that it is difficult to assess how many components are "chemical" and how many are due to noise. For PAF, the number of components must be accurately estimated, or the model will not be valid.

These studies only involved two types of noise. There are many other noise structures which exist, and at the moment there are few empirical estimation methods which have been proposed that try to simultaneously estimate the errors while also

modeling the data. Such methods have the potential to be extremely useful, as they can potentially offer similar accuracy as known-error methods such as MLPCA in circumstances where the measurement error characteristics are unknown.

Bibliography

[1] Melchor C Pasikatan, James L Steele, Charles K Spillman, and Ekramul Haque. Near infrared reflectance spectroscopy for online particle size analysis of powders and ground materials. *Journal of Near Infrared Spectroscopy*, 9(3):153–164, 2001.

[2] Charles E Miller and Tormod Naes. A pathlength correction method for near-infrared spectroscopy. *Applied Spectroscopy*, 44(5):895–898, 1990.

[3] RJ Barnes, Mewa Singh Dhanoa, and Susan J Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied spectroscopy*, 43(5):772–777, 1989.

[4] P Geladi, D MacDougall, and H Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied spectroscopy*, 39(3):491–500, 1985.

[5] Federico Marini. Classification methods in chemometrics. *Current Analytical Chemistry*, 6(1):72–79, 2010.

[6] Rasmus Bro, Jerome J Workman JR, Paul R Mobley, and Bruce R Kowalski. Review of chemometrics applied to spectroscopy: 1985-95, part 3—multi-way analysis. *Applied Spectroscopy Reviews*, 32(3):237–261, 1997.

[7] Rasmus Bro. Review on multiway analysis in chemistry—2000–2005. *Critical reviews in analytical chemistry*, 36(3-4):279–293, 2006.

[8] Harald Martens, Tormod Naes, and Tormod Naes. *Multivariate calibration*. John Wiley & Sons, 1992.

[9] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[10] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[11] Kurt Varmuza and Peter Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*. CRC press, 2016.

[12] María J Culzoni, Héctor C Goicoechea, Gabriela A Ibañez, Valeria A Lozano, Nilda R Marsili, Alejandro C Olivieri, and Ariana P Pagani. Second-order advantage from kinetic-spectroscopic data matrices in the presence of extreme spectral overlapping: A multivariate curve resolution—alternating least-squares approach. *Analytica chimica acta*, 614(1):46–57, 2008.

[13] Romà Tauler and Damià Barceló. Multivariate curve resolution applied to liquid chromatography—diode array detection. *TrAC Trends in Analytical Chemistry*, 12(8):319–327, 1993.

[14] David W Osten and Bruce R Kowalski. Multivariate curve resolution in liquid chromatography. *Analytical Chemistry*, 56(6):991–995, 1984.

[15] R Tauler, M Viana, X Querol, A Alastuey, RM Flight, PD Wentzell, and PK Hopke. Comparison of the results obtained by four receptor modelling methods in aerosol source apportionment studies. *Atmospheric Environment*, 43(26):3989–3997, 2009.

[16] William H Lawton and Edward A Sylvestre. Self modeling curve resolution. *Technometrics*, 13(3):617–633, 1971.

[17] Anna de Juan, Joaquim Jaumot, and Romà Tauler. Multivariate curve resolution (mcr). solving the mixture analysis problem. *Analytical Methods*, 6(14):4964–4976, 2014.

[18] Joaquim Jaumot, Raimundo Gargallo, Anna de Juan, and Roma Tauler. A graphical user-friendly interface for mcr-als: a new tool for multivariate curve resolution in matlab. *Chemometrics and intelligent laboratory systems*, 76(1):101–110, 2005.

[19] David Hong, Laura Balzano, and Jeffrey A Fessler. Towards a theoretical analysis of pca for heteroscedastic data. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 496–503. IEEE, 2016.

[20] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[21] Peter D Wentzell, Tobias K Karakach, Sushmita Roy, M Juanita Martinez, Christopher P Allen, and Margaret Werner-Washburne. Multivariate curve resolution of time course microarray data. *BMC bioinformatics*, 7(1):343, 2006.

[22] Peter D Wentzell, Darren T Andrews, David C Hamilton, Klaas Faber, and Bruce R Kowalski. Maximum likelihood principal component analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(4):339–366, 1997.

[23] Peter D Wentzell. Measurement errors in multivariate chemical data. *Journal of the Brazilian Chemical Society*, 25(2):183–196, 2014.

[24] Claus Borggaard and Hans Henrik Thodberg. Optimal minimal neural interpretation of spectra. *Analytical chemistry*, 64(5):545–551, 1992.

[25] Yves Roggo, Pascal Chalus, Lene Maurer, Carmen Lema-Martinez, Aurélie Edmond, and Nadine Jent. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of pharmaceutical and biomedical analysis*, 44(3):683–700, 2007.

[26] Jerry Workman Jr and Lois Weyer. *Practical guide to interpretive near-infrared spectroscopy*. CRC press, 2007.

[27] Tom Fearn, Cecilia Riccioli, Ana Garrido-Varo, and José Emilio Guerrero-Ginel. On the geometry of snv and msc. *Chemometrics and Intelligent Laboratory Systems*, 96(1):22–26, 2009.

[28] Tobias K Karakach, Robert M Flight, Susan E Douglas, and Peter D Wentzell. An introduction to dna microarrays for gene expression analysis. *Chemometrics and Intelligent Laboratory Systems*, 104(1):28–52, 2010.

[29] Zuzanna Malyjurek, Dalene de Beer, Elizabeth Joubert, and Beata Walczak. Working with log-ratios. *Analytica chimica acta*, 2019.

[30] Pentti Paatero and Unto Tapper. Analysis of different modes of factor analysis as least squares fit problems. *Chemometrics and Intelligent Laboratory Systems*, 18(2):183–194, 1993.

[31] Gilles Rabatel, Federico Marini, Beata Walczak, and Jean-Michel Roger. Vsn: Variable selection for normalization. *Journal of Chemometrics*, page 3333updatethis, 2019.

[32] Yiming Bi, Liang Tang, Peng Shan, Qiong Xie, Yong Hu, Silong Peng, Jie Tan, and Changwen Li. Interference correction by extracting the information of interference dominant regions: Application to near-infrared spectra. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 129:542–550, 2014.

[33] Åsmund Rinnan, Frans Van Den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10):1201–1222, 2009.

[34] K Norris and P Williams. Optimization of mathematical treatments of raw near-infrared signal in the. *Cereal Chem*, 61(2):158–165, 1984.

[35] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[36] Harald Martens and Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *Journal of pharmaceutical and biomedical analysis*, 9(8):625–635, 1991.

[37] Harald Martens, Jesper Pram Nielsen, and Søren Balling Engelsen. Light scattering and light absorbance separated by extended multiplicative signal correction. application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, 75(3):394–404, 2003.

[38] Inge S Helland, Tormod Næs, and Tomas Isaksson. Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 29(2):233–241, 1995.

[39] Dorthe Kjær Pedersen, Harald Martens, Jesper Pram Nielsen, and Søren Balling Engelsen. Near-infrared absorption and scattering separated by extended inverted signal correction (eisc): analysis of near-infrared transmittance spectra of single wheat seeds. *Applied spectroscopy*, 56(9):1206–1214, 2002.

[40] Willem Windig, Jeremy Shaver, and Rasmus Bro. Loopy msc: a simple way to improve multiplicative scatter correction. *Applied spectroscopy*, 62(10):1153–1159, 2008.

[41] Q Guo, W Wu, and DL Massart. The robust normal variate transform for pattern recognition with near-infrared data. *Analytica chimica acta*, 382(1-2):87–103, 1999.

[42] Tomas Isaksson and Bruce Kowalski. Piece-wise multiplicative scatter correction applied to near-infrared diffuse transmittance data from meat products. *Applied spectroscopy*, 47(6):702–709, 1993.

[43] Yiming Bi, Kailong Yuan, Weiqiang Xiao, Jizhong Wu, Chunyun Shi, Jun Xia, Guohai Chu, Guangxin Zhang, and Guojun Zhou. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. *Analytica chimica acta*, 909:30–40, 2016.

[44] Emily Grisanti, Maria Totska, Stefan Huber, Christina Krick Calderon, Monika Hohmann, Dominic Lingenfelser, and Matthias Otto. Dynamic localized snv, peak snv, and partial peak snv: Novel standardization methods for preprocessing of spectroscopic data used in predictive modeling. *Journal of Spectroscopy*, 2018, 2018.

[45] Neal B Gallagher, Thomas A Blake, and Paul L Gassman. Application of extended inverse scatter correction to mid-infrared reflectance spectra of soil. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(5-7):271–281, 2005.

[46] Yifan Wu, Silong Peng, Qiong Xie, Quanjie Han, Genwei Zhang, and Haigang Sun. An improved weighted multiplicative scatter correction algorithm with the use of variable selection: Application to near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 2019.

[47] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[48] Christopher D Brown and Robert L Green. Critical factors limiting the interpretation of regression vectors in multivariate calibration. *TrAC Trends in Analytical Chemistry*, 28(4):506–514, 2009.

[49] Karl S Booksh and Bruce R Kowalski. Theory of analytical chemistry. *Analytical Chemistry*, 66(15):782A–791A, 1994.

[50] David M Haaland and Edward V Thomas. Partial least-squares methods for spectral analyses. 1. relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical chemistry*, 60(11):1193–1202, 1988.

[51] Ronald D Snee. Validation of regression models: methods and examples. *Technometrics*, 19(4):415–428, 1977.

[52] Charles Spearman. " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292, 1904.

[53] Charles Spearman. *The abilities of man.* Macmillan, 1927.

[54] Derrick Norman Lawley and Albert Ernest Maxwell. *Factor analysis as statistical method*. Butterworths, 1971.

[55] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[56] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[57] Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, and Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, 4(3):272, 1999.

[58] Jason W Osborne. *Best practices in exploratory factor analysis*. CreateSpace Independent Publishing Platform Louisville, KY, 2014.

[59] Edmund R Malinowski. *Factor analysis in chemistry*. Wiley, 2002.

[60] Marcel Maeder. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Analytical chemistry*, 59(3):527–530, 1987.

[61] Richard A Harshman et al. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84, 1970.

[62] J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3):283–319, 1970.

[63] Barbara Campisi, Giovanni Dugo, Antonella Cotroneo, and Luciano Favretto. Chemometric analysis and extraction processes of mandarin essential oils. *Analytica chimica acta*, 312(2):199–205, 1995.

[64] Igor V Pletnev and Vladimir V Zernov. Classification of metal ions according to their complexing properties: a data-driven approach. *Analytica Chimica Acta*, 455(1):131–142, 2002.

[65] Clemens Reimann, Peter Filzmoser, and Robert G Garrett. Factor analysis applied to regional geochemical data: problems and possibilities. *Applied geochemistry*, 17(3):185–206, 2002.

[66] P De Volder, R Hoogewijs, Roger De Gryse, Lucien Fiermans, and J Vennik. Maximum likelihood common factor analysis in auger electron spectroscopy. *Surface and Interface Analysis*, 17(6):363–372, 1991.

[67] Peter Persoone, P De Volder, and R De Gryse. The influence of cobalt on the oxysulfidation of brass. *Solid state communications*, 92(8):675–680, 1994.

[68] P Persoone, R De Gryse, and P De Volder. A new powerful transformation for maximum likelihood common factor analysis (mlcfa). *Journal of electron spectroscopy and related phenomena*, 71(3):225–232, 1995.

[69] P De Volder, R Hoogewijs, R De Gryse, Lucien Fiermans, and Joost Vennik. Chemical information obtained from auger depth profiles by means of advanced factor analysis (mlcfa). *Applied surface science*, 64(1):41–57, 1993.

[70] Peter Moens, P Devolder, R Hoogewijs, Freddy Callens, and Ronald Verbeeck. Maximum-likelihood common-factor analysis as a powerful tool in decomposing multicomponent epr powder spectra. *Journal of Magnetic Resonance, Series A*, 101(1):1–15, 1993.

[71] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014.

[72] Joost CF De Winter and Dimitra Dodou. Common factor analysis versus principal component analysis: a comparison of loadings by means of simulations. *Communications in Statistics-Simulation and Computation*, 45(1):299–321, 2016.

[73] Edmund R Malinowski. Theory of error in factor analysis. *Analytical Chemistry*, 49(4):606–612, 1977.

[74] Peter D Wentzell, Darren T Andrews, and Bruce R Kowalski. Maximum likelihood multivariate calibration. *Analytical chemistry*, 69(13):2299–2311, 1997.

[75] BR Kowalski, TF Schatzki, and FH Stross. Classification of archaeological artifacts by applying pattern recognition to trace element data. *Analytical Chemistry*, 44(13):2176–2180, 1972.

[76] DP Stevenson, FH Stross, and RF Heizer. An evaluation of x-ray fluorescence analysis as a method for correlating obsidian artifacts with source location. *Archaeometry*, 13(1):17–25, 1971.

[77] MA Sharaf, DL Illman, and BR Kowalski. Chemometrics. *New York*, pages 127–128, 1986.

[78] DL Duewer, BR Kowalski, and JL Fasching. Improving the reliability of factor analysis of chemical data by utilizing the measured analytical uncertainty. *Analytical Chemistry*, 48(13):2002–2010, 1976.