Improved Projection Methods for Exploratory Data Analysis in Chemistry

by

Siyuan Hou

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

at

Dalhousie University
Halifax, Nova Scotia
August 2012

© Copyright by Siyuan Hou, 2012

DALHOUSIE UNIVERSITY
DEPARTMENT OF CHEMISTRY

The undersigned hereby certify that they have read and recommended to the Faculty of Graduate Studies for acceptance a thesis entitled "Improved Projection Methods for Exploratory Data Analysis in Chemistry", by Siyuan Hou in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: August 15, 2012

External Examiner: _____

Research Supervisor: _____

Examining Committee: _____

_____

_____

Departmental Representative: _____

DALHOUSIE UNIVERSITY

DATE: August 15, 2012

AUTHOR: Siyuan Hou

TITLE: Improved Projection Methods for Exploratory Data Analysis in Chemistry

DEPARTMENT OR SCHOOL: Department of Chemistry

DEGREE: Ph. D. CONVOCATION: October YEAR: 2012

Permission is herewith granted to Dalhousie University to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions. I understand that my thesis will be electronically available to the public.

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

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

_____

Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

With the rapid development of modern instruments, chemical data have become more complex in both volume and structure, which imposes more demanding requirements for advanced data analysis tools. As a highly interfacial subject, chemometrics plays an important role in the extraction of information from chemical data. One of the applications of chemometrics is in exploratory data analysis, which aims to reveal structures present in the data prior to or in place of the formal testing of a hypothesis.

Among the different methods for exploratory data analysis, principal component analysis (PCA) may be the one most widely used in chemistry. When PCA is viewed as a subspace modeling technique from the perspective of maximum likelihood, it essentially assumes homoscedastic measurement errors. However, heteroscedastic errors are common in multivariate chemical data. Thus, PCA often fails to extract useful information in cases of significantly heteroscedastic errors. Maximum likelihood principal component analysis (MLPCA) has been developed to address heteroscedastic errors in multivariate data, but its application in exploratory data analysis has not been examined. Chapter 2 of this thesis describes strategies for exploratory data analysis in situations with highly heteroscedastic errors, including the application of MLPCA. A partial transparency projection (PTP) technique is also introduced to improve the visualization by using the measurement error information. Following from the work in Chapter 2, Chapter 3 proposes a new optimization algorithm for MLPCA model with non-zero intercepts.

Projection pursuit (PP) is another important method for exploratory data analysis. PP is less widely used compared with PCA, but is more powerful than PCA in many cases. One major reason for the limited applications of PP is the difficulty in implementing PP efficiently. Chapter 4 describes new algorithms, referred to as quasi-power methods, for the optimization of kurtosis that is used as an objective function for projection pursuit. As an extension to the work in Chapter 4, regularized projection pursuit (RPP), designed to deal with data that have a small sample-to-variable ratio, is proposed in Chapter 5. This method is particularly relevant in chemical applications because chemical data typically have few samples but many variables.

List of Abbreviations Used

ALS	Alternating least squares
FA	Factor analysis
GLS	Generalized least squares
GUI	Graphical user interface
HCA	Hierarchical cluster analysis
HPLC	High performance liquid chromatography
ICA	Independent component analysis
IR	Infrared
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
Minres	Minimum residuals
MLPCA	Maximum likelihood principal component analysis
MS	Mass spectrometry
NIPALS	Non-linear iterative partial least squares
NMR	Nuclear magnetic resonance
PCA	Principal component analysis
PDF	Probability density function
P-P plot	Probability-probability plot
PP	Projection pursuit
PTP	Partial transparency projection
QDA	Quadratic discriminant analysis
RR	Ridge regression
SVD	Singular value decomposition
SVM	Support vector machine
ULS	Unweighted least squares
YPDA	Yeast peptone dextrose adenine

Acknowledgments

I first want to give thanks to my wife Xiuhong Ji (Ph. D.) for her longtime support and understanding during my pursing this Ph. D. degree. Whenever I struggled with difficulties, she always gave me strengths.

In a friendly research environment, I have wonderful experiences to interact with other members in Peter D. Wentzell's group, including Hannes Hochreiner, Robert M. Flight, Joe Boutilier, and Rukhshinda Jabeen.

I would like to express my thanks to my committee members, Dr. Robert Beiko, Dr. Alan A. Doucette, and Dr. Hong Gu for their help during my program progress. Dr. Heather Andreas was my committee member in the early stage. I am grateful to her for her help as well. My thanks are also given to the external examiner, Dr. David H. Burns from McGill University, for his review of this thesis.

The work reported in this thesis was carried out under the instruction of Dr. Peter D. Wentzell. I am grateful to him for his professionalism, open-mindedness, and positive altitude. I want to give special thanks to Dr. Peter D. Wentzell for his great supervision.

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Chapter 1: Introduction

1.1 Multivariate Data

The rapid development of analytical techniques in chemistry enables today's analysts to obtain a larger volume of data in a much shorter time than the pioneers in the era of "wet chemistry", who employed techniques based on weighting, titration, and single channel spectroscopic and electrochemical measurements. Modern techniques such as high performance liquid chromatography (HLPC), infrared (IR) spectroscopy, mass spectrometry (MS), and nuclear magnetic resonance (NMR) spectroscopy can measure a single sample at multiple variables, such as retention time, wavelength, mass-to-charge ratio, and chemical shift, respectively. The data obtained by measuring multiple variables for a sample are referred to as multivariate data. Multivariate data obtained for multiple samples are often arranged in a matrix form where rows denote samples and columns represent variables, although in some situations they can be placed in a vector, cube, or hypercube. Multivariate data have become very common, not only in chemistry, but also in other areas such as physics, engineering, biology, and social science.

Multivariate data, in contrast to univariate data that are obtained by measuring a single scalar variable for each sample, are more complex and generally contain more information due to the intricate relationships among different variables. As different properties of a sample are measured, useful information that is not contained in one property may be retained in other properties. It is common that the combinations of the different variables can reveal important information that cannot be obtained through individual variables. As the combinations of observable variables cannot be measured directly, they are often referred to as latent variables. In different methods of multivariate analysis, the latent variables may also be referred to as principal components or factors. The latent variables may be able to reveal useful information contained in multivariate data. However, due to the complexity of multivariate data, the relationship between the variables and the information sought are generally not obvious and difficult to obtain without using advanced data analysis methods. The body of methodology for extracting information from multivariate data is called multivariate analysis [1,2], which is the generalization of univariate statistics and relies on the principles of multivariate statistics.

1.2 Exploratory Data Analysis

In multivariate data analysis, exploratory data analysis is often an important step used to discover useful information from data. The phrase "exploratory data analysis" was introduced by John W. Tukey, an American statistician with a Ph. D. degree in mathematics, a M. Sc. degree in chemistry, and a B. A. degree, in 1977 [3]. The primary objective of exploratory data analysis is to search for hypotheses worthy of testing without prior knowledge of the data structure. Tukey held that "it is important to understand what you can do before you learn to measure how well you seem to have done it" [3]. A simple example of exploratory data analysis is a two-dimensional plot of the relationship between two variables. Exploratory data analysis can be juxtaposed against confirmatory data analysis with the latter focusing on statistical hypothesis tests. Commonly used methods such as the *t*-test and F-test belong to confirmatory data analysis. Tukey also held that "neither exploratory nor confirmatory is sufficient alone" and "finding the question is often more important than finding the answer" [4], and both exploratory and confirmatory data analyses are important [4,5].

Exploratory data analysis is often performed through unsupervised methods. Unsupervised methods (unsupervised leaning), are in contrast with supervised methods (supervised learning) [6]. Unsupervised methods extract useful information from data without using any sample class information. In other words, without knowing the characteristic of each sample (*e.g.* class) in advance, unsupervised methods explore how the data are naturally organized. For example, suppose there are a group of tests of blood samples collected from normal and diseased subjects. In an unsupervised method, the origin of the samples is "unknown", but the results of the analysis may show the samples naturally separate into two clusters. In contrast, supervised methods use the sample class information to look for a mathematical function to separate samples of different classes in a low-dimensional space. For the current example, supervised methods search for variables in blood tests that can distinguish the normal and diseased subjects and establish a model that can be used for future prediction. Supervised methods generally need a set of samples called the training set to build the model, and another set of samples called the test set to validate the model. When the "supervisor" is discrete, the purpose of supervised methods is discrimination and classification [1,6] (discriminant analysis). When the

"supervisor" is a continuous variable, supervised methods largely become a problem in multivariate calibration and prediction (regression problem) [7] although plenty of other statistical approaches are also supervised methods. In exploratory data analysis, sample class information is often unavailable and useful information needs to be extracted by unsupervised methods. As unsupervised methods do not use class information in analyzing the data, the information obtained through unsupervised methods generally can be interpreted to be less biased than that from supervised methods, which are prone to overfitting.

In exploratory data analysis, data visualization is a commonly used technique for analysts to obtain useful information [3,6]. Unfortunately, human beings do not have the ability to visualize data in four or higher dimensions. Thus, it is necessary to map high-dimensional data to a low-dimensional space so that the useful information can be visualized. The mapping function may be non-linear or linear. Compared with non-linear mapping methods, linear mapping methods are generally simpler and widely used for dimensionality reduction. In chemistry, the most common method for dimensionality reduction is perhaps principal component analysis (PCA) [1,6,8,9,10]. Other methods such as factor analysis (FA) [1,6,11] and projection pursuit (PP) [12,13] are also used in different applications.

1.3 Motivation

Despite the versatility of the methods for exploratory data analysis, they do not address all the problems. Due to the so-called "curse of dimensionality" [14], dimensionality reduction of high-dimensional data to reveal salient data features in a low-dimensional space is far from trivial. It is acknowledged that every data analysis method has strengths and weaknesses, and no one method can be a panacea that works well for all situations. The complexity of data in different applications can easily make a method fail to extract useful information. Development of new methods and improvement of existing methods for exploratory data analysis are on-going areas of research activity that are further motivated by the emergence of new kinds of data.

From the perspective of data complexity, there are many factors that can prevent currently available data analysis methods from extracting useful information. One cause is that data may be contaminated by errors with complicated error structures. Data in

chemistry are generally obtained through experiments that involve measurements and measurements always have errors. Measurement errors in multivariate data are often heteroscedastic [15,16]. Heteroscedastic errors refer to a group of measurements that have different measurement error variances, in contrast to homoscedastic errors, where all the measurements have the same error variance. In the literature, although many multivariate data analysis methods exist, most of the methods do not take account of heteroscedastic measurement error structure. In other words, they assume homoscedastic measurement errors. When the measurement errors are almost homoscedastic, such an assumption is reasonable. However, it is widely observed that in multivariate data, heteroscedastic measurement errors are the general rule rather than an exception [17]. Heteroscedastic measurement errors in multivariate data often lead to the failure of conventional multivariate data analysis tools to give useful information.

From the perspective of data analysis methods, inefficiency of the current methods or lack of suitable data analysis methods can lead to the result that useful information is not extracted from data. A multivariate analysis method generally includes two major components: the objective function and the algorithm to optimize the objective function. On one hand, if the objective function is not well-defined, useful information may not be obtained. On the other hand, if the optimization algorithm is not efficient, the objective function may not be optimized, which impairs the utility of the method even if the objective function is well-defined. Thus, a good objective function and an efficient optimization algorithm are both important for a multivariate analysis method.

This thesis reports work on the development of new approaches to solve the two problems mentioned above in exploratory data analysis. The work focuses on the theoretical aspects over applications in chemistry, though applications are provided for demonstration purposes.

This thesis is divided into six chapters. Chapter 1 gives an introduction to multivariate data, exploratory data analysis, motivation of the work, and background information on error structures in multivariate data, as well as an overview of some projection methods for exploratory data analysis and some background information for numerical optimization in the literature.

Chapter 2 discusses maximum likelihood principal component analysis (MLPCA), a technique used to deal with data with heteroscedastic errors in multivariate data, for exploratory data analysis. The work of this chapter has been published in Journal of Chemometrics [18]. The limitations of principal component analysis (PCA) for exploratory data analysis in the presence of heteroscedastic errors are described and strategies to address these weaknesses that incorporate MLPCA are proposed. A new method to improve data visualization by incorporating measurement error information, referred to as the partial transparency projection (PTP), is developed. To illustrate the utility of these improvements, simulated data and DNA microarray experimental data are used.

Chapter 3 describes an efficient algorithm to optimize the objective function of MLPCA to deal with the case where data are not centered around the origin in multi-dimensional space; *i.e.*, intercepts are present. The original MLPCA was developed under the assumption that the data were not offset from the origin. However, in practice, experimental data often have intercepts for different variables. Assuming zero intercepts for multivariate data is not in accordance with the real situation and may negatively affect the result of MLPCA. In the original work describing MLPCA, it was proposed that the intercepts be included in the maximum likelihood estimation. However, the lack of an efficient algorithm for the optimization of the objective function has limited its application. The work reported in this chapter proposes a quick and simple algorithm for the optimization of the objective function of MLPCA to deal with data that exhibit intercepts. The efficiency of the proposed algorithm is demonstrated by simulated data.

Chapter 4 is based on a paper published as a featured article [19]. In this chapter, new algorithms, referred to as "quasi-power methods", are proposed to optimize kurtosis as an objective function (called the projection index) for projection pursuit. Projection pursuit (PP) is a powerful method for exploratory data analysis, but its utility has been greatly impeded, largely due to the difficulty in the optimization of projection indices. The new algorithms proposed in this work are simple, fast, and stable, which are expected to lead to more widespread use of PP in chemistry and other areas to extract useful information from multivariate data. The performance of the algorithms is evaluated using simulated and experimental data sets.

In Chapter 5, a new projection pursuit method, referred to as "regularized projection pursuit" is proposed. This method is designed to deal with data that have a small sample-to-variable ratio. Today's chemical data often have fewer samples but more variables, so the sample-to-variable ratio is small. When the normal projection pursuit method (using kurtosis as a projection index) is applied, samples may be still separated into clusters, but the separation is often meaningless. This is a limitation of PP. This limitation is mitigated by the proposed method. The principle of the proposed method and its optimization algorithms are described, and its utility is demonstrated with simulated and experimental data.

The conclusions of the work presented in this thesis are given in Chapter 6.

1.4 Measurement Error Structures in Multivariate Data

A measurement error is defined as the difference between the measured value and the true value. Measurement errors can arise from many different sources such as the sample, the operator, the instrument, or the environment. For a specific measurement, the measurement error associated with it may not be predictable, but it is generally assumed to follow a normal distribution. The standard deviation of measurement error is normally used as a measure to evaluate the reliability of the measurement. Measurement error standard deviation (uncertainty) is an important concept for analysts because it describes the extent to which the experimental data can be trusted. Most chemists are familiar with the concepts of systematic and random errors, but in this thesis, measurement errors are viewed from another perspective, which focuses on the uniformity of measurement error variances.

Multivariate data normally consist of multiple samples measured on different variables. The measurement errors for different measurements are generally heteroscedastic. This may arise from proportional error sources such as shot noise or source flicker noise [20] or from variations in noise characteristics across different channels of a detector, such as different wavelengths in spectroscopy. Measurement errors in multivariate data may be uncorrelated or correlated. Uncorrelated errors are independent of one another, while correlated errors are related. A good description of the error structure in multivariate data can be found in reference [16], in which the error structures are

divided into six cases. The error structures in multivariate data based on the literature are briefly recounted here.

The simplest case representing uncorrelated homoscedastic errors is pictorially shown in Figure 1.1 (a). For some multivariate data, this error structure is a good approximation to the real error structure. In multivariate analysis, many of the methods have assumed this type of error structure because of its simplicity.

Figure 1.1 (b) shows uncorrelated heteroscedastic errors in a multivariate data matrix, but the errors are still homoscedastic within a row or a column. An example of this case is when variables are of fundamentally different types (*e.g.* pH, temperature) or magnitudes (*e.g.* concentrations of different elements). In this case, the different variables have different measurement error variances, but the measurement variances within each variable for different samples are the same.

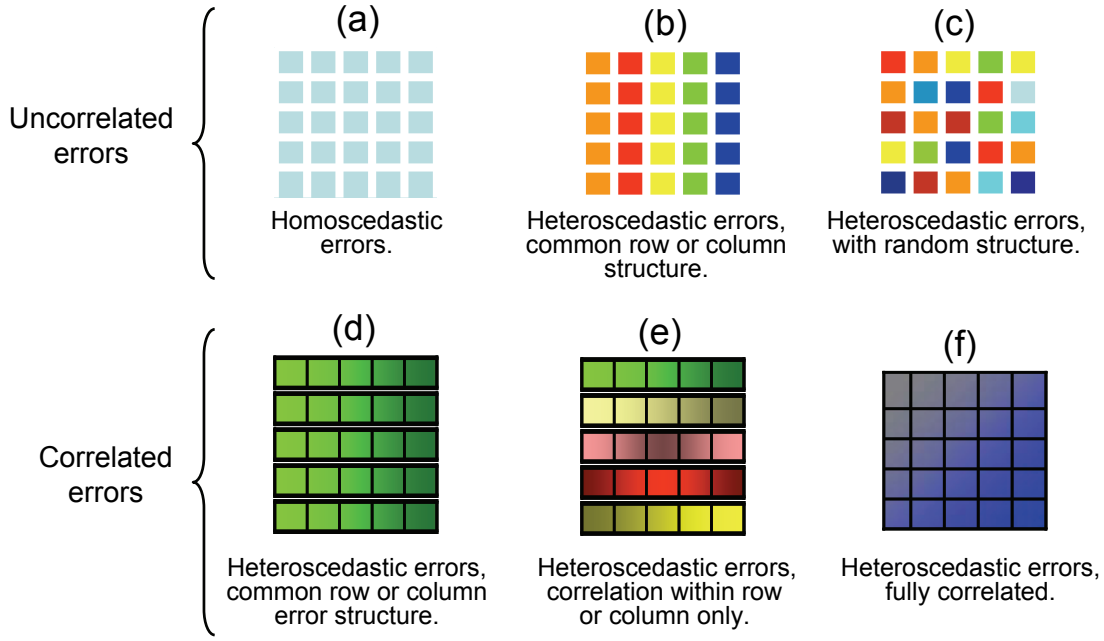

Figure 1.1 Pictorial representations of measurement error structures in multivariate data. Different colors indicate different error variances and the connectivity of the blocks indicates correlations of measurement errors (Adapted from reference [16]).

The third type of error structure, in which errors are uncorrelated but heteroscedastic randomly across the measurements, is shown in Figure 1.1 (c). For some types of multivariate data (*e.g.* DNA microarrays), this type of error structure is a good estimate of real situation. Compared with the correlated error structure, this case is still

relatively simple. In the later chapters of this thesis, this type of measurement error structure is assumed in most cases of heteroscedasticity.

Figure 1.1 (d) represents the error structure in which errors are correlated and heteroscedastic, but errors are correlated within rows or columns only and different rows or columns share the same error structure.

The fifth case is similar to the fourth. Shown in Figure 1.1 (e), it describes the situation when heteroscedastic errors are correlated within rows or columns only, but different rows or columns do not share the same error structure.

The fully correlated and heteroscedastic error structure is shown in Figure 1.1 (f). This denotes the most complicated situation and perhaps describes the most general error structure. All other cases can be viewed as special cases of this, but due to its complexity, it is less commonly assumed in multivariate data analysis.

1.5 Measurement Error Covariance Matrix

For univariate data, the standard deviation or variance is used to describe the measurement uncertainty. As a generalization, a covariance matrix is employed to describe the error structure for multivariate data. A covariance matrix, denoted by $\mathbf{\Sigma}$, can be written as

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \cdots & \cdots & \ddots & \cdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}. \tag{1.1}$$

The meaning of this covariance matrix can be explained as follows. Suppose a chemical sample is measured at *p* different channels (variables) and a measurement vector is obtained. Each channel has a measurement error variance. The measurement error variances of different channels may be uniform or non-uniform, and errors across the channels may be uncorrelated or correlated. A covariance matrix can be used to describe the measurement error structure for this vector of measurements. The diagonal elements in this covariance matrix give the measurement error variances for different channels, which are denoted by $\sigma_1^2, \sigma_2^2, \ldots,$ and σ_p^2 for channels 1 to *p*. If heteroscedastic errors are assumed, $\sigma_1^2, \sigma_2^2, \ldots,$ and σ_p^2 are not equal. The off-diagonal elements give measurement error

covariances between two different channels which are defined as, for example, $\sigma_{12} = \mathrm{E}(\varepsilon_1 \varepsilon_2)$, where ε_1 and ε_2 represent the errors for channel 1 and channel 2, respectively, and E denotes the expectation operator. Unlike variance, which is always non-negative, covariance can be negative. Covariance reflects the statistical correlation between different measurement errors. Independent errors have zero covariance, but correlated errors have non-zero covariance, either positive or negative. The error covariance matrix is an important concept in multivariate data analysis.

As an example of heteroscedastic correlated errors, consider a fluorescence emission spectrum collected from a fluorescent sample. Errors can arise from sources like the thermal noise in the detector, as well as fundamental noise such as shot noise. Because the latter is proportional to the square root of the signal, this will lead to heteroscedastic noise, resulting in variation along the diagonal of $\mathbf{\Sigma}$. Additionally, correlations in measurement errors can be introduced by variations in the source intensity (flicker) during scanning (photomultiplier detection) or crosstalk in the channels (array detection). Baseline offset and signal processing (filtering, smoothing) can also lead to correlated errors, resulting in a complex structure for the off-diagonal elements of $\mathbf{\Sigma}$.

The error structures shown in Figures 1.1 (a)-(e) can be described by different error covariance matrices corresponding to the rows or columns of the original data matrix. However, for the last case, there must be a larger covariance matrix to describe the error structure for all of the measurements as a whole. This is a complicated situation and will not be discussed in detail. It is important to note that, for uncorrelated errors (Figures 1.1 (a)-(c)), the covariance matrix is diagonal in the sense that the off-diagonal elements are zeros, while for correlated errors they are not.

1.6 Projection Methods for Exploratory Data Analysis

In multivariate analysis, there are many methods that can be used for exploratory data analysis. This section does not aim to give a complete review of all of the methods, but focuses only on some projection methods commonly used in chemistry that are related to the work reported in this thesis. The methods reviewed in this section include factor analysis (FA), principal component analysis (PCA), maximum likelihood principal component analysis (MLPCA), and projection pursuit (PP).

The notations in this thesis follow a commonly used convention in mathematics. A lower case bold letter is used to denote a column vector. A row vector is always expressed as the transpose of a column vector with the superscript "T" denoting the transpose operator. An upper case bold letter is used to represent a matrix. A scalar or scalar function is designated by an italic non-bold letter.

1.6.1 Factor Analysis (FA)

Factor analysis, like many multivariate methods, can trace its origins to the early work in the social sciences, where the importance of multiple variables was recognized at an earlier stage than other sciences. The origin of FA is credited to the work of Spearman [21 , 22] in 1904 when he published his article "General intelligence, objectively determined and measured" in the field of psychology [11]. Later researchers extended his work and created many variants of FA, such as multiple factor analysis [23], alpha factor analysis [24], maximum likelihood factor analysis [25,26], canonical factor analysis [27], and image factor analysis [28]. A few books discussing factor analysis can be found in references [29,30,31]. The general model of FA can be written as

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \tag{1.2}$$

where $\mathbf{x}$ is a p x 1 column vector denoting a group of observable random variables, $\mathbf{z}$ is a d x 1 column vector representing a group of latent (unobservable) random variables (called factors in FA), $\mathbf{W}$ denotes a p x d ($p > d$) matrix generally called the loading matrix, which linearly transforms the latent variables ($\mathbf{z}$) to observable variables ($\mathbf{x}$) , $\boldsymbol{\mu}$ is a p x 1 column vector denoting the means of observable random variables, and $\boldsymbol{\varepsilon}$ is a p x 1 column vector representing the random errors or residuals.

In case of multiple samples (realizations), a data matrix is obtained. This model can then be expressed as

$$\mathbf{X} = \mathbf{Z}\mathbf{W}^{\mathrm{T}} + \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{E}, \tag{1.3}$$

where $\mathbf{X}$ is an n x p observed data matrix, with n denoting the number of samples and p representing the number of variables. The two expressions (Equations (1.2) and (1.3)) may be confusing to chemists, but follow the convention in mathematics. Note that in Equation (1.2), rows in $\mathbf{x}$ denote variables, but in Equation (1.3), columns in $\mathbf{X}$ represent variables and rows denote samples. The unknown n x d matrix $\mathbf{Z}$ is generally called the

scores matrix. The quantities **W** and $\boldsymbol{\mu}$ retain the same definitions as in Equation (1.2). The boldface "**1**" is an n x 1 column vector with all its elements being 1's, and **E** is an n x p matrix of errors.

The model in Equation (1.2) or (1.3) itself is not enough to define FA since there are an infinite number of solutions, but with additional assumptions and constraints on the parameters, the FA model can be uniquely determined. Depending on the assumptions and constraints imposed, the model in Equation (1.2) or (1.3) can evolve into different decompositions, including PCA. For FA, the latent variables **z** (in Equation (1.2)) are assumed to follow a multivariate normal distribution,

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}). \tag{1.4}$$

The error term ($\boldsymbol{\varepsilon}$) (in Equation (1.2)) is also assumed to be multivariate normal, with errors that are independent but not identically distributed, expressed mathematically as

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi}), \tag{1.5}$$

where $\boldsymbol{\Psi}$ is diagonal with unequal diagonal elements. It is also assumed that **z** and $\boldsymbol{\varepsilon}$ are independent, so

$$\operatorname{cov}(\mathbf{z}, \boldsymbol{\varepsilon}) = \mathbf{0}. \tag{1.6}$$

With these assumptions and constraints, the observable variables (**x**) should follow a multivariate normal distribution as well,

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}). \tag{1.7}$$

Based on the FA model, there are three parameters to be estimated: $\boldsymbol{\mu}$, **W**, and $\boldsymbol{\Psi}$. The best estimate for $\boldsymbol{\mu}$ is the sample mean vector. There are several different estimation methods for **W** and $\boldsymbol{\Psi}$ existing in the literature and three of these are the most important [32,33]. The first one is the unweighted least squares (ULS) method, which minimizes the sum of the squares of the differences between the observed sample covariance matrix and the underlying covariance matrix ($\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}$), mathematically expressed as

$$U = \frac{1}{2}\operatorname{tr}\left[\mathbf{S} - \left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right)\right]^2. \tag{1.8}$$

S is the sample covariance matrix with

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^{\mathrm{T}}, \tag{1.9}$$

where $\mathbf{x}_i$ denotes the p x 1 measurement vector for a single sample, $\bar{\mathbf{x}}$ represents the sample mean, and n is the number of samples. Optimization of the objective function in Equation (1.8) gives the principal component (a.k.a. principal factor) solution [1] because the loadings of FA are essentially the same as those of PCA. This might be the reason why some people think that PCA is the same as FA.

The second estimation method is based on the generalized least squares (GLS), which minimizes

$$G = \frac{1}{2}\mathrm{tr}\left[\mathbf{I} - \mathbf{S}^{-1}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right)\right]^2, \tag{1.10}$$

where $\mathbf{I}$ is the identity matrix. Optimization of the objective function in Equation (1.10) gives the so-called *Minres* (minimum residuals) solution.

The third estimation method is based on the principle of maximum likelihood. Colloquially, likelihood is a synonym for probability, but they are distinctly different in statistical usage. When the probability (or probability density) of an observed outcome is considered as a function of a set of underlying parameters of a statistical model, it is called likelihood. Estimating the unknown parameters by maximizing the likelihood function is called maximum likelihood estimation, and is widely used in statistical inference. The well-known least squares method in regression analysis (*e.g.* calibration in chemistry) can be regarded as an application of maximum likelihood estimation. For the FA model, the likelihood function of the observed data can be expressed as

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}}\left|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right)^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\right], \tag{1.11}$$

where the notation $|\bullet|$ denotes the determinant operator. Maximizing the likelihood function is equivalent to minimizing

$$M = \mathrm{tr}\left[\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right)^{-1}\mathbf{S}\right] + \log\left|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right| - p, \tag{1.12}$$

where p is the number of observable variables. Optimization of the likelihood function in Equation (1.12) gives the maximum likelihood solution. As this estimation is based on the principle of maximum likelihood, FA with this estimation method is called maximum likelihood factor analysis, which might be the most important variant of FA.

The maximum likelihood FA result can be regarded as the solution for $\boldsymbol{\mu}$, **W**, and $\boldsymbol{\Psi}$ that is most likely to give rise to the observed data in **X** given that the assumptions of the model are valid. These assumptions include the number of latent variables (or subspace dimensionality), *d*, as well as the multivariate normality and independence of the latent variables (**z**) and the measurement errors. The PCA solution (Equation (1.8)) can be considered as a special case of this providing the maximum likelihood solution, when it can also be assumed that the diagonal elements of $\boldsymbol{\Psi}$ are equal (*i.e.*, measurement uncertainties are the same, $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$).

It is important to note that the underlying covariance matrix $\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}$ is not well-defined because this has inherent ambiguity. It can be mathematically shown that

$$\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi} = \mathbf{W}\mathbf{R}\mathbf{R}^{-1}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi} = \mathbf{W}\mathbf{R}\mathbf{R}^{\mathrm{T}}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi} = \left(\mathbf{W}\mathbf{R}\right)\left(\mathbf{W}\mathbf{R}\right)^{\mathrm{T}} + \boldsymbol{\Psi}, \tag{1.13}$$

where **R** is an arbitrary rotation matrix with $\mathbf{R}^{\mathrm{T}}=\mathbf{R}^{-1}$. In other words, the loading vectors in **W** define the *d*-dimensional subspace that contains the latent variables, but any *d* different vectors within the subspace can do this equally well, so there is a rotational ambiguity that also must be addressed by the method. This rotational ambiguity provides the rationale for "factor rotation". In practice, it is usual to rotate the loadings until some "simpler structure" is obtained. It is also worth noting that subspace spanned by **W** is generally not nested within higher dimensional solutions. In other words, the (*d*-1)-dimensional subspace is generally not included in the *d*-dimensional subspace.

As for how to obtain the factor scores, different approaches have been proposed in the literature, but two methods are recounted here [1]. The first one is the weighted least squares methods, which minimizes

$$\min\left[\left(\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_i\right)^{\mathrm{T}} \boldsymbol{\Psi}^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_i\right)\right], \tag{1.14}$$

where $\mathbf{z}_i$ is the scores vector to be determined for the *i*th sample. The solution of this minimization gives

$$\mathbf{z}_i = \left(\hat{\mathbf{W}}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{W}}\right)^{-1} \hat{\mathbf{W}}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}\right), \tag{1.15}$$

where $\hat{\mathbf{W}}$ represents the estimate of **W**, $\hat{\boldsymbol{\Psi}}$ denotes the estimate of $\boldsymbol{\Psi}$, and $\bar{\mathbf{x}}$ designates the sample mean vector which is an unbiased estimate of $\boldsymbol{\mu}$. The second approach is the regression method, which use the properties of multivariate normal

distribution. Based on the assumptions in Equations (1.2), (1.4), and (1.5), the conditional probability density function of the observed data, given the latent variables (factor scores), follows a normal distribution

$$p(\mathbf{x} \mid \mathbf{z}) \sim N(\mathbf{W}\mathbf{z} + \mathbf{u}, \boldsymbol{\Psi}). \tag{1.16}$$

Based on Bayes' theorem [6], the posterior distribution is also multivariate normal,

$$p(\mathbf{z} \mid \mathbf{x}) \sim N\left(\mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right)^{-1}(\mathbf{x} - \mathbf{u}),\ \mathbf{I} - \mathbf{W}^{\mathrm{T}}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \boldsymbol{\Psi}\right)^{-1}\mathbf{W}\right). \tag{1.17}$$

Thus, the factor scores for *i*th sample can be estimated by

$$\mathbf{z}_i = \hat{\mathbf{W}}^{\mathrm{T}}\left(\hat{\mathbf{W}}\hat{\mathbf{W}}^{\mathrm{T}} + \hat{\boldsymbol{\Psi}}\right)^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}) = \left(\hat{\mathbf{W}}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{W}} + \mathbf{I}\right)^{-1}\hat{\mathbf{W}}^{\mathrm{T}}\hat{\boldsymbol{\Psi}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \tag{1.18}$$

where $\hat{\mathbf{W}}$, $\hat{\boldsymbol{\Psi}}$, and $\bar{\mathbf{x}}$ follow the same definitions for Equation (1.15). The major difference between the two methods is that the former follows the principle of maximum likelihood, while the latter is in accordance with maximum *a posteriori* probability. In other words, the *a priori* probability for $\mathbf{z}$ is incorporated in the latter method.

1.6.2 Principal Component Analysis (PCA)

PCA is probably the most widely used method in multivariate analysis because it is relatively simple to use and understand, and it provides useful results in many cases. Understanding the principles behind PCA is also helpful in appreciating other methods in multivariate analysis. PCA was originally developed independently of factor analysis and preceded it somewhat. General consensus holds that the method describing PCA was first used by Pearson in 1901 [8], but some claim that it was used even earlier in physics in 1829 [34] and in chemistry in 1878 [34,35]. In 1933, Hotelling [9,10] independently developed a method which was proven to be essentially the same as Pearson's method and proposed the name of "principal components". Like many other methods, PCA also has many variants and extensions such as robust PCA [36,37,38], two different versions of Bayesian PCA [39,40], kernel PCA [41,42], and two different versions of MLPCA [15,16,43,44,45].

The general model in Equation (1.2) or (1.3) is also applicable to PCA. The definitions of the parameters remain the same, but different assumptions and constraints are imposed. In Pearson's definition [8] the observable data (**X**) are orthogonally projected into a subspace spanned by **W** such that the mean squared distance between the original observed data and the projected data is minimized,

$$\min\left\{\mathrm{tr}\left[\left(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)-\left(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)\mathbf{W}\mathbf{W}^{\mathrm{T}}\right]^{\mathrm{T}}\left[\left(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)-\left(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)\mathbf{W}\mathbf{W}^{\mathrm{T}}\right]\right\}, \quad (1.19)$$

where **W** is assumed to be an orthonormal basis in the sense that **W** is composed of a set of orthogonal unit vectors, $\bar{\mathbf{x}}$ is the sample mean vector, and tr is the trace operator. In Hotelling's definition [9,10], the observable data are orthogonally projected into a subspace such that the variance of the projected data is maximized,

$$\max\left\{\mathrm{tr}\left[\mathbf{W}^{\mathrm{T}}\left(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}\right)\mathbf{W}\right]\right\}, \quad (1.20)$$

where the notations **W** and $\bar{\mathbf{x}}$ remain the same as those in Equation (1.19). Pearson's objective function minimizes the sum of squares of the residuals while Hotelling's objective function maximizes the sum of squares of the projections. Since the sum of the two objective functions is the total variance of the data, it is not surprising that minimizing one is equivalent to maximizing the other, and thus the two definitions are mathematically equivalent.

PCA can also be viewed from the perspective of maximum likelihood estimation. Unlike FA, PCA does not assume the latent variables (**z**) follow a multivariate normal distribution, but the matrix **W** is composed of a set of orthonormal column vectors and **Z** consists of a set of orthogonal column vectors. If the errors are normal, independent, and identically distributed,

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \quad (1.21)$$

where σ^2 is the error variance, then the likelihood (L) of the observed data, based on the model in Equation (1.2) or (1.3), can be expressed as

$$L=\prod_{i=1}^{n}\frac{1}{(2\pi)^{\frac{p}{2}}\left|\mathbf{I}\sigma^2\right|^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}-\mathbf{W}\mathbf{z}_i\right)^{\mathrm{T}}\left(\mathbf{I}\sigma^2\right)^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}-\mathbf{W}\mathbf{z}_i\right)\right], \quad (1.22)$$

where $\mathbf{x}_i$ represents the p x 1 sample measurement vector and $\mathbf{z}_i$ denotes the d x 1 scores vector (The transposed $\mathbf{z}_i$ is a row of **Z**) for sample i. Maximizing this likelihood function with respect to **W**, **Z**, and $\boldsymbol{\mu}$ essentially gives the same solution as those of Pearson's and Hotelling's definitions.

In PCA, the columns of **W** are loading vectors that are generally called principal components, and the estimated values for **Z** are generally called scores. In different situations, the scores matrix and loadings matrix may be denoted by different notations. It

is worth noting that solutions of PCA are nested, which means that the subspace estimated from a low-dimensional subspace model is contained within the solutions for higher-dimensional model. Thus, the loading vectors (principal components) hierarchically account for the variance of the data. It can be mathematically shown that the loading vectors of PCA are the eigenvectors of the sample covariance matrix. If all the loading vectors are used, PCA can be regarded as a rotation of the original coordinate system, which gives an interpretation of PCA from the geometrical perspective.

PCA is very closely related to singular value decomposition (SVD), which was proposed by Eckart and Young in 1936 [46]. They showed that any matrix can be factorized as $\mathbf{X} = \mathbf{USV}^T$, where **U** and **V** are orthonormal matrices and **S** is a diagonal matrix. It can be shown that if the data matrix **X** is column mean-centered, the PCA scores matrix can be obtained by $\mathbf{Z} = \mathbf{US}$ and the loadings matrix is the same as **V**. A good review of PCA with related methods can be found in reference [47].

1.6.3 Maximum Likelihood Principal Component Analysis (MLPCA)

In the literature, two different methods, both named as maximum likelihood principal component analysis (MLPCA), were developed independently. One was proposed by Wentzell *et al*. in 1997 [15], and the other was developed by Tipping and Bishop in the same year [43,44]. The latter was shown to be the same as Roweis' work published in 1998 [45]. Both versions of MLPCA can fall into the general model framework defined in Equation (1.2) or (1.3) with different assumptions and constraints, and both are subspace modeling techniques as well (although a variant of Wentzell *et al*.'s MLPCA version includes both a column vector offset and a row vector offset, and does not fall into the general model framework).

Wentzell *et al*.'s definition of MLPCA originated from the context of multivariate analysis in chemistry and has been applied to a variety of problems [48,49,50]. In contrast to FA but consistent with PCA, the latent variables (**z**) are not assumed to follow a multivariate normal distribution, but the scores matrix (**Z**) consists of a set of orthogonal column vectors, and the loading matrix (**W**) is composed of a set of orthonormal column vectors. Unlike PCA, the measurement errors may be hetersocedastic (correlated or uncorrelated), and the measurement error variances (or covariance matrices) are assumed to be known. With these assumptions and constraints,

maximizing the likelihood of the observed data defines this version of MLPCA. Strictly speaking, since the true population error variances are generally unknown and thus are estimated through experiments or a theoretical model, the result using estimated error variances (or covariance matrices) is not really maximum likelihood estimation. However, in practice, it is sufficient enough in many cases to improve the data analysis results. Wentzell *et al*. proposed different error structures (recounted in Section 1.4) for MLPCA. If the errors are not fully correlated over the measurements in **X**, the likelihood function can be written column-wise or row-wise. For example, if errors are only correlated within rows, the likelihood function for the observed data can be expressed as

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}} \left|\boldsymbol{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu} - \mathbf{W}\mathbf{z}_i\right) \right], \tag{1.23}$$

where $\boldsymbol{\Sigma}_i$ is the error covariance matrix for row *i*, and is assumed to be known. Depending on the application, the mean vector $\boldsymbol{\mu}$ may or may not be included. Maximizing the likelihood function (or, in practice, minimizing the negative of the log-likelihood) gives the maximum likelihood estimates for **W**, **Z**, and $\boldsymbol{\mu}$.

Tipping and Bishop's definition of MLPCA [6,43,44] is different from that of Wentzell *et al*., but is very close to FA. Like FA, this version of MLPCA also assumes the latent variables (**z)** follow a multivariate normal distribution,

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}), \tag{1.24}$$

but the errors are assumed to be normal, independent, and identically distributed (i.i.d. normal),

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{1.25}$$

In factor analysis, the errors are assumed independent but different variables may have different error uncertainties. In other words, $\boldsymbol{\Psi}$ is diagonal but the diagonal elements are not equal. In Tipping and Bishop's MLPCA, it is assumed that $\boldsymbol{\Psi} = \sigma^2 \mathbf{I}$. Thus, this version of MLPCA could be regarded as a special case of FA. The likelihood function of the observed data can be expressed as

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}} \left|\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2 \mathbf{I}\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}} \left(\mathbf{W}\mathbf{W}^{\mathrm{T}} + \sigma^2 \mathbf{I}\right)^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}\right) \right]. \tag{1.26}$$

Maximization of the likelihood function gives the solutions for $\mathbf{W}$, $\mathbf{\Psi}$, and $\boldsymbol{\mu}$. The scores can be obtained by Equation (1.18). It is shown that the maximum likelihood estimate of $\boldsymbol{\mu}$ is equal to the sample mean, and the subspace spanned by $\mathbf{W}$ is the same as that of PCA. The scores of this version of MLPCA are the same as those of standard PCA except for a multiplicative factor. Thus, the results are essentially equivalent to those of PCA, but this MLPCA version is theoretically important.

1.6.4 Projection Pursuit (PP)

Compared with PCA or FA, projection pursuit (PP) is a relatively new technique. The term "projection pursuit" was first proposed by Friedman and Tukey in 1974 [12]. The primary purpose of PP is to linearly map high-dimensional data into a low-dimensional space so that some salient data features can be revealed, which is generally interpreted as looking for "interesting" projections in a low-dimensional subspace. Certainly, the notion of "interestingness" may have different interpretations in different applications, but it is often construed as clusters or outliers. PP does not have an unambiguous objective function (referred to as projection index), and there are many different projection indices in the literature [13,51,52,53,54,55]. In a broad sense, many other methods such as PCA can be regarded as special cases of PP.

PP is a dimensionality reduction technique, but the residuals cannot be interpreted as random errors. Unlike other methods mentioned earlier, PP generally cannot be viewed from the perspective of maximum likelihood estimation. Most of the projection indices are designed to measure the non-normality of the low-dimensional data. This is actually in accordance with the central limit theorem [56]. The central limit theorem states that a linear combination of a group of random variables tends to be normal. Thus, low-dimensional data with salient features should be far from a normal distribution and a normal distribution is interpreted to be "uninteresting". PP is closely related to another independently developed technique called independent component analysis (ICA) [57,58,59]. ICA searches for independent components that are conceptually "stronger" than uncorrelated components, in the sense that the requirement for independent components is higher than that for uncorrelated components. To many, "uncorrelatedness" is synonymous with "independence", but actually this is not true. Statistical independence for two random variables requires that the joint probability

density function (PDF) (or probability function for discrete distributions) is equal to the product of individual PDF's of the two variables, while "uncorrelatedness" means the covariance between the two variables is zero. Independent variables are always uncorrelated, but the converse is not true. More discussion about projection pursuit is given in Chapter 4.

1.7 Optimization Methods

As mentioned earlier, the optimization algorithm of a multivariate data analysis method is another important component. Probably because the optimization steps in most of the software packages are hidden and users only need to click menus and buttons to get results, one might have thought that optimization of an objective function is trivial, but it is not. Optimization of an objective function plays an important role and the efficiency of an optimization algorithm can determine if a method is successful. Optimization aims to find the solutions for the estimated parameters when an objective function defined in a data analysis method is optimized and it involves a variety of issues such as the convexity of the objective function and the convergence speed of the algorithm. Optimization is an old topic that can be dated back to the times of Isaac Newton (1643-1727) [60], but it became an independent subject in the mid-20th century due to the contributions of many researchers, such as Dantzig who coined the term "linear programming" for optimization in 1947 and Neumann who proposed the duality theory in the same year [61,62]. Now, a number of methods and their variants have been developed, from the simplex algorithm [61] to the genetic algorithm [63]. Mathematical optimization has become an important subject in different areas, from mathematics and computer science to chemometrics.

An optimization problem may be constrained or unconstrained. As the names imply, a constrained optimization problem has some constraints imposed on the objective function. For a constrained optimization problem, a Lagrange multiplier [64] is often introduced to avoid explicitly solving the constraints imposed on the objective function.

An objective function may be convex or non-convex. Simply speaking, a real-valued convex function is a continuous function for which the value at the mid-point in any interval in its domain is not larger (or smaller) than the arithmetic mean of values at the two interval ends. A non-convex function is just the opposite of a convex function.

A convex objective function has only one maximum (and/or one minimum), while a non-convex function has multiple maxima (and/or minima).

For the objective functions of multivariate data analysis methods, it is usual that there are no closed-form solutions and the solutions need to be found through iterative means. Among the various optimization methods, Newton's method and the gradient descent/ascent method may be the most familiar to data analysts. Newton's method generally converges quickly, but it involves the computation of the second-order derivatives, which is often difficult to implement for complicated functions, especially in high-dimensional spaces. The gradient descent/ascent method is conceptually simple, but it often converges slowly and is less stable. It is important to note that, when an objective function is convex, the commonly used methods such as Newton's method and the gradient descent/ascent method can guarantee the global solution, but when an objective function is non-convex, these optimization methods may hit a local optimum, since it has multiple local optima. In other words, the global optimum is not guaranteed. Thus, there is generally a requirement to start from many different initial guesses to increase the probability of finding the global solution. This imposes a high demand on the speed of an optimization algorithm.

For the multivariate data analysis methods used in chemometrics, the objective functions are generally complicated and optimization of the objective functions is a big challenge. Simple and efficient optimization algorithms for some current methods and new methods are in great demand. Thus, chemometricians often need to develop new optimization algorithms. Much of the work presented in this thesis focuses on such algorithms.

1.8 Bibliography

1. R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, Inc., New Jersey, 1998.
2. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Second Edition, John Wiley & Sons, Inc., New York, 1984.
3. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Inc., 1977.
4. J. W. Tukey, We Need Both Exploratory and Confirmatory, *The American Statistician*, 34 (1980) 23-25.
5. L. T. Fernholz, and S. Morgenthaler, A Conversation with John W. Tukey and Elizabeth Tukey, *Statistical Science*, 15 (2000) 79-94.
6. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, 2006.
7. D. C. Montgomery, and E. A, Peck, *Introduction to Linear Regression Analysis*, Second Edition, John Wiley & Sons, Inc., New York, 1992.
8. K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2 (1901) 559–572.
9. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part I), *The Journal of Educational Psychology*, 24 (1933) 417-441.
10. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part II), *The Journal of Educational Psychology*, 24 (1933) 498-520.
11. C. Spearman, “General Intelligence”, Objectively Determination and Measured, *The American Journal of Psychology*, 15 (1904) 201-292.
12. J. H. Friedman, and J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions and Computers*, 23 (1974) 881-890.
13. P. J. Huber, Projection Pursuit, *The Annals of Statistics*, 13 (1985) 435-475.
14. R. E. Bellman, *Dynamic Programming* (Republished version), Dover Publications, Inc., New York, 2003.
15. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, Maximum Likelihood Principal Component Analysis, *Journal of Chemometrics*, 11 (1997) 339-366.
16. P. D. Wentzell, Chapter 2.25: Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods, in S. D. Brown, R. Tauler, and B. Walczak (Editors): *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Elsevier Ltd., 2009.
17. W. A. Fuller, *Measurement Error Models*, Wiley, New York, 1987.

18. P. D. Wentzell, and S. Hou, Exploratory Data Analysis with Noisy Measurements, *Journal of Chemometrics,* 26 (2012) 264-281.

19. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.

20. J. D. Ingle, and S. R. Crouch, *Spectrochemical Analysis*; Prentice-Hall: Englewood Cliffs, New Jersey, 1988.

21. D. J. Bartholomew, Spearman and the Origin and Development of Factor Analysis, *British Journal of Mathematical and Statistical Psychology*, 48 (1995) 211-220.

22. S. A. Mulaik, Factor Analysis and Psychometrika: Major Developments, *Psychometrika*, 51 (1986) 23-33.

23. L. L. Thurstone, Multiple Factor Analysis, *Psychological Review*, 38 (1931) 406-427.

24. H. F. Kaiser, and J. Caffrey, Alpha Factor Analysis, *Psychometrika*, 30 (1965) 1-14.

25. R. I. Jennrich, and S. M. Robinson, A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 111-123.

26. K. G. Jöreskog, A General Approach to Confirmatory Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 183-202.

27. T. S. Rao, Canonical Factor Analysis and Stationary Times Series Models, *Sankhyā: The Indian Journal of Statistics, Series B*, 38 (1976) 256-271.

28. L. Stankov, Hierarchical Factoring Based on Image Analysis and Orthoblique Rotations, *Multivariate Behavioral Research*, 14 (1979) 339-353.

29. H. H. Harman, *Modern Factor Analysis*, Second Edition, The University of Chicago Press, Chicago, 1967.

30. E. R. Malinowski, *Factor Analysis in Chemistry*, Third Edition, John Wiley and Sons Inc., New York, 2002.

31. R. Cudeck, and R. C. MacCallum (Editors), *Factor Analysis at 100: Historical Development and Future Directions*, Lawrence Erlbaum Associates, Publishers, New Jersey, 2007.

32. K. G. Jöreskog, Factor Analysis by Least-Squares and Maximum-likelihood Methods, in K. Enslein, A. Ralston, and H. S. Wilf (Editors): *Statistical Methods for Digital Computers*, Volume III, John Wiley & Sons, Inc., 1977, pp. 125-153.

33. H. H. Harman, Minres Method of Factor Analysis, in K. Enslein, A. Ralston, and H. S. Wilf (Editors): *Statistical Methods for Digital Computers*, Volume III, John Wiley & Sons, Inc., 1977, pp. 154-165.

34. R. G. Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons Ltd., Chichester, England, 2007.

35. R. J. Adcock, A Problem in Least Squares, *The Analyst*, 5 (1878) 53-54.

36. G. Li, and Z. Chen, Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo, *Journal of American Statistical Association*, 80 (1985) 759-766.

37. M. Hubert, P. J. Rousseeuw, and S. Verboven, A Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60 (2002) 101-111.

38. M. Hubert, and S. Engelen, Robust PCA and Classification in Bioscience, *Bioinformatics*, 20 (2004) 1728-1736.

39. M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen, Bayesian Principal Component Analysis, *Journal of Chemometrics*, 16 (2002) 576-595.

40. C. M. Bishop, Bayesian PCA, in M. S. Kearns, S. A. Solla, and D. A. Cohn (Editors), *Advances in Neural Information Processing Systems*, MIT press, 1999, pp. 328-388.

41. B. Schölkopf, A. Smola, and K. R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, 10 (1998) 1299-1319.

42. B. Schölkopf, A. Smola, and K. R. Müller, Kernel Principal Component Analysis, *Lecture Notes in Computer Science*, 1327 (1997) 583-588.

43. M. E. Tipping, and C. M. Bishop, Probabilistic Principal Component Analysis, *Technical Report NCRG/97/010*, Neural Computing Research Group, Aston University, 1997.

44. M. E. Tipping, and C. M. Bishop, Probabilistic Principal Component Analysis, *Journal of the Royal Statistical Society, Series B*, 21 (1999) 611-622.

45. S. Roweis, EM Algorithm for PCA and SPCA, in M. I. Jordan, M. J. Kearns, and S. A. Solla (Editors): *Advances in Neural Information Processing Systems*, MIT press, 1998, pp. 626-632.

46. C. Eckart, and G. Young, The Approximation of One Matrix by Another of Lower Rank, *Psychometrika*, 1 (1936) 211-218.

47. P. Horst, Sixty Years with Latent Variables and Still More to Come, *Chemometrics and Intelligent Laboratory System*, 14 (1992) 5-21.

48. D. T. Andrews, and P. D. Wentzell, Applications of Maximum Likelihood Principal Component Analysis: Incomplete Data Sets and Calibration Transfer, *Analytica Chimica Acta*, 350 (1997) 341-352.

49. P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Maximum Likelihood Multivariate Calibration, *Analytical Chemistry*, 69 (1997) 2299-2311.

50. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.

51. M. C. Jones, and R. Sibson, What is Projection Pursuit? *Journal of the Royal Statistical Society, Series A (General)*, 150 (1987) 1-37.

52. J. H. Friedman, Projection Pursuit, *Journal of the American Statistical Association*, 82 (1987) 249-266.

53. P. Hall, On Polynomial-Based Projection Indices for Exploratory Projection Pursuit, *The Annals of Statistics*, 17 (1989) 589-605.

54. C. Posse, An Effective Two-Dimensional Projection Pursuit Algorithm, *Communications in Statistics - Simulation and Computation*, 19 (1990) 1143-1164.

55. I. S. Yenyukov, Indices for Projection Pursuit, in E. Diday (Editors): *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science Publishers, New York, 1989, pp. 181-189.

56. J. N. Miller, and J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, Fifth Edition, Pearson Education Limited, 2005.

57. P. Common, Independent Component Analysis, A New Concept? *Signal Processing*, 36 (1994) 287-314.

58. J. V. Stone, *Independent Component Analysis: A Tutorial Introduction*, Cambridge, The MIT Press, 2004.

59. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., New York, 2001.

60. P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*; Springer, Berlin, 2004.

61. G. B. Dantzig, and M. N. Thapa, *Linear Programming 1: Introduction*, Springer, New York, 1997.

62. G. B. Dantzig, and M. N. Thapa, *Linear Programming 2: Theory and Extensions*, Springer, New York, 2003.

63. M. Mitchell, *An Introduction to Genetic Algorithms*, The MIT Press, Cambridge, 1998.

64. J. Stewart, *Calculus: Early Transcendentals*, Fourth Edition, Brooks/Cole Publishing Company, New York, 1999.

Chapter 2: Exploratory Data Analysis with Noisy Measurements: An Application of Maximum Likelihood Principal Component Analysis*

2.1 Introduction

As introduced in Chapter 1, exploratory data analysis [1] is a widely used method in chemistry and many other areas. The term "exploratory data analysis" is a rather vague description of a variety of procedures that can be used to gain some understanding of the general characteristics of a data set and guide further investigation through more refined techniques. These characteristics can include features like the range of measurements in each variable, the nature of the noise, the relationships among samples or variables, and the presence of outliers. A variety of data analysis methods can be employed at this stage, from basic statistical and visualization techniques, to correlation maps and hierarchical clustering methods. However, there is perhaps no technique as widely applied to multivariate data analysis as principal component analysis (PCA) [2,3,4]. This is particularly true in chemistry, where PCA plays many different roles. These include dimensionality reduction, rank estimation, data visualization, and modeling, to name but a few. One of the most common applications often takes place during the early stages of data analysis, where PCA is used as a tool for exploratory data analysis.

To many, exploratory data analysis is nearly synonymous with PCA because it is such a powerful tool to accomplish multiple objectives. By performing an efficient dimensionality reduction that preserves the maximum amount of meaningful variance, PCA readily allows the visualization of the relationships among samples and variables through the use of scores and loadings plots. The use of eigenvalues and other metrics can also give an indication of the so-called pseudo-rank of the data, which is useful in assessing the inherent information content of the data or the number of components present in mixture analysis problems. Another important outcome of this type of analysis is often the visual recognition of clusters within the data. It is often the case that samples

* This chapter is based on the published article: P. D. Wentzell, and S. Hou, Exploratory Data Analysis with Noisy Measurements, *Journal of Chemometrics,* 26 (2012) 264-281.

or variables form clusters, and such separation suggests some internal relationships in multivariate data. The coupling of data compression with human perception is a powerful combination to look for information from data. Unlike supervised methods, such as discriminant analysis in its various forms [5,6,7,8,9,10], PCA does not use class information, so the appearance of clusters at this stage of analysis is more likely to reflect a true partitioning of the data. It should be pointed out that PCA is not the only method used to find clusters; many other methods such as clustering analysis [11,12,13] and projection pursuit (PP) [14,15,16] may also be employed.

Despite the versatility and power of PCA, it is not without problems and is by no means a panacea. Practitioners quickly learn that PCA is sensitive to outliers and robust PCA methods have been proposed in the literature [17,18]. PCA is also affected by preprocessing methods such as mean centering and scaling, and the failure to implement these when needed, or their application when unnecessary, leads to meaningless results. Mean centering is almost always used in exploratory data analysis, since it removes the mean effects of all of the variables and allows their interactions to be more efficiently visualized, but it may be undesirable in some cases. Scaling is a particularly sensitive issue. As PCA attempts to model the total variance of the data hierarchically as orthogonal principal components and scaling changes the variance in some directions, scaling can greatly affect the PCA results.

PCA can be viewed from the perspective of maximum likelihood estimation. For an $n \times p$ data matrix, PCA models measurements in a d-dimensional subspace ($d < n$ and $d < p$). When d is chosen to be the pseudo-rank of the data and all of the measurement errors are independently and identically distributed with a normal distribution (referred to as i.i.d. normal), then PCA estimation of the subspace will be optimal in a maximum likelihood sense [19,20]. However, when measurement error variances become non-uniform (heteroscedastic), the PCA estimation of the subspace becomes sub-optimal. Heteroscedastic errors may be roughly proportional to signals, such as shot noise or source flicker noise [21] or measurement-specific noise variance, such as noise in DNA microarrays [22,23,24]. When errors are heteroscedastic, but have fixed variance within a variable, scaling may make the errors more homogeneous (*i.e.* uniform measurement uncertainties) across all measurements, but it has long been known that optimal scaling is

only possible under certain conditions, namely when the matrix of measurement standard deviations is a rank one matrix [25]. Additionally, the assumption of independence in the measurement errors is often violated. In fact, as demonstrated in the literature, heteroscedastic and correlated measurement errors are the general case in multivariate data [26]. Small deviations from these conditions are not likely to cause serious problems and many routine analytical measurements are not problematic. Increasingly, however, analytical chemistry is stepping outside of traditional boundaries into areas such as proteomics, metabolomics, image analysis, surface science and environmental monitoring, where the data are not always as well-behaved. Finally, the topic of missing measurements in multivariate data sets is also related, since these can be considered as measurements with very large degree of uncertainties.

Maximum likelihood principal component analysis (MLPCA)* is one technique that has been used for several years to address the issue of heteroscedastic measurement errors in multivariate analysis [20,27] and has been applied to different problems [22,28,29,30,31,32,33,34,35,36]. This technique is related to similar approaches such as total least squares [37,38,39,40,41] and positive matrix factorization [42,43], and is a generalization of PCA to non-ideal error structures that range from simple heteroscedascity and within-sample correlation, to more complex structures that can affect multiple orders. However, to date its application in exploratory data analysis has not been examined. This has not been due to a limitation of MLPCA itself, which is readily applied, but rather as a consequence of the data, which impairs the visualization of scores plots. In this work, the limitations of conventional PCA for exploratory data analysis when applied to data exhibiting a high degree of heteroscedascity are described. This is followed by an examination of strategies designed to address these weaknesses using MLPCA. To illustrate the effects of both methods, simulated data with controlled error structures are used. An application of MLPCA to noisy DNA microarray data is provided.

* MLPCA in this thesis refers to the method developed in this group (see reference [27]) unless specified otherwise.

2.2 Theoretical Aspects

It is necessary to make it clear that the type of data sets being discussed in this chapter are those in which there is a high degree of heteroscedasticity, both in terms of magnitude and proportion. In other words, while the methods described are applicable to any data sets where measurement errors are not i.i.d. normal, they will have the most impact when the range of measurement error variances is large and a significant percentage of the measurements have high uncertainties. Also, it is assumed that the principal objective is to identify clusters, as this will make it easier to visually assess the effectiveness of various methods. Although such clustering is typically done on samples in the scores space, it might also be done on variables in the loadings space. Further, in the context of this work, it is assumed that the measurement error variances are known or can be estimated. Finally, while measurement errors are assumed to be independent, the extension of these ideas to correlated measurement errors should be relatively straightforward.

As introduced in Chapter 1, MLPCA is a subspace modeling technique and follows the general model shown in Equation (1.3). In this chapter, by following the convention of MLPCA, the notations **Z** and **W** shown in the general model in Equation (1.3) are replaced by **T** and **V**, respectively and thus this model can be re-written as

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{1}\boldsymbol{\mu}^{\mathrm{T}} + \mathbf{E}. \tag{2.1}$$

In this chapter, the discussion will focuses on MLPCA for data without intercept term, which means that $\boldsymbol{\mu} = \mathbf{0}$ or the hyperplane in which the error-free data are located passes through the origin. Thus, this general model for MLPCA can be written as

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{E}. \tag{2.2}$$

In this model, one might notice that the vectors defining the subspaces **T** and **V** have a rotational ambiguity issue just as in the case of factor analysis. To make the scores and loadings unique, MLPCA actually imposes a restriction on **T** and **V** by applying traditional PCA or singular value decomposition (SVD) to $\mathbf{T}\mathbf{V}^{\mathrm{T}}$ so that the components hierarchically account for the maximum of the estimated data variance.

When MLPCA is used for exploratory data analysis to deal with data that have significant heteroscedastic errors, there are several issues that affect the discovery of

meaningful information from the data, especially the visualization of clusters within the scores plots:

1. The number of principal components selected for MLPCA.
2. Estimation of the optimal projections (scores) and subspace (loadings).
3. Processing the estimated data by methods such as row normalization, column mean centering, and column scaling.
4. Visualization of clusters in the scores plots at the presence of noisy measurements that may obscure the cluster separation.

It should be noted that, for a given data set, not all of these issues will be equally prevalent, nor their solutions equally necessary, but each is a potentially complicating issue. In the rest of this chapter, each of the issues will be discussed and a general procedure for using MLPCA in exploratory data analysis is recommended.

2.2.1 Number of Principal Components Selected for MLPCA

Unlike PCA, MLPCA solutions are not nested; therefore, this raises the question of how many principal components should be extracted. In other words, what is the dimensionality of the subspace? As a subspace modeling technique, MLPCA assumes that the error-free data are located in a subspace of the observed variable space with a rank d (called pseudo-rank, chemical rank, or intrinsic rank). The observed data, which are contaminated by errors, have a rank p (called mathematical rank) higher than the pseudo-rank d. In the sense of maximum likelihood, it is only when the number of components is chosen to be the same as the pseudo-rank, that the estimated subspace and scores are maximum likelihood estimates. In conventional PCA, while certain techniques exist to estimate the pseudo-rank d, they are not entirely reliable and become even less so for MLPCA.

Although it is agreed that the number of principal components should ideally be equal to the pseudo-rank, it has been observed that, as long as the number of principal components does not greatly deviate from the pseudo-rank, the utility of MLPCA in extracting useful information for exploratory data analysis will not be heavily affected. If the number of principal components is chosen to be smaller than the pseudo-rank, some chemical variation will be treated as measurement errors, and the direction of the subspace determined and the maximum likelihood projections can be unpredictable. On

the other hand, if the number of principal components is chosen to be larger than the pseudo-rank, some variation due to measurement errors in the observed data will be treated as underlying components. As the number of principal components increases beyond the pseudo-rank d, more noisy measurements will contaminate the projected data and begin to obscure relationships among the objects, but these changes generally occur gradually. The situation is analogous to multivariate calibration, where there is initially a rapid decrease in prediction error with model dimension, but a slower rise after the optimum dimension. The number of principal components may cover a range with which useful information can be obtained. In other words, the result of MLPCA is not very sensitive to the selection of the number of principal components when it is not smaller than the pseudo-rank. Certainly, the effect of selection of different numbers of principal components depends on the characteristics of the data set and error structure for real data, because real data often do not have a definite pseudo-rank. The choice of the number of principal components should represent a balance between the estimated proportion of "bad" measurements for a given sample and reasonable guess for the pseudo-rank of the data. In practice, different numbers of principal components can be tried for a specific data set.

2.2.2 Estimation of Subspace and Scores

As a subspace modeling technique, MLPCA estimates the error-free data in a subspace by using the information contained in the measurement uncertainties. In cases of measurements with very large error uncertainties, an alternative approach may be to simply exclude "bad" measurements from further analysis. While this should be done if all of the measurements for a given variable or sample are unreliable, there are a number of drawbacks to this approach when the quality of data varies within a row or column of a data matrix. First, most methods require the elimination of an entire row or column of data in order to reject a single measurement, so one risks decimation of what may already be a limited data set based on a relatively small number of unreliable measurements. Moreover, any useful information the censored sample or variable contributes to the interpretation of the results will be lost. Second, a binary classification of "good" and "bad" requires setting a somewhat arbitrary threshold for uncertainty and does not reflect the fact that the information content of measurements follows a continuum. Therefore, it would be better to

employ methods that reflect this characteristic of the data. Finally, multivariate data typically have a high degree of redundancy, and this can be exploited in the analysis. This means that large errors in one measurement will not necessarily propagate to lower dimensions, especially if appropriate strategies are used to minimize the effects of uncertain measurements.

Compared with PCA, the estimates of error-free data by MLPCA may be greatly improved given reasonably accurate characterization of measurement noise. PCA intrinsically assumes homoscedastic errors and the result of applying PCA directly to the data with significant heteroscedastic errors will be suboptimal. MLPCA, however, gives a more accurate estimate of the error-free data than PCA by making optimal use of measurement uncertainty estimates. It does this by optimizing the subspace and using a maximum likelihood projection to obtain the scores. It is worth noting that estimations of the subspace and scores are simultaneous and not two separate steps.

For a given p x 1 column vector $\mathbf{x}_i$ representing a sample measured on p variables (i is the sample index), if a subspace is defined by a p x d ($d < p$) orthonormal basis $\mathbf{V}$, the scores of this sample ($\mathbf{t}_i$) with respect to the subspace can be obtained by the maximum likelihood projection

$$\mathbf{t}_i = (\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{x}_i\,, \tag{2.3}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}_i}$ is the measurement error covariance matrix for sample $\mathbf{x}_i$. Typically, when the measurement errors can be considered to be independent, $\boldsymbol{\Sigma}_{\mathbf{x}_i}$ is a diagonal matrix containing measurement error variances, but correlated measurement errors can also be represented. In theory, the same subspace can be described by an infinite number of orthornormal bases; therefore, PCA or SVD is applied to $\mathbf{TV}^{\mathrm{T}}$ so that $\mathbf{V}$ becomes unique, as mentioned in Equation (2.2). The estimation of $\mathbf{t}_i$ is an oblique projection of the original data into the subspace, which is different from the orthogonal projection method used in PCA. The oblique projection used in MLPCA is a maximum likelihood projection, meaning that the observed data are the most possibly obtained if the underlying error-free data are the projected data, given the known error structure. Note that if homoscedastic errors are assumed, the orthogonal projection used in PCA is the maximum likelihood projection. However, in this context, the term "maximum likelihood projection" implies the oblique projection by MLPCA unless specified otherwise.

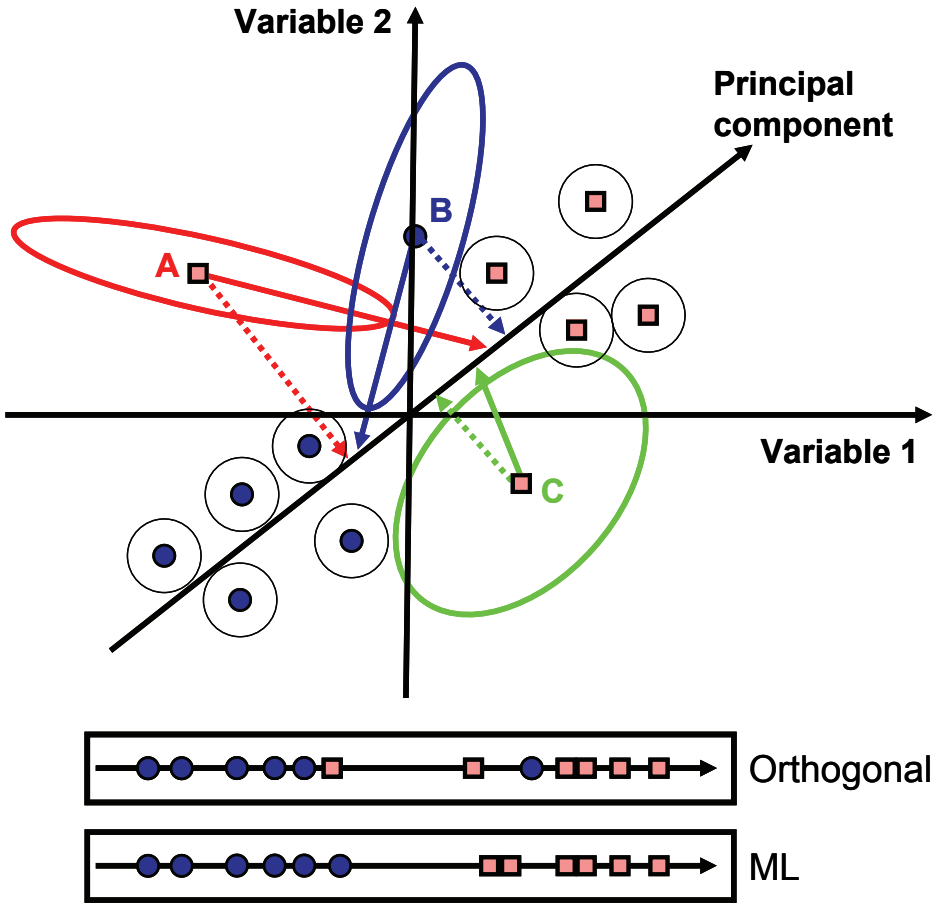

Figure 2.1 Schematic representations of orthogonal projection method in PCA and maximum likelihood projection method in MLPCA.

To illustrate the concept of the maximum likelihood projection in MLPCA, a simple artificial data set is used. This data set consists of twelve samples (objects) evenly divided into two classes with the marker shapes (squares and dots) and colors (blue and pink) indicating classes, and the data have an intrinsic rank of unity (*i.e.* a pseudo-rank of one), as shown in Figure 2.1. To make the illustration simple, it is assumed that data points A, B and C have heteroscedastic errors and all other points have homoscedastic errors. The ellipses of constant probability density for A, B, and C (and circles for other points) schematically illustrate the error covariance matrices. The orientations of the ellipses indicate if errors are correlated. The size of the ellipse or circle shows the magnitude of the measurement error variance in the corresponding direction. For A, B, and C, the dashed and solid lines indicate the orthogonal and maximum likelihood projections onto the subspace of rank one (the straight line labeled as principal

component), respectively. The results of the projection into one-dimensional space by the two methods are shown under the main figure. It can be seen that if the orthogonal projection method is used (dashed line), each class has one object (indicated in Figure 2.1 as A and B) projected into a domain close to the other class, and the class separation is not distinct. On the other hand, if the maximum likelihood projection method is used (solid line), the points A and B are both projected into their own regions, respectively and the samples of the two classes are clearly separated.

As mentioned earlier, the estimation of scores and subspace is not two separate steps. To estimate the scores, a random initial guess of the subspace, denoted by **V,** is typically provided and the scores are estimated by maximum likelihood projection. Once the scores are estimated based on the given subspace defined by **V**, the scores can be used to update the subspace in a maximum likelihood way. This is done by transposing the original data matrix and adjusting the error covariance matrices accordingly. Whereas the first step projected the measurements (as sample vectors) into the variable space, the second step projects the measurements (as variable vectors) into the sample space [20,27]. This procedure, known as alternating least squares (ALS) is repeated until convergence, at which point, **T** and **V** are determined by applying PCA or SVD to the projected data. Like other methods, this iterative method cannot guarantee to hit the global optimum of the MLPCA objective function and thus multiple initial guesses of the subspace are generally used to increase the chance to find the global optimum.

Once the subspace and scores are estimated, it may be necessary to present the estimated data in the original space (observed variable space). The estimates of the error-free data in the original space ($\hat{\mathbf{x}}_i$) can be expressed as

$$\hat{\mathbf{x}}_i = \mathbf{V}(\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{x}_i \,. \tag{2.4}$$

This transformation does not affect the spatial relationship of the objects in scores space, but only uses a different coordinate system to express the projected data and no information is changed.

An important aspect of the maximum likelihood projection is that the measurement error information in the original measurements can be propagated to the projected data. This measurement error information can be associated with the original variables or with the scores, but the error information associated with the scores is more

important in the present work. As heteroscedastic errors have been assumed for the data in the original variable space, the scores for each data point will have a different error covariance matrix. It has been shown that if an original data point has an error covariance matrix denoted by $\mathbf{\Sigma}_{\mathbf{x}_i}$, the error covariance matrix for its scores ($\mathbf{\Sigma}_{\mathbf{t}_i}$) can be obtained by [20,27]

$$\mathbf{\Sigma}_{\mathbf{t}_i} = \left(\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{V}\right)^{-1}, \tag{2.5}$$

where $\mathbf{V}$ is the basis for the estimated subspace. Because the estimation of the subspace and scores is not two separate processes, the error covariance matrix for the scores (obtained by Equation (2.5)) is an approximation for the true error covariance matrix and does not account for errors in the estimation of the model. Note that, even if the original measurement errors were independent ($\mathbf{\Sigma}_{\mathbf{x}_i}$ is diagonal), projection of the original data into the subspace generally produces correlated errors ($\mathbf{\Sigma}_{\mathbf{t}_i}$ is not diagonal). For the estimated data in the original space, $\hat{\mathbf{x}}_i$, the corresponding error covariance matrix can be expressed as

$$\mathbf{\Sigma}_{\hat{\mathbf{x}}_i} = \mathbf{V}\mathbf{\Sigma}_{\mathbf{t}_i}\mathbf{V}^{\mathrm{T}} = \mathbf{V}\left(\mathbf{V}^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{x}_i}^{-1}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}. \tag{2.5}$$

In general, $\mathbf{\Sigma}_{\hat{\mathbf{x}}_i}$ is also not diagonal. In addition, $\mathbf{\Sigma}_{\hat{\mathbf{x}}_i}$ becomes singular because it is essentially a matrix with a lower rank expressed in a higher-dimensional space. These equations are important in subsequent steps that make use of uncertainty estimates.

2.2.3 Processing of the Estimated Data

Once the error-free data are estimated based on the principle of maximum likelihood, a projection method such as PCA or PP is applied to the estimated data (either the scores with respect to the subspace or the estimated data in the original space) to look for useful information, especially clusters in a lower-dimensional space. Prior to this, however, some preprocessing methods such as row normalization, mean centering and/or column scaling, may be applied and can, dramatically affect the results.

In the context of the discussion here, “row normalization” refers to the procedure where each row vector of the data for a particular object or sample is normalized to unit length or unit sum (area) prior to mean centering. The reason that such a pretreatment is

common is that the class membership of an object is often determined by the overall shape of the measurement profile (*e.g.* spectrum, chromatogram); that is, the relative rather than the absolute magnitude of the measurements. If an object is normalized to unit area (refer to Equation (2.9)), the estimated data in the original space should be used since the scores may have positive or negative values. In the original space, it is more typical for data to be positive. If normalization to unit length is performed (refer to Equation (2.7)), use of the scores in the subspace or the estimated data in the original space leads to equivalent results. Thus, it does not matter whether the scores or the estimated data in the original space are used. In the discussion here, the analysis is based on using the estimated data in the original space, but these steps may be applied to the scores as well. Row normalization changes the magnitude of an object's measurements, but does not alter the profile. Depending on the purpose and requirement, row normalization can be performed before other projection methods are applied to the estimated data.

Mean centering of the variables, also referred to as column mean centering, is generally recommended when using projection methods for exploratory data analysis, since removal of the mean vector will tend to increase the information content of the data when represented in low dimensions. The sample mean is an unbiased estimate of the population mean. If a normal distribution and homoscedastic errors are assumed, the sample mean is also the maximum likelihood estimate of the population mean. When measurement errors become increasingly heteroscedastic, however, the reliability of this estimate becomes worse, and a few measurements with large uncertainties could degrade the quality of the estimate of the population mean and adversely affect the estimation of the subspace and scores.

One possible solution is to use the weighted mean with the measurement covariance matrices used to give the weights, which can be expressed as

$$\bar{\mathbf{x}} = \left(\sum_{i=1}^{n} \boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{\Sigma}_{\mathbf{x}_i}^{-1} \mathbf{x}_i \right), \tag{2.6}$$

where $\bar{\mathbf{x}}$ is a p x 1 vector representing the overall mean, $\mathbf{x}_i$ is the vector for object i, and $\boldsymbol{\Sigma}_{\mathbf{x}_i}$ is its corresponding covariance matrix. It is worth noting that this equation shows a general formula and is not specific to the work here. The use of a weighted mean is not recommended, however, since the covariance estimates available are those of the

measurement errors and not of the overall distribution. As a consequence, the weighted calculation will likely over-emphasize a few measurements with small errors, leading to an unreliable estimate of the mean (high variance).

One more possible solution is to estimate the mean, subspace, and scores simultaneously, as defined in Equation (2.1). However, the difficulty is that no algorithm is readily available. In Chapter 3, a new algorithm is proposed to estimate a point to remove the data intercepts and this point is chosen as the mean of the estimated data. In the absence of this here, however, it is recommended to use the mean of estimated data (based on Equation (2.4)) to perform mean centering.

Column scaling is another preprocessing step that is commonly employed for projection methods for exploratory data analysis. This is especially important for PCA, because variance, which is used as the objective function of PCA, is sensitive to the variable scales. Column scaling makes different variables have compatible scales and thus reduces the possibility that the few largest principal components are dominated by the variables with extremely large scales. Since measurement error uncertainties are often roughly proportional to the magnitudes of the data, column scaling essentially makes the measurement errors more homoscedastic in all variables. Column scaling is an optional data preprocessing step, but if row normalization is performed, column scaling is not advisable since the row normalization has changed the magnitudes of the measurements and their uncertainties.

Row normalization, column mean centering and/or column scaling change the data structure of the estimated data and the measurement error information. For an object, if row normalization to unit length is performed, the normalized data ($\mathbf{y}_i$) can be expressed as

$$\mathbf{y}_i = \frac{\hat{\mathbf{x}}_i}{\left\|\hat{\mathbf{x}}_i\right\|}, \tag{2.7}$$

where $\hat{\mathbf{x}}_i$ is the vector for estimated object *i* and the notation $\|\bullet\|$ denotes the operator of Euclidean norm ($\left\|\hat{\mathbf{x}}_i\right\| = \left(\hat{\mathbf{x}}_i^{\mathrm{T}}\hat{\mathbf{x}}\right)^{\frac{1}{2}}$). If the measurement error covariance matrix associated with $\hat{\mathbf{x}}_i$ is represented by $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_i}$, the measurement error covariance matrix for $\mathbf{y}_i$ can be obtained by

$$\boldsymbol{\Sigma}_{\mathbf{y}_i} = \left(\frac{\mathbf{I}}{\|\hat{\mathbf{x}}_i\|} - \frac{\hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^{\mathrm{T}}}{\|\hat{\mathbf{x}}_i\|^3} \right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_i} \left(\frac{\mathbf{I}}{\|\hat{\mathbf{x}}_i\|} - \frac{\hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^{\mathrm{T}}}{\|\hat{\mathbf{x}}_i\|^3} \right), \tag{2.8}$$

where $\mathbf{I}$ is the identity matrix of the same size as $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_i}$. If the object is normalized to unit area, the normalized data ($\mathbf{y}_i$) can be obtained by

$$\mathbf{y}_i = \frac{\hat{\mathbf{x}}_i}{S_i}, \tag{2.9}$$

where S_i is the sum of the elements in vector $\hat{\mathbf{x}}_i$. The error covariance matrix for $\mathbf{y}_i$ can be calculated by

$$\boldsymbol{\Sigma}_{\mathbf{y}_i} = \left(\frac{\mathbf{I}}{S_i} - \frac{\mathbf{1}\hat{\mathbf{x}}_i^{\mathrm{T}}}{S_i^2} \right)^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_i} \left(\frac{\mathbf{I}}{S_i} - \frac{\mathbf{1}\hat{\mathbf{x}}_i^{\mathrm{T}}}{S_i^2} \right), \tag{2.10}$$

where $\mathbf{1}$ is a column vector with all its elements being 1's. Equations (2.8) and (2.10) can be derived based on the rule of error propagation [44] by using some calculus. Note that after normalization to unit length or unit area, the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{y}_i}$ has a rank of one less than $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_i}$.

When column mean centering is performed, the error covariance matrix of each object can be regarded to be unchanged. When column scaling is performed, the measurement error covariance matrix for each variable can be obtained by following the same principle, but this case will not be dealt with in this work.

The matrix obtained from the previous steps will be designated as $\mathbf{Y}$. When a projection method such as PCA is applied to $\mathbf{Y}$, the scores of $\mathbf{Y}$ can be obtained by

$$\mathbf{W} = \mathbf{YQ}, \tag{2.11}$$

where $\mathbf{W}$ denotes the scores and $\mathbf{Q}$ represents the basis of the space. The first few columns in $\mathbf{W}$ can be examined in two- or three-dimensional plots to look for clusters. For object *i*, its scores can be expressed as

$$\mathbf{w}_i = \mathbf{Q}^{\mathrm{T}} \mathbf{y}_i . \tag{2.12}$$

Note that this is an orthogonal projection of $\mathbf{y}_i$ into the space of $\mathbf{Q}$ in contrast with the oblique maximum likelihood projection in Equation (2.3). The measurement error covariance matrix for the scores of object *i* can be calculated as

$$\boldsymbol{\Sigma}_{\mathbf{w}_i} = \mathbf{Q}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathbf{y}_i} \mathbf{Q} . \tag{2.13}$$

This equation is important because the error information contained in $\boldsymbol{\Sigma}_{\mathbf{w}_i}$ will be used in scores plots to improve the visualization of clusters.

Although PCA is typically used to process the estimated error-free data with different preprocessing steps, it is not necessarily the best choice. When PCA is used to look for clusters for exploratory data analysis, it is implicitly assumed that the between-group variance is larger than the within-group variance and the first few principal components are dominated by the between-group variance. Thus, the clusters are more likely to be revealed in the first few principal components. As PCA hierarchically maximizes the variance of the data, it is sensitive to the scales of the data, and normalization is often important. Other methods such as PP may be applied to the estimated data as well. PP optimizes different objective functions (not variance) and is designed to look for salient features of the data, generally clusters or outliers. PP has not been widely used as PCA, largely because the algorithms of PP are more complicated. If PP or other projection methods are applied to the estimated error-free data, the error covariance matrices can be obtained in the same way as Equation (2.13). However, in the work presented in this chapter, PCA is applied to the estimated error-free data because it is the most widely used method and works well with the simulated and experimental data.

2.2.4 Visualization of Clusters in Scores Plots

At this point, the information contained in the data must be represented in a space of lower-dimensionality for visual analysis and interpretation. This is generally performed by plotting the first few columns in **W** to see if the objects can naturally form clusters. Given that there is no information about object classes included in exploratory analysis, clusters revealed in the scores plots are more likely to be a valid reflection of inherent data structures, which may guide analysts for further exploration by other refined methods. The clusters may also be used to confirm the relationship of the known classes. Regardless of the subspace projection method used in the preceding step, there remains the possibility that a significant number of relatively noisy measurements in the scores plot can obscure the underlying structure of the data. This is because measurements with large errors will be projected in a random fashion throughout the subspace, thereby making the

identification or confirmation of clusters difficult. This is illustrated in Figure 2.2 which shows the projection of hypothetical measurements into a two-dimensional space. The measurements are characterized by two classes that separate along the first dimension, with the separation characterized by $\sigma_{between}$, as indicated in Figure 2.2 (a). The within-class variation follows a Gaussian distribution characterized by σ_{within}, which is substantially smaller than the separation between classes. In addition, there are three groups of objects in

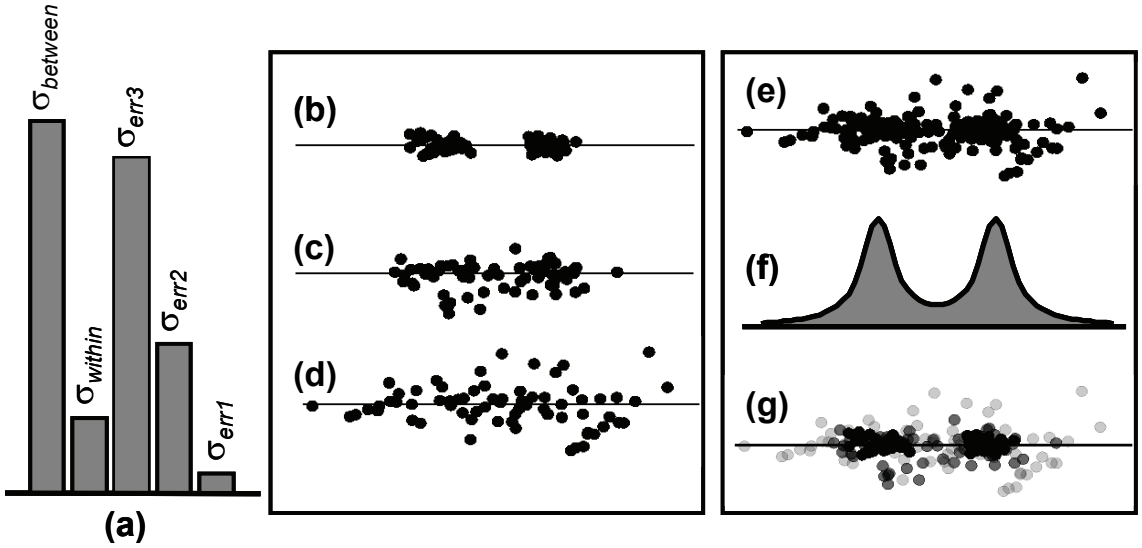

Figure 2.2 Simulated samples showing the effects of heteroscedastic errors and the resulting mixture model. The variance contributions shown in (a) represent the within and between class variance, as well as three different levels of measurement errors applied to three equal-sized subgroups, each with two classes of objects. Measurement realizations are shown in (b)-(d), where class distinction is increasingly obscured as the measurement error increases. The combined measurements are shown in (e), along with the combined distribution in (f). Class distinction is difficult until object transparency is related to uncertainty, as shown in (g).

each class characterized by different levels of measurement noise: low (σ_{err1}), moderate (σ_{err2}), and high (σ_{err3}). (Errors for each object are the same in both directions, but the axes scales are different). Figures 2.2 (b-d) show the projections of each group of objects. When errors are small (Figure. 2.2 (b)), the class separation is evident, but is obscured as the measurement uncertainty approaches the magnitude of the class separation. When all of the objects are pooled, the result is shown in Figure 2.2 (e), where it is clear that the presence of measurements with large errors obscures the underlying class structure because all measurements are given equal visual weight. A mixture model applies here, where each object can follow a different distribution that results from a convolution of all of the

underlying distributions. The overall distribution, calculated from the simulation parameters, is shown in Figure 2.2 (f). While this suggests the presence of two classes, such distributions are difficult to determine based on the data alone, especially if the number of samples is limited, and most rely on the visual interpretation.

A guiding principle through the preceding steps of this procedure has been the preservation of the measurement error information through each stage in the form of the error covariance matrix, $\mathbf{\Sigma}$. This information can be used at the current step to improve the visualization of clusters. The most obvious approach would be to simply exclude those objects with large measurement errors from display. As already stated, however, this involves setting a threshold for good/bad measurements and potential loss of information. A better approach would be to apply a procedure that exploits the continuous nature of measurement quality. One possibility would be to display the error information along with the objects in the form of error bars or error ellipses. Although this would make the uncertainty information available, it would quickly add clutter to the projections and not improve the visualization of relationships. As an alternative, summary metrics based on $\mathbf{\Sigma}$ (*e.g.* the volume of the ellipsoid or its maximum dimension) could be used to obtain a quantitative measure of the quality of each object. This could in turn be used to modify the appearance of those objects in the scores plot. This is the approach recommended in this work.

The method described here is referred to as the partial transparency projection (PTP) method and exploits the enhanced capabilities of modern graphical display units. These units allow attributes such as the color and size of displayed objects to be readily modified. One attribute that has not been fully exploited is the object transparency. In a typical mapping, an object with a transparency of zero will be opaque, while one with a transparency of unity will be completely transparent, effectively making it invisible. By mapping measurement quality to the gradation of transparency values between zero and unity, the appearance of objects with low quality measurements can be modified in an interactive fashion so that the underlying structure of the data is more clearly revealed. This effect is illustrated in Figure 2.2 (g), where objects with greater uncertainty are assigned a higher level of transparency. This has the advantage of retaining all of the objects in the

data set, while still allowing data analysts to determine the influence of low quality measurements on the visual interpretation of the projection.

Mathematically, the PTP is performed by first choosing some quality measure, Q, for each object displayed in the space. This could be, for example, the volume of the error covariance ellipsoid (described by the covariance matrix $\mathbf{\Sigma}_{\mathbf{w}_i}$), or the trace, maximum eigenvalue, or maximum diagonal element of the covariance matrix. A transformation function, $f(Q)$, such as a logarithm, square, square root, or simple ordering, may then be applied to Q to modify the distributional characteristics of the quality measure if necessary. A mapping function, g, is then used to map the transformed quality measure into a corresponding transparency value, τ. This function could typically be represented by a linear or sigmoidal-type response, or, in the limiting case, a simple step function that distinguishes between "good" and "bad" measurements. The overall relationship is given by:

$$\tau_i = g\left[f(Q_i)\right]. \tag{2.14}$$

This concept is illustrated with a simple example in Figure 2.3, which consists of two clusters of data in a two-dimensional space. The measurements in the first dimension (x_1)

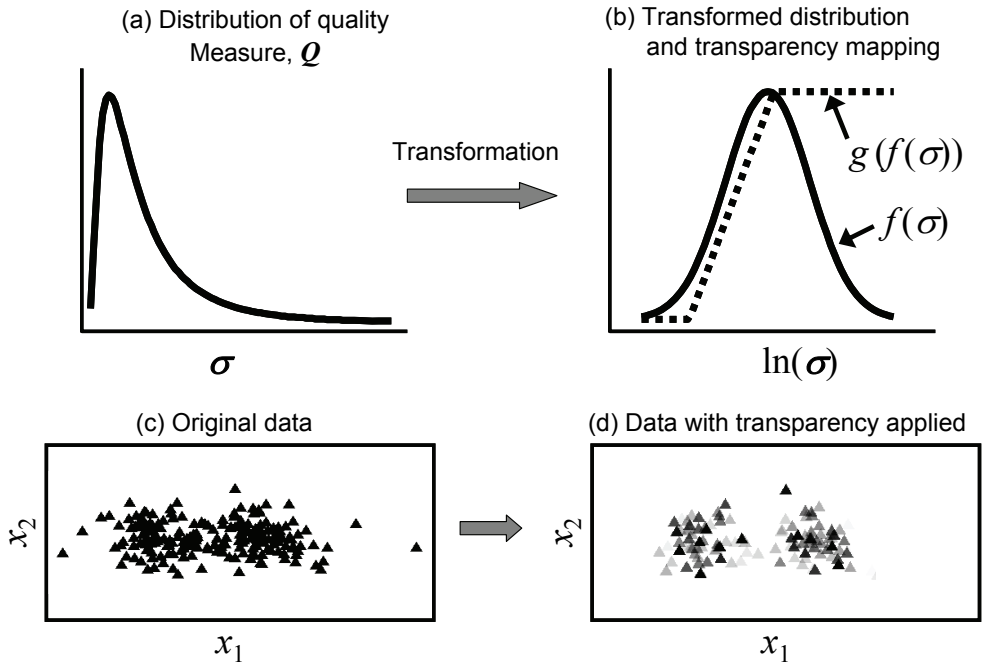

Figure 2.3 Principles of the partial transparency projection (PTP) showing the transformation and mapping steps and the effects on data visualization.

are corrupted with heteroscedastic noise such that the distribution of measurement standard deviations follows a log-normal distribution as shown in Figure 2.3 (a). The presence of measurements with large errors obscures the presence of the two clusters, as shown in Figure 2.3 (c). In this example, the quality measure, Q, is simply taken as the standard deviation of each measurement in the first dimension, σ. A logarithmic transformation function, $f(Q)$, is applied along with a simple truncated linear transparency mapping, $g[f(Q)]$, as shown in Figure 2.3 (b). The result, shown in Figure 2.3 (d), more clearly reveals the underlying cluster structure of the data by diluting the appearance of low quality measurements.

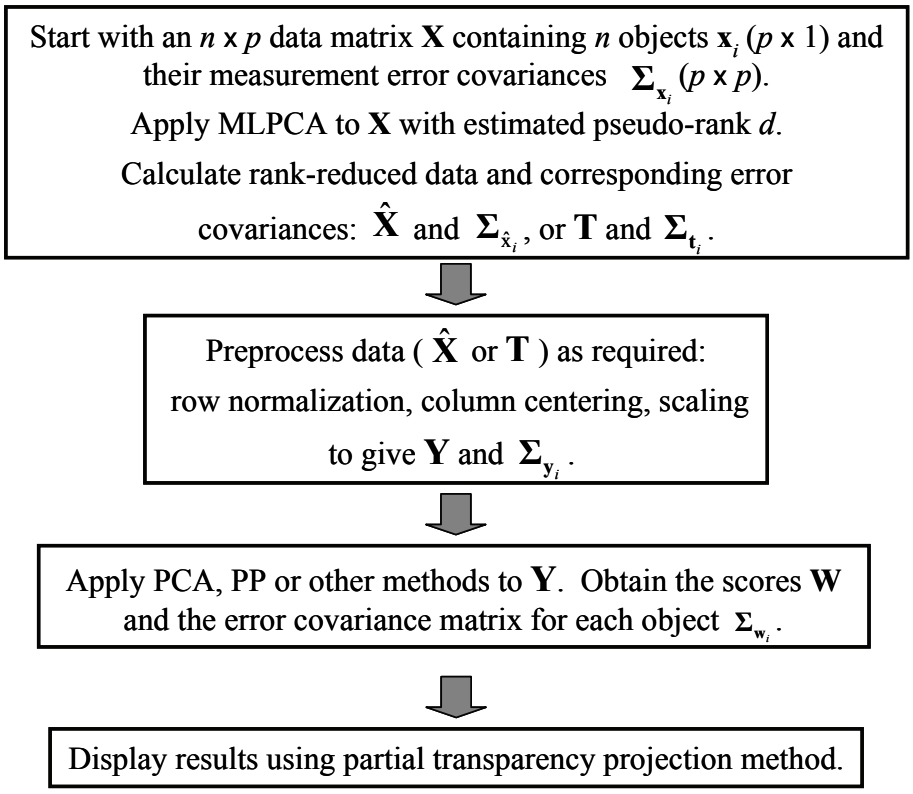

Figure 2.4 Summary of recommended procedure for visualization of multivariate data with a high degree of known heteroscedasticity in the measurement errors.

It should be noted that the application of the PTP does not depend on following the other steps recommended in this procedure and it can be used independently with any projection method. All that is required is some quantitative measure of quality associated with the objects in the projected space. Moreover, it is best applied in an interactive fashion where the quality measure, transformation and mapping can be modified to obtain the best

visual representation of the data. Software has been developed in this work to allow such manipulations to be readily carried out.

The overall procedure recommended for exploratory analysis of data with a high degree of measurement error heteroscedasticity is shown in Figure 2.4.

2.3 Experimental

2.3.1 Simulation Studies

To validate the methods described above, a variety of studies were carried out using simulated data under various conditions of heteroscedastic noise. All simulations were carried out in MatLab® v.7.4.0 (MathWorks, Natick, MA) and are described in Section 2.4.

2.3.2 DNA Microarray Data*

To illustrate the methods described for experimental data, a DNA microarray transcriptomics data set from a time course study in yeast was chosen because it exhibited a highly heteroscedastic error structure that had been previously characterized [22,23,45,46]. The experimental details of this data set have been reported in the literature [47] and will be only briefly summarized here. Spotted two-color microarrays were used to study gene expression levels of yeast cells (*S. cerevisiae*) exiting stationary phase as a function of time. The DNA microarray experiments were conducted as follows. Wild-type MAT*a* S288C yeast cells were cultured in yeast peptone dextrose adenine (YPDA) for 7 days. Due to the lack of nutrition after 7 days, the yeast cells went into the stationary phase. The yeast cells were then transferred into fresh YPDA medium. As the cells came into an environment with nutrients, they exited out of the quiescent state and started to grow. The cells were then harvested at the starting time point and subsequent time points. In the work presented here, 18 microarray experiments were used with the time points as: 0, 0, 0, 1, 5, 10, 10, 15, 20, 20, 25, 30, 35, 35, 40, 45, 50, and 55 min (note triplicates at time zero and duplicates at 10, 20, and 35 min). The reference mRNA was provided by MAT*a* S288C cells grown in YPDA medium and harvested when the optical density at 600 nm was equal to unity. The RNA from the test and reference cells was extracted and converted to cDNA with differential labelling for the test (Cy3) and reference cells (Cy5). The cDNA was

* Preprocessing of the raw DNA microarray data was performed by Dr. Robert M. Flight.

hybridized with the DNA microarrays and scanned at two different wavelength channels. Ratios (test/reference) were calculated for about 6300 genes using the regression method, and uncertainty estimates were obtained using methods previously described [23]. Spots (genes) with a comparatively large proportion of measurements with high uncertainty relative to the mean signal intensity were removed prior to more comprehensive analysis. These were typically genes where the reference channel was close to background, where the test channel did not vary significantly from background across the time course, or where the quality of spot images was generally poor. This resulted in the retention of 3695 genes across 18 time points, as well as a corresponding 3695 x 18 matrix of measurement uncertainties. Measurements (ratios) that were classified as missing or recorded as negative were assigned a value of zero and the associated standard deviation was set to a large value (100 in this case) to reduce their influence on the analysis.

2.4 Results and Discussion

2.4.1 Simulation Studies

To illustrate the principles and advantages of applying MLPCA to exploratory data analysis, a simple data set was generated and subsequently modified in terms of its error structure. The original data set consisted of 300 objects equally divided into three classes of 100 objects each. The centers of these classes were initially located at the vertices of an equilateral triangle (centered at the origin, sides of unit length) in a two-dimensional space. Individual objects were randomly clustered around each of these centers according to a symmetric bivariate normal distribution with $\sigma = 0.15$. A scatter plot for this initial data set is shown in Figure 2.5 (a). To create a multivariate data set, this original data set was rotated into a 20-dimensional space using a randomly generated rotation matrix. The resulting 300 x 20 data matrix represented the error-free measurements. The two-dimensional scores plot generated by applying PCA to these data is shown in Figure 2.5 (b) and, as expected, shows a separation of the clusters equivalent to that in the original two-dimensional space since no measurement errors were present in the data. Homoscedastic measurement errors were then added to the data by adding normally distributed random numbers ($N(0, \sigma^2_{noise})$). Figures 2.5 (c) and 2.5 (d) show the scores plots obtained for the data set with $\sigma_{noise} = 0.2$ and 0.8, respectively. As anticipated,

increased noise levels degrade the separation of the clusters because of an increase in total within-class variance and suboptimal estimation of the PCA subspace.

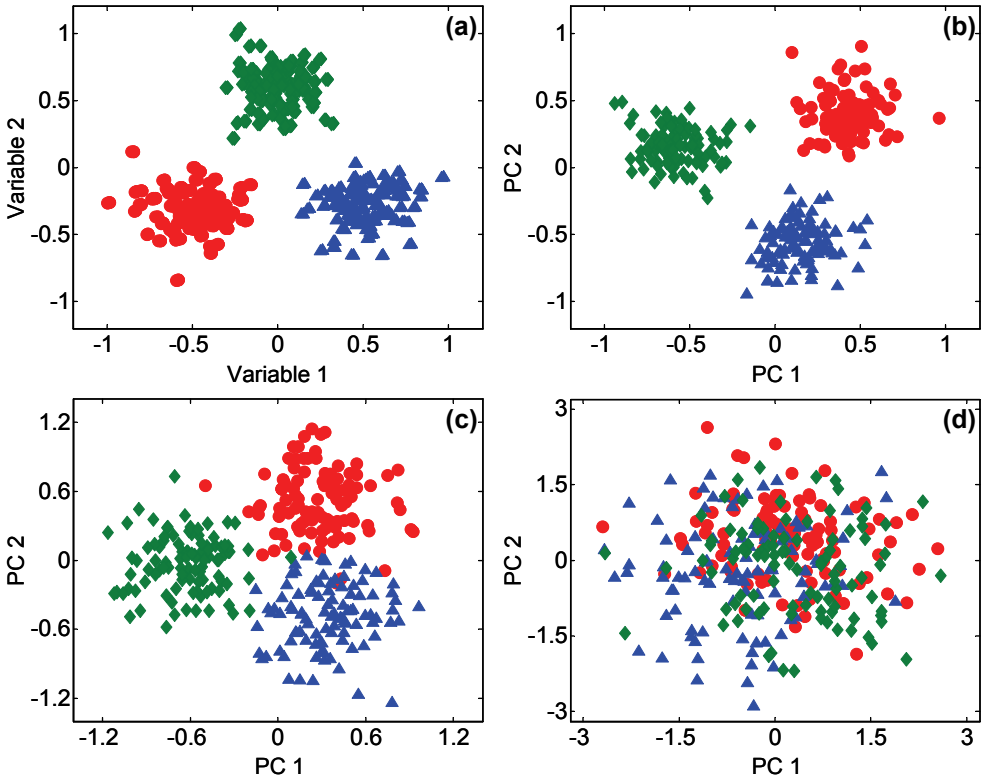

Figure 2.5 Effect of measurement errors on the visualization of PCA scores for simulated data: (a) scatter plot of the two-dimensional error-free data, (b) PCA scores plot of the two-dimensional error-free data following rotation into 20-dimensional space, (c) PCA scores plot of noisy 20-dimensional data ($\sigma = 0.2$), (d) PCA scores plot of noisy 20-dimensional data ($\sigma = 0.8$).

2.4.1.1 Measurement of Class Separation

To provide a more quantitative measure of class separation following projection of objects from a *p*-dimensional space into a *d*-dimensional subspace, a commonly used quantity is given by Equation (2.15)

$$F = \mathrm{tr}\left\{\mathbf{S}_{within}^{-1}\,\mathbf{S}_{between}\right\}. \tag{2.15}$$

This is a generalization of the Fisher's discriminant value [48], which measures the ratio of between-class variance to within-class variance. In this equation, $\mathbf{S}_{within}$ and $\mathbf{S}_{between}$ are the within-class and between-class covariance matrices for the projected objects, respectively (note that the within-class covariance matrix is assumed to be the same for all classes). In general, this applies to any projection, but in the current context, it is used

for the orthogonal projections described earlier. Of course, this equation is not normally used in exploratory data analysis where class information is not available, but is employed here to assess the quality of class separation. Assuming that there are N_c classes and N_i objects in class i, and n objects in total, the two matrices can be calculated by.

$$\mathbf{S}_{within} = \frac{1}{n}\sum_{i=1}^{N_c}\sum_{j=1}^{N_i}(\mathbf{w}_{ij} - \bar{\mathbf{w}}_i)(\mathbf{w}_{ij} - \bar{\mathbf{w}}_i)^{\mathrm{T}}, \text{ and} \tag{2.16}$$

$$\mathbf{S}_{between} = \frac{1}{n}\sum_{i=1}^{N_c} N_i(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})(\bar{\mathbf{w}}_i - \bar{\mathbf{w}})^{\mathrm{T}}, \tag{2.17}$$

respectively. In these equations, all of the **w**'s are the scores or mean scores in the final projection space. The vector $\mathbf{w}_{ij}$ represents the scores for object j in class i, $\bar{\mathbf{w}}_i$ is the vector of mean scores for objects in class i, and $\bar{\mathbf{w}}$ is the overall mean score vector for all classes. The application of Equation (2.16) makes some simplifying assumptions such as the homogeneity of within-class variance, but for the simple example here, it is applicable.

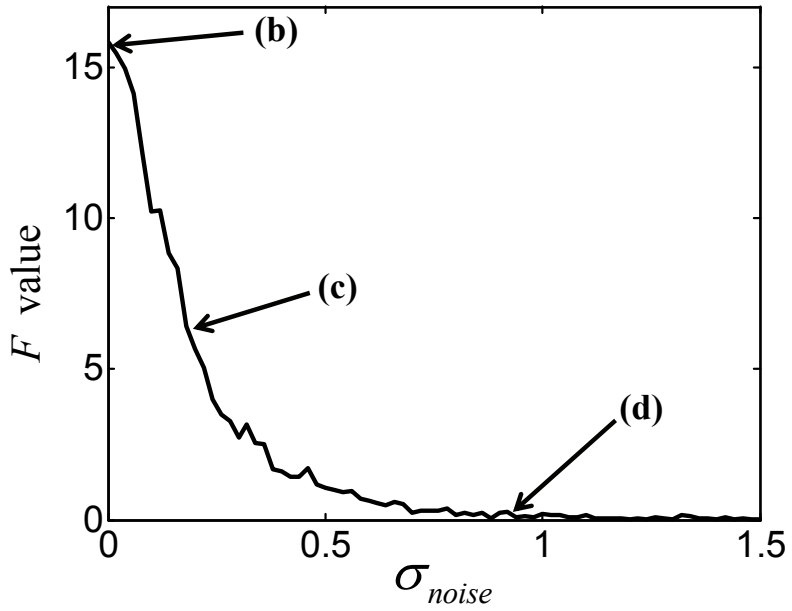

Figure 2.6 Plot of generalized Fisher's value (F) for simulated data (following PCA) as a function of the measurement error standard deviation. The points indicated by (b), (c), and (d) correspond to Figures 2.5 (b), (c), and (d), respectively.

To illustrate how Equation (2.15) relates to the visual separation of objects in a two-dimensional space, it was applied to the example described above with increasing values of σ_{noise}. Since the errors are homoscedastic and the clusters define a

two-dimensional space, PCA represents the optimum projection method and was used to generate the two-dimensional projections after simple mean centering. The value of F is plotted as a function of the noise level in Figure 2.6. For comparison, points (b)-(d) on this figure show the F values associated with the visual projections shown in Figures 2.5 (b)-(d). It is clear that the value of F decreases as the separation of classes becomes worse, indicating that this is a useful quantitative measure of class separation.

2.4.1.2 Heteroscedastic Noise Simulation

To compare the results of PCA and MLPCA for exploratory data analysis using simulated data with heteroscedastic noise, it is of course necessary to simulate the noise. Because heteroscedastic noise may exhibit a wide range of characteristics for different systems, there is no single universal model that can be applied. For these simulations, however, the principal objective was to demonstrate that MLPCA can be advantageous in certain situations without implying that it will necessarily yield improved results in all situations that might be encountered. To this end, a noise model that exhibited a log-normal distribution of measurement error standard deviations was chosen for simplicity. Log-normal distributions can produce the wide range of measurement uncertainties likely to accentuate the differences between PCA and MLPCA and are commonly observed in various systems. For example, this distribution is a good approximation to the measurement errors observed for the ratios in DNA microarray experiments, as discussed later. Based on this distribution, individual measurements in the simulated data could be randomly assigned a standard deviation. Individual measurement errors were then applied by generating a random number from a normal distribution with the corresponding standard deviation. For the purpose of the simulation, the population standard deviation for each measurement was exactly known. While this would not be the case for real data, this limitation was not explored in this study. For a more realistic simulation, the log-normal distribution was truncated at the low end such that no measurement was allowed to have a standard deviation less than 0.001 (values below this were set to the limit). This was considered to be more accurate, since few practical measurements can be infinitely precise.

The strategy for the simulations was to compare how MLPCA and PCA perform under different noise scenarios. The log-normal distribution has two adjustable parameters, its mean and standard deviation, which will be designated here as μ_{log} and σ_{log}, not to be

confused with the parameters for the measurements. For the purposes of the simulations, however, it is more useful to replace μ_{log} with another parameter, α, which is defined as follows. In the case of homoscedastic noise, a certain noise threshold ($\sigma_{noise} = \sigma_{thresh}$) can be chosen, where the separation of the classes begins to become blurred. In this case, this was arbitrarily chosen as $\sigma_{noise} = 0.2$, which corresponds to the case illustrated in Figure 2.5 (c). If this is considered as an arbitrary dividing line between "good" and "bad" measurements, then α represents the fraction of measurements in the data that can be considered "bad". Although there is a constrained relationship between α, μ_{log} and σ_{log}, α was considered to be a more intuitive metric than using μ_{log} directly. These relationships are illustrated in Figure 2.7.

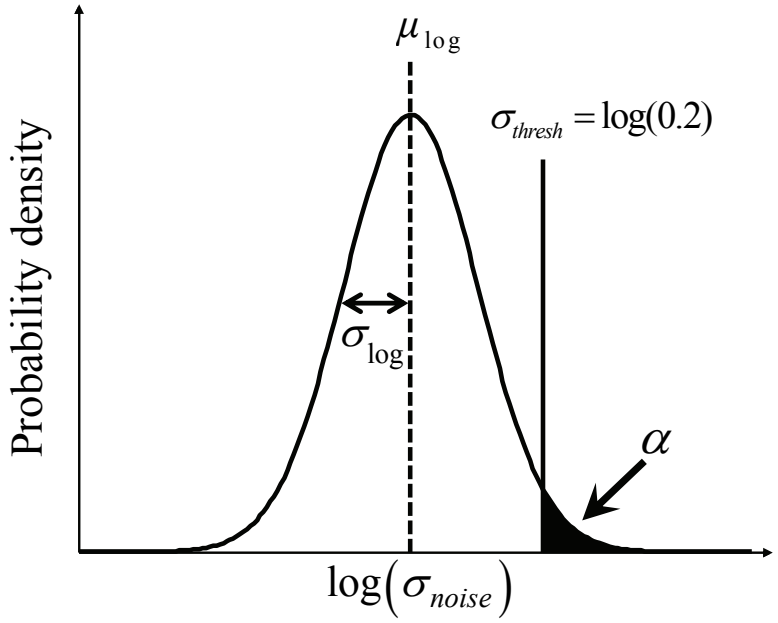

Figure 2.7 Illustration of the relationship among the parameters of the log-normal distribution used to generate the measurement error standard deviations for heteroscedastic errors in simulations. The values of σ_{log} (the degree of heteroscedasticity) and α (fraction of "bad" measurements) were independently varied, with σ_{thresh} fixed and μ_{log} varying as a function of the other two variables.

Figure 2.8 shows representative scores plots obtained under various conditions of noise heteroscedasticity by applying PCA directly and PCA following preprocessing by MLPCA, designated as MLPCA/PCA. To make the application of the latter approach more consistent with real implementations, the rank selected for MLPCA was five, in excess of the actual pseudo-rank of two. Four cases were selected, representing two degrees of

heteroscedasticity (*i.e.* ranges of measurement uncertainty, as expressed by σ_{log}) and two levels of percentage "bad" measurements (as expressed by α).

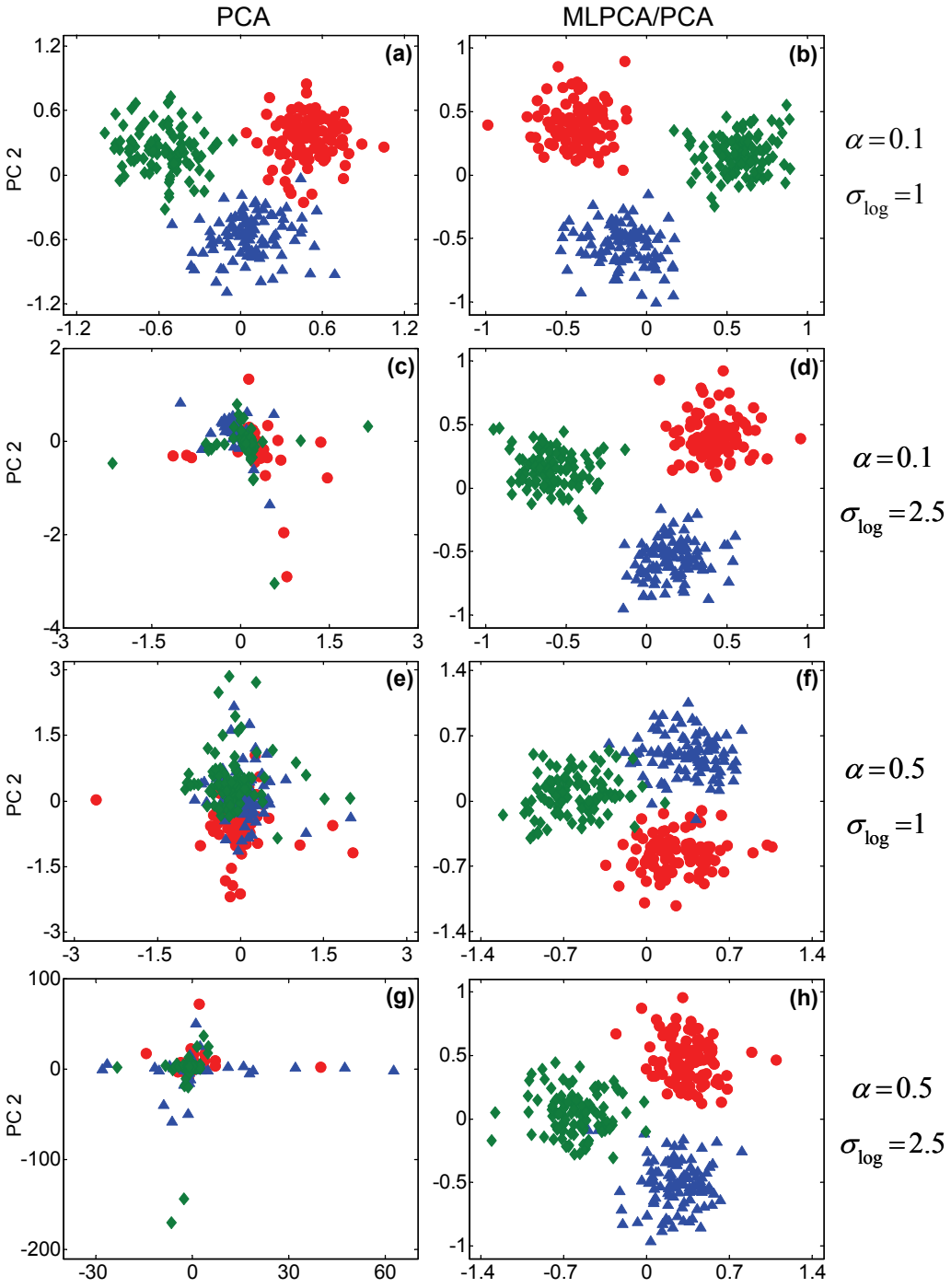

Figure 2.8 PCA (left column) and MLPCA/PCA (right column) scores plots for data with different patterns of heteroscedastic noise, as indicated by α and σ_{log} to the right of each row. See text for further discussion. Note that some of the points in (c), (e), and (g) are outside of the displayed region.

2.4.1.3 Subspace Estimation

As noted earlier, the first difficulty encountered in using PCA for exploratory data analysis is that subspace estimated will be suboptimal due to the influence of noisy measurements. These problems should be mitigated with MLPCA since it will weight the more precise measurements more highly in estimating the subspace. To demonstrate this, it is necessary to compare the subspaces estimated by the two methods. One way to compare two subspaces is to measure the angle between them. This angle can be calculated as follows. First imagine that the first subspace is described by an orthonormal basis $\mathbf{U}_A$ and the second subspace is denoted by an orthonormal basis $\mathbf{U}_B$. If the two bases are of the same size, the angle (θ) between the subspaces can be calculated by [49,50,51]

$$\theta = \cos^{-1}(\left|\det(\mathbf{U}_A^{\mathrm{T}}\mathbf{U}_B)\right|), \tag{2.18}$$

where the notation det is the determinant operator and $|\bullet|$ denotes the absolute value. To compare the ability of PCA and MLPCA/PCA to estimate the true subspace under different conditions of heteroscedasticity, the angle between the estimated subspace and the true subspace was calculated as a function of α and σ_{log}. The true subspace in this context is taken to be the two-dimensional subspace represented by the error-free data. The results are shown in Figure 2.9. Note that these contour plots are the result of 300 individual realizations of the random noise structure at each grid point used (noise-free data remained constant). Since it is most desirable to have the smallest angles, it is immediately clear from the two figures that MLPCA is able to extract a more accurate subspace over a much wider range of conditions. Superimposed on the plot are the letters A-D, corresponding to the conditions shown in Figure 2.8. PCA is most effective at estimating the correct subspace when the degree of heteroscedasticity, σ_{log}, is relatively low, or in other words, when the measurement errors approach homoscedastic conditions. This is not surprising, since these are the conditions under which PCA is designed to perform optimally. As the degree of heteroscedasticity increases, measurements with large errors begin to dominate the overall variance structure and influence the direction of the subspace. The performance of PCA is influenced to a lesser extent by α, the fraction of bad measurements. This is anticipated, since the number of measurements with large errors is less critical than the magnitude of those errors with PCA. With MLPCA, the

estimation of the subspace is much more stable, since it weights the measurements by their uncertainties and is able to obtain a good estimate of the subspace based on high quality measurements.

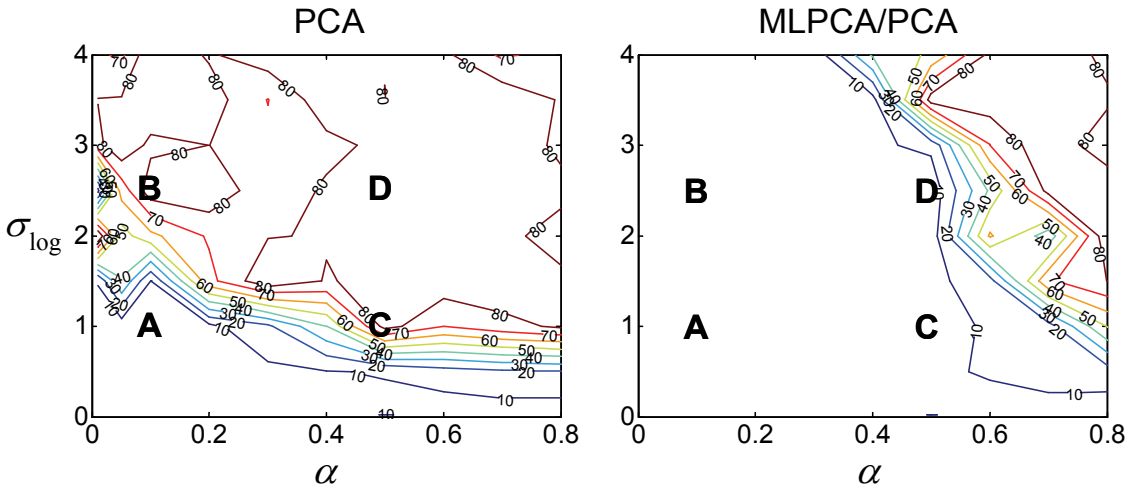

Figure 2.9 Contour plots of the angle (in degrees) between the true subspace and subspace estimated by PCA (left) and PCA following MLPCA preprocessing (right) as a function of heteroscedastic error parameters. The superimposed letters correspond to the four cases in Figure 2.8: A = (a), (b); B = (c), (d); C = (e), (f); D = (g), (h).

Two comments can be made concerning these results. First, known error variances were used in these simulations to represent a best case scenario. Obviously, performance will diminish with the quality of the measurement uncertainty estimates for experimental data. Second, the heteroscedastic error structure in this case was distributed across measurements rather than samples; that is, each object could have both "good" and "bad" measurements in its measurement vector. This scenario may be a good approximation in some circumstances, such as DNA microarrays, but it is perhaps more likely that there are "good" and "bad" objects; that is, cases where the heteroscedasticity is distributed across the samples. MLPCA actually performs better over a wider range under these conditions, since a minimum number of good objects is guaranteed, but the projections of the "bad" objects is worse.

2.4.1.4 Class Separation

Exploration of the estimated subspace is one way to evaluate the utility of MLPCA, examination of the scores obtained by maximum likelihood projection is another aspect compared with the orthogonal projection used by PCA. The cluster

separation in the scores plots is characterized by the discriminant parameter, F, given in Equation 2.15. Figure 2.10 shows the value of the discriminant parameter as a function of α and σ_{log} as before. Also as before, the locations of the examples shown in Figure 2.8 are overlaid on the plot area. In contrast to the angle in Figure 2.9, the value of F should be large to reflect a good class separation. As with the subspace estimation, it is seen that MLPCA offers a better class separation over a much wider range than PCA. To a large extent, this is a consequence of the subspace estimation, but is also related to the quality of the projection. Interestingly, for α values less than about 0.5, the separation for MLPCA improves as the degree of heteroscedasticity (σ_{log}) increases. Although this might seem counter-intuitive, it is an effect of having more measurements with better precision since, for a fixed value of α in this region, the mean of the log-normal distribution is pushed to lower values. This effect is also observed to a small extent in the PCA plot, where a ridge is observed for low α values.

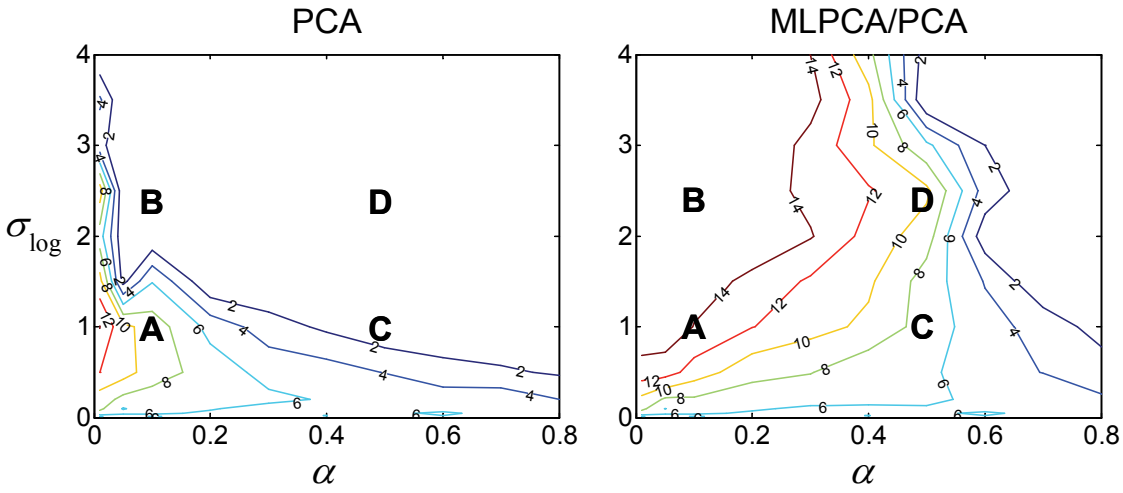

Figure 2.10 Contour plots of the generalized Fisher's discriminant value (Equation 2.15) as a measure of the class separation resulting from PCA (left) and PCA following MLPCA preprocessing (right) as a function of heteroscedastic error parameters. Superimposed letters are as indicated in Figure 2.9.

2.4.1.5 Effect of the Number of Principal Components

As discussed in Section 2.2.1, the number of principal components used in MLPCA is a factor that affects the result, but the effect of the number of principal components should change gradually and useful information can be still extracted if the number of principal components is not very different from the pseudo-rank of a data set.

To verify this hypothesis, the data set corresponding to Figure 2.8 (h) was processed by MLPCA with the number of principal components chosen as 4, 6, 8, and 10 (larger than the pseudo-rank two), respectively. The scores plots are shown in Figure 2.11. When the number of principal components is chosen as 4, 5 (Figure 2.8 (h)), or 6, clear separations of the three clusters are observed. This shows that the number of principal components does not necessarily have to be the same as the pseudo-rank. When the number of principal components is increased to 8 or 10, the separation of the clusters becomes blurred because the number of principal components becomes too far from the pseudo-rank two and noise contaminates the projection. This validates the anticipated effect of the number of principal components. For real experimental data that often do not

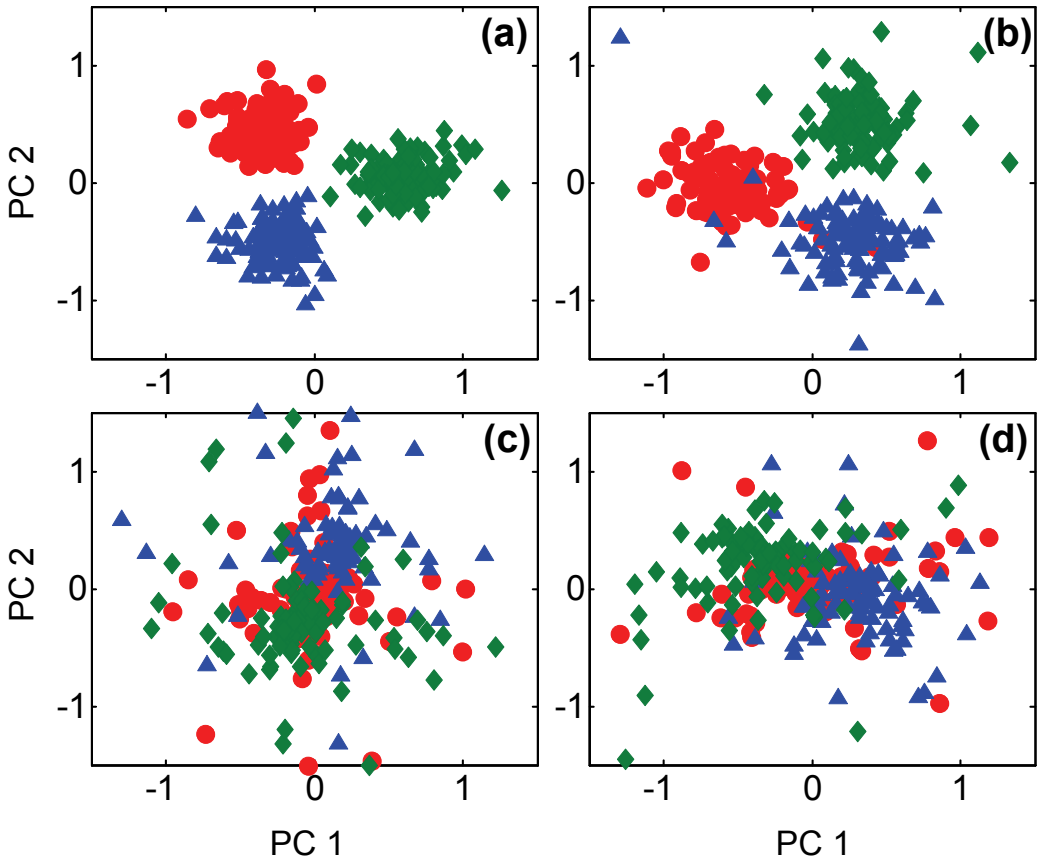

Figure 2.11 MLPCA/PCA scores plots with different numbers of principal components. (a) 4 principal components, (b) 6 principal components, (c) 8 principal components, and (d) 10 principal components.

have a well-defined pseudo-rank, trying different numbers of principal components can be helpful. Theoretically, if the underlying error-free data contain clearly separated

clusters, a good separation of clusters in the obtained scores plot could be an indication of a suitable choice for the number of principal components. However, a poor separation or no separation does not necessarily mean a bad choice for the number of principal components because the underlying error-free data may not contain well-separated clusters at all.

2.4.1.6 Partial Transparency Projection

In the simulations described so far, the heteroscedastic error structure has been applied equitably across all measurements, which means that most objects will have at least a few good measurements to obtain a reliable projection into the scores space except when α is large. Alternatively, if the heteroscedasticity is applied to objects such that each object has uniformly good or bad precision in the measurements, then nothing can be done to improve the projection of objects with poor quality measurements. As the proportion of noisy objects increases, their projection will obscure the true class relationships among objects since they will be projected more or less randomly in the subspace. Although such objects could be removed, this raises the issue of what exclusion threshold should be used. In fact, the question is one of the quantity of information contained in each object, and whether this enhances or degrades the visualization of the underlying relationships. Given the continuum of information content and its complex relationship with the visualization space, it would seem more logical to provide an interactive environment in which the value of objects for establishing class relationships could be visually assessed. This is the reason for the development of the PTP. At one extreme, this projection method can be used to simply exclude the measurements whose uncertainty exceeds some predefined threshold, while at the other it can provide a mapping of an object's relative information content to the transparency of its symbolic representation on the display.

There are many factors that can influence the visualization of dimensionally compressed data with the PTP, including the size and color of symbols used to represent the objects, the parameter (Q) used to represent the measurement uncertainty from the error covariance matrix (largest eigenvector, volume, etc.), the transformation (f) applied to this parameter (logarithm, square root, etc.), and the mapping function (g) for the symbol's transparency (linear, sigmoidal, etc.). For this reason, these parameters are best adjusted in an interactive graphical user interface (GUI). An example of a simple GUI generated for

the purposes of this work is shown in Figure 2.12. In addition to allowing adjustment of the above-mentioned parameters, this interface also has provisions to display object profiles and other features, but its details will not be described here.

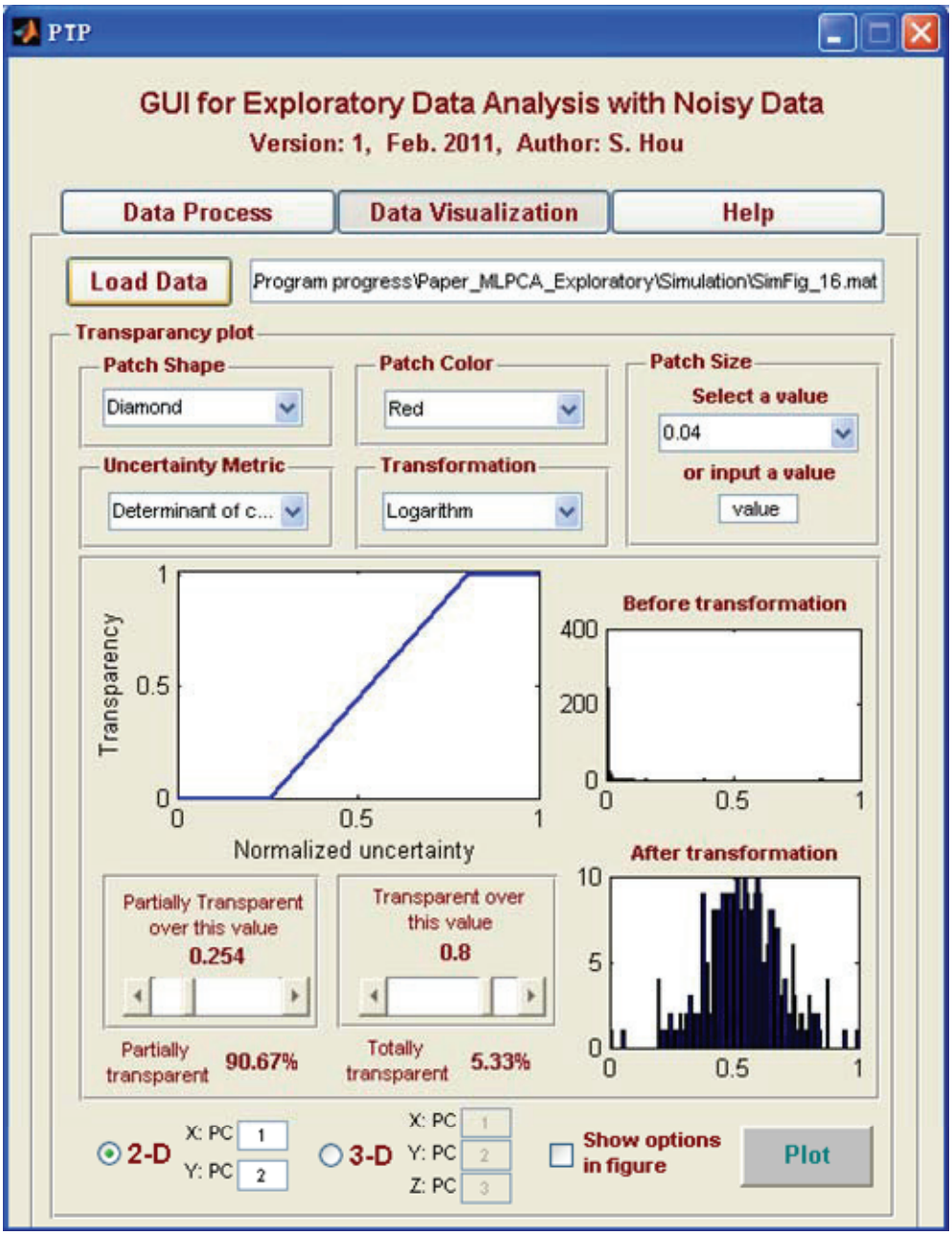

Figure 2.12 A screen shot of the graphical user interface developed to interactively adjust the settings for the partial transparency projection.

To illustrate the effect of the PTP, data were generated in a manner similar to the earlier simulations, with $\alpha = 0.5$ and $\sigma_{log} = 0.5$. In this case, however, a common error standard deviation was applied across all measurements for a given object (*i.e.* the

measurement errors were homoscedastic within a measurement vector). This ensured a range of quality in the objects generated. Figure 2.13 shows the effects of various settings of the PTP, which incrementally reveal the underlying structure of the data. Note that the classes have not been differentiated by color in this case, as generally that information would not be available to evaluate the class structure. For this simulation, a pseudo-rank of five was assumed for the MLPCA preprocessing (overestimating the true value of two), the quality measure was taken as the largest eigenvalue of the projected error covariance matrix, and a logarithmic transformation was applied to the quality measures. The mapping function is shown in the upper left corner of each subfigure.

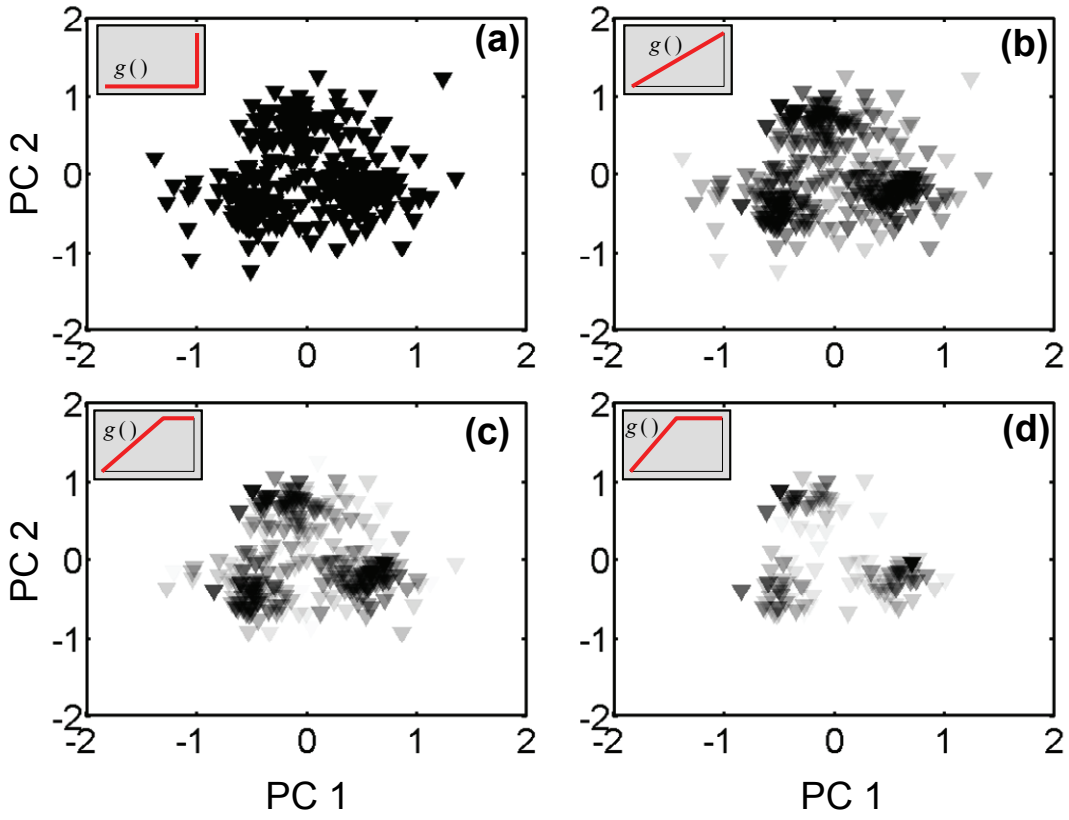

Figure 2.13 Illustration of partial transparency technique to reveal clusters in simulated data: (a) no transparency adjustment, (b) transparency adjusted linearly across the full range, (c) about one-third of the low quality objects rendered transparent, (d) about one-half of low quality objects rendered transparent. The transparency mapping function is shown in the upper left corner in each case.

2.4.2 DNA Microarray Data

The DNA microarray data set was chosen to demonstrate the method here because it has been widely observed that heteroscedastic errors are common in this type of

experimental technique. By following the proposed procedure in Figure 2.4, the DNA microarray data were processed. MLPCA was first applied to the data set, with 8 principal components chosen. After the estimation of the subspace and the scores, the estimates of the gene expression levels in the original space were performed by rotating the scores back to the original space. From the biological perspective, the absolute gene expression levels are less important, and thus the estimated gene expression profile of each gene is normalized to unit length. The normalized estimated gene expression data were then column mean-centered and normal PCA was applied. The scores obtained from PCA are used for visualization. It is worth noting that in each step of the process, the measurement uncertainties were propagated by following Equations (2.5), (2.8) and (2.13).

The scores plots with adjusted transparencies are shown in Figure 2.14. Figure 2.14 (a) shows the scores plots of the first two principal components without applying transparency. It can be seen that the points spread across the plot area and no clusters can be seen clearly. In Figure 2.14 (b), the 2×2 measurement error covariance matrix of each gene for the first two components was first transformed to a scalar by taking the largest eigenvalue of the covariance matrix. The transformed scalar was further transformed by taking the natural logarithm. The degree of transparency of each gene was adjusted based on the transformed value. If the value is large, the point will be more transparent. The transparency function is schematically shown in the lower right corner of the subfigures of Figure 2.14. In this figure, although some points are partially transparent, no clear clusters can be seen. In Figure 2.14 (c), the same transformation of the measurement uncertainties as that in Figure 2.14 (b) was followed, but some points were set to be totally transparent, which is indicated in the lower right corner of this figure. It can be seen that clusters emerge even though the clusters are not very clear. Figure 2.14 (d) follows that same transformation of error uncertainties, but transparency was further adjusted, indicated in the lower right corner. In this figure, three large clusters and one small cluster can be clearly seen, showing that the transparency technique has helped to reveal clusters. Depending on applications, the boundaries of the clusters may or may not be clear, but if clusters are seen, it might indicate some important information and further study is desired.

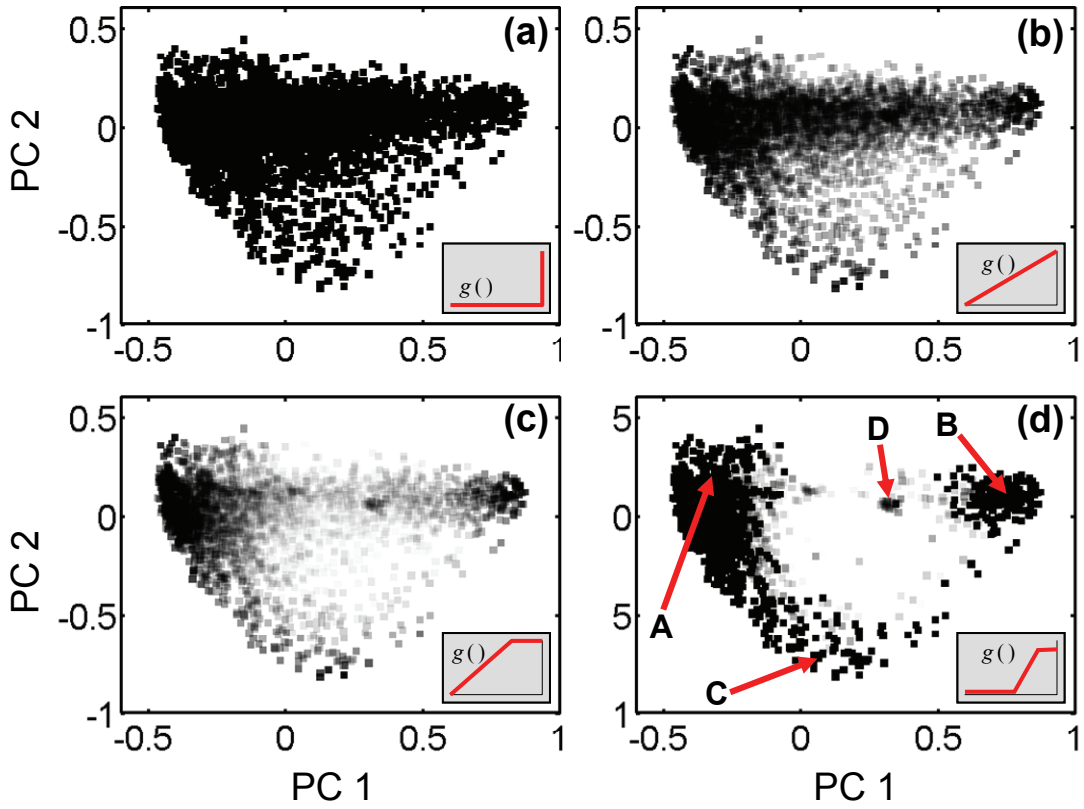

Figure 2.14 Scores plots of experimental DNA microarray data with transparency adjusted to reveal clusters: (a) no transparency adjustment, (b) linear transparency mapping, (c) linear transparency mapping with a transparency cutoff (plateau), (d) linear transparency mapping with transparency and opacity cutoffs (pseudo-sigmoid). The transparency mapping function is shown in the lower right corner.

In Figure 2.14, although clusters can be revealed by the adjustment of transparency of the data points, it is useful to examine the degree of correlation of genes in the same clusters. To do this, the original gene expression profiles were normalized to unit length. The genes in each cluster shown in Figure 2.14 (d) were extracted and their normalized profiles are shown in Figure 2.15. Figure 2.15 (a) shows the normalized profiles of 100 genes extracted in cluster A shown in Figure 2.15 (d). It can be seen that the genes present consistent profiles, although a few genes have high values at time point 8 (15 min). These high values are likely caused by large measurement errors, as suggested by large uncertainties associated with these points. It is encouraging that these objects are included in the cluster even though some measurements have large errors, and this is likely a consequence of MLPCA preprocessing, which mitigates those effects. When the yeast cells

exit from stationary phase, these genes are up-regulated after a short initial lag period. In contrast, Figure 2.15 (b) shows the normalized gene expression profiles of 100 genes extracted in cluster B in Figure 2.14 (d) that are rapidly down-regulated within a period of about five minutes upon exit from stationary phase. It can be seen that these genes are also highly correlated and contain some measurements with large errors. The normalized expression profiles of 100 genes extracted from cluster C of Figure 2.14 (d) are shown in Figure 2.15 (c) and are characterized by a transient up-regulation of these genes between 1 min and 15 min after exit from the quiescent state. While the pattern is clear, these gene profiles are not as highly correlated as the first two sets, probably because the cluster is small and 100 genes encompasses a wider region than the other two. Finally, Figure 2.15 (d) shows 20 normalized gene expression profiles from the small cluster D indicated in Figure 2.14 (d). These gene profiles most closely resemble those in cluster B, as anticipated from

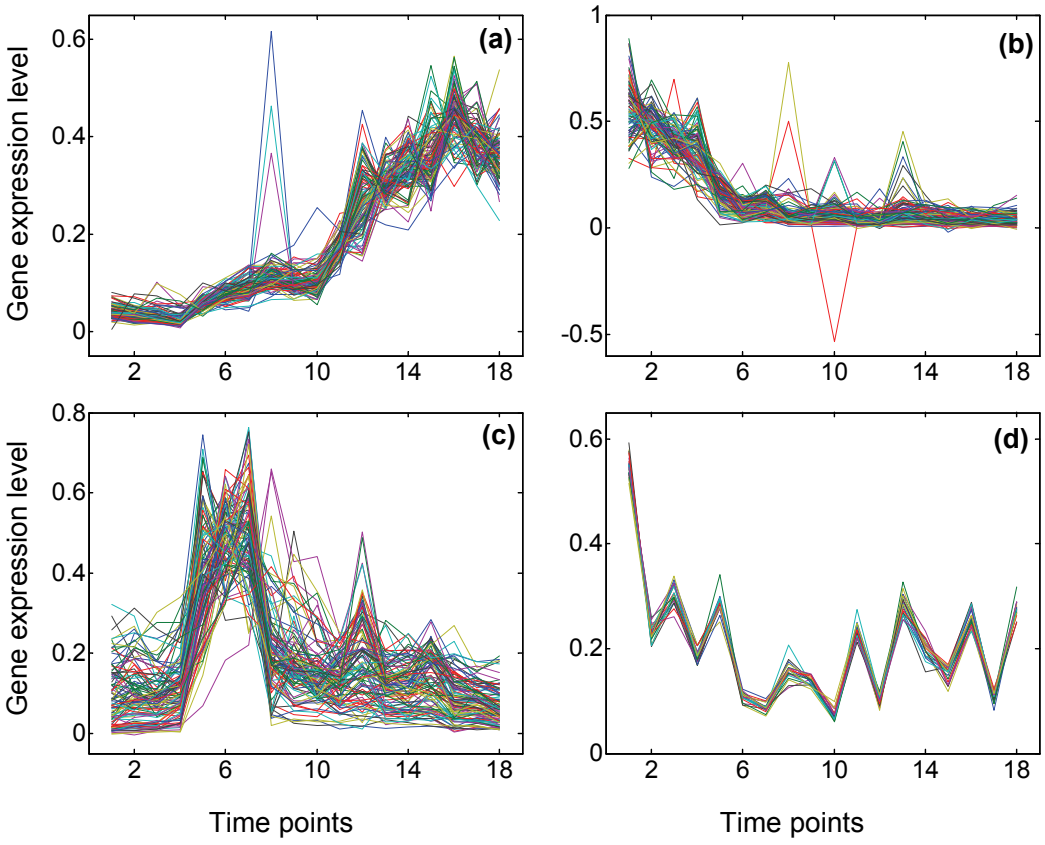

Figure 2.15 Normalized gene expression profiles for genes associated with the clusters shown in Figure 2.14 (d) (A = (a), etc.). Profiles for the 100 genes nearest the points indicated by arrows in scores plot are shown, except for (d), which includes only 20 profiles. The x-axis indicates the order of the experiment in the time course study, but is not linear in time.

their spatial location, and are very highly correlated. This high correlation is anticipated because these profiles represent a complete set of measurements from 20 cross-hybridization replicates on the microarray; *i.e.*, the same DNA 70-mer sequence replicated at 20 different spots on the array. As such, the appearance of this cluster is artificial and a consequence of oversampling of the measurement, but it serves as a validation of the methodology which allowed the identification of a cluster that otherwise would have been difficult to distinguish with other methods. These observations are consistent with profiles previously identified for this experiment using other methods [47].

2.5 Conclusions

Exploratory data analysis involving visualization in a lower-dimensional space is unique among applications of multivariate data analysis in that there is no identifiable model assumed for the data, and therefore no optimal solution that is clearly defined. The best solution is the one that is interesting to analysts and unbiased. Methods such as PCA and PP have been applied successfully because they optimize parameters (variance, non-normality) that often correlate to interesting projections, but when observations show significant heteroscedasticity in their measurement errors, both in magnitude and extent, the statistical underpinnings of these methods are eroded and they are less effective. When measurement error variance can be estimated, however, some of the methods described here may help to separate the information from the noise. The three principles guiding the procedures described in this work have been: (1) a reduction in the magnitude of measurement errors by maximum likelihood modeling and projection in a lower-dimensional space, (2) propagation of measurement uncertainties throughout the process, and (3) visualization in a manner that allows the propagated uncertainties to be incorporated into the low-dimensional projection through transparency mapping to place emphasis on higher quality data. These principles can be applied individually or in tandem and have been shown to improve the quality of information displayed for simulated and experimental data sets. In many cases, such measures may not be necessary, but as data sets become more complex and uncertainty information becomes integrated into the measurements, the application of such strategies is likely to become more important.

2.6 Bibliography

1. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Inc., 1977.
2. K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2 (1901) 559–572.
3. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part I), *The Journal of Educational Psychology*, 24 (1933) 417-441.
4. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part II), *The Journal of Educational Psychology*, 24 (1933) 498-520.
5. R. A. Fisher, The Statistical Utilization of Multiple Measurements, *Annals of Eugenics*, 8 (1938) 376-386.
6. I. E. Frank, and J. H. Friedman, Classification: Oldtimers and Newcomers, *Journal of Chemometrics*, 3 (1989) 463-475.
7. M. Barker, and W. Rayens, Partial Least Squares for Discrimination, *Journal of Chemometrics*, 17 (2003) 166-173.
8. R. Rosipal, and N. Kramer, Overview and Recent Advances in Partial Least Squares, *Lecture Notes in Computer Science*, 3940 (2006) 34-51.
9. R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, Inc., New Jersey, USA, 1998.
10. C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2 (1998) 121-167.
11. M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, USA, 1973.
12. B. S. Everitt, *Cluster Analysis*, Third Edition, Edward Arnold, London, UK, 1993.
13. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, USA, 2006.
14. J. H. Friedman, and J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions and Computers*, 23 (1974) 881-890.
15. P. J. Huber, Projection Pursuit, *The Annals of Statistics*, 13 (1985) 435-475.
16. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.
17. M. Hubert, P. J. Rousseeuw, and S. A. Verboven, Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60 (2002) 101-111.

18. M. Hubert, and S. Engelen, Robust PCA and Classification in Bioscience, *Bioinformatics*, 20 (2004) 1728-1736.

19. M. A. Girshick, Principal Components, *Journal of the American Statistical Association*, 31 (1936) 519-528.

20. P. D. Wentzell, Chapter 2.25: Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods, in S. D. Brown, R. Tauler, and B. Walczak (Editors): *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Elsevier Ltd., 2009.

21. J. D. Ingle, and S. R. Crouch, *Spectrochemical Analysis*, Prentice-Hall, New Jersey, 1988.

22. P. D. Wentzell, T. K. Karakach, S. Roy, M. J. Martinez, C. P. Allen, and M. Werner-Washburne, Multivariate Curve Resolution of Time Course Microarray Data, *BMC Bioinformatics*, 7 (2006) 343.

23. T. K. Karakach, R. M. Flight, and P. D. Wentzell, Bootstrap Method for the Estimation of Measurement Uncertainty in Spotted Dual-Color DNA Microarrays, *Analytical and Bioanalytical Chemistry*, 389 (2007) 2125-2141.

24. T. K. Karakach, and P. D. Wentzell, Methods for Estimating and Mitigating Errors in Spotted, Dual-color DNA Microarrays, *OMICS: A Journal of Integrative Biology*, 11 (2007) 186-199.

25. P. Paatero , and U. Tapper, Analysis of Different Modes of Factor Analysis as Least Squares Fit Problems, *Chemometrics and Intelligent Laboratory Systems*, 18 (1993) 183-194.

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

27. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, Maximum Likelihood Principal Component Analysis, *Journal of Chemometrics*, 11 (1997) 339-366.

28. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.

29. D. T. Andrews, and P. D. Wentzell, Applications of Maximum Likelihood Principal Component Analysis: Incomplete Data Sets and Calibration Transfer, *Analytica Chimica Acta*, 350 (1997) 341-352.

30. P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Maximum Likelihood Multivariate Calibration, *Analytical Chemistry*, 69 (1997) 2299-2311.

31. S. K. Schreyer, M. Bidinosti, and P. D. Wentzell, Application of Maximum Likelihood Principal Components Regression to Fluorescence Emission Spectra, *Applied Spectroscopy*, 56 (2002) 789-796.

32. M. N. Leger, and P. D. Wentzell, Maximum Likelihood Principal Components Regression on Wavelet-Compressed Data, *Applied Spectroscopy*, 58 (2004) 855-862.

33. M. R. Keenan, Maximum Likelihood Principal Component Analysis of Time-of-Flight Secondary Ion Mass Spectrometry Spectral Images, *Journal of Vacuum Science & Technology A*, 23 (2005) 746-750.

34. L. Vega-Montoto, and P. D. Wentzell, Maximum Likelihood Parallel Factor Analysis, *Journal of Chemometrics*, 17 (2003) 237-253.

35. L. Vega-Montoto, H. Gu, and P. D. Wentzell, Mathematical Improvements to Maximum Likelihood Parallel Factor Analysis: Theory and Simulations, *Journal of Chemometrics*, 19 (2005) 216-235.

36. L. Vega-Montoto, and P. D. Wentzell, Mathematical Improvements to Maximum Likelihood Parallel Factor Analysis: Experimental Studies, *Journal of Chemometrics*, 19 (2005) 236-252.

37. M. Schuermans, I. Markovsky, P. D. Wentzell, and S. Van Huffel, On the Equivalence between Total Least Squares and Maximum likelihood PCA, *Analytica Chimica Acta*, 544 (2005) 254-267.

38. G. H. Golub, and C. F. Van Loan, An Analysis of the Total Least Squares Problem, *SIAM Journal on Numerical Analysis*, 17 (1980) 883-893.

39. S. Van Huffel, and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*; The Society for Industrial and Applied Mathematics, Philadelphia, 1991.

40. S. Van Huffel (Editors), *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, The Society for Industrial and Applied Mathematics, Philadelphia, 1997.

41. S. Van Huffel, and P. Lemmerling (Editors), *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, 2002.

42. P. Paatero, and U. Tapper, Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values, *Environmetrics*, 5 (1994) 111-126.

43. P. Paatero, Least Squares Formulation of Robust Non-negative Factor Analysis, *Chemometrics and Intelligent Laboratory Systems*, 37 (1997) 23-35.

44. J. Kragten, Calculating Standard Deviations and Confidence Intervals with a Universally Applicable Spreadsheet Technique, *The Analyst*, 119 (1994) 2161-2165.

45. Y. Chen, E. R. Dougherty, and M. L. Bittner, Ratio-base Decisions and the Quantitative Anlaysis of cDNA Microarray Images, *Journal of Biomedical Optics*, 2 (1997) 364-374.

46. D. M. Rocke, and B. Durbin, A Model for Measurement Error for Gene Expression Arrays, *Journal of Computational Biology*, 8 (2001) 557-569.

47. M. J. Martinez, S. Roy, A. B. Archuletta, P. D. Wentzell, S. S. Anna-Arriola, A. L. Rodriguez, A. D. Aragon, G. A. Quinones, C. Allen, and M. Werner-Washburne,

Genomic Analysis of Stationary-Phase and Exit in Saccharomyces Cerevisiae: Gene Expression and Identification of Novel Essential Genes, *Molecular Biology of the Cell*, 15 (2004) 5295-5305.

48. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.
49. A. Björck, and G. H. Golub, Numerical Methods for Computing Angles between Linear Subspaces, *Mathematics of Computation*, 27 (1973) 579-594.
50. S. Jiang, Angles between Euclidean Subspace, *Geometriae Dedicata*, 63 (1996) 113-121.
51. J. Miao, and A. Ben-Israel, On Principal Angles between Subspace in R^n, *Linear Algebra and Its Applications*, 171 (1992) 81-98.

Chapter 3: Development of an Optimization Algorithm for Maximum Likelihood Principal Component Analysis Model with Intercepts

3.1 Introduction

Data preprocessing is a crucial part of multivariate analysis and often can be a critical factor in determining the success or failure of a procedure. Many data treatment methods fall into this category, from the simple (*e.g.* baseline subtraction) to the complex (*e.g.* wavelet transformation). Perhaps the most widely applied methods, and the most fundamental, are mean centering and scaling. Despite their simplicity, these adjustments can have a profound impact on the quality of results. Scaling is generally defined here as the multiplication of rows or columns of data by a scalar quantity associated with, for example, the inverse of the range of data. The role of scaling, particularly in the case of variable or column scaling, is often to normalize the range of variables prior to treatment by principal component analysis (PCA), but it can have detrimental effects when variables consist only of measurement noise. It has been pointed out that a more fundamental motivation for scaling is to normalize the measurement error variance so that the subspace modeled by PCA is optimal in the maximum likelihood sense [1,2]. An alternative way to treat non-uniform measurement noise variances (*i.e.* heteroscedastic errors), and one that is essential for unstructured or correlated noise, is maximum likelihood principal component analysis (MLPCA). MLPCA uses measurement error information to select the optimal subspace for a given set of data and has been employed in a wide variety of applications where scaling is insufficient to provide optimal subspace estimation [3,4,5,6,7,8,9,10].

While the effects of mean centering are generally not as dramatic or complex as scaling, this step can still influence the quality of results. Here, mean centering is defined as the subtraction of column means and/or row means from a data matrix. Where both are used, one is employed following the application of the other. The rationale behind mean centering is that the chemical information is carried in the variance of the data, so that the mean is of little value. In the application of PCA, for example, it is well known that the first eigenvector (loading) is simply the scaled column averages. In removing the mean, more information may, in certain cases, be compressed into fewer latent variables. Some of the

implications of mean centering have been examined in the literature [11,12,13,14,15,16,17]. In multivariate calibration, mean centering is widely used, but may also suffer from disadvantages in certain cases [12,14,15,17]. In other applications, such as multivariate curve resolution (MCR), mean centering is generally not applied because it excludes the use of a non-negativity constraint. For exploratory data analysis, specifically in data visualization after projection by PCA or other methods, mean centering is particularly useful. This is because the removal of the mean often reduces the dimensionality (*i.e.* pseudo-rank) of a data set by one. While this is not of major importance in applications such as multivariate calibration where the addition of one more factor is usually inconsequential, visualization methods require projection of the data into one to three dimensions and therefore require efficient retention of information.

While PCA can be regarded as a variance modeling method and mean centering as a way to examine variance around the mean as opposed to variance around the origin, it can also be regarded as a way to optimally estimate a subspace of the original space that contains the measurements. The latter definition is more consistent with MLPCA, which assumes the error-free data define a hyperplane of dimensionality lower than the mathematical rank of the observed data matrix. In this interpretation, mean centering can be regarded as one way to remove the intercepts of the hyperplane in the row or column space, thereby reducing the number of vectors in a basis for the subspace. This is illustrated in Figure 1, which shows a data set in a one-dimensional hyperplane located in an observed two-dimensional space. The presence of the intercept in Figure 1 (a) means that the data could only be accurately represented using two principal components. Mean centering, as illustrated in Figure 1 (b), forces the model to pass through the origin, permitting the subspace to be estimated with a single principal component. It is important to note, however, that there are an infinite number of other transformations, such as those shown in Figures 1 (c) and (d), that will also reduce the intercepts of the hyperplane. Mean centering is simply a convenient way of forcing a zero intercept and its application in the case of homoscedastic errors is consistent with PCA as a maximum likelihood method for subspace estimation. Thus, mean centering for PCA can be interpreted from two perspectives: (1) it can remove the intercepts of a data set, and (2) the subspace estimated by PCA always passes through the mean point.

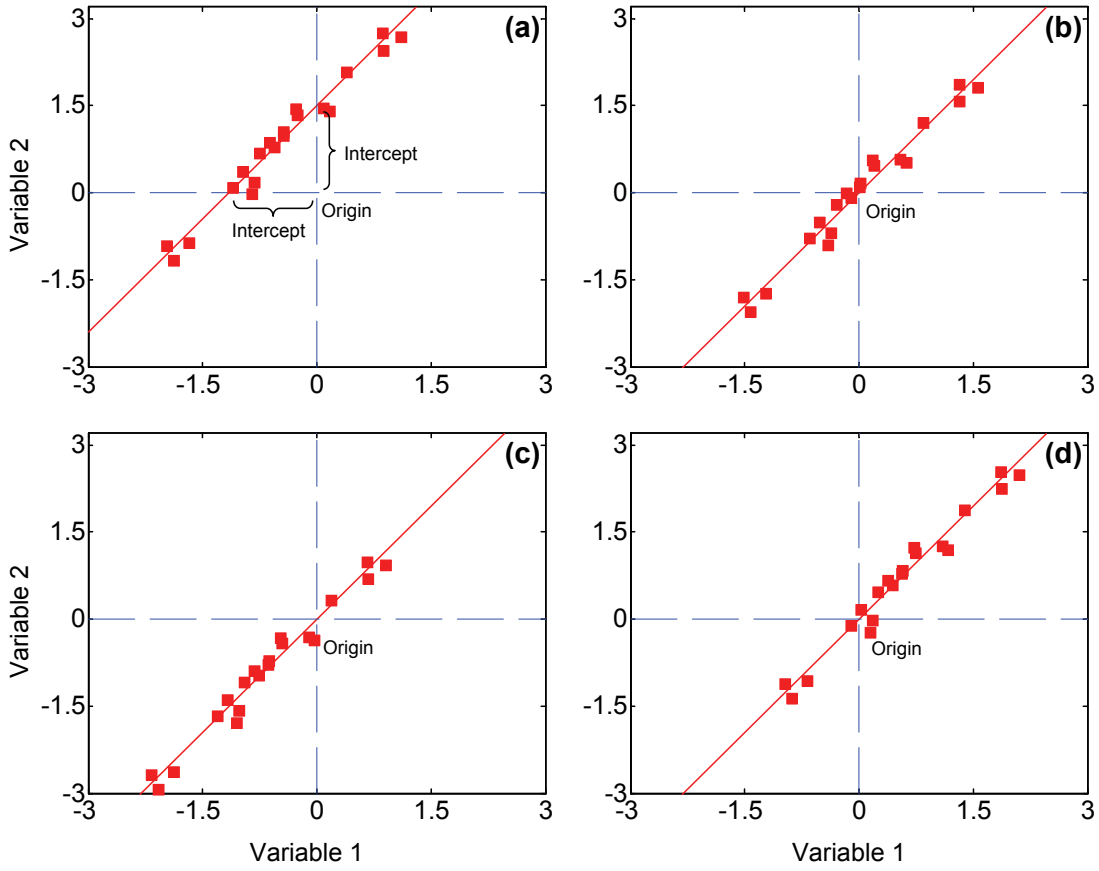

Figure 3.1 Representation of the effect of mean centering and other transformations on the removal of intercepts. (a) Original data. (b) Mean centering. (c)-(d) Other transformations.

When measurement errors in a data set are sufficiently heteroscedastic as to justify the application of MLPCA for subspace estimation, mean centering is problematic because it does not ensure the removal of intercepts that would be consistent with a maximum likelihood estimation of the subspace. In other words, the conventional column mean vector is not necessarily located in the subspace estimated from maximum likelihood because of the presence of heteroscedastic errors. This problem was noted in the original development of MLPCA [1], but at that time no efficient iterative algorithm was available to treat this situation. The present work reports the development of a variant of the MLPCA alternating least squares (ALS) algorithm to treat the case where the subspace of the data does not exhibit zero intercepts, thereby creating an analogue to mean centering in PCA that will be useful in treating data where row and/or column offsets are important. The proposed optimization algorithm is assessed by simulated data.

3.2 Background

Consider a data matrix, **X** (n x p), where the error-free measurements occupy a d-dimensional subspace in the n-dimensional row space (sample space) and the p-dimensional column space (variable space). A general mathematical model can be written in the form:

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{1}_n\mathbf{c}^{\mathrm{T}} + \mathbf{r}\mathbf{1}_p^{\mathrm{T}} + \mathbf{E}\,. \tag{3.1}$$

Here **T** represents the n x d matrix of factor scores, **V** is the p x d matrix of factor loadings, **c** is a p x 1 vector of column offsets, **r** is a n x 1 vector of row offsets, **1**'s represents a vector of ones of appropriate length (n or p) and **E** is an n x p matrix of residuals consistent with a defined error variance/covariance structure. It should be noted that there exist ambiguities in rotation, translation, and scale for the terms **T**, **V**, **c** and **r** in this model such that different constraints will produce equivalent solutions satisfying maximum likelihood conditions, but this is not important in considering the general case. Four subsets of this general model can be considered. The simplest of these is when the row and column offsets are zero. The second case is where the row offsets are zero, but the column offsets are not, generally leading to a d-dimensional hyperplane with non-zero intercepts in the column space, and a (d+1)-dimensional hyperplane with zero intercepts in the row space. In PCA, this case is handled by column mean centering, which makes the intercepts zero and restores the dimensionality to d in the row space. The opposite is true for the third case, where the column offsets are zero and the row offsets are not, and this case is handled by row mean centering. The fourth and most general case, where both column and row offsets are non-zero, leads to (d+1)-dimensional hyperplanes with non-zero intercepts in both column and row spaces, with an overall pseudo-rank of (d+2). This case can be reduced to a simple bilinear model by successive subtraction of row and column means.

The origin of row and column offsets varies with the type of data set under investigation. In the context of spectroscopy, for example, where spectra of individual samples are represented as rows, column offsets can be thought of as a fixed background spectrum present in all samples. A row offset could be considered to arise from a background signal that was fixed at all wavelength channels, but which varied from sample to sample, such as a baseline shift or cell positioning error. Whatever the source, the elimination of these effects is a desirable outcome. While the offsets can be accommodated

by increasing the dimensionality of the model, this is generally undesirable because of increased complexity and the loss of degrees of freedom.

In the following sections, an optimization algorithm is developed to include row or column offsets in MLPCA decomposition. Because only one of these treatments is usually applied, and because the problem of rows and columns is symmetric through a transpose, the work presented here will emphasize column offsets in the results and discussion, but this may be extended to other cases. For completeness, the derivation of the optimization algorithm for the MLPCA model with both column and row offsets is presented in Appendix (Section 3.7). The two terms intercept and offset are used interchangeably in this chapter.

3.3 Theory

When the MLPCA model consisting only of column intercepts is considered, the general model in Equation (3.1) can be simplified as

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{1}\mathbf{c}^{\mathrm{T}} + \mathbf{E}, \tag{3.2}$$

where the subscript n for $\mathbf{1}$ is dropped to simplify the notation. For the convenience of the derivation in the next steps, it is good to clearly delineate the notations as follow:

$$\underset{(n\times 1)}{\mathbf{1}} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \end{bmatrix}^{\mathrm{T}}, \tag{3.3}$$

$$\underset{(p\times 1)}{\mathbf{c}} = \begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_p \end{bmatrix}^{\mathrm{T}}, \tag{3.4}$$

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{x}}_1 & \cdots & \underline{\mathbf{x}}_p \end{bmatrix}, \tag{3.5}$$

$$\underset{(n\times d)}{\mathbf{T}} = \begin{bmatrix} t_{11} & \cdots & t_{1d} \\ \vdots & \ddots & \vdots \\ t_{n1} & \cdots & t_{nd} \end{bmatrix} = \begin{bmatrix} \mathbf{t}_1^{\mathrm{T}} \\ \vdots \\ \mathbf{t}_n^{\mathrm{T}} \end{bmatrix}, \text{ and} \tag{3.6}$$

$$\underset{(p\times d)}{\mathbf{V}} = \begin{bmatrix} v_{11} & \cdots & v_{1d} \\ \vdots & \ddots & \vdots \\ v_{p1} & \cdots & v_{pd} \end{bmatrix} = \begin{bmatrix} \underline{\mathbf{v}}_1^{\mathrm{T}} \\ \vdots \\ \underline{\mathbf{v}}_p^{\mathrm{T}} \end{bmatrix}. \tag{3.7}$$

It is worth noting that $\mathbf{x}_i$ denotes a single sample measured on a group of variables, the underlined notation $\underline{\mathbf{x}}_j$ represents a group of samples measured on a single variable, $\mathbf{t}_i$ is the score vector for a single sample, and $\underline{\mathbf{v}}_j^{\mathrm{T}}$ designates a row in **V**. The subscript *i* is reserved for the index of samples (*i*=1, ..., *n*), and the subscript *j* is used for the index of variables (*j*=1, …, *p*). In ordinary expression of singular value decomposition (SVD) [18] or PCA, the columns of **V** are the basis vectors of the subspace, but here they are not assigned notations. Instead, each row in **V** is assigned an underlined notation $\underline{\mathbf{v}}_j^{\mathrm{T}}$.

One might have realized that this MLPCA model is very similar to that used for maximum likelihood factor analysis [19,20], but substantial differences exist. In maximum likelihood factor analysis, it is assumed that the scores (or called factors or latent variables) follow a multivariate normal distribution [21,22]. In this MLPCA model, there is no such assumption. Also, in this MLPCA model, the measurement error variances are assumed to be known while in the maximum likelihood factor analysis, the error variances are not.

In this development, the measurement errors are assumed to be independent across the measurements in a data matrix. Based on the principle of maximum likelihood, the objective is to estimate **T**, **V** and **c** such that the likelihood function is maximized. The likelihood function (*L*) for the observed data **X** can be written, with respect to the rows of **X**, as

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{n}{2}} \left|\mathbf{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right)^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right) \right], \tag{3.8}$$

where $\mathbf{\Sigma}_i$ is the measurement error covariance matrix for sample *i* (each row in **X**) and other notations are defined in Equations (3.4) to (3.7). Since independent errors are assumed, it is a diagonal matrix. The likelihood function (*L*) can also be written, with respect to the columns of **X**, as

$$L = \prod_{j=1}^{p} \frac{1}{(2\pi)^{\frac{p}{2}} \left|\mathbf{\Psi}_j\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\underline{\mathbf{x}}_j - c_j\mathbf{1} - \mathbf{T}\underline{\mathbf{v}}_j\right)^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j\mathbf{1} - \mathbf{T}\underline{\mathbf{v}}_j\right) \right], \tag{3.9}$$

where $\mathbf{\Psi}_j$ is the error covariance matrix (a diagonal matrix as well) for variable *j* (each column in **X**), c_j is the *j*th element of vector **c** defined in Equation (3.4), and other notations are defined in Equations (3.3) to (3.7).

As it is generally easier to work on the natural logarithm of the likelihood function, the log-likelihood function ($\ln L$) is employed and written as

$$\ln L = \text{constant} - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right), \text{ or} \tag{3.10}$$

$$\ln L = \text{constant} - \frac{1}{2}\sum_{j=1}^{p}\left(\underline{\mathbf{x}}_j - c_j\mathbf{1} - \mathbf{T}\underline{\mathbf{v}}_j\right)^{\mathrm{T}} \boldsymbol{\Psi}_j^{-1}\left(\underline{\mathbf{x}}_j - c_j\mathbf{1} - \mathbf{T}\underline{\mathbf{v}}_j\right). \tag{3.11}$$

Note the constant terms in Equation (3.10) and (3.11) are not necessarily the same, but they do not affect the maximum likelihood estimation. Based on Equation (3.10), the partial derivatives of $\ln L$ with respect to $\mathbf{t}_i$ can be written as

$$\frac{\partial \ln L}{\partial \mathbf{t}_i} = -\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{c}\right) + \left(\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}\right)\mathbf{t}_i, \tag{3.12}$$

based on vector calculus [23,24,25]. Setting this to zero leads to

$$\mathbf{t}_i = \left(\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{c}\right). \tag{3.13}$$

If $\mathbf{c}$ is a zero vector, Equation (3.13) reduces to the equation in the original algorithms for MLPCA model without intercepts [1]. Based on Equation (3.10), the partial derivatives of $\ln L$ with respect to $\mathbf{c}$ can be written as

$$\frac{\partial \ln L}{\partial \mathbf{c}} = -\sum_{i=1}^{n}\left[\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{V}\mathbf{t}_i\right)\right] + \sum_{i=1}^{n}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{c}\right). \tag{3.14}$$

Note that $\sum_{i=1}^{n}$ is the summation notation and $\boldsymbol{\Sigma}_i^{-1}$ denotes the inverse of the error covariance matrix. Setting this to zero gives

$$\mathbf{c} = \left(\sum_{i=1}^{n}\boldsymbol{\Sigma}_i^{-1}\right)^{-1}\left\{\sum_{i=1}^{n}\left[\boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{V}\mathbf{t}_i\right)\right]\right\}. \tag{3.15}$$

Based on Equation (3.11), the partial derivatives of $\ln L$ with respect to $\underline{\mathbf{v}}_j$ can be written as

$$\frac{\partial \ln L}{\partial \underline{\mathbf{v}}_j} = -\mathbf{T}^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\left(\underline{\mathbf{x}}_j - c_j\mathbf{1}\right) + \left(\mathbf{T}^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{T}\right)\underline{\mathbf{v}}_j. \tag{3.16}$$

Setting this to zero yields

$$\underline{\mathbf{v}}_j = \left(\mathbf{T}^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\boldsymbol{\Psi}_j^{-1}\left(\underline{\mathbf{x}}_j - c_j\mathbf{1}\right). \tag{3.17}$$

Equations (3.13), (3.15), and (3.17) give an iterative method to calculate $\mathbf{t}_i$, $\mathbf{c}$, and $\underline{\mathbf{v}}_j$, respectively. Iteratively applying the three equations until convergence will result in the maximum likelihood estimates of **T**, **V** and $\mathbf{c}$.

Examination shows that **T** and **V** have rotational ambiguity in the sense that

$$\mathbf{T}\mathbf{V}^{\mathrm{T}} = (\mathbf{T}\mathbf{R})(\mathbf{R}^{-1}\mathbf{V}^{\mathrm{T}}) = (\mathbf{T}\mathbf{R})(\mathbf{V}\mathbf{R})^{\mathrm{T}}, \tag{3.18}$$

where **R** is an arbitrary rotation matrix with $\mathbf{R}^{-1}=\mathbf{R}^{\mathrm{T}}$. To remove this ambiguity, SVD can be applied to $\mathbf{T}\mathbf{V}^{\mathrm{T}}$ to make the columns of **T** be mutually orthogonal and **V** be an orthonormal basis. Based on Equation (3.2), the estimated data can be expressed as $\mathbf{X}_{est} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{1}\mathbf{c}^{\mathrm{T}}$. Referring to the vector form of the equation of a plane discussed in basic linear algebra, $\mathbf{c}$ should be a point located in the hyperplane for $\mathbf{X}_{est}$, and any point in the hyperplane can be used. Thus, $\mathbf{c}$ can be chosen as any point in the hyperplane. This indicates that the intercept term $\mathbf{c}$ has ambiguity as well. To make it unique, $\mathbf{c}$ can be forced to be orthogonal to **V**, but a better choice is to let $\mathbf{c}$ be the mean vector of the estimated data $\mathbf{X}_{est}$. Then $\mathbf{T}\mathbf{V}^{\mathrm{T}}$ has a zero mean vector. In each iteration, when $\mathbf{c}$ is forced to be the mean vector of the estimated data, the scores need to be adjusted accordingly so that $\mathbf{T}\mathbf{V}^{\mathrm{T}}$ has a zero mean vector. Since $\mathbf{X}_{est}$ is the maximum likelihood estimate of the error-free data, it is conceived that the mean vector of the estimated data is a good estimate of that of the error-free data, which should give a better estimate than the conventional sample mean vector. By adding these constraints to **T**, **V**, and $\mathbf{c}$, the ambiguity problems are solved and they become unique*. However, the uniqueness of three terms does not mean the iterative search algorithm can guarantee the global optimum. Like many other algorithms, this algorithm may hit a local optimum. Thus, different initial guesses should be tried to increase the probability of finding the global optimum.

Similar to the original MLPCA algorithms, the new algorithm also employs an iterative method derived by taking the partial derivatives of the objective function with respect to the parameters to be estimated, and setting them to zeros. The proposed algorithm estimates the three terms **T**, **V**, and $\mathbf{c}$ simultaneously, and is a generalization

* Strictly speaking, applying SVD to $\mathbf{T}\mathbf{V}^{\mathrm{T}}$ cannot make **T** and **V** unique because the signs can change, *e.g.* $\mathbf{T}\mathbf{V}^{\mathrm{T}} = \mathbf{(-T)(-V)}^{\mathrm{T}}$. Here variants due to the sign change are treated to be the same.

of the optimization algorithm for MLPCA model without intercepts [1]. The proposed algorithm can be summarized as follows:

1. Give a random guess each to **V** and **c**, respectively;
2. Use Equation (3.13) to calculate $\mathbf{t}_i$ and obtain **T**;
3. Update **c** based on Equation (3.15);
4. Adjust **c** by replacing it with the mean vector of the estimated data ($\mathbf{TV}^{\mathrm{T}} + \mathbf{1c}^{\mathrm{T}}$) and adjust **T** accordingly;
5. Apply SVD to $\mathbf{TV}^{\mathrm{T}}$ and make the columns of **T** be mutually orthogonal and **V** be an orthonormal basis;
6. Calculate $\underline{\mathbf{v}}_j$ based on Equation (3.17) and obtain **V**;
7. Repeat steps 2 to 6 until convergence;
8. If different initial guesses are needed, repeat steps 1 to 7 and choose the best solution - the one that gives the smallest sum of squares of the weighted residuals, calculated by $\sum_{i=1}^{n}\left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right)$.

3.4 Experimental

3.4.1 Computational Aspects

All data processing was carried out using programs written by the author in MatLab® v.7.4.0 (MathWorks, Natick, MA).

3.4.2 Data Simulation

3.4.2.1 Data Set 1

To evaluate the performance of the proposed algorithms, simulated data were used. Data set 1 consisted of 40 objects evenly divided into two classes in a two-dimensional space with the aim to plot them in a two-dimensional Cartesian coordinate system for visualization. The measurement error-free data have a pseudo-rank of one and are located on a straight line with an angle of 45° from the positive abscissa with an intercept of 5 on the y-axis. Twenty samples were drawn from each of the two classes. The first coordinates of the error-free data were randomly generated from two normal distributions

$N(5,\ 1)$ and $N(15,\ 1)$, respectively. The second coordinates were obtained by adding 5 to the first coordinates.

The measurement errors were assumed to be independent and heteroscedastic. The standard deviations of measurement errors were assumed to follow a log-normal distribution. A 40 x 2 matrix containing the measurement standard deviations was created by randomly drawing 40 x 2 numbers from a normal distribution $N(0,1)$, followed by exponential transformation (base e). As the measurement error standard deviations for real experiments cannot be infinitely small, 0.1 was added to each of them to make them more realistic. Measurement errors were simulated by randomly drawing 40 x 2 numbers from a normal distribution $N(0,1)$ and multiplying these numbers by the corresponding standard deviations obtained previously. This created the heteroscedastic measurement errors. The simulated errors were added to the error-free data to generate the measured data.

3.4.2.2 Group Data Set 1

To further assess the algorithm for data in higher-dimensional spaces, a group consisting of 100 data sets, designated as group data set 1, was created. Each of the data sets contained 300 objects from three classes in a 12-dimensional space, with 100 from each class. The 100 data sets were created by using the same method and simulation parameters; therefore they were 100 different random realizations. The three class centers were first simulated in a six-dimensional space with the three mean vectors set as

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -3 & -3 & 0 & 0 & 3 & 3 \end{bmatrix}^T,$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 0 & 3 & -3 & 3 & 0 & -3 \end{bmatrix}^T, \text{ and}$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} 3 & 0 & 3 & -3 & -3 & 0 \end{bmatrix}^T,$$

respectively. The three class centers were located at the vertices of an equilateral triangle in the six-dimensional space with the overall mean vector as a zero vector. The objects in each class were assumed to follow a multivariate normal distribution and the covariance matrix for the three classes was the same, which was set as a 6 x 6 identity matrix. For each class, 100 objects were generated based on the class center and the covariance matrix. This created a 300 x 6 matrix containing the error-free data. The overall mean

vector of the data in the six-dimensional space is expected to be a zero vector, but generally it is not, due to the sampling variation. Thus, the data were mean-centered. This matrix was transformed into a 12-dimensional space by a 12 x 6 matrix (containing an orthonormal basis) obtained by applying SVD to a randomly generated matrix. This gave a 300 x 12 data matrix containing the error-free data. Note that the overall mean vector of the data in the 12-dimensional space is still a zero vector and the dimensionality of the data (pseudo-rank) is still six. The subspace in the 12-dimensional space, where the error-free data were located, was used as the true subspace.

To simulate the column intercepts for the 12 variables, the numbers of 1, 2, ...,12 with an increment of one were added to the 12 variables, respectively. This intercept vector was treated as the true mean vector of the error-free data.

Measurement errors were simulated by following the similar manner as for data set 1. The measurement errors were still assumed to be independent and heteroscedastic and measurement error standard deviations were assumed to follow a log-normal distribution. A 300 x 12 matrix was generated by drawing 300 x 12 numbers from a normal distribution $N(0,4)$ followed by exponential transformation (base e). As before, 0.1 was added to each of the elements to make them more realistic. This matrix contained the simulated measurement error standard deviations. Another 300 x 12 matrix was created by drawing 300 x 12 numbers from a normal distribution $N(0,1)$. The element-by-element products of the two 300 x 12 matrices resulted in the simulated errors. Adding the simulated errors to the error-free data yielded the observed data.

3.4.2.3 Group Data Set 2

For the purpose of validation of the proposed algorithm from the statistical perspective, a second group consisting of 100 data sets, referred to as group data set 2, was simulated. For each of the 100 data sets, a 6 x 3 matrix and a 3 x 5 matrix were first generated with the elements of the matrices drawn randomly from the standard normal distribution $N(0,1)$. The outer product of the two matrices gave a 6 x 5 matrix with the rank of 3. An 1 x 5 vector with its elements drawn randomly from the standard normal distribution $N(0,1)$ was generated to simulate the column intercepts. Adding the intercept vector to the 6 x 5 matrix resulted in the measurement error-free data matrix.

The measurement errors were assumed to be heteroscedastic. A 6 x 5 matrix was generated by randomly drawing numbers from a normal distribution $N(0, 0.04)$. Taking the absolute values of the elements of this matrix and then adding 0.01 to each of them gave a 6 x 5 matrix consisting of the simulated measurement error standard deviations. Another 6 x 5 matrix was generated with its elements randomly drawn from the standard normal distribution $N(0,1)$. The element-by-element products of the two matrices led to the heteroscedastic measurement errors. The observed data were simulated by adding the simulated measurement errors to the measurement error-free data. For the 100 data sets, the error-free data matrix was retained to be the same. The measurement errors were generated by following the same simulation parameters and procedure. For the 100 data sets, the measurement error variances were intentionally set to be relatively small because this reduces the number of local minima or even makes each data set have only one minimum (global minimum). Also, the degrees of freedom approximated by using the literature formula for the PCA and MLPCA residuals [1,26] are closer to the true ones in the cases of small measurement errors.

3.5 Results and Discussion

3.5.1 Data Set 1

The proposed algorithm was applied to process this data set with the number of principal components chosen to be unity. Twenty different initial guesses were tried and the solution that gave the largest likelihood was chosen. The simulated data are plotted in Figure 3.2 with solid upward- and down-pointing triangles to distinguish the objects from the two classes. The dashed lines show the x-axis and y-axis of the coordinate system. The mean point of the error-free data (true mean point) is indicated by P_0 (dot), the estimated mean point based on the proposed algorithm is represented by P_1 (circle), and the conventional mean point of the error-contaminated data (measured data) is denoted by P_2 (square). It can be seen that the estimated mean point is quite close to that of the error-free data, indicating the proposed algorithm has achieved its objective to estimate the mean vector of the error-free data. The conventional sample mean point is also close to that of the error-free data. This is not unanticipated because of relatively mild

heteroscedasticity, but it is expected that it may deviate much from the true mean vector in case of more significantly heteroscedastic errors.

The estimated subspace by the proposed algorithm, denoted by $\mathbf{v}_1$, is shown in Figure 3.2. For this data set, the dimensionality of the true hyperplane is one and its direction is 45° from the positive x-axis, which is denoted by $\mathbf{v}_0$ (solid line). $\mathbf{v}_1$ makes an angle of 44.4° with respect to the positive x-axis and is basically parallel to $\mathbf{v}_0$, which is a good agreement with the true hyperplane. For comparison purpose, the original algorithm for MLPCA model without intercepts was applied to this data set as well, but the data set was mean-centered based on the conventional sample mean vector. The estimated subspace is indicated by $\mathbf{v}_2$, which holds an angle of 44.3° with respect to the positive x-axis and basically overlaps with $\mathbf{v}_1$. However, it is expected that the proposed algorithm will give a better estimate when the error heteroscedasticity increases. The agreement of the estimated mean vector and direction of the subspace with the true values show that the proposed algorithm was successful in the optimization of the objective function of the MLPCA model with intercepts.

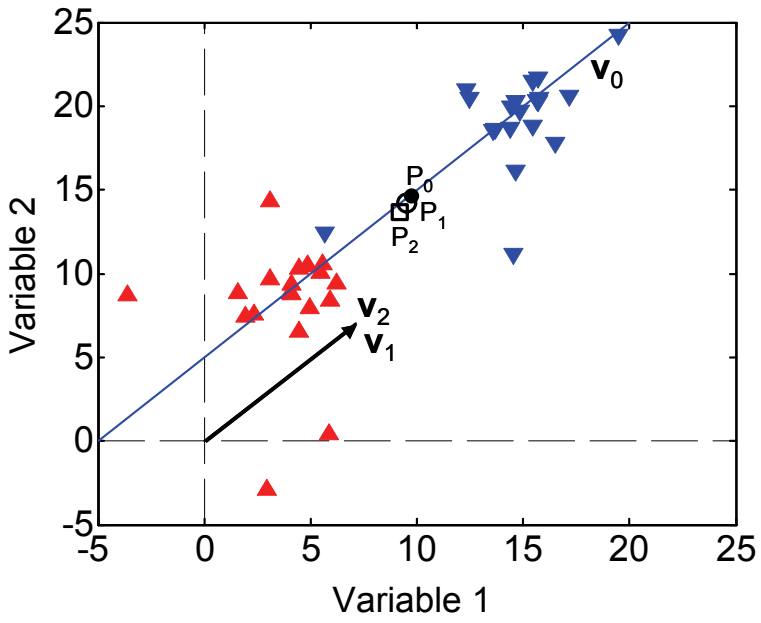

Figure 3.2 Plots of the two-dimensional simulated data and the results of the proposed algorithm and the original algorithm for MLPCA model without intercepts.

3.5.2 Group Data Set 1

The proposed algorithm was applied to each of the 100 data sets with the number of principal components chosen as 6 and 50 different initial guesses to search for the best solution. As the numbers of objects and variables were increased compared with those of data set 1, the number of local optima might increase, so the number of initial guesses was increased. For experimental data, the pseudo-rank is generally unknown, but for this simulation study, the number of principal components was known and chosen as the pseudo-rank to focus only on the evaluation of the algorithm.

One way to evaluate the proposed algorithm is to examine if it can effectively estimate the true mean vector of the measurement error-free data, which are the intercepts in this simulation study. For each data set, the estimated mean vector was compared with the intercepts added to the error-free data (true mean vector) and the sum of squares of the differences between the two vectors' elements was used as a measure for their matching status. Mathematically, this can be expressed as

$$SS_{mean} = \left(\mathbf{c}_{true} - \mathbf{c}\right)^{\mathrm{T}} \left(\mathbf{c}_{true} - \mathbf{c}\right), \tag{3.19}$$

where SS_{mean} denotes the sum of squares of the differences, $\mathbf{c}_{true}$ represents the true mean vector, and $\mathbf{c}$ is the estimated mean vector. This is an analogue of the sum of the squares of the residuals in regression analysis [27]. SS_{mean} is dependent on the data structure and error structure and thus its value only is not a good measure. Therefore, the conventional mean vector of the error-contaminated data was calculated as well and was compared with the true mean vector. The sum of squares of the differences between the two vectors' elements was also obtained by using Equation (3.19) except that $\mathbf{c}$ is replaced by the conventional sample mean vector. If the proposed algorithm performs well, the estimated mean vector based on it should give a smaller SS_{mean} value than the conventional mean vector.

Figure 3.3 show the logarithms (base 10) of the sums of squares of the differences between the estimated mean vectors by the proposed algorithm and the true mean vectors for the 100 data sets, indicated as solid triangles. The sums of squares of the differences between the conventional mean vectors and true mean vectors are indicated by dots. It can be seen that the estimated mean vectors by the proposed algorithm give smaller

values than the conventional mean vectors. The results show that the proposed algorithm is effective in estimating the true mean vectors of error-free data in high-dimensional space.

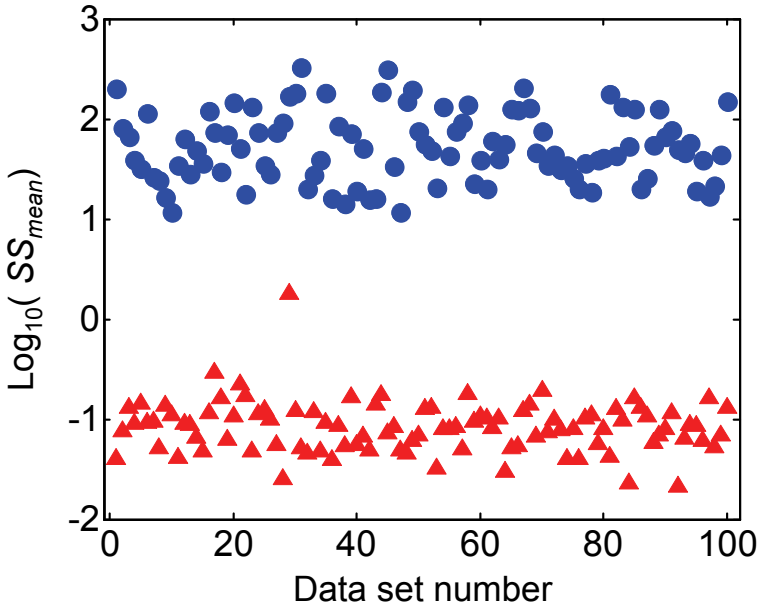

Figure 3.3 Plots of the logarithms of the sums of squares of the differences between the true and estimated mean vectors by two different algorithms. Solid triangles: the proposed algorithm. Dots: the original algorithm for MLPCA model without intercepts.

Another perspective to evaluate the proposed algorithm may be to explore the subspaces estimated by the proposed algorithm. To do that, the angles between the estimated subspaces and the subspaces for the error-free data (true subspaces) were used to evaluate the quality of the estimated subspaces obtained by the proposed algorithm. A small angle implies a good estimate of the true subspace. The method to calculate the angle between two subspaces is available in the literature [28,29,30] and has been employed in Chapter 2 (Equation (2.18)). Again, the angles are dependent on the structures of the data and errors. Thus, the angles between the true subspaces and the subspaces estimated by the original algorithm for MLPCA model without intercepts were calculated as well. For the original algorithm that assumes zero intercepts, the data were mean-centered based on the conventional sample mean vector.

Figure 3.4 shows the subspace angles calculated as above. As expected, the proposed algorithm gives smaller subspace angles for most of the data sets, demonstrating its utility in extracting more accurate subspaces. It is not surprising that in

some cases, good estimates of the subspaces were obtained by performing mean centering using the conventional mean vectors; however, the overall better estimation based on the proposed algorithm indicates it has performed successfully.

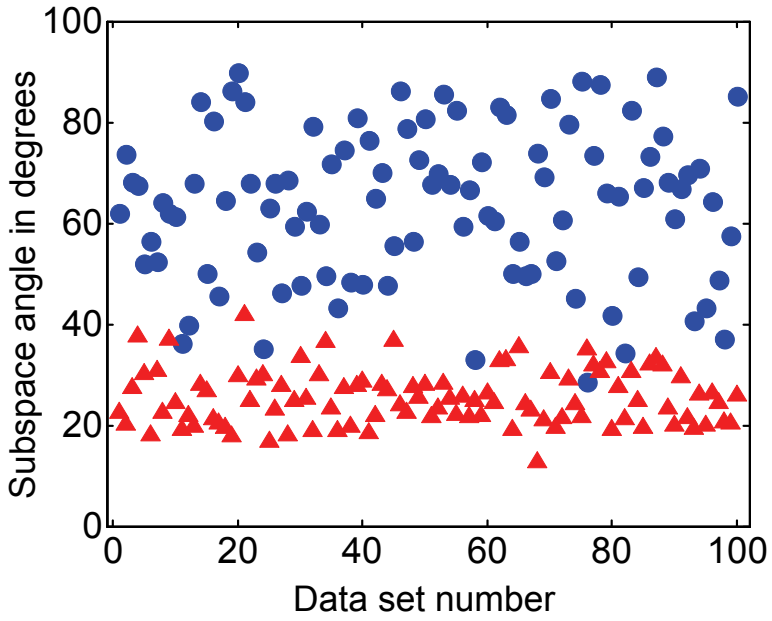

Figure 3.4 Plots of the angles (in degrees) between the true subspaces and the subspaces estimated by two different algorithms. Solid triangles: the proposed algorithm. Dots: the original algorithm for MLPCA model without intercepts.

The performance of the proposed algorithm can also be evaluated by viewing the separation of clusters from different classes in the scores plots. As it is not practical to show all the scores plots for the 100 data sets, the cluster separation in a two-dimensional plane based on the first two scores was measured by the generalized Fisher's discriminant value. A larger generalized Fisher's discriminant value means a better separation of the clusters. The method to calculate this value is described in the literature [31], and has been used in Chapter 2 (Equations (2.15) to (2.17)). The generalized Fisher's discriminant values were calculated for the scores obtained by the proposed algorithm and the original algorithm for MLPCA without intercepts. For the latter, the data were mean-centered based on the conventional mean vector, which has been mentioned in the calculation of subspace angles.

Figure 3.5 shows the generalized Fisher's discriminant values for the 100 data sets. It can be seen that the generalized Fisher's discriminant values obtained by the

proposed algorithm are overall larger. The improved separation of clusters indicates that the proposed algorithm has optimized the objective function effectively.

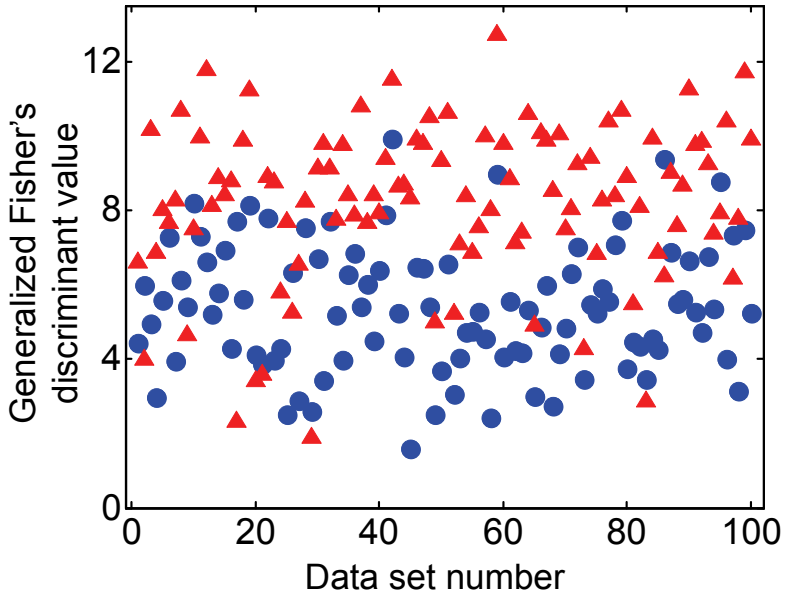

Figure 3.5 Plots of the generalized Fisher's discriminant values as a measure of the cluster separation resulting from two different algorithms. Solid triangles: the proposed algorithm. Dots: the original algorithm for MLPCA model without intercepts.

As an example, the scores plots for the first data set in group data set 1 obtained by the two methods as above are shown in Figures 3.6 (a) and (b), respectively. In Figure 3.6 (a), the three clusters are roughly located at the vertices of an equilateral triangle, which is in accordance with the three simulated class centers. In Figure 3.6 (b), however, the relative location of the three clusters is changed. Two clusters (red dots and blue solid triangles) are closer and they are far from the third cluster. This might be because the conventional mean vector is not a good estimate of the mean vector of the error-free data. Subtracting the conventional mean vector does not effectively remove the intercepts and thus the subspace is not optimally estimated. The mean value for the first principal component is far from zero, but this is not found in the scores plot obtained by using the proposed algorithm. These results give further support that the proposed optimization algorithm for MLPCA model with intercepts has performed successfully.

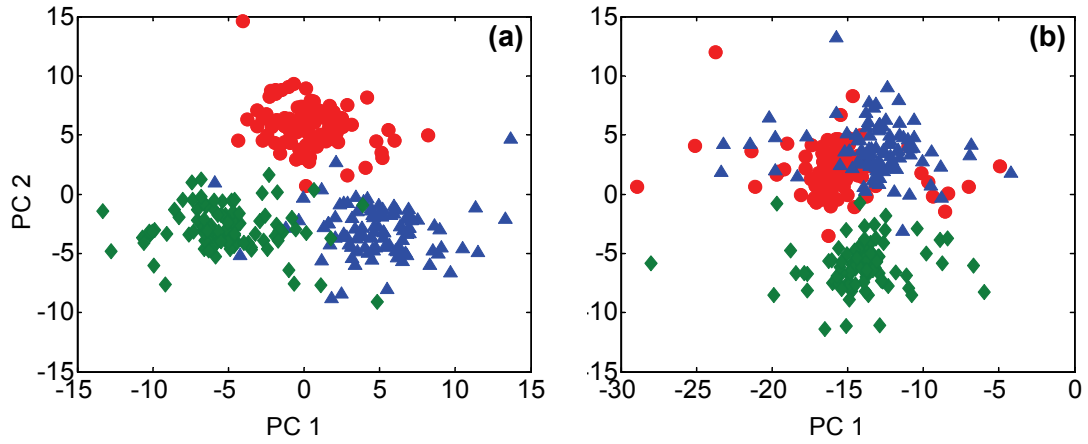

Figure 3.6 Scores plots of MLPCA obtained from two methods. (a) The proposed algorithm. (b) The original algorithm for MLPCA model without intercepts, for which the data were mean-centered by the conventional mean vector. Some points for (a) and (b) are outside the display region.

3.5.3 Group Data Set 2

To further test if the proposed optimization algorithm achieves its objective, a validation from statistical perspective was performed using group data set 2. It is known that the sum of squares of weighted residuals of MLPCA should follow a χ^2-distribution with the appropriate degrees of freedom if the error-free data have been estimated by correctly maximizing the likelihood function. The sum of squares of the weighted residuals (S^2) can be calculated by

$$S^2 = \sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{x}_i - \mathbf{c} - \mathbf{V}\mathbf{t}_i\right), \tag{3.20}$$

where $\mathbf{x}_i$ denotes the observed measurement vector for each sample i, $\mathbf{c}$ is the estimated mean vector, $\mathbf{t}_i$ represents the estimated scores, and $\mathbf{V}$ designates the estimated loadings matrix. The measurement error covariance matrix $\boldsymbol{\Sigma}_i$ for each sample is known in this case.

When the measurement errors are small, the degrees of freedom for PCA with mean centering and MLPCA with column intercepts can be approximated by $(n-d-1)(p-d)$ [1,26], where n is the number of samples, p denotes the number of observed variables, and d represents the dimension of the subspace. If the measurement errors are large, the degrees of freedom calculated using this formula will not be accurate enough and thus affects the choice of an appropriate χ^2-distribution.

For each of the 100 data sets in group data set 2, the proposed MLPCA optimization algorithm was employed with 10 random initial guesses and with the dimensionality of the subspace chosen as 3. Based on different initial guesses, the solution giving the lowest S^2 value for each data set was selected. Examination of the S^2 values from different initial guesses for several data sets did not show multiple local minima, indicating 10 random initial guesses should be sufficient to find the global solution. To examine if the obtained S^2 values follow a χ^2- distribution, the probability-probability plot (P-P plot) was used. The obtained S^2 values were sorted and then the cumulative probability was calculated based on the S^2 values and 4 degrees of freedom (= (6-3-1) x (5-3)). The theoretical cumulative probability of the χ^2- distribution with 4 degrees of freedom was also calculated. For comparison purpose, the S^2 values for the scores and loadings obtained by applying PCA to the 100 data sets (mean-centered by the conventional mean vectors) were calculated as well. Equation (3.20) was still used, but $\mathbf{c}$ was replaced by the conventional mean vector, and $\mathbf{t}_i$ and $\mathbf{V}$ were estimated by PCA instead of MLPCA.

The P-P plots for MLPCA and PCA results are shown in Figure 3.7. It can be seen that, for MLPCA, the plot largely falls on the straight line with a slope of unity, indicating the calculated S^2 values have no deviation from a χ^2- distribution. Theoretically, if the proposed algorithm fails to find the solution corresponding to the maximum likelihood, the S^2 value will be larger than it should be and the calculated cumulative probability will be larger. When the cumulative probability calculated based on the observed data is plotted against the theoretical cumulative probability, the P-P plot will lie above the straight line with a slope of unity. The P-P plot for the PCA result in Figure 3.7 shows this phenomenon. It deviates heavily from the straight line, showing the S^2 values calculated using the PCA scores and loadings do not follow a χ^2- distribution with the appropriate degrees of freedom. Although the P-P plot for the result from the proposed MLPCA optimization algorithm is not a direct mathematical proof, it is an indication that the proposed algorithm has given maximum likelihood estimates of the error-free data and achieved its objective.

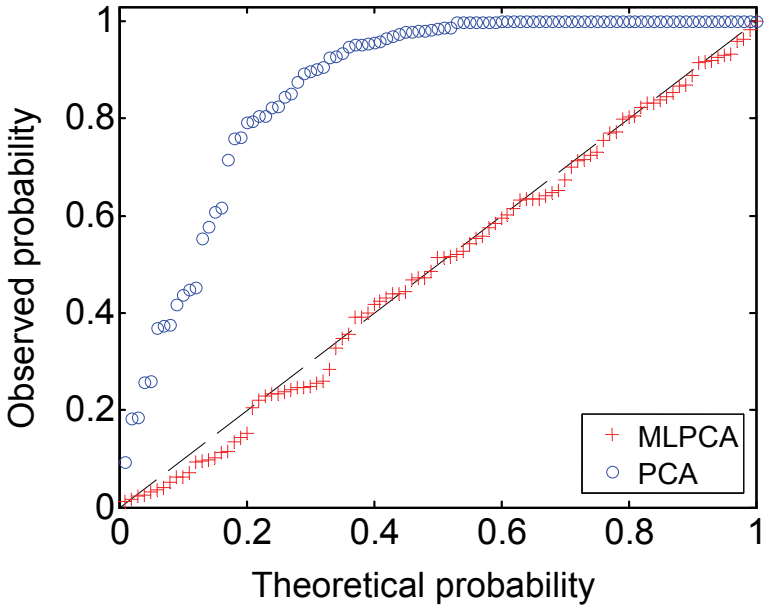

Figure 3.7 P-P plots of S^2 values for group data set 2, obtained by the proposed MLPCA optimization algorithm and PCA.

3.5.4 Convergence Speed Issue

Convergence speed is another important aspect for an optimization algorithm. The utility of an optimization algorithm is highly dependent on how quickly it can give a solution. The gradient descent method, for example, is not very useful in practice because of its slow convergence speed, although it is theoretically important. When an objective function is not convex and an algorithm cannot guarantee the global optimum of the objective function, many different initial guesses are required to increase the probability of finding the global optimum. Thus, an optimization algorithm with fast convergence speed is desired.

The proposed optimization algorithm converges quickly overall, but since multiple initial guesses are needed to increase the likelihood of finding the global optimum, it is good to gain some insight into its speed. Certainly, the convergence speed is dependent on the numbers of objects and variables of the data sets, and the structures of the data and errors, but the computation time for the data sets used in work can give an indication of the speed of the proposed algorithm. For a typical data set in group data set 1, the time to perform 50 random initial guesses was about 20 min, which was carried out in MatLab® v.7.4.0 by a computer with 1.8 GHz of CPU speed and 3 Gb of memory. This

gives an average time of 24 seconds for a single initial guess. This time may sound a little long, but for the purpose of processing of most experimental data in chemistry, it should not be a big problem because it is not critical for chemists to wait for one or two hours to get the results in contrast with the time spent to carry out the experiments that may take several weeks or months.

3.6 Conclusions

A data analysis method generally consists of two major components: the objective function and the optimization algorithm. Both components are important for a method to be successfully used to extract useful information from data. MLPCA is a technique that has been developed to deal with bilinear data that have significant heteroscedastic and/or correlated errors and has been applied to different scenarios. It has a well-defined objective function and several refined variants to deal with different data and error structures. One situation that can arise is that the underlying data have non-zero intercepts for different variables and the conventional sample mean vector is a poor estimate of the mean vector for the error-free data due to significant measurement errors. In the original MLPCA work, it was proposed that the intercept term be included and estimated together with the subspace and scores, but this was not accomplished because of the lack of efficient optimization algorithms.

The new optimization algorithm presented in this chapter has been developed to overcome the difficulty in optimizing the objective function of MLPCA model with intercepts. The algorithm is theoretically simple and is essentially a generalization of the original optimization algorithm for MLPCA model without intercepts. The simulation study shows the proposed optimization algorithm performs well. Although the simulation study was centered on the application of MLPCA for exploratory data analysis, this is not a requirement for the proposed optimization algorithm. If MLPCA is used for other purposes such as multivariate calibration, the proposed algorithm is still applicable. Like many other optimization algorithms, it cannot guarantee that the global optimum will be found and therefore there is a requirement to start from multiple initial guesses to increase the probability of finding the global optimum, but the relatively fast convergence speed makes this possible. The performance of this optimization algorithm in processing the simulated data sets has demonstrated it is effective and efficient. It is hoped that the

development of this algorithm can make the MLPCA method more useful in multivariate data analysis.

3.7 Appendix

3.7.1 Optimization Algorithm for MLPCA Model with Column and Row Intercepts

Starting with the general MLPCA model in Equation (3.1)

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{1}_n \mathbf{c}^{\mathrm{T}} + \mathbf{r}\mathbf{1}_p^{\mathrm{T}} + \mathbf{E}, \tag{3.21}$$

and following the notations in Section 3.3 (assume measurement errors are independent across measurements in the matrix), the likelihood function for the observed data can be expressed as

$$L = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{n}{2}} \left|\mathbf{\Sigma}_i\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\mathbf{x}_i - \mathbf{c} - r_i \mathbf{1}_p - \mathbf{V}\mathbf{t}_i \right)^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - \mathbf{c} - r_i \mathbf{1}_p - \mathbf{V}\mathbf{t}_i \right) \right], \text{ or} \tag{3.22}$$

$$L = \prod_{j=1}^{p} \frac{1}{(2\pi)^{\frac{p}{2}} \left|\mathbf{\Psi}_j\right|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{r} - \mathbf{T}\underline{\mathbf{v}}_j \right)^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{r} - \mathbf{T}\underline{\mathbf{v}}_j \right) \right]. \tag{3.23}$$

Note that $\mathbf{\Sigma}_i$ and $\mathbf{\Psi}_j$ are diagonal. The log-likelihood function ($\ln L$) can be written as

$$\ln L = \text{constant} - \frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{x}_i - \mathbf{c} - r_i \mathbf{1}_p - \mathbf{V}\mathbf{t}_i \right)^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - \mathbf{c} - r_i \mathbf{1}_p - \mathbf{V}\mathbf{t}_i \right), \text{ or} \tag{3.24}$$

$$\ln L = \text{constant} - \frac{1}{2} \sum_{j=1}^{p} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{r} - \mathbf{T}\underline{\mathbf{v}}_j \right)^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{r} - \mathbf{T}\underline{\mathbf{v}}_j \right). \tag{3.25}$$

Based on Equation (3.24), the partial derivatives of $\ln L$ with respect to $\mathbf{t}_i$ can be expressed as

$$\frac{\partial \ln L}{\partial \mathbf{t}_i} = -\mathbf{V}^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - \mathbf{c} - r_i \mathbf{1}_p \right) + \left(\mathbf{V}^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \mathbf{V} \right) \mathbf{t}_i. \tag{3.26}$$

Setting this to zero followed by rearrangement leads to

$$\mathbf{t}_i = \left(\mathbf{V}^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \mathbf{V} \right)^{-1} \mathbf{V}^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - \mathbf{c} - r_i \mathbf{1}_p \right). \tag{3.27}$$

Also, based on Equation (3.24), the partial derivatives of $\ln L$ with respect to $\mathbf{c}$ can be written as

$$\frac{\partial \ln L}{\partial \mathbf{c}} = -\sum_{i=1}^{n} \left[\mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - r_i \mathbf{1}_p - \mathbf{V}\mathbf{t}_i \right) \right] + \sum_{i=1}^{n} \left(\mathbf{\Sigma}_i^{-1} \mathbf{c} \right). \tag{3.28}$$

Setting this to zero yields

$$\mathbf{c} = \left(\sum_{i=1}^{n} \mathbf{\Sigma}_i^{-1} \right)^{-1} \left\{ \sum_{i=1}^{n} \left[\mathbf{\Sigma}_i^{-1} \left(\mathbf{x}_i - r_i \mathbf{1}_p - \mathbf{V}\mathbf{t}_i \right) \right] \right\}. \tag{3.29}$$

Similarly, based on Equation (3.25), the partial derivatives of $\ln L$ with respect to $\underline{\mathbf{v}}_j$ can be obtained as

$$\frac{\partial \ln L}{\partial \underline{\mathbf{v}}_j} = -\mathbf{T}^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{r} \right) + \left(\mathbf{T}^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \mathbf{T} \right) \underline{\mathbf{v}}_j . \tag{3.30}$$

Setting this to zero gives

$$\underline{\mathbf{v}}_j = \left(\mathbf{T}^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \mathbf{T} \right)^{-1} \mathbf{T}^{\mathrm{T}} \mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{r} \right). \tag{3.31}$$

The partial derivatives of $\ln L$ with respect to **r**, based on Equation (3.25), is

$$\frac{\partial \ln L}{\partial \mathbf{r}} = -\sum_{j=1}^{p} \left[\mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{T}\underline{\mathbf{v}}_j \right) \right] + \sum_{j=1}^{p} \left(\mathbf{\Psi}_j^{-1} \mathbf{r} \right). \tag{3.32}$$

Setting this to zero gives results in

$$\mathbf{r} = \left(\sum_{j=1}^{p} \mathbf{\Psi}_j^{-1} \right)^{-1} \left\{ \sum_{j=1}^{p} \left[\mathbf{\Psi}_j^{-1} \left(\underline{\mathbf{x}}_j - c_j \mathbf{1}_n - \mathbf{T}\underline{\mathbf{v}}_j \right) \right] \right\}. \tag{3.33}$$

Equations (3.27), (3.29), (3.31), and (3.33) give an iterative method to optimize the likelihood function of MLPCA model with both column and row intercepts.

It has been found in this study that when both intercept terms are included in the MLPCA model, the likelihood function seems to become a convex function. This means that there is no need for the optimization algorithm to start from multiple different initial guesses and the global minimum can be guaranteed from any initial guess. This is interesting, but has not been examined thoroughly. As mentioned earlier, the ambiguities in rotation, translation and scale for the terms **T**, **V**, **c** and **r** still exist, although the estimated error-free data expressed in the original space $\mathbf{X}_{est} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{1}_n \mathbf{c}^{\mathrm{T}} + \mathbf{r}\mathbf{1}_p^{\mathrm{T}}$ remains the same. Depending on the purposes of different applications, the estimated data $\mathbf{X}_{est}$ can be decomposed in different manners, but this will not be discussed here.

3.8 Bibliography

1. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, Maximum Likelihood Principal Component Analysis, *Journal of Chemometrics*, 11 (1997) 339-366.
2. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.
3. P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Maximum Likelihood Multivariate Calibration, *Analytical Chemistry,* 69 (1997) 2299-2311.
4. D. T. Andrews, and P. D. Wentzell, Applications of Maximum Likelihood Principal Component Analysis: Incomplete Data Sets and Calibration Transfer, *Analytica Chimica Acta*, 350 (1997) 341-352.
5. S. K. Schreyer, M. Bidinosti, and P. D. Wentzell, Application of Maximum Likelihood Principal Components Regression to Fluorescence Emission Spectra, *Applied Spectroscopy,* 56 (2002) 789-796.
6. L. Vega-Montoto, and P. D. Wentzell, Maximum Likelihood Parallel Factor Analysis, *Journal of Chemometrics*, 17 (2003) 237-253.
7. L. Vega-Montoto and P. D. Wentzell, Approaching the Direct Exponential Curve Resolution Algorithm from a Maximum Likelihood Perspective, *Analytica Chimica Acta*, 556 (2006) 383-399.
8. P. D. Wentzell, T. K. Karakach, S. Roy, M. J. Martinez, C. P. Allen, and M. Werner-Washburne, Multivariate Curve Resolution of Time Course Microarray Data, *BMC Bioinformatics*, 7 (2006) 343.
9. R. Tauler, M. Viana, X. Querol, A. Alastuey, R. M. Flight, P. D. Wentzell, and P. K. Hopke, Comparison of the Results Obtained by Four Receptor Modeling Methods in Aerosol Source Apportionment Studies, *Atmospheric Environment*, 43 (2009) 3989-3997.
10. P. D. Wentzell, and S. Hou, Exploratory Data Analysis with Noisy Measurements, *Journal of Chemometrics,* 26 (2012) 264-281.
11. R. J. Pell, M. B. Seasholtz, and B. R. Kowalski, The Relationship of Closure, Mean Centering and Matrix Rank Interpretation, *Journal of Chemometrics*, 6 (1992) 57-62.
12. M. B. Seasholtz, and B. R. Kowalski, The Effect of Mean Centering on Prediction in Multivariate Calibration, *Journal of Chemometrics*, 6 (1992) 103-111.
13. T. Iwata, and J. Koshoubu, Pretreatment of Spectral Data in PLS1 Quantitative Analysis, *Bunseki Kagaku*, 45 (1996) 85-89.
14. A. Lorber, K. Faber, and B. R. Kowalski, Local Centering in Multivariate Calibration, *Journal of Chemometrics*, 10 (1996) 215-220.

15. N. M. Faber, Mean Centering and Computation of Scalar Net Analyte Signal in Multivariate Calibration, *Journal of Chemometrics*, 12 (1998) 405-409.

16. T. J. Thurston, R. G. Brereton, D. J. Foord, and R. E. A. Escott, Principal Components Plots for Exploratory Investigation of Reactions Using Ultraviolet-Visible Spectroscopy: Application to the Formation of Benzophenone Phenylhydrazone, *Talanta*, 63 (2004) 757-769.

17. J. H. Kalivas, Learning from Procrustes Analysis to Improve Multivariate Calibration, *Journal of Chemometrics*, 22 (2008) 227-234.

18. D. Poole, *Linear Algebra: A Modern Introduction*, Brooks/Cole, 2003.

19. R. I. Jennrich, and S. M. Robinson, A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 111-123.

20. K. G. Jöreskog, A General Approach to Confirmatory Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 183-202.

21. R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, Inc., New Jersey, 1998.

22. A. C. Rencher, *Methods of Multivariate Analysis*, John Wiley & Sons, Inc., New York, 1995.

23. J. Stewart, *Calculus: Early Transcendentals*, Fourth Edition, Brooks/Cole Publishing Company, New York, 1999.

24. J. R. Magnus, and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Inc., New York, 1988, pp. 177.

25. K. B. Petersen, and M. S. Pedersen, *The Matrix Cookbook*, Version: Nov. 14, 2008, http://matrixcookbook.com. Last access on Oct. 1, 2011.

26. N. M. Faber, Degrees of Freedom for the Residuals of Principal Component Analysis – A Clarification, *Chemometrics and Intelligent Laboratory System*, 93 (2008) 80-86.

27. M. A. Golberg, and H. A. Cho, *Introduction to Regression Analysis*, WIT Press, 2004.

28. S. Jiang, Angles between Euclidean Subspace, *Geometriae Dedicata*, 63 (1996) 113-121.

29. J. Miao, and A. Ben-Israel, On Principal Angles between Subspace in R^n, *Linear Algebra and Its Applications*, 171 (1992) 81-98.

30. A. Björck, and G. H. Golub, Numerical Methods for Computing Angles between Linear Subspaces, *Mathematics of Computation*, 27 (1973) 579-594.

31. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.

Chapter 4: Development of Quasi-Power Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index*

4.1 Introduction

Exploratory data analysis and classification methods have always been important tools of multivariate data analysis in chemistry. The application of these methods has expanded in recent years due to, among other things, an increased emphasis on high throughput biological analysis, where researchers are often interested in differentiating among different biological states of organisms. PCA has dominated as a method to visualize high-dimensional data in lower-dimensional spaces, but suffers from the drawback that it is based on maximizing the variance along the projection vectors, which is not always the best way to separate classes. This problem can be circumvented through the use of projection pursuit (PP) analysis, which uses different criteria to identify projection vectors. While there are examples of the application of this technique to chemistry [1,2,3], it is not nearly as widely applied as PCA, probably because the algorithms are fairly complex and not readily accessible in many standard packages.

The term "projection pursuit" was firstly coined by Friedman and Tukey [4], but the concept of PP can be tracked back to the work of Kruskal [5,6] who proposed the term "index of condensation". PP generally refers to an unsupervised technique for exploratory data analysis, but some researchers have used this term for discriminant analysis [7]. The primary purpose of PP is to look for "interesting" projections in a low-dimensional subspace that can reveal the natural structure of the data. The notion of "interestingness" may have different interpretations in different applications, but in the present context, interesting projections are those where the data projected in the low-dimensional space can reveal clusters or outliers.

Because the description of PP does not unambiguously define how to determine what is interesting, any linear projection method, including PCA, could be regarded as a special case of PP. An objective function that characterizes the "interestingness" is called a

* This chapter is based on the published article: S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.

"projection index". In the literature, various projection indices have been developed, leading to many PP variants. The original projection index was proposed by Friedman and Tukey [4], but this was followed by proposals for other projection indices in the literature [8,9,10,11,12,13,14,15]. Most of the projection indices are designed to measure the non-normality of a distribution. Deviations from normality in the projected data are considered interesting because, for multivariate data, the observed variables are often the linear combinations of a small number of latent variables. By the central limit theorem, even if the latent variables reveal important elements of data structure, such as clusters or outliers, the observed variables often cannot directly disclose meaningful information because they tend towards normality. The latent variables that reveal useful information deviate from a normal distribution, so projections that deviate strongly from normality may uncover this structure.

In theory, any function that relates directly to the normality of a distribution can be used as a projection index, but a good index should be a simple measure and easy to optimize. Several functions have been used, with entropy and kurtosis being the most familiar. Kurtosis was one of the early functions proposed [8] and has the advantage of conceptual simplicity. Peña and Prieto showed that maximization of the kurtosis can be used to detect outliers [16], although this is not always effective and other methods may be preferred [17]. On the other hand, projections with bimodality tend to have a small kurtosis, and minimization of kurtosis can therefore be used as a criterion to search for clusters [15]. Kurtosis is also used to measure the non-normality in independent component analysis (ICA) [18,19], which is a technique closely related to PP. In univariate statistics, a normal distribution has a kurtosis of 3. A super-gaussian (peaked, or leptokurtic) distribution has a larger kurtosis, while a sub-gausssian (flat, or platykurtic) distribution has a smaller kurtosis. Either maximization or minimization of kurtosis can give useful information. Kurtosis satisfies the condition of the Class III objective functions set by Huber [8] for good projection indices; that is, scaling and translation do not change the values of the functions. One more appealing property of kurtosis is that the univariate case can be easily generalized to multivariate kurtosis, which not only has the useful properties of univariate kurtosis, but is independent of the choice of the basis for a subspace. Therefore, kurtosis appears to be an ideal statistic for the projection index.

The projection index acts as the heart of PP, but its utility is mostly dependent on computational aspects. Optimization of the projection index, which greatly determines whether a projection index is successful, plays a crucial role in PP. Because of the quartic nature of kurtosis, optimization is a difficult problem. Kurtosis can have multiple local maxima and minima, and commonly used optimization algorithms cannot guarantee the global extrema. Therefore, it is generally necessary to start from different initial guesses to search for the global optimum, or better local optima, and so the speed of an optimization algorithm is critical. Gradient descent or ascent methods are ubiquitous in optimization problems, but gradient methods have the well-known shortcoming of slow convergence rates and the choice of optimal step size is difficult. Gradient methods have been used for the optimization of kurtosis [20,21], but other algorithms have also been developed in the literature. Peña and Prieto [15] proposed iterative methods for optimization of kurtosis by applying a modified Newton's method, which is complicated, or by solving first-order optimality conditions, similar to one method proposed in this work (relationship to the algorithms proposed in this work is noted in Section 4.8.4). Croux's algorithm [22] has also been used for the optimization of kurtosis as a projection index [3]. This algorithm calculates the objective function for many projections based on the sample space and works well when the number of variables is relatively small, but will perform poorly if the dimensionality of the data becomes too high. Hyvärinen and Oja proposed a fast fixed-point algorithm to optimize the kurtosis [19,23] based on sphered data (relationship to the algorithms proposed in this work is noted in Section 4.8.5). Sphering, which differs from autoscaling, is a transformation that ensures the data have unit variance when projected in any direction [24]. This algorithm is one of the most widely used because of its fast convergence. It has several variants [25,26,27,28,29] and can be viewed to be a continuum between gradient methods and Newton's method. As with other such methods, the determination of the optimal step size for the fixed-point algorithm is computationally involved, but this has been described [27,28].

In the present work, new algorithms, referred to as "quasi-power methods", to optimize kurtosis are proposed. The algorithms use the well-known conclusion in calculus that if all the partial derivatives are zeros at a point, the point may be a maximum or a minimum. By setting all the derivatives of kurtosis to be zeros followed by rearrangements,

equations emerge that allow the principle of the power method and its variants (used to solve eigenvalue problems) to be employed. Because the algorithms are developed from the perspective of the power method instead of gradient methods, they are simple, fast, and stable. Commonly required preprocessing steps, such as sphering or whitening of the data, are not necessary. The algorithms can search for maxima or minima according to user's requirements, without the need to optimize step size, and they can be used for both univariate and multivariate kurtosis with little modification.

4.2 Theory

4.2.1 Univariate Kurtosis

For univariate data, the sample kurtosis (K) is defined as

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2\right)^2}, \tag{4.1}$$

where n is the number of samples, z_i is the individual sample value, and $\bar{z}$ is the sample mean. The numerator is the fourth central moment and denominator is the square of the second central moment or the biased sample variance (as opposed to the unbiased variance which has n-1 degree of freedom). For the purpose of optimization, the offset of "-3" that is included in some definitions of kurtosis to give the normal distribution a kurtosis of zero is not included. The current definition ensures that the kurtosis is always positive.

For multivariate data, if there are n samples measured on p variables, the entire data can be arranged in an $n \times p$ matrix:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_n^{\mathrm{T}} \end{bmatrix}. \tag{4.2}$$

Each column of $\mathbf{X}$ represents a set of samples measured on a single variable and each row contains the measurements on different variables for a single sample, denoted by the notation $\mathbf{x}_i^{\mathrm{T}}$, where the subscript "$i$" is the sample index. In the following, the data matrix $\mathbf{X}$ is assumed to have been column mean-centered to simplify the derivation. PP tries to

search for a unit length projection vector $\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_p \end{bmatrix}^{\mathrm{T}}$ such that, when the p-dimensional data $\mathbf{X}$ are projected onto this projection vector, the kurtosis of the projected data reaches a maximum or a minimum. If a projected data point is denoted by z_i with $z_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{v}$ and $\bar{z} = 0$, the kurtosis defined in Equation (4.1) of the projected data can be written as

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2\right]^2}. \tag{4.3}$$

Rearrangement of this equation yields

$$K = \frac{n\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^2}. \tag{4.4}$$

To find a projection vector $\mathbf{v}$ that maximizes or minimizes the kurtosis, a simple method from calculus is to set all the first-order partial derivatives of kurtosis with respect to the variables in $\mathbf{v}$ to be zeros. The application of standard vector calculus [30,31,32] leads to the following equations (see Section 4.8.1 in Appendix for detailed derivation):

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v} = \rho\mathbf{v}, \text{ and} \tag{4.5}$$

$$\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v} = \frac{1}{\rho}\mathbf{v}, \tag{4.6}$$

where

$$\rho = \frac{\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2}{\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}}. \tag{4.7}$$

The expressions in Equations (4.5) and (4.6) are similar in form to an eigenvalue problem given by $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, although an important difference in this case is that multipliers on the left and right sides are both functions of $\mathbf{v}$. Nevertheless, this suggests that the solution may be obtained through an iterative procedure, or learning algorithm, that is embodied in the following equations:

$$\mathbf{v}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v}_k \text{, and} \tag{4.8}$$

$$\mathbf{v}_{k+1} \leftarrow \left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}_k \text{.} \tag{4.9}$$

Here the symbol "$\leftarrow$" means that the right-hand terms are calculated and assigned to the left-hand side at each iteration, and k stands for the iteration number. In Equation (4.4), the projection vector $\mathbf{v}$ is not constrained to unit length. However, in the learning algorithms, the projection vector $\mathbf{v}_{k+1}$ is generally normalized to unit length in each iteration. The purpose of normalization is not to impose a constraint on the optimization problem, but to make the test for convergence more straightforward. Given an initial guess $\mathbf{v}_1$, the learning algorithm in Equation (4.8) converges to a kurtosis maximum, while Equation (4.9) leads to a kurtosis minimum. It can be seen that the matrix in Equation (4.8) is the inverse of the matrix in Equation (4.9), and the two iterative learning algorithms are analogues of the power method and the inverse power method that are discussed in many linear algebra textbooks [33]. It can be shown (see Section 4.8.2 in Appendix) that the iterative algorithms defined by Equations (4.8) and (4.9) are convergent on maxima and minima, respectively, of the kurtosis, but it is important to note that, as with any algorithm based on kurtosis, these are not guaranteed to be globally optimum values.

4.2.2 Shifted Algorithms

It is known that one of the variants of the power method is the shifted power method [33]. Similar variants for the proposed algorithms can be developed by the same principle. If there is a projection vector $\mathbf{v}$ that satisfies Equations (4.5) and (4.6), the following equations also hold for any scalar c:

$$\left\{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]+c\mathbf{I}\right\}\mathbf{v} = \left(\rho + c\right)\mathbf{v} \text{, and} \tag{4.10}$$

$$\left\{\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)+c\mathbf{I}\right\}\mathbf{v} = \left(\frac{1}{\rho}+c\right)\mathbf{v} \text{,} \tag{4.11}$$

where $\mathbf{I}$ is the identity matrix. Introduction of scalar c does not change the solutions of $\mathbf{v}$, *i.e.*, the direction of the projection vector $\mathbf{v}$ does not change. The two equations indicate

that the learning algorithms defined by Equations (4.8) and (4.9) can have the following variants, respectively:

$$\mathbf{v}_{k+1} \leftarrow \left\{ \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right)^{-1} \left[\sum_{i=1}^{n} \left(\mathbf{x}_i^{\mathrm{T}} \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}} \mathbf{x}_i \right) \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \right] + c\mathbf{I} \right\} \mathbf{v}_k \text{, and} \tag{4.12}$$

$$\mathbf{v}_{k+1} \leftarrow \left\{ \left[\sum_{i=1}^{n} \left(\mathbf{x}_i^{\mathrm{T}} \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}} \mathbf{x}_i \right) \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \right]^{-1} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right) + c\mathbf{I} \right\} \mathbf{v}_k \,. \tag{4.13}$$

In principle, the scalar c can be any number, but a positive number is recommended for the optimization (see Section 4.8.2 in Appendix for more details). As the matrices in both Equations (4.8) and (4.9) have only positive eigenvalues, introduction of a positive number c will not change the relative sequence of the magnitudes of the eigenvalues, but introduction of a negative number c may alter this order. As the learning algorithm tends to converge to the dominant eigenvector corresponding to the largest eigenvalue in magnitude, introduction of negative number c may disrupt the optimization, *i.e.*, finding a maximum instead of minimum, or vice versa. In this work, c was set equal to the trace of the first term in the brace brackets divided by p, the number of variables, although the algorithm is not sensitive to this value.

Based on simulated data, it has also been found that if the learning algorithm in Equation (4.9) is used to search for a minimum of kurtosis, the algorithm becomes less stable when the number of samples is only slightly larger than the number of variables. This problem can be solved by using the shifted algorithm in Equation (4.13). The reason may be that, when the number of samples is close to the number of variables, the first matrix in Equation (4.9) is not well-conditioned and the computational round-off errors cause the inverse of this matrix to be unstable. When the shifted algorithm in Equation (4.13) is used, a positive number is added to the diagonal elements of product of the two matrices. This actually changes the condition number of the matrix and makes the algorithm more stable. Optimization by the algorithm in Equation (4.8) is quite stable in this work but alternatively, the shifted algorithm in Equation (4.12) can be used.

4.2.3 Stepwise Univariate Kurtosis

In PP, it is often necessary to extract two or more projection vectors for visualization or other purposes. The projection vectors are generally chosen to be mutually

orthogonal. To find two or more mutually orthogonal projection vectors, the same algorithms described in Sections 4.2.1 and 4.2.2 can be applied in a stepwise fashion to the data after deflation. Deflation is a process that removes the projected structure from the original data set, leaving behind a residual matrix. The deflation method has been used in many other algorithms such as non-linear iterative partial least squares (NIPALS) [34] and is based on the Gram-Schmidt process that is discussed in many textbooks [33]. Expressed mathematically, the deflation process can be written as

$$\mathbf{X}_{\text{new}} \leftarrow \mathbf{X}_{\text{old}} - \mathbf{X}_{\text{old}}\mathbf{v}\mathbf{v}^{\text{T}}, \tag{4.14}$$

where $\mathbf{X}_{\text{old}}$ denotes the matrix before deflation and $\mathbf{X}_{\text{new}}$ is the matrix after deflation. After deflation, the residual matrix $\mathbf{X}_{\text{new}}$ becomes rank deficient. To overcome this problem, PCA can be applied to reduce the dimensionality of the residual matrix $\mathbf{X}_{\text{new}}$ and the same algorithms can be applied to the scores of PCA. Once a new projection vector is found by the same algorithms, it can be rotated back to the original space. Deflation in this way can guarantee the projection vectors are mutually orthogonal. The deflation can be applied until the required number of projection vectors is found. It should be mentioned that deflation of the matrix is not the only method to make the projection vectors mutually orthogonal and other methods can be used as well [19,23].

4.2.4 Multivariate Kurtosis

In PP, the stepwise search for the projection vectors is widely used, but multidimensional approaches have been proposed as well [4,9,13,35,36,37]. Projection vectors extracted by multidimensional approaches are generally different from those obtained by stepwise univariate searches because multidimensional approaches search for the subspace (a plane or hyperplane, normally two or three dimensions) as a single entity and optimize different criteria. Projection vectors obtained by stepwise one-dimensional approaches are generally not nested in the solutions of multidimensional approaches. Huber [8] has conceived that “stepwise one-dimensional approaches may miss structure that a direct *k*-dimensional search would find easily”. The algorithms developed in the work presented here can be easily generalized, without substantial change, for the optimization of multivariate kurtosis. This is introduced as the projection index for multidimensional approaches.

For multivariate data, a definition of multivariate kurtosis [38] analogous to that for univariate kurtosis has been proposed as

$$K = \mathrm{E}\left[(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{z}-\boldsymbol{\mu}) \right]^2, \tag{4.15}$$

where E is the expectation operator, $\mathbf{z}$ is the vector of sample measurements, $\boldsymbol{\mu}$ is the population mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix which is defined as

$$\boldsymbol{\Sigma} = \mathrm{E}\left[(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}} \right]. \tag{4.16}$$

Based on this, the sample multivariate kurtosis can be written as

$$K = \frac{1}{n} \sum_{i=1}^{n} \left[(\mathbf{z}_i - \overline{\mathbf{z}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{z}_i - \overline{\mathbf{z}}) \right]^2, \tag{4.17}$$

where $\mathbf{z}_i$ stands for the measurement vector for a single sample, $\overline{\mathbf{z}}$ is the sample mean vector, and $\mathbf{S}$ is the biased sample covariance matrix which can be written as

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{z}_i - \overline{\mathbf{z}})^{\mathrm{T}}. \tag{4.18}$$

Multivariate kurtosis has been applied to measure homogeneity of planar point-patterns [39]. In PP, when multivariate data are projected onto a subspace (a plane or a hyperplane), the projected data point, denoted by $\mathbf{z}_i$, can be expressed as

$$\mathbf{z}_i^{\mathrm{T}} = \mathbf{x}_i^{\mathrm{T}} \mathbf{V}, \tag{4.19}$$

where $\mathbf{V}$ is a $p \times q$ ($q < p$) matrix whose columns form an orthonormal basis for the subspace (a plane or a hyperplane). For visualization purposes, q is generally chosen to be two or three. To simplify the derivation that follows, the matrix $\mathbf{X}$ is assumed to have been column mean-centered so that $\overline{\mathbf{z}} = \mathbf{0}$ and the multivariate kurtosis for the projected data can be written as

$$\begin{aligned} K &= n \sum_{i=1}^{n} \left[\mathbf{x}_i^{\mathrm{T}} \mathbf{V} \left(\mathbf{V}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{V} \right)^{-1} \mathbf{V}^{\mathrm{T}} \mathbf{x}_i \right]^2 \\ &= n \sum_{i=1}^{n} \left\{ \mathrm{tr} \left[\left(\mathbf{V}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{V} \right)^{-1} \left(\mathbf{V}^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{V} \right) \right] \right\}^2, \end{aligned} \tag{4.20}$$

where tr is the trace operator and the relationship $\mathbf{u}^{\mathrm{T}} \mathbf{M} \mathbf{u} = \mathrm{tr}(\mathbf{M} \mathbf{u} \mathbf{u}^{\mathrm{T}})$ for a symmetric matrix $\mathbf{M}$ and a vector $\mathbf{u}$ has been used.

Multivariate kurtosis is independent of the choice of the basis for the subspace. This means that the value of multivariate kurtosis does not change if the basis vectors are

rotated within the subspace. In a two-dimensional projection, this is consistent with the fact that rotation of the coordinate axes within the plane does not give a different visual interpretation of the data.

The optimization of multivariate kurtosis proceeds in a manner similar to that for univariate kurtosis, requiring that partial derivative be obtained with respect to the matrix **V** and then set to zero. The details of this derivation are presented in Appendix (see Section 4.8.3) and lead to the following learning algorithms:

$$\mathbf{V}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left\{\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\mathbf{V}_k\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}_k\right)^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{x}_i\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right\}\mathbf{V}_k\text{, and} \tag{4.21}$$

$$\mathbf{V}_{k+1} \leftarrow \left\{\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\mathbf{V}_k\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}_k\right)^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{x}_i\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right\}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}_k\,. \tag{4.22}$$

Similar to the learning algorithms in Equations (4.8) and (4.9), the learning algorithms in Equations (4.21) and (4.22) can be used to search for maxima and minima of multivariate kurtosis, respectively. At each iteration, $\mathbf{V}_{k+1}$ is normally adjusted to an orthonormal basis to simplify the test for convergence. In this work, this is done by using singular value decomposition (SVD) to generate orthonormal vectors from $\mathbf{V}_{k+1}$ prior to the next iteration, but other methods could also be used. The convergence criterion for **V** can be an element-by-element comparison between $\mathbf{V}_k$ and $\mathbf{V}_{k+1}$. When the element differences are very small, convergence is reached. Alternatively, the convergence criterion could be the angle between the two subspaces spanned by $\mathbf{V}_k$ and $\mathbf{V}_{k+1,}$ respectively. Discussion about the angle between two subspaces can be found in [40,41]. Once the algorithm has converged, the data are projected into **V** and SVD is carried out on the projected data to give a new set of orthonormal projection vectors that are defined in terms of the directions of decreasing variance in the projected data. While this new (rotated) **V** does not change the spatial relationships among the objects, it provides a consistent and interpretable orientation of the PP subspace.

Similar to the learning algorithms in Equations (4.12) and (4.13), shifted algorithms can be developed for the optimization of multivariate kurtosis by following the same principle as in the shifted power method:

$$\mathbf{V}_{k+1} \leftarrow \left\{\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left\{\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\mathbf{V}_k\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}_k\right)^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{x}_i\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right\}+c\mathbf{I}\right\}\mathbf{V}_k\text{, and} \tag{4.23}$$

$$\mathbf{V}_{k+1} \leftarrow \left\{ \left\{ \sum_{i=1}^{n} \left[\mathbf{x}_i^{\mathrm{T}} \mathbf{V}_k \left(\mathbf{V}_k^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{V}_k \right)^{-1} \mathbf{V}_k^{\mathrm{T}} \mathbf{x}_i \right] \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \right\}^{-1} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right) + c\mathbf{I} \right\} \mathbf{V}_k , \tag{4.24}$$

where c is a scalar and $\mathbf{I}$ is the identity matrix. The learning algorithms in Equations (4.23) and (4.24) are the generalizations of in Equations (4.12) and (4.13) and can be used to search for maxima and minima of multivariate kurtosis, respectively. For the same reason as before, c should be positive and can be assigned as a fraction of the trace of the matrices involved.

The algorithms developed in this work, either for univariate or multivariate kurtosis, are in accordance with the nature of, and can be regarded as generalizations of, the power method and its variants. Because the matrices are updated at each iteration by the updated projection vectors, however, the algorithms are not real power methods and thus they are referred to here as "quasi-power methods".

4.3 Experimental

4.3.1 Computational Aspects

All calculations were carried out using programs in MatLab® v.7.4.0 (MathWorks, Natick, MA) under Windows XP® on a 1.8 GHz computer with 3 Gb of memory. The MatLab codes for the proposed algorithms along with the simulated and experimental data sets can be obtained in the supplementary materials associated with reference [42].

4.3.2 Simulated Data

Three sets of simulated data were used to evaluate the algorithms developed in this work. Data set 1 was a simple two-dimensional data set consisting of 100 objects divided evenly into two classes and was intended to allow direct visualization of the results. Data set 2 included a total of 200 objects in two classes with 10 variables. This was used to examine the performance of the algorithm in higher dimensions. Data set 3 extended this case to three classes, with 100 objects in each class, to observe the effect of adding more classes. More details on the simulation parameters are included in Section 4.4.

4.3.3 Experimental Data

Three experimental data sets were employed in this work to demonstrate the algorithms. The first data set, which will be referred to as the yogurt data, was downloaded

from the website of the Department of Food Science, University of Copenhagen [43]. Details are included in the original reference [44] and are only briefly recounted here. The experiments studied the fluorescence of plain yogurt stored under different conditions for several weeks. Four factors were considered: batch number (1 or 2), container material (polylactate and polystyrene), light exposure (dark or light) and storage time (1, 2, 3, 4 and 5 weeks). This resulted in 40 experiments with triplicate measurements, plus two samples (one from each batch) measured in triplicate at time zero, for a total of 126 experiments. For the purposes of this study, not all of the available measurements were used, only the fluoresce emission spectra at 15 wavelength channels between 310 and 590 nm (20 nm increments) excited at the 270 nm. One set of measurements in the second batch was excluded as an outlier by the original authors, leaving a total of 125 emission spectra at 15 channels.

The second data set, which will be referred to as the salmon data, was obtained from a metabolomics study of plasma from Atlantic salmon [45]. The original study used a nested design to examine sources of variance in ^{1}H NMR spectra obtained on 500 MHz spectrometer, but only part of the data set was used, consisting of technical replicate spectra from each of five individual fish (replicate sample preparations). For each fish, 15 replicates were obtained, except in one case where only 14 were measured. Binned data in the chemical shift range of 0.1725 to 5.7525 ppm in steps of 0.005 ppm were used in the analysis, resulting in a 74×1117 matrix. For the purposes of this study, each fish was assumed to represent a different class, even though they could be considered as biological replicates.

The third data set consists of fatty acid profiles from 572 olive oil samples collected from nine regions of Italy [46,47] and will be designated as the olive oil data. The regions sampled and the number of samples collected were: Northern Apulia (NA-25), Calabria (CA-56), Southern Apulia (SA-206), Sicily (SI-36), Inland Sardinia (IS-65), Coastal Sardinia (CS-33), Eastern Liguria (EL-55), Western Liguria (WL-50), and Umbria (UM-51). The concentrations of eight fatty acids were determined (palmitic, palmitoleic, stearic, oleic, linoleic, arachidic, linolenic, and eicodenoic) and these were scaled in the range of 0 to 100 based on the highest and the lowest values for each fatty acid. The olive oil data has been used to demonstrate the performance of various clustering and

classification methods including projection pursuit [3,46,47,48,49,50] and thus can be regarded as kind of benchmark data set. It was obtained from reference [51].

4.4 Simulation Results

4.4.1 Data Set 1

The first data set contained two-dimensional data drawn from two populations following bivariate normal distributions. The covariance matrices were first set to be the same for the two populations:

$$\mathbf{\Sigma} = \begin{bmatrix} 0.2 & 0 \\ 0 & 1.5 \end{bmatrix}.$$

The population means were set as $\boldsymbol{\mu}_1 = \begin{bmatrix} -1 & 0 \end{bmatrix}^T$ and $\boldsymbol{\mu}_2 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$, respectively and 50 samples were drawn randomly from each population. These distributions were chosen so that the direction of greatest variance did not correspond to the direction for optimal class separation, thereby intentionally impeding the ability of PCA to distinguish classes. The sample data were then rotated by 30^o clockwise to introduce correlation in the measurements. The resulting data points are shown in Figure 4.1 (a).

For two-dimensional data, the projection vector can be rotated to explore how the kurtosis of the projected data varies with the direction of the projection vector. The projection vector was rotated between 0^o and 180^o with respect to the positive abscissa and the result is shown in Figure 4.1 (b). It can be seen that there are two maxima, indicated by $\mathbf{v}_1$ and $\mathbf{v}_3$, and two minima, indicated by $\mathbf{v}_2$ and $\mathbf{v}_4$. Following the univariate mapping of kurtosis as a function of angle, the algorithms defined by Equations (4.8) and (4.9) were used to search for maxima and minima, respectively, in one-dimensional space with ten initial guesses. Vectors $\mathbf{v}_1$ and $\mathbf{v}_3$ were found by the maximum search and $\mathbf{v}_2$ and $\mathbf{v}_4$ resulted from the minimum search, and the corresponding projection vector directions are shown in Figure 4.1 (a). The global minimum was found at 152.9^o ($\mathbf{v}_4$), in good agreement with the theoretical population value of 150^o (or -30^o).

The projection vector corresponding to the first principal component of PCA is indicated by PC 1 in Figure 4.1 (a). It can be seen that the two classes would not be separated if the data were projected onto this projection vector. However, projection on the global minimum of kurtosis obtained by PP ($\mathbf{v}_4$) will result in clustering of the two classes.

Figure 4.1 (c) represents the kurtosis of the projected data versus the direction of projection vector in the form of a radial plot, where the distance between a point on the curve to the origin corresponds to the magnitude of the kurtosis and the line connecting the two points indicates the direction of the projection vector. The two maxima and two minima are clearly visible on this plot. For comparison, Figure 4.1 (d) shows the locus of the variance

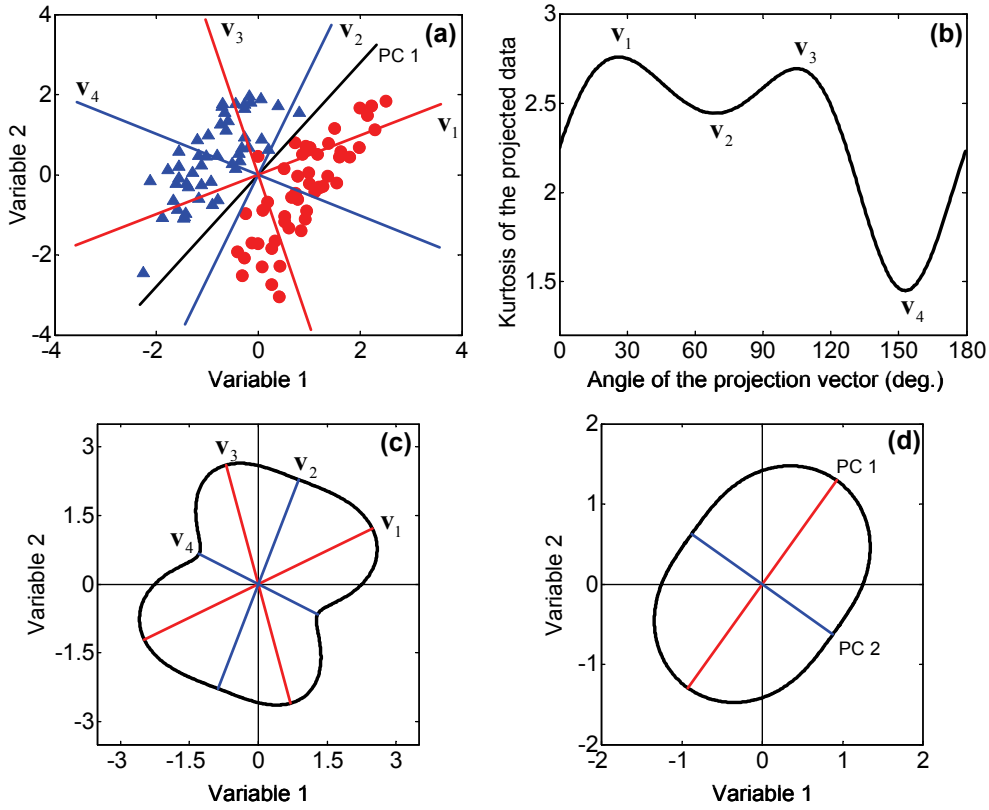

Figure 4.1 Plots of the two-dimensional simulated data and kurtosis and variance with respect to the projection vector. (a) Simulated data and projection vectors found by the proposed algorithms. (b) Kurtosis versus the angle of the projection vector. (c) Representation of kurtosis with respect to the projection vector in a two-dimensional plane, where the magnitude of kurtosis is represented by the distance from a point in the curve to the origin. (d) Representation of variance with respect to the projection vector in a two-dimensional plane, where the magnitude of variance is represented by the distance from a point in the curve to the origin.

of the projected data with respect to the projection vector (*i.e.* the objective function for PCA). It can be seen the variance curve has only one maximum and one minimum, illustrating geometrically why the power method for PCA can find the global optimum, while the proposed quasi-power methods for PP cannot guarantee that the global optima are found. Although the quasi-power methods follow the same principles of the power

method and its variants, due to the nature of kurtosis, multiple initial guesses are needed to increase the chances of finding the global optimum and success cannot be guaranteed by any algorithm. In the present example, it is clear that vector $\mathbf{v}_2$, which is a local minimum in the PP search, would not be able to distinguish the clusters, but in other cases, some local minima may achieve this goal.

4.4.2 Data Set 2

The purpose of PP is to compress the information in higher dimensions into a lower-dimensional subspace, typically two dimensions when the relationship of the objects is to be visually assessed. The second data set was generated in a ten-dimensional space, with 200 samples evenly divided between two multivariate normal populations. As before, the initial covariance matrices for the two populations were set to be the same and the diagonal elements were

$$diag(\mathbf{\Sigma}) = \begin{bmatrix} 0.2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The off-diagonal elements were set to zero. The two population means were set as:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \text{ and}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

Based on this, the separation of the classes occurred only along the first dimension, but the variance is approximately the same in all directions. Correlation was introduced by post-multiplying this 200 x 10 matrix by a 10 x 10 rotation matrix, randomly generated by carrying out SVD on a random 10 x 10 matrix. In the final step, the rotated data were mean-centered (although the population means are zero, the sample means are not). PCA and PP analyses using the proposed algorithms were applied to these data to generate two-dimensional projections of the samples. For PP, searches for both the maximum and minimum kurtosis were employed, with two different approaches: the stepwise univariate approach (Equations (4.8) and (4.9), with deflation as defined by Equation (4.14)) and the multivariate approach (Equations (4.21) and (4.22)). In all cases, 100 random initial guesses were used.

Figure 4.2 (a) shows the profiles of the mean-centered data and Figure 4.2 (b) shows the PCA scores plot for the first two principal components. It is clear that PCA fails to provide a clean separation of the two classes, as anticipated. On the other hand, Figures

4.2 (c) and 4.2 (d) show the PP results for minimization of kurtosis in two dimensions using the stepwise univariate and multivariate approaches, respectively. Several points are

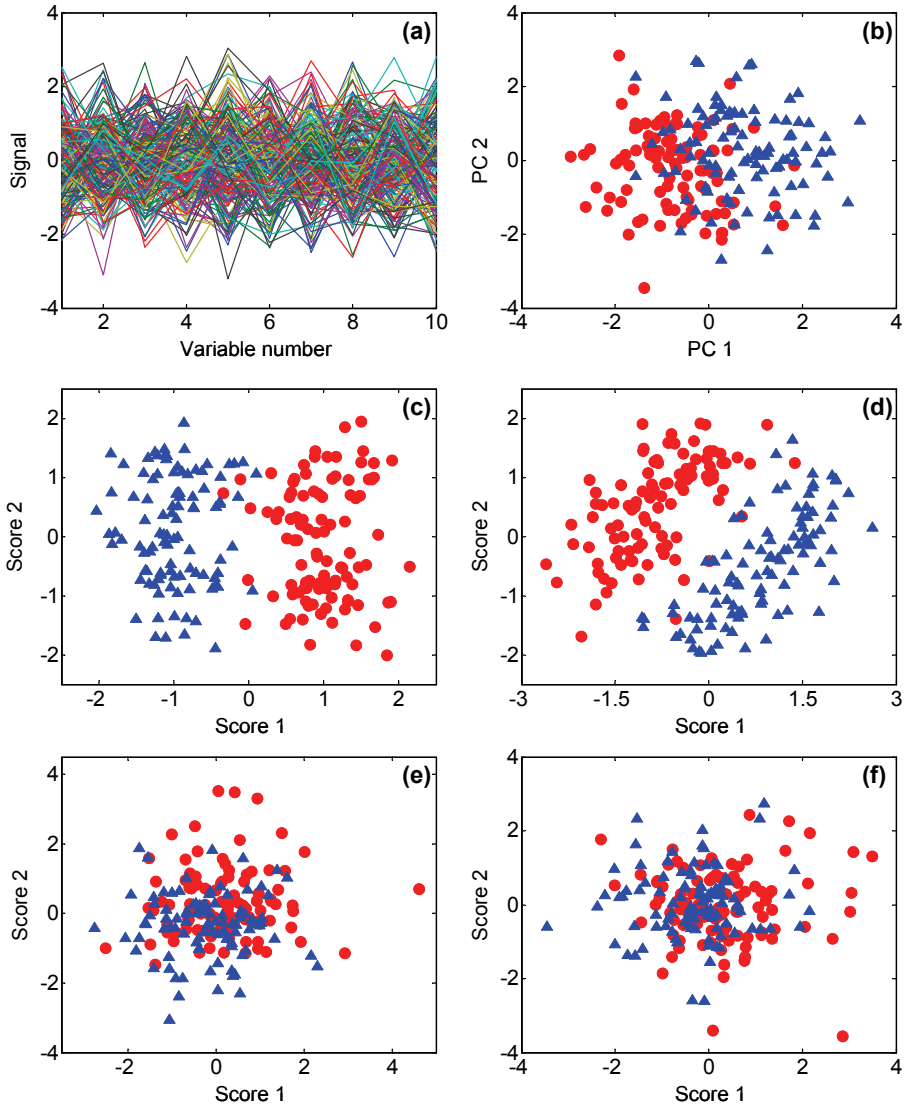

Figure 4.2 Projection results for data set 2. Samples from the two populations are distinguished by the shapes and colors of the symbols. (a) Sample data profiles. (b)-(f) Scores plots on first two projection vectors for: (b) PCA, (c) PP, minimized stepwise univariate kurtosis, (d) PP, minimized bivariate kurtosis, (e) PP, maximized stepwise univariate kurtosis, and (f) PP, maximized bivariate kurtosis.

worth noting for the PP minimizations. First, Figures 4.2 (c) and (d) both show a good separation of the two classes, indicating that PP has achieved its objective. Second, the subspaces identified by the two algorithms are clearly different, even though each separates the classes along one direction. For Figure 4.2 (c), the two basis functions were found in successive steps to identify projection vectors forming the plane, so it is not

surprising that class separation occurs primarily along the x-axis. For Figure 4.2 (d), the projection plane is found in a single step using one objective function which is rotationally invariant, so in principle there are an infinite number of equivalent solutions corresponding to different rotations of the vectors in the solution plane. In practice, the method used to orthogonalize the basis vectors constrains the result to a single solution. In this case, because SVD was used for the orthogonalization, the first basis vector will be in the direction of the greatest variance within the projection plane. Therefore, cluster separation may occur along any direction in the plane. The results in Figures 4.2 (c) and (d) show both algorithms give good separations of clusters, indicating that both algorithms work well to find good subspaces for this data set.

Minimization of kurtosis, as described above searches for distributions that are broad and flat (platykurtic), or even multimodal, and as such tend to force objects into clusters. Maximization of kurtosis, on the other hand, is usually able to locate projected distributions that are peaked with long tails (leptokurtic), and this will be more likely to isolate outliers in the data and produce less information about clusters, except when classes are disproportionately populated. Figures 4.2 (e) and (f) show the PP scores plots from the maximization of kurtosis using the stepwise univariate and multivariate algorithms, respectively. As before, the subspaces identified by the two algorithms are different, but both tend to compress the majority of samples into the center and highlight a few samples that are more distant from the rest. Since no outliers were included in this simulation, there is nothing of particular note in the maximization results.

4.4.3 Data Set 3

The third simulated data set, intended to examine the case of more than two classes, contained samples drawn randomly from three populations in a ten-dimensional space. The three populations are assumed to follow multivariate normal distributions and the covariance matrices were the same for the three populations, with diagonal elements given by

$$diag(\boldsymbol{\Sigma}) = \begin{bmatrix} 0.1 & 0.2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix},$$

and off-diagonal elements equal to zero. The three population means were set at the vertices of an equilateral triangle centered at the origin, given by

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -1 & -0.58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T,$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 1 & -0.58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \text{ and}$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} 0 & 1.15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.$$

From each population, 100 samples were drawn randomly. The samples were transformed by a randomly generated 10 x 10 rotation matrix and then column mean-centered. PCA and the proposed algorithms were applied, with 100 random initial guesses used for all PP searches.

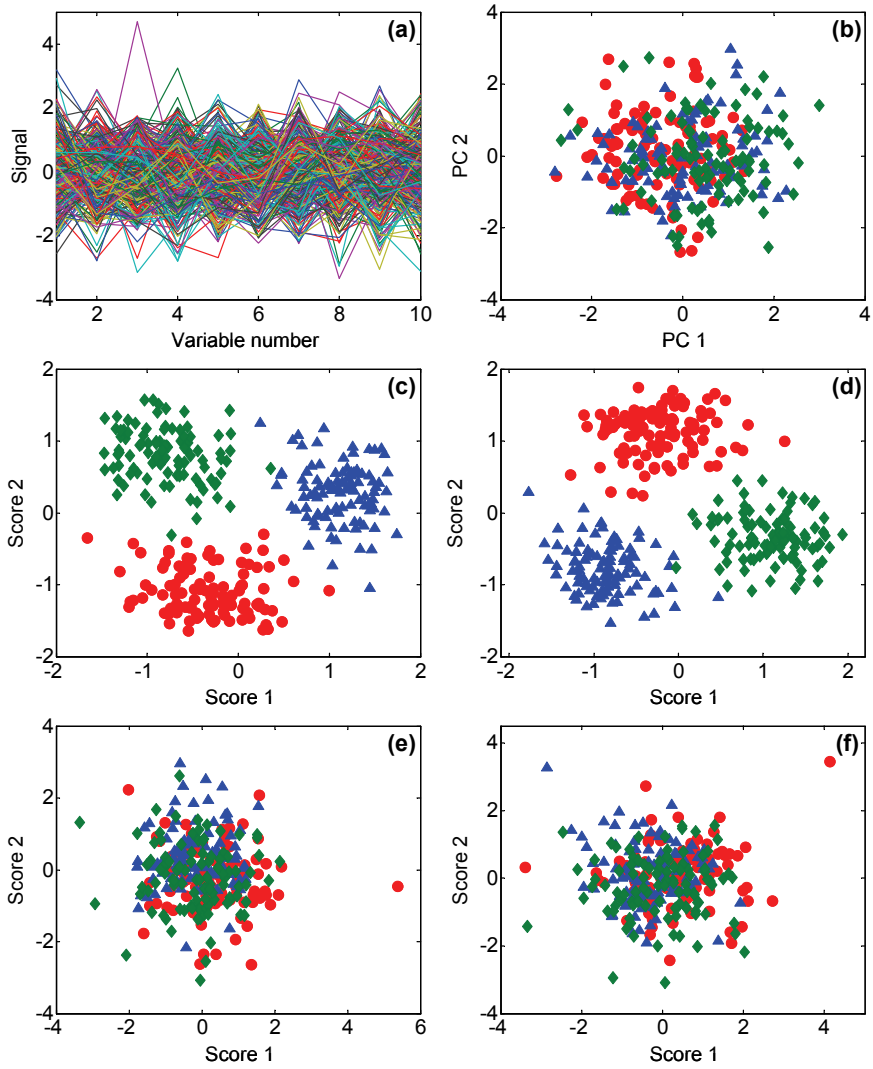

Figure 4.3 Projection results for data set 3. Samples from the three populations are distinguished by the shapes and colors of the symbols. (a) Sample data profiles. (b)-(f) Scores plots on first two projection vectors for: (b) PCA, (c) PP, minimized stepwise univariate kurtosis, (d) PP, minimized bivariate kurtosis, (e) PP, maximized stepwise univariate kurtosis, and (f) PP, maximized bivariate kurtosis.

Figure 4.3 (a) shows the profiles of the mean-centered data and Figure 4.3 (b) shows the PCA scores plot for the first two principal components. As expected, PCA is unable to provide any separation of the clusters because of the high variance in directions other than those separating the classes. Figures 4.3 (c) and (d) show the scores plots for the first two components of PP found by minimum searches with the stepwise univariate and multivariate algorithms, respectively. Although the spaces found by the two algorithms are clearly not identical, both minimizations clearly produce a good separation of the three populations in two dimensions with no prior knowledge of the class structure. Figures 4.3 (e) and (f) show the results of maximization with the same two algorithms. As expected, they produce no class separation and only serve to isolate points that are numerically distant from the majority of the data as potential outliers.

4.5 Experimental Results

The simulation results in Section 4.4 serve to illustrate the efficacy of the proposed algorithms for providing interesting and meaningful projections of multivariate data under carefully controlled conditions for which PCA failed to give informative results. In all cases, reproducible results were obtained quickly and reliably. However, the simulations do not guarantee the utility of the algorithms in the case of real experimental data with more complex class structures, high dimensionality, and possible outliers. Ideally, PP should give results that are no worse than PCA, and in certain circumstances, it may give superior performance. In this section, three different studies found in the literature are used to demonstrate the utility of the proposed algorithms. The first is a data set with relatively low dimensionality but an intricately nested class structure. The second data set has a simpler structure, but a large number of variables that require dimensionality reduction through PCA. The third data set has been widely used as a benchmark for exploratory data analysis and classification studies.

4.5.1 Yogurt Data

The yogurt data were chosen to illustrate the algorithms presented here because it consisted of a high sample-to-variable ratio and several factors upon which separation of the samples might be based. The fluorescence measurements from the yogurt samples under different conditions were placed in a 125 x 15 matrix. The original 125 spectra are

shown in Figure 4.4 (a). For analyses by PCA and PP, the data were column mean-centered, but no scaling or sphering was performed. In all applications of the PP algorithms, 100 random initial guesses were used and the solutions were taken as the ones with the lowest (minimization) and highest (maximization) kurtoses. To minimize convergence time in the presence of a high degree of multicollinearity, the shifted algorithms (Equations (4.12), (4.13), (4.23) and (4.24)) were used in this study.

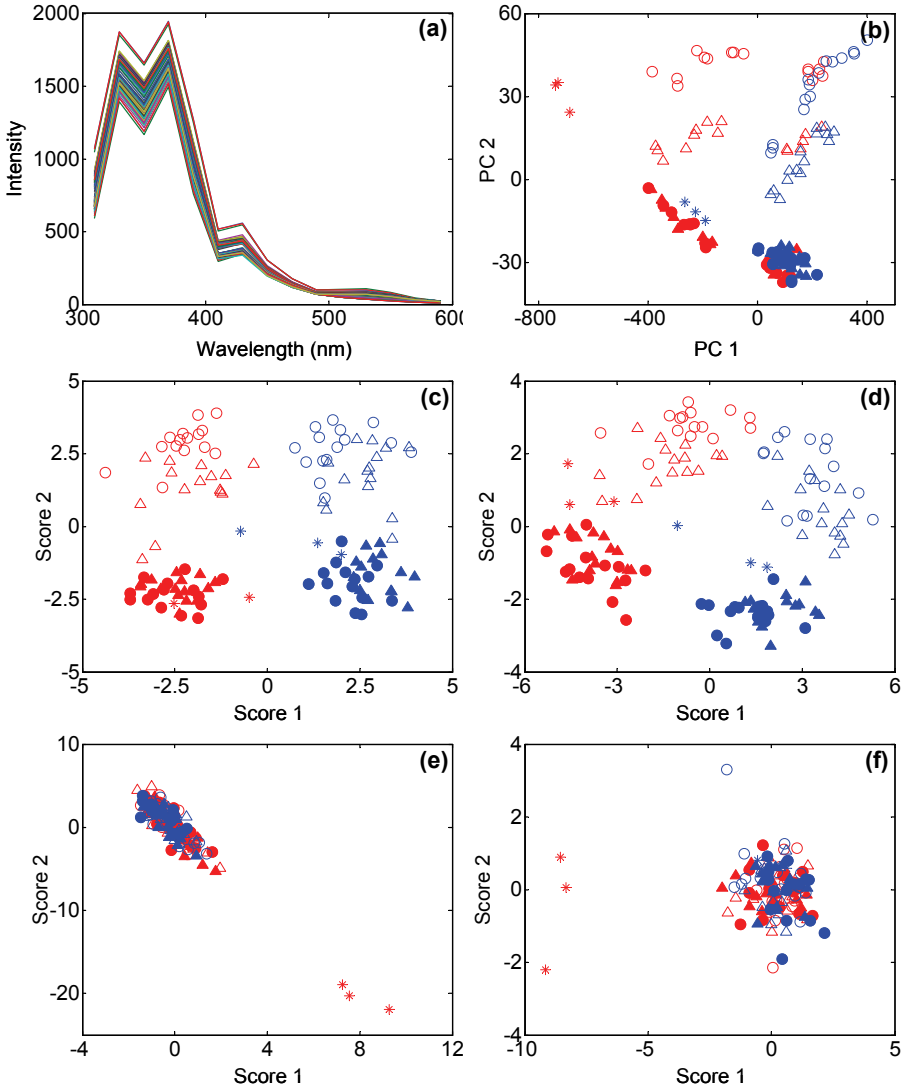

Figure 4.4 (a) Fluorescence emission spectra for yogurt data set. (b)-(f) Scores plots for PCA and PP: (b) PCA, (c) PP, minimized stepwise univariate kurtosis, (d) PP, minimized bivariate kurtosis, (e) PP, maximized stepwise univariate kurtosis, and (f) PP, maximized bivariate kurtosis. The color of the symbols indicates the batch number (red=B1, blue=B2); the shape indicates the packing material (circles=P1, triangles=P2, asterisks=before packaging); the fill indicates the light exposure during storage (solid=dark, open=light).

The PCA scores plot for the first two eigenvectors is shown in Figure 4.4 (b), and some groupings of the samples are immediately obvious. The most apparent separation is between the samples stored in the dark (solid symbols) and samples stored in the light (open symbols). This separation occurs mainly along the second principal component and is well-defined. The first PC correlates primarily with differences in the two batches (red and blue symbols), although this separation is not complete. All of the samples from batch two (blue), with the exception of those taken at time zero (asterisks) are on the right hand side of Figure 4.4 (b), but these are mixed with a significant number of samples from batch one. There is some separation based on packing material (circles vs triangles) for samples stored in the light, but this is not apparent for samples stored in the dark.

Figure 4.4 (c) shows the scores plot of the first two components obtained by the minimum search using the stepwise univariate algorithm (shifted form, Equation (4.13)). The first distinctive feature apparent in Figure 4.4 (c) is that the samples divide into four distinct quadrants, with the batches (red/blue) separated on the first projection vector and the light exposure (open/solid) separated on the second projection vector. This is consistent with the results obtained by PCA, but the spatial relationships are quite different and the groups more uniformly defined. Unlike PCA, the batches are completely separated in this case, with the exception of one of the time zero samples (asterisk) from batch two. The separation by light exposure is also largely complete, with the exception of perhaps four borderline samples. As with PCA, there is some marginal separation by packing material for the light-exposed samples along the *y*-axis, but none for the samples stored in the dark.

Figure 4.4 (d) shows the scores resulting from the application of the shifted multivariate algorithm for minimum kurtosis (Equation (4.24)) in two dimensions. The result is similar to that for the stepwise univariate algorithm in Figure 4.4 (c), with the samples separating into four quadrants based on batch and storage conditions, but in this case the quadrants are not aligned with the projection axes. This is not surprising because multivariate kurtosis is independent of rotation within the subspace and the final axes are arbitrarily selected on the basis of maximized variance by SVD. Although the spatial relationships resulting from the two algorithms are similar, the subspaces are not identical.

Figures 4.4 (e) and (f) show scores plots resulting from maximum kurtosis searches by the stepwise univariate and multivariate approaches, respectively. As previously noted,

maximum searches will tend to isolate outlying data points. These may be individual samples that are true statistical outliers arising from erroneous measurements or non-representative sampling, or they may represent a collection of samples belonging to a class whose membership is small relative to the other classes in the data. In this case, both maximum searches isolate the three time zero samples associated with batch one. This is not surprising, since the samples measured at time zero might be expected to differ substantially from others in the data set.

4.5.2 Salmon Data

The salmon data set was chosen as a second example because it is perhaps more typical for chemical data sets in that there are many more variables than samples. Because the algorithms here involve inversion of $\mathbf{X}^T\mathbf{X}$, which will be singular under these conditions, it is necessary to carry out some kind of variable compression prior to PP. Even in cases where the matrix is not singular but the ratio of samples to variables is relatively low, PP can be problematic because the space is sparsely populated by objects and solutions are likely to result in opportunistic clustering of the samples that are not consistent with meaningful classes. In both of these cases, SVD is perhaps the simplest and most reliable means to achieve variable compression, since it should retain the maximum amount of information in the scores. In this case, the original data matrix was first column autoscaled and then decomposed by SVD, retaining the scores on the first seven eigenvectors for analysis by PP. The end result of the analysis is somewhat dependent on the number of factors retained, but this is discussed in more detail in Section 4.6.1 and will not be dealt with here. As before, the shifted algorithms were used for the results reported.

Figure 4.5 (a) shows the 74 NMR spectra used in this study after binning and truncation of the range to remove the reference peak. Figure 4.5 (b) shows the scores plots for the first two principal components resulting from PCA on the centered data, with replicates from each fish shown with different colors and symbols. Some grouping of replicates for each fish is evident, but only Fish 1 (red triangles) separates cleanly from the others. The results of kurtosis minimization by stepwise univariate and multivariate searches are shown in Figures 4.5 (c) and (d), respectively. In contrast to the PCA result, both of these figures show a clear separation of the replicates for each fish, even though no class information is provided to the algorithm. In this case, the stepwise univariate method

provides somewhat better separation. Figures 4.5 (e) and (f) show the results of the maximization search for the two methods. The univariate search isolates a few samples as potential outliers, while the multivariate approach distinguishes the replicates for Fish 1 as distinct from the others.

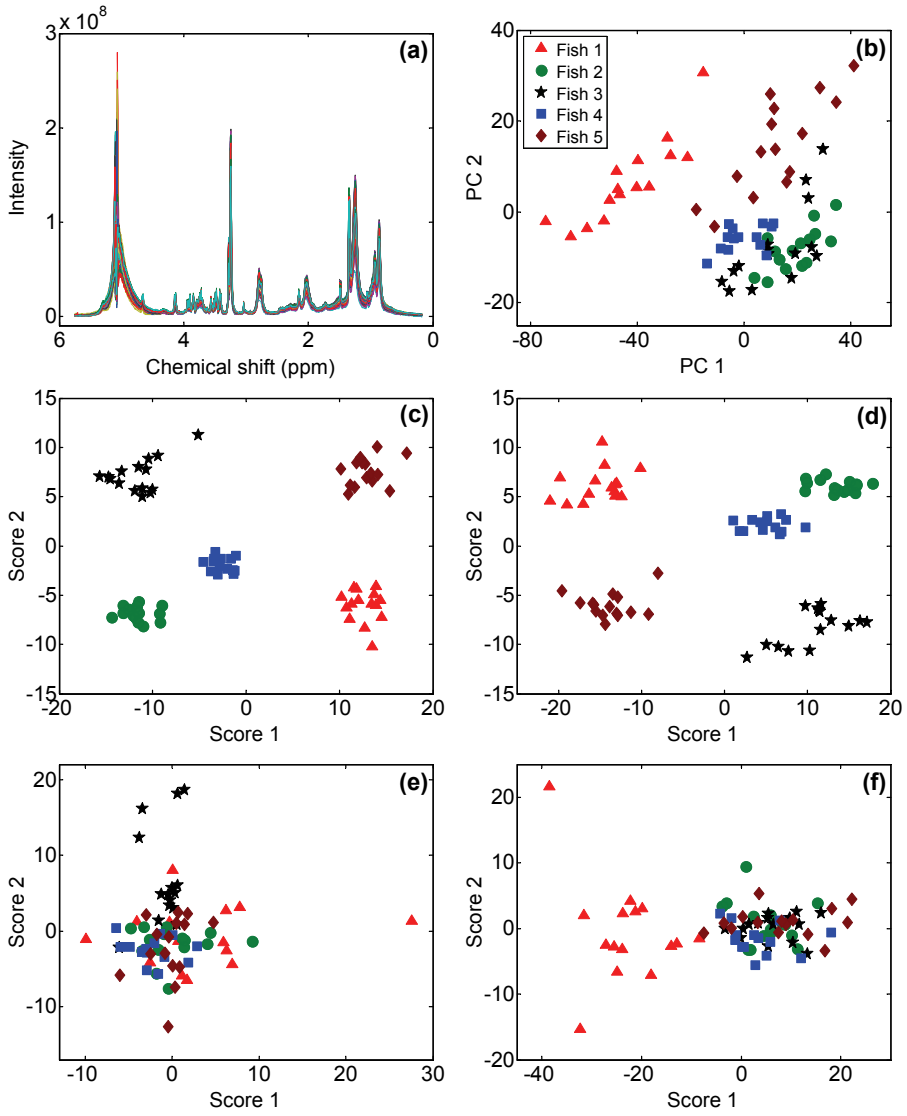

Figure 4.5 (a) ^{1}H NMR spectra for salmon data set. (b)-(f) Scores plots for PCA and PP: (b) PCA, (c) PP, minimized stepwise univariate kurtosis, (d) PP, minimized bivariate kurtosis, (e) PP, maximized stepwise univariate kurtosis, and (f) PP, maximized bivariate kurtosis. The colors and shapes of the symbols indicate five different fish as indicated in the legend.

4.5.3 Olive Oil Data

The olive oil data set, like the yogurt data, represents another situation where the ratio of samples to variables is high, so no variable compression was necessary. This data

set was included in the study because it has been widely used as a benchmark in the literature [3,46,47,48,49,50]. As before, the data were column autoscaled and the normal algorithms (Equations (4.8), (4.9), (4.21), and (4.22)) were employed.

Figure 4.6 (a) indicates generally on the map of Italy where the samples were collected, with boundaries shown in red and blue that will be used below to reference the locations of samples. The scores plot from PCA is shown in Figure 4.6 (b), where, for convenience, the symbols used are the same as those employed in reference [50] and colors have been employed to differentiate samples associated with more northerly regions (blue) from those in the southern locales (red). It is apparent from Figure 4.6 (b) that the samples exhibit a clear geographical correlation, both in terms of their location and their characteristic dispersion in the scores space. The figure does not show well-separated clusters, but some classes (*e.g.* S. Apulia and W. Liguria) segregate into fairly pure groups, while mixing is apparent among other classes. Some classes (*e.g.* Coastal Sardinia) are tightly grouped, while others (*e.g.* Sicily) are more dispersed. While this information is useful and higher PCs will augment this picture, it has been demonstrated elsewhere [3,49,50] that other projection methods can provide a clearer distinction among the groups, so the proposed PP algorithm was applied to the data.

Figure 4.6 (c) shows the first two scores resulting from the stepwise univariate minimization of kurtosis. Immediately obvious is a clean separation of the samples from region 1 (blue/north) from those in region 2 (red/south) (see Figure 4.6 (a)). However, there is little else that can be surmised from this plot and it is clear that the information contained on the two projection axes is redundant. This redundancy is an artifact that can sometimes occur with the stepwise approach. One would anticipate that the first projection axis should be the one running diagonally through the figure, but it is suspected that the orthogonal projections of the samples leads to a marginally smaller kurtosis in the direction of this axis. As a consequence, the two major classes are partitioned along two directions in the stepwise procedure and no new information is generated. However, new information does appear along the third projection vector, as is evidenced in Figure 4.6 (d), which shows the scores on projection vectors 1 and 3 from the stepwise procedure. Here, projection vector 1 retains the separation of north/south (blue/red) samples, while projection vector 3 generates additional separation in the classes. In particular, among the northern samples,

the two groups of samples from Sardinia (Inland and Coastal), separate from the other three samples, as indicated by the inner boundaries in Figure 4.6 (a). The separation of classes within the southern samples is less distinct, but the samples from S. Apulia are clearly different from the other three regions in this group, also indicated by the inner boundaries in Figure 4.6 (a).

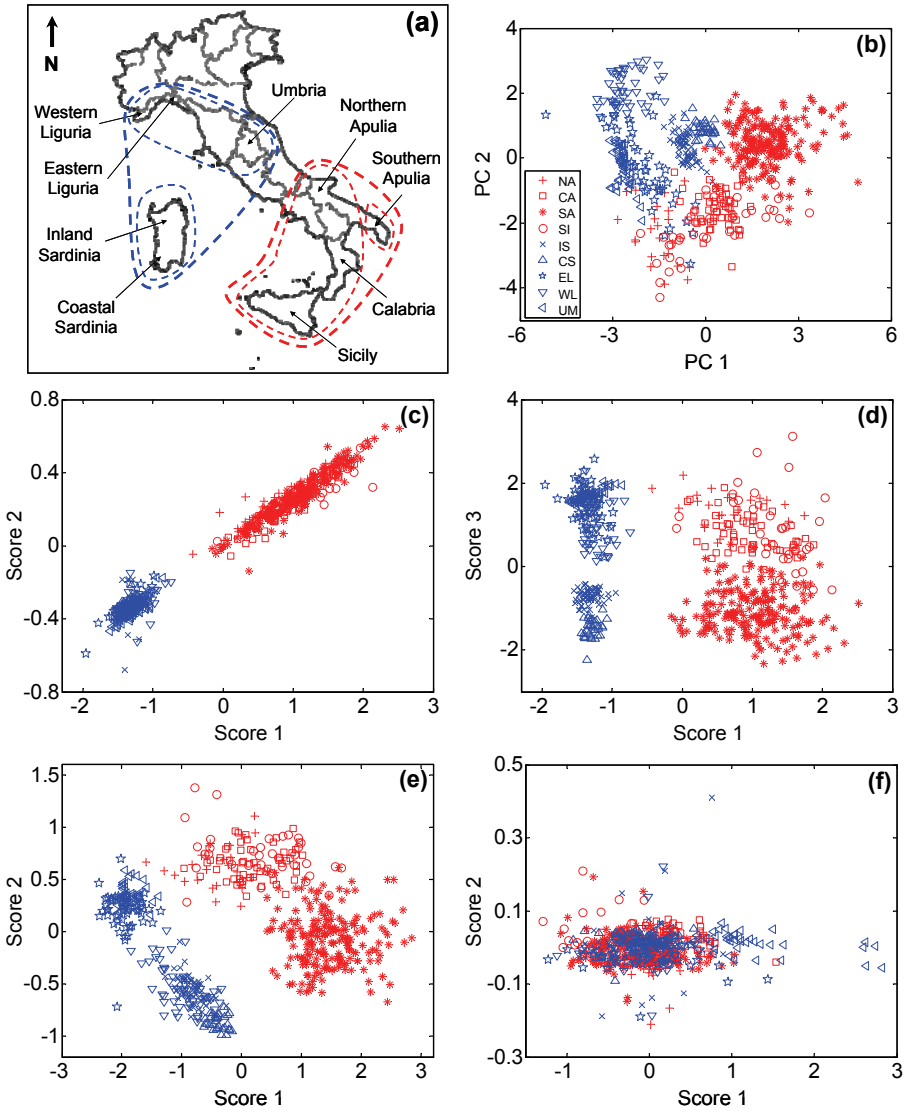

Figure 4.6 (a) Map of Italy showing approximate locations where olive oil samples were collected. (b)-(f) Scores plots for PCA and PP: (b) PCA, (c) PP, minimized stepwise univariate kurtosis (2 vs 1), (d) PP, minimized stepwise univariate kurtosis (3 vs 1), (e) PP, minimized bivariate kurtosis, and (f) PP, maximized bivariate kurtosis. The legend in (b) applies to (c)-(f).

The minimization of multivariate kurtosis in two dimensions, as shown in Figure 4.6 (e), produces results similar to Figure 4.6 (d). The problems of the stepwise algorithm

do not occur in this case as kurtosis is minimized across both dimensions simultaneously. While the separation of samples for Figure 4.6 (e) shows similarities to Figure 4.6 (d), the two are certainly not identical. One first notices that the general arrangement of samples appears rotated in one with respect to the other, but this is not surprising because the multivariate kurtosis is independent to rotation in the plane and the axes are arbitrarily aligned along directions of decreasing variance. The most noticeable change in the grouping of the samples occurs in the northern (blue) classes, where the samples from W. Liguria are now more closely associated with the samples from Sardinia.

The results of kurtosis maximization using the multivariate algorithm are shown in Figure 4.6 (f) for completeness. As noted previously, maximization will attempt to isolate outlying samples, but the results in this case do not show any features that are particularly noteworthy. Similar results were generated by univariate maximization, but are not shown here.

Generally speaking, the groupings of samples obtained by the methods presented here are qualitatively similar to other approaches in the literature [3,49,50], especially those obtained using a bottleneck neural network approach [50]. An important difference, however, is that the projection pursuit approach produces a linear mapping, which is not generally the case with neural network techniques.

4.6 Discussion

The results above show that the proposed PP algorithms can be applied quite successfully to experimental data sets to yield useful information. However, some aspects of their implementation, including variable compression, speed and reliability warrant further comment. These aspects will be considered here.

4.6.1 Variable Compression

It is common in chemistry for data sets to be obtained where the number of variables is greater than the number of samples, sometimes by a wide margin. When this is the case, inversion of matrices in the algorithms presented here becomes problematic because of singularity. This necessitates a reduction in the number of variables to a level where they not greater than the number of samples. Although several approaches can be used for this, SVD is probably the most straightforward, with the subsequent PP analysis

being carried out on a truncated scores matrix rather than the original variables. This was the method used here for the salmon data. However, even when the singularity is eliminated, results from PP may be unsatisfactory unless the number of variables is further reduced to the point where the ratio of samples to variables is substantially greater than unity. If this condition is not met, clustering of the samples may be observed that is not meaningful.

The initial objective in the analysis of experimental data sets such as those described in this study is most often to determine if there is a meaningful separation of the samples according to a known or unknown class structure. This is a preliminary step towards other objectives that could include classification of new samples or developing a fundamental understanding of the underlying physical reasons for the separation. In the initial exploration of the data, PCA is widely used because it is unsupervised. Because it makes no use of class information, any separation observed that correlates to class structure can be interpreted as meaningful. However, the converse is not true, so failure of PCA to support the class structure does not mean it is not embedded in the data. At the other extreme, supervised methods such as linear discriminant methods use the class information directly to optimize projection axes and are virtually guaranteed to provide a separation of classes when the ratio of variables to samples is large. Therefore, careful validation of the results is necessary to ensure that they are meaningful. PP falls somewhere between these two approaches. It can be considered unsupervised because it does not make use of class information, but it does search the projection space to find the one that forces the samples into clusters, so unlike PCA, it imposes a criterion based on the separation of the data. When the dimensionality of the space is large and the number of samples is relatively small, undersampling is a problem and PP is more likely to find opportunistic clusters that are not correlated with any real class structure.

Because PP can exhibit this kind of artificial clustering of the data, the selection of the appropriate number of variables is important. Selection of an insufficient number of variables will not provide PP with enough information to separate the data in a meaningful way, while the use of too many variables can lead to the generation of spurious groupings. These concepts are illustrated in Figure 4.7, which shows the results of the analysis of the salmon data with stepwise univariate minimization of kurtosis. Figures 4.7 (a) and (b)

show the results of where four and five scores, respectively, have been retained from SVD of the original data. Although some segregation of objects in different classes is apparent, it is clear that there is insufficient information to completely separate the data. However, when six and eight variables are retained, the separation is quite clear, as shown in Figures 4.7 (c) and (d) (the case of seven variables was shown in Figure 4.5 (c)). As the dimensionality of the space increases, the separation is less reliable, and for nine and fifteen retained variables there is some mixing of the samples from different classes, as shown in Figures 4.7 (e) and 4.7 (f). Finally, in the extreme case when the ratio of samples to variables is much too small, undersampling will lead to random clustering of the data that minimizes the kurtosis. This is illustrated in Figures 4.7 (g) and 4.7 (h), where fifty latent variables have been used for stepwise univariate and multivariate minimizations, respectively. For the stepwise univariate approach, clustering approaches the limiting case where objects will be placed symmetrically at the corners of a square, with little or no retention of class information. The limiting case in the multivariate minimization is to place the objects evenly around the edges of a circle, again with little correlation to the real data structure.

This raises the question as to what is the optimum sample-to-variable ratio to employ in projection pursuit. Of course, the answer to this will depend on the characteristics of the data set, but the author's experience with a number of simulated and experimental data sets suggests that a minimum value for this ratio is around ten. It is likely that there is a window within which acceptable results can be obtained, as in the case of the salmon data, but this window may not exist in cases where the number of samples requires a level of variable compression that is insufficient to retain the information necessary for class separation.

If class information is known for the original data, the clustering exhibited through PP can be immediately validated against the known class structure. It is unlikely in most cases that clustering based on overfitting of objects to minimize kurtosis will correlate with the known class structure of the data. This is a distinct advantage of PP over supervised methods, where validation through permutation methods or external samples is generally necessary to ensure the validity of the results. It is also possible in the case of PP to carry out Monte Carlo simulations to test the null hypothesis that, for a given number of samples

and variables, the measurements are drawn from a multivariate normal distribution. This can be done by evaluating the *p*-value of the optimized kurtosis as a function of the number of variables used. Note, however, that such a test cannot verify how a distribution differs from multivariate normality or be used to select the optimum number of variables to retain.

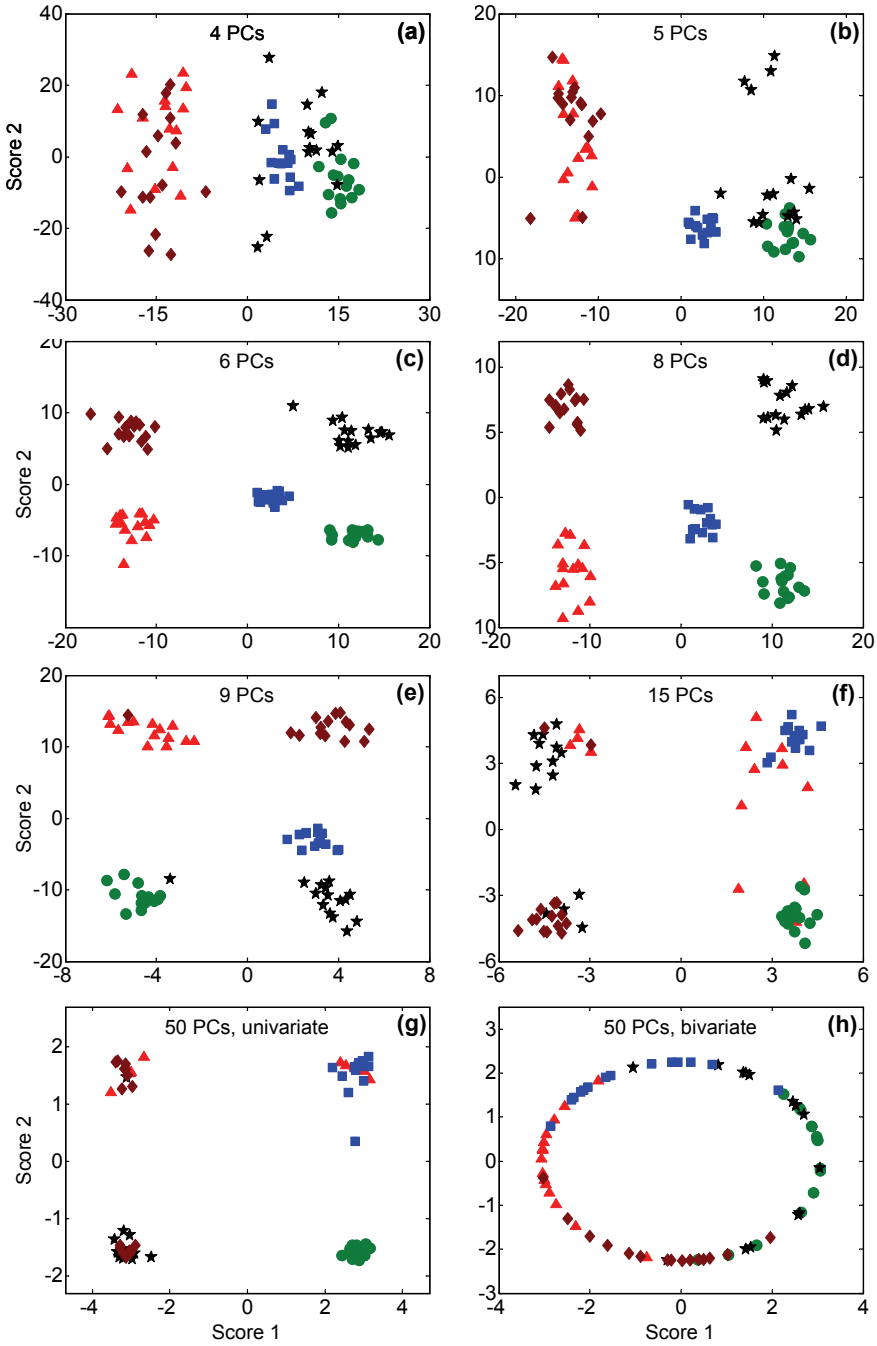

Figure 4.7 Scores plots for PP applied to different numbers of principal components extracted from the salmon data set. (a)-(g) Univariate minimizations of kurtosis: (a) four PCs, (b) five PCs, (c) six PCs, (d) eight PCs, (e) nine PCs, (f) fifteen PCs, and (g) fifty PCs. (h) Bivariate minimization of kurtosis, fifty PCs. (See legend in Figure 4.5.)

4.6.2 Speed and Convergence Reliability

The algorithms presented in this work are simple and converge relatively quickly, but because kurtosis can exhibit multiple maxima and minima for a given data set, this raises the question of how confident one can be that global optima have been located and how many initial guesses should be used to increase this confidence to a reasonable level. Again, this will depend on the characteristics of the data set, but some insight can be obtained based on the data sets examined in this work. Figure 4.8 shows histograms of kurtosis values resulting from 1000 bivariate minimizations for the three experimental data

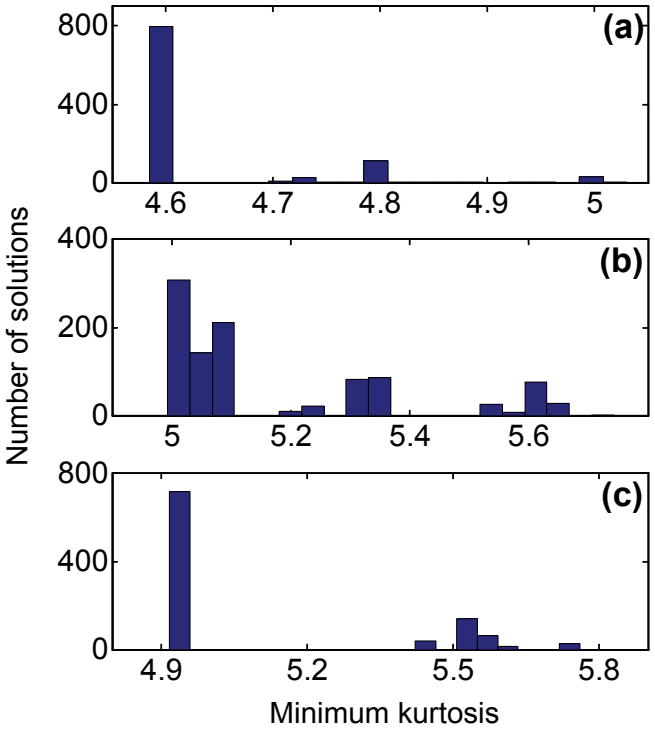

Figure 4.8 Distribution of solutions from minimization of bivariate kurtosis with 1000 random initial guesses: (a) yogurt data, (b) salmon data, and (c) olive oil data.

sets presented in this work, with the original variables used for the yogurt and olive oil data (Figures 4.8 (a) and (c), respectively) and seven latent variables used for the salmon data (Figure 4.8 (b)). Assuming that the smallest kurtosis value represents the global minimum (a reasonable assumption, but never guaranteed), the frequency at which this solution was located ranged from about 30% to about 80%. For the stepwise univariate method, the probability for the global minimum was taken to be the combined probabilities of finding the global minimum on the first step and the second step, and values were 5% (yogurt), 23% (salmon) and 4% (olive oil). Note that lowest value, that for the olive oil, was

computed based on three steps, since successful separation required the first and third projection vectors. Results for maximizations were in similar ranges. If these results are taken as typical, and the probability of finding the global solution is pushed even lower to 1%, the probability of not finding the global minimum is less than 37% with 100 initial guesses and less than 0.005% with 1000 initial guesses. A complete analysis of the convergence characteristics as they depend on the dimensionality of the space and the sample-to-variable ratio is beyond the scope of this work, but it is expected that as long as the latter is kept reasonably high, similar results should be obtained.

Since multiple initial guesses are required with these algorithms, the convergence time for each guess is an important consideration. A comprehensive analysis of this would need to consider the dimensionality of the space, the number of points, the nature of the data, and the type of optimization. In lieu of this, representative times are given in Table 4.1 for the three experimental data sets examined here, using the number of variables in the results presented. The times given are the total time for 1000 initial guesses, exclusive of loading and preprocessing of the data. The shifted algorithms were used for yogurt and salmon data and the ordinary algorithm was used for the olive oil data.

Table 4.1 Comparison of computation times for 1000 initial guesses.

Data set	Optimization	Algorithm	Time (s)
Yogurt (125x15)	min	univariate (2)	28.6
		bivariate	239.2
	max	univariate (2)	7.2
		bivariate	62.8
Salmon (74x7)	min	univariate (2)	10.8
		bivariate	87.1
	max	univariate (2)	25.9
		bivariate	52.4
Olive oil (572x8)	min	univariate (3)	82.4
		bivariate	65.7
	max	univariate (3)	27.1
		bivariate	79.6

Of course, these times will be reduced by a factor of ten for the 100 initial guesses used for the results presented in this work. The times reported are reasonable for routine applications that do not involve extensive repetitions for cross-validation or Monte-Carlo studies. Generally, the convergence time is shorter for the stepwise univariate approach

than that for the bivariate algorithm, an exception being the minimization for the olive oil data. In this case, however, three projection vectors were extracted by the univariate algorithm, requiring additional time.

4.6.3 Other Considerations

A limitation of the results reported here is that all of the examples contained roughly the same number of samples in each class. Because of this, minimization of kurtosis is more likely to partition the data by class. As was demonstrated, the limiting case for the stepwise univariate approach is to push samples into narrow equally populated groups at the corners of a square, while the bivariate approach ultimately tries to project the samples evenly around a narrow ring to minimize kurtosis. Thus, the optimum solutions have to balance the number of samples within each class with the distribution of those classes. In cases where the number of samples per class is substantially unbalanced, this may produce unsatisfactory results. In such cases, the application of the maximization algorithms may provide more useful results, since the goal of maximization is to isolate a small number of samples that are relatively distant from the distribution represented by a majority of samples. Alternatively, using the stepwise algorithm, it should be possible to alternate minimization and maximization in subsequent steps to more effectively partition the samples in unbalanced studies. Finally, the use of skewness instead of kurtosis as a projection index may fit these situations and the algorithms presented can be readily adapted to optimize this index.

For most of the cases presented here, the stepwise univariate and bivariate minimizations produced similar, but not identical, results. In any given case, one method may provide a somewhat cleaner or more rational separation of the data than the other, so the application of both is probably useful. If only one method is applied, however, the bivariate approach is probably more reliable since it considers the two projection directions simultaneously and is therefore less likely to generate redundant information, as was the case with the olive oil data.

4.7 Conclusions

Unsupervised exploratory data analysis has been dominated by PCA and hierarchical clustering methods which, despite their utility, do not always partition objects

in a manner that is interesting in the context of the problem at hand. PP has existed for many years and has the advantage of using distributional parameters as the criterion for choosing optimal projections. However, it is much less widely applied in chemistry largely because of the difficulty in implementing algorithms to effectively search the projection space. In this work, simple and efficient PP algorithms, referred to as quasi-power methods, for the optimization of kurtosis as a projection index have been proposed. The algorithms consist of a simple iterative procedure that requires only a few lines of code and no optimization of search parameters. The speed of convergence allows multiple runs with different initial guesses to increase the confidence that global optima have been found in quartic search space. In addition to simplicity and speed, the algorithms offer other options to increase their utility. Both stepwise univariate and multivariate approaches can be employed for projections into multiple dimensions, permitting slightly different perspectives on the same data. Moreover, kurtosis can be maximized or minimized with only small changes to the algorithm. Minimization of kurtosis is most effective for separating uniformly populated classes of objects, while maximization can be used to identify possible outliers or separating classes with unbalanced populations. Finally, except for mean centering, no preprocessing of the data, such as sphering or whitening, is required, unlike some other methods.

The results presented here have demonstrated that, for both simulated and experimental data, PP with the quasi-power methods is effective for separating classes within the data and providing information that may not be evident from the application of PCA alone. Because PP is an unsupervised method, there can be more confidence that the resulting class separation reflects the true underlying data structure and is not the result of overfitting some model. However, because PP optimizes a distributional parameter that favors the appearance of clusters, one must be careful not to over-interpret the organization of the data in the absence of known classes. In particular, it is important to ensure that the ratio of samples to variables be kept relatively high for results to be meaningful, and some compression of variables by SVD or other methods may be necessary. In this work, useful results were obtained when this ratio was about ten or more, but this has not been rigorously examined. It is hoped that the demonstrated utility of PP based on kurtosis,

coupled with the availability of the simple and efficient algorithms reported here, will lead to the more widespread use of this method in the analysis of multivariate chemical data.

4.8 Appendix

4.8.1 Derivation of Learning Algorithms for Univariate Kurtosis

Starting with the definition of univariate kurtosis for the projections of measurement vectors, **x**, onto the projection vector defined by **v**,

$$K = \frac{n\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^2}, \tag{4.25}$$

the objective is to minimize or maximize K with respect to **v** by setting the partial derivatives to zeros. In many optimization problems, projection vectors are constrained to unit length and Lagrange multipliers are often introduced [30]. In PP, the projection vector **v** is generally chosen to be unit length as well. However, examination of Equation (4.25) shows that the length of the projection vector **v** does not affect the value of kurtosis. Therefore, the optimization of kurtosis in Equation (4.25) can be treated as an unconstrained optimization problem. Although the expression in Equation (4.25) seems complex, it actually simplifies the optimization because a Lagrange multiplier is not needed. Applying the Quotient Rule and Chain Rule to Equation (4.25) yields

$$\frac{\partial K}{\partial \mathbf{v}} = \begin{pmatrix} \frac{\partial K}{\partial v_1} \\ \frac{\partial K}{\partial v_2} \\ \vdots \\ \frac{\partial K}{\partial v_p} \end{pmatrix} = \frac{2n\sum_{i=1}^{n}\left[\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\frac{\partial\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)}{\partial \mathbf{v}}\right]}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^2} - \frac{2n\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2\right]\frac{\partial\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)}{\partial \mathbf{v}}}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^3}, \tag{4.26}$$

where $\frac{\partial K}{\partial \mathbf{v}}$ is the shorthand for the partial derivatives arranged in the vector form, which in some calculus textbooks is called the gradient vector and is denoted by the symbol "∇". It is also known from vector calculus [30,31,32] that

$$\frac{\partial(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v})}{\partial \mathbf{v}} = 2\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\text{, and} \tag{4.27}$$

$$\frac{\partial\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)}{\partial\mathbf{v}} = 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\,. \tag{4.28}$$

Substitution of Equations (4.27) and (4.28) back into Equation (4.26) gives

$$\frac{\partial K}{\partial\mathbf{v}} = \frac{4n\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^2} - \frac{4n\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2\right]\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^3}\,. \tag{4.29}$$

Setting this equal to zero leads to

$$\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{v} = \frac{\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2}{\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}\,, \tag{4.30}$$

where it has been assumed that the scalar term $\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}$ is not zero. Replacement of the ratio on the right of Equation (4.30) with the scalar variable ρ (a function of **v)** yields

$$\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v} = \rho\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}\,. \tag{4.31}$$

In this work, it is assumed that $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ are both invertible. The assumption requires that in the data matrix **X**, the number of samples cannot be less than the number of variables. In case that there are more variables than samples, PCA can be applied and a small number of principal components can be used in place of the original data matrix. If $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, the matrix $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ is generally invertible because if each row in **X** is multiplied by a scalar, the scaled matrix **X** multiplied by its transpose from the left will give $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ and such scaling in general does not reduce the rank of the matrix. The optimization of kurtosis is the result of solving Equation (4.31) to find the solutions for **v**.

It can be seen that there are no closed-form solutions for **v** in Equation (4.31), and **v** must be sought through iterative methods. With the assumption of invertibility, Equation (4.31) can be re-written in two ways:

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v} = \rho\mathbf{v}\text{, and} \tag{4.32}$$

$$\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}=\frac{1}{\rho}\mathbf{v}\,. \tag{4.33}$$

The above expressions are similar to the standard eigenvalue problem and indicate that **v** can be found by the following learning algorithms that are reported in Section 4.2.1.

$$\mathbf{v}_{k+1}\leftarrow\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v}_k\text{, and} \tag{4.34}$$

$$\mathbf{v}_{k+1}\leftarrow\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}_k\,. \tag{4.35}$$

4.8.2 Convergence Interpretation

The convergence to a maximum or a minimum for Equations (4.34) and (4.35) can be interpreted by following the same principle as in the power method. It can be seen that if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is positive definite since for any vector **v**, $\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}=\left(\mathbf{X}\mathbf{v}\right)^{\mathrm{T}}\left(\mathbf{X}\mathbf{v}\right)>0$. Similarly, if the term $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ is invertible, it is positive definite. Since both $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ are positive definite, they both have only positive eigenvalues. Obviously, both $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ and $\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}$ are positive definite and have only positive eigenvalues as well. It is also known that the product of two positive definite matrices has only positive eigenvalues [52]. This means that the matrices in Equations (4.34) and (4.35) have only positive eigenvalues.

If the learning algorithm in Equation (4.34) is used for optimization, it can be assumed that, for any $\mathbf{v}_k$ in the iteration, the matrix $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]$ has p positive eigenvalues $\lambda_1,\lambda_2,\cdots,\lambda_p$, with $\lambda_1>\lambda_2\geq\cdots\geq\lambda_p$ and p linearly independent eigenvectors: $\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_p$ corresponding to the eigenvalues. The projection vector $\mathbf{v}_k$ can be expressed as $\mathbf{v}_k=c_1\mathbf{u}_1+c_2\mathbf{u}_2+\cdots+c_p\mathbf{u}_p$ with $c_1,c_2,\cdots,c_p$ being scalar constants. The iteration process can be written as

$$
\begin{aligned}
\mathbf{v}_{k+1} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v}_k \\
&= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\left(c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_p\mathbf{u}_p\right) \\
&= c_1\lambda_1\mathbf{u}_1 + c_2\lambda_2\mathbf{u}_2 + \cdots + c_p\lambda_p\mathbf{u}_p \\
&= \lambda_1\left(c_1\mathbf{u}_1 + c_2\frac{\lambda_2}{\lambda_1}\mathbf{u}_2 + \cdots + c_p\frac{\lambda_p}{\lambda_1}\mathbf{u}_p\right).
\end{aligned}
\tag{4.36}
$$

where the relationship $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{u}_j = \lambda_j\mathbf{u}_j$ ($j = 1, 2, ..., p$) has been used. Since $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_p$, it can be seen that the updated projection vector $\mathbf{v}_{k+1}$ (scaled to unit length) gets closer to the dominant eigenvector $\mathbf{u}_1$ than $\mathbf{v}_k$. This means that after the iteration, the updated projection vector $\mathbf{v}_{k+1}$ moves towards the dominant eigenvector of the matrix. This trend is the same as in the power method. The difference from power method is that the matrix in this method is updated by the updated projection vector in each iteration, while the matrix in the power method is not. Kurtosis is a continuous function and the algorithm will lead to convergence to the dominant eigenvector of a matrix in which the projection vector itself is included. When convergence is reached, the projection vector corresponds to the largest eigenvalue of $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]$. This is verified by applying the algorithm to many data sets, but a rigorous mathematical proof is highly desired. As the matrix contains the projection vector and both are updated in each iteration, the iterative algorithm cannot guarantee convergence to the projection vector corresponding to the global maximum, but to a local maximum of kurtosis. Which maximum is reached depends on the initial guess of the projection vector.

Examination of Equation (4.34) reveals that if a maximum is reached by this learning algorithm, a maximum of kurtosis is found since $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$ reflects the fourth moment and $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$ reflects the reciprocal of the variance. A maximum means a combination of a small variance and a large fourth moment.

If the learning algorithm in Equation (4.35) is used, a similar interpretation follows as above. However, a significant difference should be mentioned. One might expect that,

with the iterative process, the projection vector will finally converge on the eigenvector corresponding to the largest eigenvalue of $\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$. However, this is not true, which might be because the projection vector is also employed in the inverse of $\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$, leading to an opposite effect different from that of the projection vector on the right of $\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$. The result is that when convergence is reached, the projection vector does not correspond to the largest, but the smallest (generally but not always), eigenvalue of $\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)$. Thus, when the convergence is reached, the projection vector based on the algorithms in either Equation (4.34) or Equation (4.35) generally corresponds to the largest eigenvalue of $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]$. However, the learning algorithm in Equation (4.35) can be used to search for a minimum rather than a maximum for the kurtosis value. Again, when convergence is reached, the global minimum of kurtosis cannot be guaranteed. In the eigenvalue problem, the power method can be used to search for the largest eigenvalue (in magnitude) and the inverse power method can be used to search for the smallest eigenvalue (in magnitude). The learning algorithms proposed in Equations (4.34) and (4.35) can be regarded as generalizations of these, respectively.

4.8.3 Derivation of Learning Algorithms for Multivariate Kurtosis

Starting with the definition of multivariate kurtosis for the projections of measurement vectors, **x**, into the subspace defined by **V**,

$$K = n\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}\right)^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right]^2 = n\sum_{i=1}^{n}\left\{\mathrm{tr}\left[\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}\right)^{-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]\right\}^2, \quad (4.37)$$

the objective is to optimize K with respect to the projection subspace **V**. It can be seen from Equation (4.37) that the lengths and angles among the vectors in **V** do not affect the value of multivariate kurtosis, so the optimization of multivariate kurtosis can be treated as an unconstrained problem. To simplify the expression, let

$$\mathbf{A} = \mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}. \tag{4.38}$$

Applying the Chain Rule [32] yields

$$\frac{\partial K}{\partial \mathbf{V}} = \begin{pmatrix} \frac{\partial K}{\partial v_{11}} & \cdots & \frac{\partial K}{\partial v_{1q}} \\ \frac{\partial K}{\partial v_{21}} & \cdots & \frac{\partial K}{\partial v_{2q}} \\ \vdots & \ddots & \vdots \\ \frac{\partial K}{\partial v_{p1}} & \cdots & \frac{\partial K}{\partial v_{pq}} \end{pmatrix} = 2n\sum_{i=1}^{n}\left\{\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\frac{\partial \mathrm{tr}\left[\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}\right)^{-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]}{\partial \mathbf{V}}\right\}. \tag{4.39}$$

Matrix calculus results [32] lead to

$$\frac{\partial \mathrm{tr}\left[\left(\mathbf{V}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}\right)^{-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]}{\partial \mathbf{V}} = -2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}\mathbf{A}^{-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\mathbf{A}^{-1} + 2\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{V}\mathbf{A}^{-1}. \tag{4.40}$$

Substitution of Equation (4.40) back to Equation (4.39) gives

$$\frac{\partial K}{\partial \mathbf{V}} = 4n\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left[-\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}\mathbf{A}^{-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\mathbf{A}^{-1} + \left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{V}\mathbf{A}^{-1}\right]. \tag{4.41}$$

Setting this to be equal to zeros gives the result

$$\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\mathbf{V}\mathbf{A}^{-1} = \sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}\mathbf{A}^{-1}\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\mathbf{A}^{-1}. \tag{4.42}$$

Post-multiplying by **A** on both sides of the equation and rearranging the scalars yield

$$\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{V} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}\mathbf{A}^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]. \tag{4.43}$$

It is apparent that Equation (4.43) reduces to Equation (4.31) if **V** is a vector, so the latter is actually a special case of the former, as anticipated. Although the deviation for univariate kurtosis in Section 4.8.1 is different, the derivatives of univariate kurtosis can be obtained in the same way as for multivariate kurtosis.

Obviously, there are no closed-form solutions for **V** in Equation (4.43), and **V** needs to be found through iterative methods. With the assumption of invertibility of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, rearrangement of Equation (4.43) gives the following two formulas:

$$\mathbf{V} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{V}\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]^{-1}\mathbf{A}, \tag{4.44}$$

$$\mathbf{V} = \left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}\mathbf{A}^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]. \quad (4.45)$$

Examination of these two equations reveals that the term $\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]^{-1}\mathbf{A}$ in Equation (4.44) and the term $\mathbf{A}^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)\left(\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\right)\right]$ in Equation (4.45) only change the basis for the subspace (a plane or a hyperplane) and the subspace itself does not change. As multivariate kurtosis is independent of the choice of the basis for a subspace, these two terms in the two equations are not necessary. The learning algorithms can be simplified to

$$\mathbf{V}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left\{\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\mathbf{V}_k\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}_k\right)^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{x}_i\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right\}\mathbf{V}_k, \text{ and} \quad (4.46)$$

$$\mathbf{V}_{k+1} \leftarrow \left\{\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\mathbf{V}_k\left(\mathbf{V}_k^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}_k\right)^{-1}\mathbf{V}_k^{\mathrm{T}}\mathbf{x}_i\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right\}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{V}_k, \quad (4.47)$$

where k is the iteration number and $\mathbf{A}$ is substituted by $\mathbf{V}_i^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{V}_k$ by Equation (4.38).

4.8.4 Relationship to Peña and Prieto's Algorithm

The shifted algorithm proposed in this work for univariate kurtosis optimization has some similarities to the algorithm proposed by Peña and Prieto [15]. The two methods share some common features, such as the use of the first-order derivative in the derivation and iterative search methods in the optimization. At the core of the algorithms, the equations to be solved are essentially the same (Equation (4.32) in this work and their unnumbered equation following Equation (12)). However there are some important differences, as described below.

The derivation of Peña and Prieto starts with sphered data, leading to a constrained problem requiring the projection vector to have a unit length. This removes the denominator in Equation (4.25) in this work and simplifies the problem. The solution in this work is unconstrained and does not require the projection vector to have unit length. This makes the problem more complicated, but more general. Sphering of the data changes the projections. Sphering may be used for the algorithms in this work, but *must* be used for the algorithm of Peña and Prieto.

For maximization of univariate kurtosis, Peña and Prieto solve the problem using Lagrange multipliers, requiring an iterative decomposition of a weighted covariance matrix to find the eigenvector with the largest eigenvalue. The method in this work imposes a direct iterative solution of the unconstrained problem based on a generalization of the power method.

Peña and Prieto do not explicitly describe methods for kurtosis minimization and the author of this thesis was not able to successfully implement their method for this purpose based on their description. Also, they do not extend their method to the optimization of multivariate kurtosis.

Finally, the methods proposed in this work allow the use of shifted algorithms for numerical stabilization.

4.8.5 Relationship to the Fixed-Point Algorithm

The algorithms proposed in this work are also close to the fixed-point algorithm and its variants [19,23,25,26,27,28]. The fixed-point algorithm and its variants are also based on the derivatives and are iterative methods. However some important differences are worth noting.

The fixed-point algorithm generally works on sphered data [19,23] although some of its variants for non-sphered data [25,26,28] have been proposed. Sphering seems to simplify the expression of kurtosis, but changes the unconstrained optimization problem to a constrained optimization problem, which actually makes the optimization more complicated. The algorithms proposed in this work do not require that the data are sphered and sphered data can be regarded as a simplified case in this work.

The fixed-point algorithm includes the "-3" in the definition of univariate kurtosis, which is not included in the kurtosis definition in this work. The fixed-point algorithm proposed in the original work [23] can be written as

$$\mathbf{v}_{k+1} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\right)^3 - 3\mathbf{v}_k\,, \tag{4.48}$$

which is similar to the shifted learning algorithm in Equation (4.12) in this work. In Equation (4.12), c is required to be a positive number, but in their algorithm (Equation (4.48)), "-3" is negative. Due to the introduction of the negative number "-3", the search for a maximum or a minimum of kurtosis may be subject to complications. In other

words, if one wants to search for a maximum of kurtosis, a minimum may be found. This has led to the modifications to choose a step size for the fixed-point algorithm [27,28]. However, optimization of step size makes the problem more difficult.

The third major difference between the fixed-point algorithm and the proposed algorithms in this work relates to the simultaneous extraction of more than one projection vector. In the iteration process, the fixed-point algorithm searches for individual projection vectors and then uses symmetric orthogonalization to make them mutually orthogonal [19,23,25,26]. If the data are sphered and there are two projection vectors, denoted by $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix}$, to be extracted, the fixed-point algorithm actually optimizes

$$\begin{aligned} K &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\mathbf{x}_i\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\mathbf{x}_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\mathbf{x}_i\right)^2 + \left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\mathbf{x}_i\right)^2\right]. \end{aligned} \tag{4.49}$$

For the algorithms proposed in this work, multivariate kurtosis is used as the objective function. If the data are sphered, the objective function (multivariate kurtosis) can be expressed as

$$K = n\sum_{i=1}^{n}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{V}\mathbf{V}^{\mathrm{T}}\mathbf{x}_i\right)^2, \tag{4.50}$$

which can be re-written as

$$\begin{aligned} K &= n\sum_{i=1}^{n}\left[\mathbf{x}_i^{\mathrm{T}}\left(\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}} + \mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\right)\mathbf{x}_i\right]^2 \\ &= n\sum_{i=1}^{n}\left[\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_1\mathbf{v}_1^{\mathrm{T}}\mathbf{x}_i\right) + \left(\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_2\mathbf{v}_2^{\mathrm{T}}\mathbf{x}_i\right)\right]^2. \end{aligned} \tag{4.51}$$

Disregarding the constant n, it is obvious that the objective functions optimized by the fixed-point algorithm and multivariate kurtosis in this work are different. The two objective functions might give similar results in some cases, but differences are expected in general. It can be seen that the objective function optimized by the fixed-point algorithm is essentially the sum of two univariate kurtoses. The objective function is not rotationally invariant; that is, rotation of the coordinate axes will lead to a different value. Multivariate kurtosis is rotationally invariant and is expected to give a different result from the sum of two univariate kurtoses.

4.9 Bibliography

1. D. M. Glover, and P. K. Hopke, Exploration of Multivariate Chemical Data by Projection Pursuit, *Chemometrics and Intelligent Laboratory Systems*, 16 (1992) 45-59.
2. M. Hubert, P. J. Rousseeuw, and S. Verboven, A Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60 (2002) 101-111.
3. M. Daszykowski, I. Stanimirova, B. Walczak, and D. Coomans, Explaining a Presence of Groups in Analytical Data in Terms of Original Variables, *Chemometrics and Intelligent Laboratory Systems*, 78 (2005) 19-29.
4. J. H. Friedman, and J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions and Computers*, 23 (1974) 881-890.
5. J. B. Kruskal, Toward a Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation Which Optimizes a New "Index of Condensation", in R. C. Milton, and J. A. Nelder (Editors): *Statistical Computation*, Academic Press, New York, 1969.
6. J. B. Kruskal, Linear Transformation of Multivariate Data to Reveal Clustering, in R. N. Shepard, A. K. Romney, and S. B. Nerlove (Editors): *Multidimensional Scaling: Theory and Applications in the Behavioral Science*, Seminar Press, New York, 1972.
7. C. G. Posse, Projection Pursuit Discriminant Analysis for Two Groups, *Communications in Statistics - Theory and Methods*, 21 (1992) 1-19.
8. P. J. Huber, Projection Pursuit, *The Annals of Statistics*, 13 (1985) 435-475.
9. M. C. Jones, and R. Sibson, What is Projection Pursuit? *Journal of the Royal Statistical Society, Series A* (*General*), 150 (1987) 1-36.
10. J. H. Friedman, Projection Pursuit, *Journal of the American Statistical Association*, 82 (1987) 249-266.
11. P. Hall, On Polynomial-Based Projection Indices for Exploratory Projection Pursuit, *The Annals of Statistics*, 17 (1989) 589-605.
12. S. C. Morton, *Interpretable Projection Pursuit*, SLAC Report-355, Stanford Linear Accelerator Center, Standford University, California, 1989.
13. C. Posse, An Effective Two-Dimensional Projection Pursuit Algorithm, *Communications in Statistics - Simulation and Computation*, 19 (1990) 1143-1164.
14. I. S. Yenyukov, Indices for Projection Pursuit, in E. Diday (Editors): *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science Publishers, New York, 1989, pp. 181-189.
15. D. Peña, and F. J. Prieto, Cluster Identification Using Projections, *Journal of American Statistical Association*, 96 (2001) 1433-1445.

16. D. Peña, and F. J. Prieto, Multivariate Outlier Detection and Robust Covariance Matrix Estimation, *Technometrics*, 43 (2001) 286-310.

17. M. Hubert, Multivariate Outlier Detection and Robust Covariance Matrix Estimation: Discussion, *Technometrics*, 43 (2001) 303-306.

18. J. V. Stone, *Independent Component Analysis: A Tutorial Introduction*, Cambridge, The MIT Press, 2004.

19. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., New York, 2001.

20. N. Delfosse, and P. Loubaton, Adaptive Blind Separation of Independent Sources: A Deflation Approach, *Signal Processing*, 45 (1995) 59-83.

21. A. Hyvärinen, and E. Oja, A Neuron That Learns to Separate One Signal from a Mixture of Independent Sources, in *IEEE International Conference on Neural Networks,* Washington DC, 1996, pp. 62-67.

22. C. Croux, and A. Ruiz-Gazen, A Fast Algorithm for Robust Principal Components Based on Projection Pursuit, *COMPSTAS: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 1996, pp. 211-217.

23. A. Hyvärinen, and E. Oja, A Fast Fixed-Point Algorithm for Independent Component Analysis, *Neural Computation*, 9 (1997) 1483-1492.

24. P. A. Tukey, and J. W. Tukey, Preparation; Prechosen Sequences of Views, in V. Barnett (Editors), *Interpreting Multivariate Data*, John Wiley, New York, 1981, pp. 189-213.

25. A. Hyvärinen, A Family of Fixed-Point Algorithms for Independent Component Analysis, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp. 3917-3920.

26. A. Hyvärinen, Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, *IEEE Transactions on Neural Network*s, 10 (1999) 626-634.

27. P. A. Regalia, and E. Kofidis, Monotonic Convergence of Fixed-Point Algorithms for ICA, *IEEE Transactions on Neural Network*s, 14 (2003) 943-949.

28. V. Zarzoso, and P. Comon, Robust Independent Component Analysis by Iterative Maximization of the Kurtosis Contrast with Algebraic Optimal Step Size, *IEEE Transactions on Neural Network*s, 21 (2010) 248-261.

29. G. Brys, M. Hubert, and P. J. Rousseeuw, A Robustification of Independent Component Analysis, *Journal of Chemometrics*, 19 (2005) 364-375.

30. J. Stewart, *Calculus: Early Transcendentals*, Fourth Edition, Brooks/Cole Publishing Company, New York, 1999.

31. J. R. Magnus, and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Inc., New York, 1988, pp. 177.

32. K. B. Petersen, and M. S. Pedersen, *The Matrix Cookbook*, Version: Nov. 14, 2008, http://matrixcookbook.com. Last access on April 1, 2011.

33. D. Poole, *Linear Algebra: A Modern Introduction*, Brooks/Cole, 2003.

34. A. Höskuldsson, PLS Regression Methods, *Journal of Chemometrics*, 2 (1998) 211-228.

35. C. Posse, Tools for Two-Dimensional Exploratory Projection Pursuit, *Journal of Computational and Graphical Statistics*, 4 (1995) 83-100.

36. G. Nason, Three-Dimensional Projection Pursuit, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 44 (1995) 411-430.

37. G. Nason, *Design and Choice of Projection Indices*, Ph. D. thesis, University of Bath, 1992.

38. K. V. Mardia, Measures of Multivariate Skewness and Kurtosis with Applications, *Biometrika*, 57 (1970) 519-530.

39. J. Johansson, Measuring Homogeneity of Planar Point-Pattern by Using Kurtosis, *Pattern Recognition Letters*, 21 (2000) 1149-1156.

40. S. N. Afriat, Orthogonal and Oblique Projections and the Characteristics of Pairs of Vector Spaces, *Mathematical Proceedings of the Cambridge Philosophy Society*, 53 (1957) 800-816.

41. J. Miao, and A. Ben-Israel, On Principal Angles between Subspaces in R^n, *Linear Algebra and Its Applications*, 171 (1992) 81-98.

42. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.

43. http://www.models.kvl.dk/. Last access on April 1, 2011.

44. J. Christensen, E. M. Becker, and C. S. Frederiksen, Fluorescence Spectroscopy and PARAFAC in the Analysis of Yogurt, *Chemometrics and Intelligent Laboratory Systems*, 75 (2005) 201-208.

45. T. K. Karakach, P. D. Wentzell, and J. A. Walter, Characterization of the Measurement Error Structure in 1D ^{1}H NMR Data for Metabolomics Study, *Analytica Chimica Acta*, 636 (2009) 163-174.

46. M. Forina, and C. Armanino, Eigenvector Projection and Simplified Non-Linear Mapping of Fatty Acid Content of Italian Olive Oils, *Annali di Chimica*, 72 (1982) 127-141.

47. M. Forina, and E. Tiscornia, Pattern Recognition Methods in the Prediction of Italian Olive Oil Origin by Their Fatty Acid Content, *Annali di Chimica*, 72 (1982) 143-155.

48. M. P. Derde, and D. L. Massart, Supervised Pattern Recognition: The Ideal Method? *Analytica Chimica Acta*, 191 (1986) 1-16.

49. J. Zupan, M. Novič, X. Li, and J. Gasteiger, Classification of Multicomponent Analytical Data of Olive Oils Using Different Neural Networks, *Analytica Chimica Acta*, 292 (1994) 219-234.

50. M. Daszykowski, B. Walczak, and D. L. Massart, A Journey into Low-Dimensional Spaces with Autoassociative Neural Networks, *Talanta*, 59 (2003) 1095-1105.

51. http://www2.ccc.uni-erlangen.de/publications/ANN-book/download/572oils.dat. Last access on April 1, 2011.

52. R. A. Horn, and C. A. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985, pp. 465.

Chapter 5: Development of Regularized Projection Pursuit for Data with a Small Sample-to-Variable Ratio

5.1 Introduction

For multivariate data analysis in chemistry and many other areas, exploratory data analysis [1,2] is often the initial and important step to extract information from the data. As discussed in Chapter 4, projection pursuit (PP) is an important method for exploratory data analysis and often outperforms principal component analysis (PCA). PP has not been widely used in chemistry compared with PCA or other methods of clustering analysis. This is largely due to the complexity and optimization difficulty of the projection indices. Among the different projection indices, kurtosis is relatively simple and has the property of affine invariance in the sense that scaling and translation do not affect its value. Chapter 4 of this thesis reported new methods to optimize kurtosis as a projection index that have been published in reference [3]. These methods improve the utility of PP and make one variant of the projection pursuit technique readily adaptable to different applications. It was demonstrated in that work that the results obtained by PP show significant improvement over PCA when the number of samples is large compared to the number of variables.

When the sample-to-variable ratio is small, minimization of kurtosis can still separate the samples into different clusters, but the separation is often meaningless because it does not reflect real underlying factors but is the result of random associations of samples in a high-dimensional variable space. In other words, because the number of variables is relatively high, the samples can be clustered in the subspace spanned by the projection vectors (projection directions), but this subspace is not related to a meaningful data structure. This is similar to the over-fitting problem in regression analysis [4] and linear discriminant analysis [5,6], where the training data fit the model very well, but the model has poor predictive performance. One reason for the over-fitting problem in regression and linear discrminant analysis is the small sample-to-variable ratio, which also causes problems for kurtosis used as a projection index.

The “over-fitting-like” problem for kurtosis as projection index requires many more samples than variables to give meaningful results. However, today’s experiments

often have fewer samples but more variables, so the sample-to-variable ratio is small. One possible solution to this problem is to choose fewer variables, but the selection of these variables is generally difficult and the variables that contain useful information may be excluded. Another possible solution is to apply PCA to reduce the dimensionality of the data and to use a small number of principal components of PCA for PP. This works for some situations [3], but is subject to potential loss of useful information that is not contained in the first few principal components. In general, the small sample-to-variable ratio limits the utility of PP and the advantage of PP over PCA is diminished.

In this work, a new alternative projection pursuit method using regularized kurtosis as a projection index, referred to as regularized project pursuit (RPP), is proposed. It is designed to deal with data that have a small sample-to-variable ratio and to solve the "over-fitting-like" problem. Its optimization algorithms are based on slight modifications of the quasi-power methods [3]. The proposed RPP is applied to simulated and experimental data and the results show that the "over-fitting-like" problem can, to a great extent, be solved to discover meaningful information that cannot be obtained through PCA or ordinary PP.

5.2 Background

Regularization [7,8] is a technique that is widely used in many methods to solve the ill-posed problems. Typical applications include ridge regression (RR) [5,9,10,11,12] and regularized discriminant analysis [13,14]. For a multiple linear regression problem

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y}$ represents the response variable, $\mathbf{X}$ denotes the regressor variables, $\boldsymbol{\beta}$ designates the regression coefficients, and $\boldsymbol{\varepsilon}$ indicates the error term. The estimator ($\hat{\boldsymbol{\beta}}$) for the regression coefficients through the least squares method is given as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

This estimator is unbiased, but the estimated regression coefficients are very unstable if high multicollinearity exists in $\mathbf{X}$. This is because $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is ill-posed in the sense that $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ is very sensitive to changes in $\mathbf{X}$. In the extreme case, when the number of variables is greater than the number of samples, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ will be singular and the inverse cannot be calculated. These problems can be mitigated by adding a positive number, $k_{\mathrm{R},}$ called

biasing or ridge parameter (normally small), to the diagonal elements of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and the regression coefficients can be estimated through the biased estimator ($\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$) called ridge estimator

$$\widehat{\boldsymbol{\beta}}_{\mathrm{R}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + k_{\mathrm{R}}\mathbf{I}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y},$$

where $\mathbf{I}$ is the identity matrix. This regularized regression method is labeled "ridge regression". The benefit of ridge regression is that the estimated regression coefficients become much more stable at a small expense that the estimator is biased. The ridge estimator has an important interpretation from the Bayesian point of view [5, 15]. Statistically, if it is assumed that there is *a priori* information for $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $(\sigma^2/k)\mathbf{I}$ ($\boldsymbol{\beta} \sim N(\mathbf{0}, (\sigma^2/k)\mathbf{I})$), maximizing the *a posteriori* probability leads to the ridge regression solution. The ridge regression is in accordance with the maximum *a posteriori* probability while the normal multiple linear regression follows the principle of maximum likelihood.

Another important application of regularization is for regularized discriminant analysis. For linear discriminant analysis (LDA) or quadratic discriminant analysis (QDA) [16], when the sample size is small compared to the number of variables, the within-group covariance matrices estimated from samples will highly variable. If the number of samples is smaller than the number of variables, the within-group covariance matrices will be singular and LDA or QDA cannot be used directly. Either case represents an ill-posed problem. A key aspect of the solution to the problems is to regularize the within-group covariance matrices by adding a positive number to the diagonal elements of the matrices, which is similar to the method in ridge regression.

Regularization is also used in other methods, such as support vector machine (SVM) [17], the least absolute shrinkage and selection operator (LASSO) [18], maximum likelihood principal component analysis (MLPCA) [19], and regularized independent component analysis (ICA) [20]. Application of regularization in supervised methods (*e.g.* regression and discriminant analyses) also helps to mitigate the over-fitting problem [5] and improves the predictive performance of models.

In general, regularization can be regarded as introduction of a penalty term to the objective function. In many situations, the regularization term can also be interpreted as imposing *a priori* information on the model from Bayesian point of view and/or the addition of random noise [21].

5.3 Theory

5.3.1 Regularized Univariate Kurtosis

For a multivariate data set **X** with *n* samples and *p* variables, PP looks for a unit length projection vector **v** such that the kurtosis (*K*) of projected data onto this projection vector is optimized, expressed mathematically as

$$K = \frac{n\sum_{i=1}^{n}\left(\mathbf{v}^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}\right)^2}{\left(\mathbf{v}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}\right)^2}, \tag{5.1}$$

where it is assumed that the data have been mean-centered and a sample measured on *p* variables is denoted by $\mathbf{x}_i$ [3]. The quasi-power methods show that maximization and minimization of kurtosis can be implemented through

$$\mathbf{v}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v}_k \text{, and} \tag{5.2}$$

$$\mathbf{v}_{k+1} \leftarrow \left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}_k, \tag{5.3}$$

respectively, where the symbol "$\leftarrow$" means that the right-hand terms are calculated and replace the left-hand terms at each iteration, and *k* stands for the iteration number. It can be seen that the algorithm in Equation (5.2) involves the inverse of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and the algorithm in Equation (5.3) involves the inverse of the matrix $\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)$. When the sample-to-variable ratio is small, both matrices tend to be ill-posed and their inverses may become very sensitive to the changes in **X**, which causes the obtained projection vector **v** to be highly uncertain and PP fails to reveal meaningful data structure. Mathematically, the ill-posed problems in PP are the same as those in ridge regression and regularized discriminant analysis. Following an idea similar to the regularization in ridge regression, the regularized kurtosis used as a projection index is proposed as

$$K_{\mathrm{R}} = \frac{n\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} + \lambda_i\mathbf{I}\right)\mathbf{v}\right]^2}{\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{v}\right]^2}, \tag{5.4}$$

where **I** is the identity matrix with

$$\lambda_i = \gamma\frac{\mathrm{tr}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)}{p}, \text{ and} \tag{5.5}$$

$$\lambda = \gamma\frac{\mathrm{tr}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)}{p}, \tag{5.6}$$

respectively. The notation, γ, denote a non-negative quantity called the "regularization coefficient" in this work, and tr is the trace operator. Introduction of the regularization terms $\lambda_i\mathbf{I}$ and $\lambda\mathbf{I}$ in Equation (5.4) sets lower bounds for the fourth moment (numerator) and the squared variance (denominator) terms and can be interpreted as additions of white noise to the measurements. For the data matrix **X**, if the sample size is small, it is very likely to find a direction such that the projected data onto this direction have very small values for the variance and the fourth moment. In theory, both the variance and the fourth moment have lower bounds and it is impossible for them to be very small. This is because there are always errors (*e.g.* sampling errors and measurement errors) associated with the observed data and these errors will be propagated into the projected data and set lower bounds for the variance and the fourth moment. When the sample size is large, it is less likely to get small values for them. When the sample size is small, it is more likely to get very small values for the fourth moment and the variance. Because the very small values are not good estimates of the true values, *i.e.*, over-estimated values, the fourth moment and variance should be offset by regularization terms which are interpreted as random noise in this work. If the errors are assumed to be independent, the error variance will be positive and the error covariance will be zero. This is reflected in $\lambda_i\mathbf{I}$ and $\lambda\mathbf{I}$, with their off-diagonal elements being zeros in Equation (5.4). Introduction of the regularization terms prevents the variance and the fourth moment of the projected data onto any direction from being too small and thus reduces the possibility of the projection vectors to be unreasonably estimated.

Optimization of the regularized kurtosis can be performed by adapting the quasi-power methods [3] with slight modifications. As $\lambda_i\mathbf{I}$ and $\lambda\mathbf{I}$ in Equation (5.4) are not

functions of **v**, the derivative of the regularized kurtosis with respect to **v** can be expressed as

$$\frac{\partial K_{\mathrm{R}}}{\partial \mathbf{v}} = \frac{4n\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}}{\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{v}\right]^2} - \frac{4n\left\{\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]^2\right\}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{v}}{\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{v}\right]^3} \quad (5.7)$$

by following the same procedure as for the quasi-power methods and using the calculus results [22,23]. Setting this to zero followed by the similar rearrangement in the quasi-power methods gives the following learning algorithms

$$\mathbf{v}_{k+1} \leftarrow \left\{\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\right\}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{v}_k \text{, and} \quad (5.8)$$

$$\mathbf{v}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\left\{\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\right\}\mathbf{v}_k \quad (5.9)$$

for minimization and maximization of kurtosis, respectively. The terms $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)$ in Equation (5.8) and $\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)$ in Equation (5.9) do not involve inversion and the solution of **v** is not very sensitive to them; therefore, regularization of the two terms is actually not very important and thus the algorithms can be simplified to

$$\mathbf{v}_{k+1} \leftarrow \left\{\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\right\}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)\mathbf{v}_k \text{, and} \quad (5.10)$$

$$\mathbf{v}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\left[\sum_{i=1}^{n}\left(\mathbf{v}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{v}_k\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]\mathbf{v}_k . \quad (5.11)$$

The simplified forms actually show that for the purpose of minimization of kurtosis, only the regularization term $\lambda_i\mathbf{I}$ in the numerator of Equation (5.4) is needed, and for the purpose of maximization of kurtosis, only the regularization term $\lambda\mathbf{I}$ in the denominator of Equation (5.4) is required. The simplification reduces the computation steps in the iteration and speeds up the convergence of the algorithms. Ignorance of the term $\lambda\mathbf{I}$ for Equation (5.10) also gives a new interpretation of the algorithm. It can be seen that if $\lambda_i\mathbf{I}$ is very large, the matrix $\sum_{i=1}^{n}\left[\mathbf{v}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{v}\right]\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)$ will be close to the identity matrix multiplied by a large scalar, and the regularized kurtosis value is largely determined

by the variance term $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. Thus, minimization of the regularized kurtosis essentially becomes maximization of the variance and RPP reduces to PCA.

Regularization of the matrices in Equations (5.10) and (5.11), to some extent, changes the condition number and makes the algorithms more stable. However, because the matrices still may not be well-posed, the shifted learning algorithms which are analogues of those in the power methods or the quasi-power methods [3] are recommended for the optimization of the regularized kurtosis. These shifted algorithms can be expressed as

$$\mathbf{v}_{k+1} \leftarrow \left\{ \left\{ \sum_{i=1}^{n} \left[\mathbf{v}^{\mathrm{T}} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \mathbf{v} \right] \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \right\}^{-1} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right) + c\mathbf{I} \right\} \mathbf{v}_k \text{, and} \tag{5.12}$$

$$\mathbf{v}_{k+1} \leftarrow \left\{ \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left[\sum_{i=1}^{n} \left(\mathbf{v}_k^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{v}_k \right) \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \right] + c\mathbf{I} \right\} \mathbf{v}_k \tag{5.13}$$

for minimization and maximization of kurtosis, respectively, where c can be chosen as a fraction of the traces of the first terms in the outmost brace brackets in Equations (5.12) and (5.13), respectively.

To extract two or more orthonormal projection vectors, the deflation method used in the quasi-power methods can be applied. After deflation, singular value decomposition (SVD) [24] can be used to reduce the dimensionality of the deflated data matrix to its rank. This is also applicable to the original data. If the original data have more variables than samples, using SVD to reduce the dimensionality to the rank of the data matrix and working on the low-dimensional data are recommended. This is because dimensionality reduction to the rank of the data matrix does not lose any information but it is computationally more efficient when a lower-dimensional matrix is used. If some prior knowledge about the original data is known, reducing the dimensionality of the original data matrix to a dimension lower than its rank is also safe to retain the useful information.

5.3.2 Regularized Multivariate Kurtosis

Multivariate kurtosis can be regularized by following the same principle as for univariate kurtosis. The regularized multivariate kurtosis is defined as

$$K_{\mathrm{R}} = n \sum_{i=1}^{n} \left\{ \mathrm{tr} \left[\mathbf{V}^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right) \mathbf{V} \right]^{-1} \left[\mathbf{V}^{\mathrm{T}} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \mathbf{V} \right] \right\}^2, \tag{5.14}$$

where $\mathbf{V}$ is an orthonormal basis for the subspace (a plane or a hyperplane) and λ_i and λ follow the same definitions as in Equations (5.5) and (5.6). It is also assumed that the data have been mean-centered. Following the same derivation steps as in the quasi-power methods [3] and using the calculus results [23], the derivatives of regularized multivariate kurtosis with respect to $\mathbf{V}$ can be expressed as

$$\frac{\partial K_{\mathrm{R}}}{\partial \mathbf{V}} = 4n\sum_{i=1}^{n}\left\{\begin{array}{l}\left\{\mathrm{tr}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}\right]\right\} \\ \times\left\{\begin{array}{l}-\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}\right]\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1} \\ +\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1}\end{array}\right.\end{array}\right\}. \quad (5.15)$$

To simplify the expression, let

$$a_i = \mathrm{tr}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}\right]. \quad (5.16)$$

Note a_i is a scalar. Setting $\frac{\partial K_{\mathrm{R}}}{\partial \mathbf{V}}$ to zero, followed by rearrangement yields

$$\sum_{i=1}^{n} a_{\mathrm{i}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V} = \sum_{i=1}^{n} a_{\mathrm{i}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}\right]. \quad (5.17)$$

To further simplify the expression, let

$$\mathbf{B}_i = \left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\right]^{-1}\left[\mathbf{V}^{\mathrm{T}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}\right]. \quad (5.18)$$

Note $\mathbf{B}_i$ is a matrix. Then Equation (5.17) becomes

$$\sum_{i=1}^{n} a_{\mathrm{i}}\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}\sum_{i=1}^{n}\left(a_{\mathrm{i}}\mathbf{B}_i\right). \quad (5.19)$$

Following the same idea in the quasi-power methods, the following learning algorithms for minimization and maximization, respectively are obtained:

$$\mathbf{V}_{k+1} \leftarrow \left[\sum_{i=1}^{n} a_i\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\right]^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)\mathbf{V}_k\sum_{i=1}^{n}\left(a_i\mathbf{B}_i\right), \text{ and} \quad (5.20)$$

$$\mathbf{V}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\sum_{i=1}^{n} a_i\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}+\lambda_i\mathbf{I}\right)\mathbf{V}_k\left[\sum_{i=1}^{n}\left(a_i\mathbf{B}_i\right)\right]^{-1}. \quad (5.21)$$

As $\sum_{i=1}^{n}\left(a_i\mathbf{B}_i\right)$ only changes the basis but does not affect the subspace, it can be dropped in the iteration. Following the same reason as for Equations (5.10) and (5.11), the terms $\lambda\mathbf{I}$ in Equation (5.20) and $\lambda_i\mathbf{I}$ in Equation (5.21) do not involve matrix inversion and are

ignored. The term a_i is replaced by Equation (5.16). Thus, the above algorithms in Equations (5.20) and (5.21) are simplified as

$$\mathbf{V}_{k+1} \leftarrow \left\{ \sum_{i=1}^{n} \left\{ \mathrm{tr}\left(\mathbf{V}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{V} \right)^{-1} \left[\mathbf{V}^{\mathrm{T}} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \mathbf{V} \right] \right\} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \right\}^{-1} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right) \mathbf{V}_k \text{, and} \quad (5.22)$$

$$\mathbf{V}_{k+1} \leftarrow \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left\{ \sum_{i=1}^{n} \left\{ \mathrm{tr}\left[\mathbf{V}^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right) \mathbf{V} \right]^{-1} \left(\mathbf{V}^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{V} \right) \right\} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \right\} \mathbf{V}_k . \quad (5.23)$$

For the same reason that the matrices may not be well-posed, the shifted algorithms corresponding to those in the power methods or quasi-power methods [3] are proposed as

$$\mathbf{V}_{k+1} \leftarrow \left\{ \left\{ \sum_{i=1}^{n} \left\{ \mathrm{tr}\left(\mathbf{V}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{V} \right)^{-1} \left[\mathbf{V}^{\mathrm{T}} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \mathbf{V} \right] \right\} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} + \lambda_i \mathbf{I} \right) \right\}^{-1} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} \right) + c\mathbf{I} \right\} \mathbf{V}_k \text{, and} \quad (5.24)$$

$$\mathbf{V}_{k+1} \leftarrow \left\{ \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \left\{ \sum_{i=1}^{n} \left\{ \mathrm{tr}\left[\mathbf{V}^{\mathrm{T}} \left(\mathbf{X}^{\mathrm{T}} \mathbf{X} + \lambda \mathbf{I} \right) \mathbf{V} \right]^{-1} \left(\mathbf{V}^{\mathrm{T}} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \mathbf{V} \right) \right\} \left(\mathbf{x}_i \mathbf{x}_i^{\mathrm{T}} \right) \right\} + c\mathbf{I} \right\} \mathbf{V}_k \quad (5.25)$$

for minimization and maximization, respectively, where c can be chosen as a fraction of the traces of the first terms in the outmost brace brackets in Equations (5.24) and (5.25), respectively.

The above optimization algorithms are the straightforward extensions of the quasi-power methods [3] that have been introduced in Chapter 4. In Section 5.3.1 and this section, although the algorithms for maximization are presented, minimization of kurtosis to reveal clusters is more important. Thus, the rest of this chapter will focus on minimization of kurtosis and ignores the results of maximization.

5.3.3 Choice of the Regularization Coefficient γ

Probably the most controversial issue related to regularization techniques is the choice of a suitable size for the regularization parameter. In this work, it relates to the determination of the regularization coefficient, γ, defined in Equations (5.5) and (5.6). For ridge regression, some methods to choose the biasing parameter involve the residual mean squares, but these cannot be used in this work because PP is an unsupervised technique. However, the *ridge trace* method that is used in ridge regression [9,10,12] can be adapted as an indicator to determine the suitable size of the regularization coefficient.

A ridge trace is a plot of the regression coefficients versus the biasing parameter in ridge regression. When the biasing parameter increases, the regression coefficients will

change as well. It is recommended in ridge regression that the biasing parameter should be chosen to correspond to the region where the regression coefficients change slowly and become stable.

By analogy, the coefficients of the projection vectors (or projection matrix for multivariate kurtosis hereinafter) versus the regularization coefficient γ can be plotted to determine the suitable size of γ. As changing the orientation of the projection vector to its opposite direction does not affect the kurtosis value but the signs of the coefficients of the projection vector found will change, it is necessary to adjust the signs of projection vector coefficients to keep consistency so that the signs will not lead to discontinuities in the curves in the ridge trace plot. For the ridge trace plots in this work, the first non-zero element of each projection vector was forced to be positive and the signs of other elements were adjusted with respect to the sign of the first non-zero element accordingly.

The introduction of the ridge trace plot to determine the regularization coefficient is based on the following interpretation for minimizing kurtosis. If samples are drawn from different populations that are separable, the direction that effectively separates different classes of samples is more likely to account for a large variance. In other words, if the data are projected onto this direction, the variance of the projected data is more likely to be large. This is actually the underlying principle explaining why PCA is often able to find directions to reveal clusters. If a direction is not effective for revealing class separation but gives a small kurtosis of the projected data, the variance accounted by it is more likely to be small. For regularized univariate kurtosis, the regularization term $\lambda_i\mathbf{I}$ is introduced in the fourth moments. If the variance accounted by a direction is large, the regularization term $\lambda_i\mathbf{I}$ will affect the kurtosis relatively less. With the increase of the regularization coefficient γ, the direction of the projection vector is more likely to vary smoothly and slowly. On the other hand, if the variance accounted by a direction is small, the regularization term $\lambda_i\mathbf{I}$ will affect the kurtosis dramatically. A small increase of the regularization parameter γ may quickly change the direction to a very different one. Reflected in the ridge trace plot, it is more likely that the coefficients of the projection vector with the small kurtosis will show a smooth pattern if the direction is a good one to reveal clusters and a disorderly pattern if the direction is not. The same interpretation can be used to understand the regularization coefficient for regularized multivariate kurtosis.

In practice, it is necessary to choose γ large enough to solve the ill-posed problem and make the projection vector more stable, but choosing a value of γ that is too large may reduce the sensitivity of the method and its ability to find good projection vectors. It is recommended that γ be varied between 0 and 1 with an increment of 0.05 and the ridge trace plot examined as a function of γ. If the coefficients of the projection vector follow a smoothly changing pattern and change slowly at some value of γ, this value is recommended for γ. Depending on the data, different values for γ may be tried to determine if an interesting data structure can be found. It is worth noting that, because kurtosis generally has multiple local optima and the algorithms cannot guarantee the global optima based on one search, it is necessary to have enough initial guesses to find the global optimum so that the ridge trace plot does not show a disorderly pattern caused by finding different local optima.

5.4 Experimental

5.4.1 Computational Aspects

All calculations were implemented by using programs written in MatLab® v.7.4.0 (MathWorks, Natick, MA) under Windows XP® on a 1.8 GHz computer with 3 Gb of memory.

5.4.2 Simulated Data

The simulation study in this work includes individual data sets and group data sets. The first individual data set, referred to as data set 1, is used to evaluate the proposed ridge trace method for determining the suitable size of the regularization coefficient, γ. This data set included a total of 30 samples evenly divided into two classes with 29 variables. The data of the two classes were assumed to follow a multivariate normal distribution. The population means for the first variable of the two classes were set to -1.5 and 1.5, respectively and the population means of all other variables for the two classes were set to 0's. The population covariance matrices for both classes were set to be the same and diagonal. The first diagonal element was set to 0.2. Other diagonal elements were simulated by generating 28 random numbers from the standard normal distribution N (0,1), taking the squares of them, and adding 0.1 to the squared values. Addition of 0.1 set a lower bound so that the variances were not too small. Based on the setting, class separation

occurred on the first variable. After the data were randomly drawn from the two populations, the measurement vectors were rotated using a 29 x 29 rotation matrix which was created by applying SVD to a randomly generated data set. The rotated data were mean-centered before analysis by PCA, PP, and RPP.

To evaluate whether the improved performance of RPP over PP and PCA is the general case, 100 data sets, designated as group data set 1, were generated by following the same procedure as for data set 1. The sample size, dimensionality of data, population parameters, rotation matrix, and preprocessing steps were all unchanged, except that samples were randomly drawn for 100 times to simulate 100 different data sets.

As an extension for two-class case, a second individual data set consisting of three classes, referred to as data set 2, was simulated. This data set also included 30 samples in a 29-dimensional space, with equal number of samples drawn from each of the three classes. The three class centers were set to

$$\boldsymbol{\mu}_1 = \begin{bmatrix} -2 & -1.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T,$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 2 & -1.16 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \text{ and}$$

$$\boldsymbol{\mu}_3 = \begin{bmatrix} 0 & 2.3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T,$$

respectively. The covariance matrices for the three classes were set to be the same and diagonal. The first two diagonal elements were set to 0.1 and the remaining 27 diagonal elements were randomly drawn from the standard normal distribution N (0,1) followed by taking the squares of them. To avoid too small values, 0.2 was added to each of the 27 generated values. Samples were then randomly drawn based on these parameters (class means and covariance matrix). The simulated data were rotated by using a 29 x 29 rotation matrix created by applying SVD to a randomly generated data set. The rotated data were then mean-centered.

A second group consisting of 100 data sets, designated as group data set 2, was simulated by using the same parameters and following the same procedure as for individual data set 2. The 100 data sets represent 100 realizations of the same population parameters and simulation steps.

5.4.3 Experimental Data

Three experimental data sets were used in this work to demonstrate the performance of the proposed RPP. The first data set, referred to as the soybean data, was downloaded from the website of UCI Machine Learning Repository [25]. The original work [26] aimed to select descriptors (variables) that contained sufficient information to diagnose soybean diseases in terms of macrosymptoms without using sophisticated mechanical assistance. The original study included 35 plant and environmental descriptors (such as plant stand, precipitation, and occurrence of hail for environmental descriptors, and leaf spot size, stem cankers, and seed size for plant descriptors), and 19 classes (diseases). A detailed list of the descriptors can be found in reference [26]. Actually, only 15 classes were used because the last four classes had too few samples. The website of UCI Machine Learning Repository also provided a small data set including four classes, which is a subset of the original data set [26]. The small data set contained 47 samples (10 samples for diseases 1, 2, and 3, respectively and 17 samples for disease 4) and 35 variables. The data included discrete variables. Some variables were not discriminatory. In other words, the values were the same for all the samples. In the work presented here, this small soybean data set was used because it had a small sample-to-variable ratio.

The second data set, referred to as the glomerulonephritis data, consisted of ^{1}H NMR spectral data of urine samples from glomerulonephritis patients and healthy people. The original study investigated the correlation between histopathologically accessed tubulointerstitial lesions and the urinary metabolites profiles by ^{1}H NMR [27]. The original data included 80 ^{1}H NMR spectra from patients with mild, moderate, and severe glomerulonephritis and 85 ^{1}H NMR spectra from healthy people. The data set used in this work was a subset of the original data which consisted of 25 spectra from patients with severe glomerulonephritis and 25 spectra from healthy people. This subset was selected in the course of a metabolomics study and was used as an example to demonstrate metabolomic data analysis [28]. The spectra were binned with equal bin width of 0.04 ppm in the chemical shift range of 0.2 to 10.0 ppm. The range between 4.38 and 6.30 ppm was excluded to remove the effects from suppression of water resonance or solvent exchanging protons, resulting in a 50 x 200 matrix. As this data set also had a small sample-to-variable ratio, it was chosen in this work to test the performance of RPP.

The third data set, designated as the cow diet data, was obtained from the metabolomics study of diary cows fed with increasing proportion of barley grains [29]. The diary cows were fed with increasing proportions of rolled barley grain (0%, 15%, 30%, and 45% on the dry matter basis) for a period of 11 days to adapt to the experimental diets and in the next 10 days, rumen fluid samples were collected. The samples were analyzed by ^{1}H NMR and 46 metabolite compounds were identified and their concentrations were measured. One more compound, endotoxin, was measured by the pyrochrom *limulus* amebocyte lysate assay [30]. The final data set included 47 variables and 39 samples for different proportions of barley grain (9 samples for 0% barley grain and 10 samples for 15%, 30%, and 45% barley grain, respectively).

For all the data sets, either simulated or experimental, if the number of variables was larger than the mathematical rank of the data matrix, SVD was applied and the scores were truncated to a dimensionality equal to the rank and used for analysis. For all the minimization searches by PP and RPP, the shifted algorithms were used with 200 random initial guesses. For the ridge trace plots of RPP, the regularization coefficient γ, either for univariate or multivariate kurtosis, varied from 0 to 1 with an equal increment of 0.05. Although SVD was applied if the dimensionality of the data was larger than the rank, the coefficients of projection vectors in the original spaces were used in the ridge trace plots.

5.5 Simulation Results

5.5.1 Data Set 1

For the first set of the simulated data, Figure 5.1 (a) shows the profiles of the mean-centered data and Figure 5.1 (b) shows the PCA scores plot. It can be seen that PCA shows some separation of the two classes of samples but the separation is not clear. The scores plot of PP with stepwise univariate approach is shown in Figure 5.1 (c). Although the samples are clustered into four quadrants, the clustering does not match the class separation, showing PP fails to give meaningful results. This is typical for data with a low sample-to-variable ratio, which is similar to the over-fitting problem in supervised methods. It is worth noting that the ranges of scores for PP, either the first or the second, are quite small, indicating that projection vector direction is possibly over-estimated. For RPP with stepwise univariate approach, it is necessary to choose the suitable size for the

regularization coefficient γ. The ridge trace plot for the first projection vector is shown in Figure 5.1 (d). It can be seen that when γ changes from 0 to 0.05, the projection coefficients change dramatically, indicating the projection vector changes from one direction to another. When $\gamma > 0.05$, the coefficient changes become smoother, and thus γ was chosen as 0.1 for

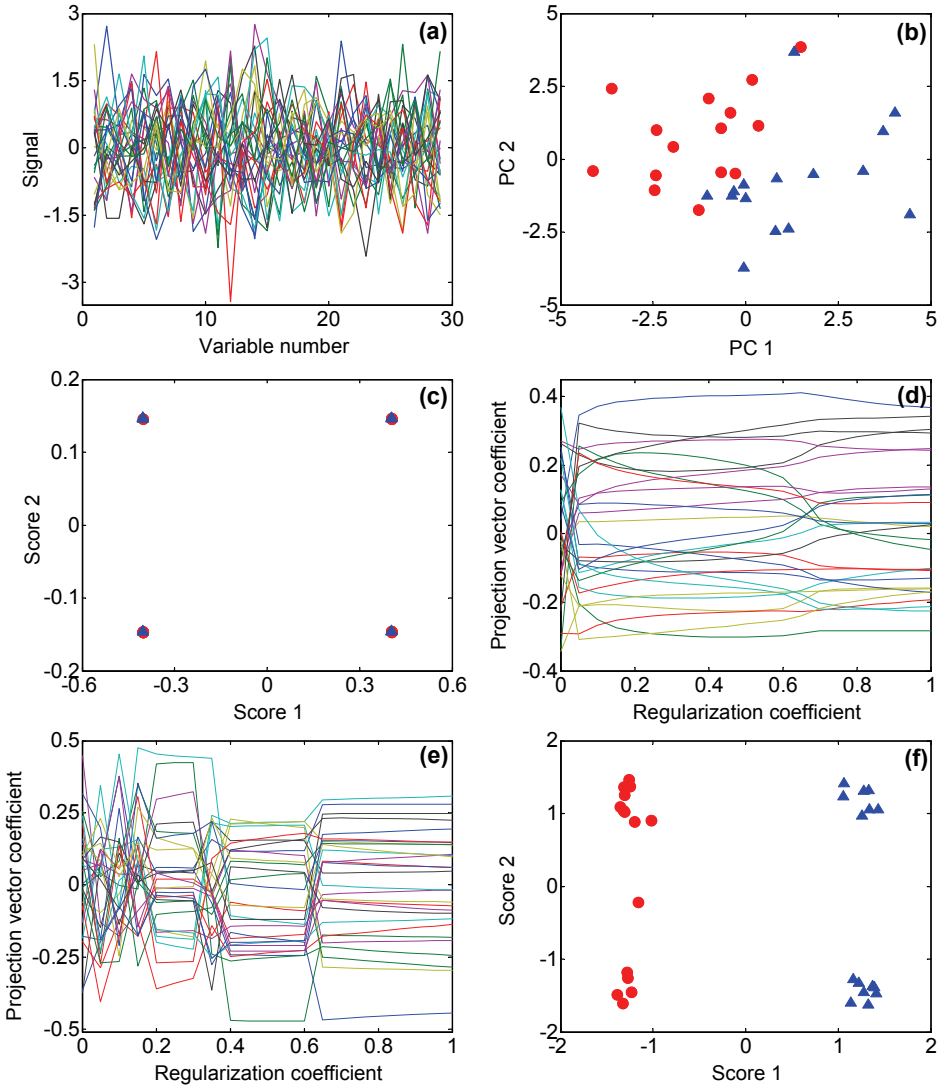

Figure 5.1 Results of data set 1 (PP methods: univariate approaches). (a) Sample data profiles, (b) PCA scores plot, (c) PP scores plot (stepwise univariate approach), (d) ridge trace plot for the 1st projection vector coefficients, (e) ridge trace plot for the 2nd projection vector coefficients, and (f) RPP scores plot (stepwise univariate approach).

the first projection vector. The ridge trace for the second projection vector (based on $\gamma = 0.1$ for the first projection vector) is shown in Figure 5.1 (e). It can be seen there are several dramatic changes happening. This is probably because the first projection vector has accounted for the major separation of the classes, and the separation in the second direction

is due to chance of sampling. In theory, it may also be because the global minimum is not hit when the number of initial guesses is not large enough. The γ value was chosen as 0.4 because the change of the second projection vector is relatively smooth when $\gamma > 0.4$. The scores plot from the stepwise univariate approach based on the above regularization coefficients is shown in Figure 5.1 (f). It can be seen that samples from the two classes are clearly separated along the first projection vector. Samples are also separated along the second projection vector, but this separation is artificial and forced by the algorithm. However, the clustering from RPP with the stepwise univariate approach does give more meaningful results than those from PCA and PP.

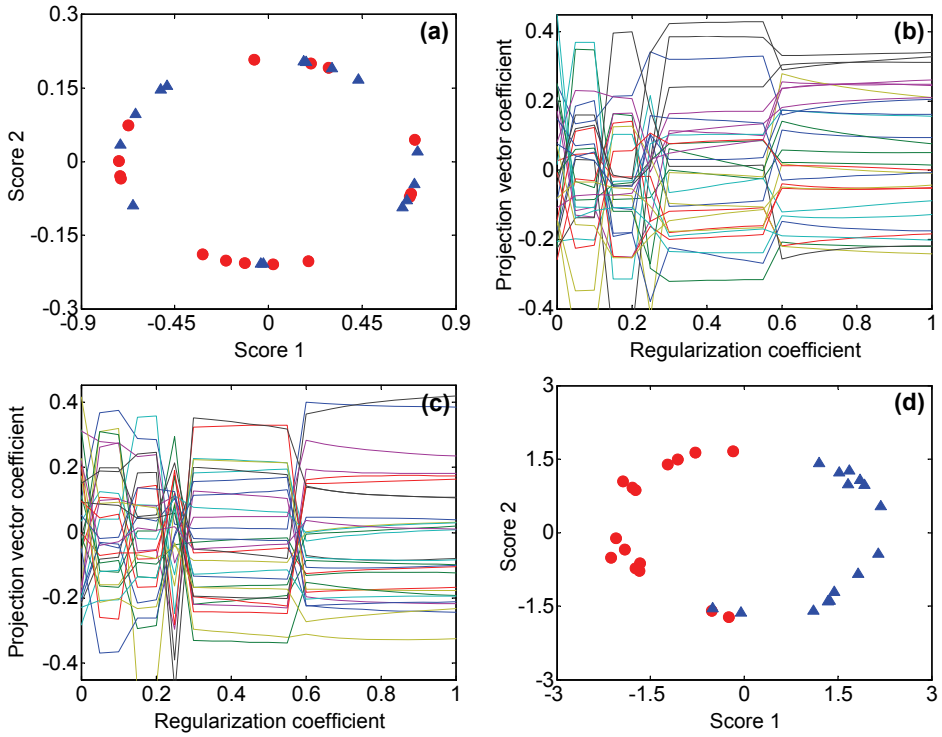

Figure 5.2 Results of data set 1 (PP methods: bivariate approaches). (a) PP scores plot (bivariate approach), (b) ridge trace plot for the 1st projection vector coefficients, (c) ridge trace plot for the 2nd projection vector coefficients, and (d) RPP scores plot (bivariate approach).

The scores plot of data set 1 from PP with the bivariate approach is shown in Figure 5.2 (a). It can be seen that samples are pushed to the edge of the circle and samples from the two classes are not well separated, as expected. For RPP of bivariate approach, the ridge trace plots for the two projection vectors are shown in Figures 5.2 (b) and (c), respectively. It can be seen that when the regularization coefficient, γ, is smaller than 0.3, the projection

vector coefficients change quickly, presenting a disorderly pattern. When the regularization coefficient is larger than 0.3, the projection vector coefficients still change, but the change shows a smooth pattern. Note that for the bivariate approach, the two projection vectors are determined simultaneously. Based on the ridge trace plots shown in Figures 5.2 (b) and (c), the regularization coefficient γ was chosen as 0.3, and the scores plot is shown in Figure 5.2 (d). Although the data points are still located on the edge of a circle, the samples from the two classes have a good separation. The ranges of scores also become much larger than those in Figure 5.2 (a). Obviously, a different subspace has been found.

5.5.2 Group Data Set 1

In Section 5.5.1, data set 1 is used to demonstrate the performance of RPP over PCA or PP. However, it is necessary to examine whether the better performance of RPP is a general case or an anomaly. Group data set 1, consisting of 100 data sets, has been used to explore the answer. For data set 1, it was found that when $\gamma > 0.05$ for the first projection vector of the stepwise univariate approach, and $\gamma > 0.3$ for the bivariate approach, the projection vector coefficients changed smoothly. For group data set 1, without choosing the regularization coefficients for the 100 data sets using the ridge trace plots individually, the regularization coefficients were arbitrarily chosen as 0.5 for the univariate and bivariate approaches for all the 100 data sets. Although this may not always be optimal, it allowed the calculation to be carried out efficiently by exploiting observed characteristics of γ for these data.

As it is not an efficient way to show hundreds of scores plots for the 100 data sets, the quality of class separation in the two-dimensional scores plot was evaluated by the generalized Fisher's discriminant value (F) [31] which can be expressed as

$$F = \mathrm{tr}(\mathbf{S}_w^{-1}\mathbf{S}_b),$$

where $\mathbf{S}_w$ and $\mathbf{S}_b$ are the within-class covariance matrix and between-class covariance matrix, respectively. This statistic has been used in Chapters 2 and 3. The larger the F value is, the better the separation. The logarithms (base 10) of the F values obtained through stepwise univariate RPP (RPP-UNI), stepwise univariate PP (PP-UNI), multivariate RPP (RPP-MULTI), multivariate PP (PP-MULTI), and PCA are shown in Figure 5.3. It can be seen that for most of the data sets, RPP with stepwise univariate

approach gives higher F values than PP (univariate and multivariate) and PCA. The RPP with multivariate approach also gives larger F values than PP and PCA in general. The results showed that the better performance of RPP is a general but not a special case.

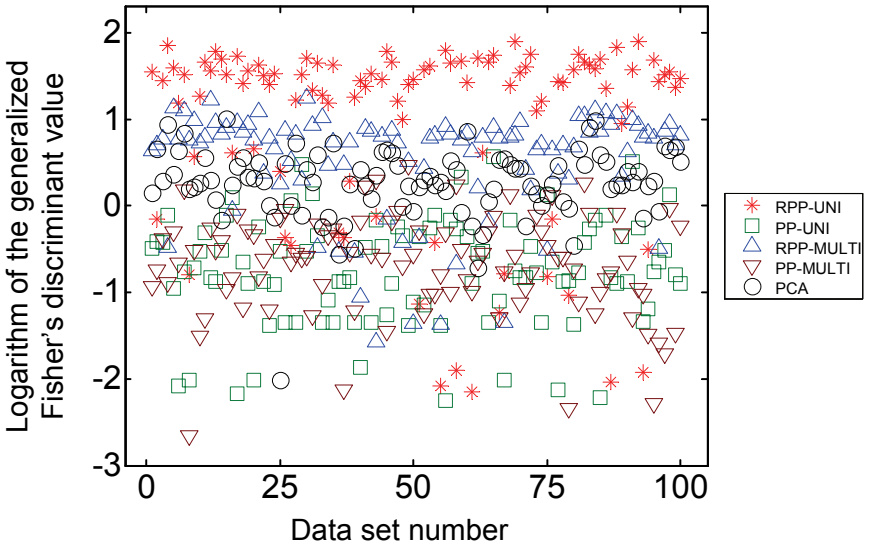

Figure 5.3 Plots of the logarithms of the generalized Fisher's discriminant values based on different methods for group data set 1.

5.5.3 Data Set 2

The results in Sections 5.5.1 and 5.5.2 indicate that the proposed RPP method overall outperform the normal PP method for two-class cases. However, it is important to examine the utility of the proposed RPP method in the case of more complex data structures. For this purpose, data set 2 consisting of samples drawn from three classes was used to test the performance of the RPP method. Figure 5.4 (a) shows the profiles of the data, and the scores plots obtained from PCA, univariate normal PP, and univariate RPP are shown in Figures 5.4 (b), (c), and (f), respectively. As the data were simulated by making the discriminatory variables have small variances, it is not surprising to see that no clear separation of clusters is observed in the PCA scores plot. The univariate normal PP fails to reveal meaningful separation, as expected. However, unfortunately the univariate RPP also fails to reveal meaningful separation. The regularization coefficients for the first and second projection vectors were chosen as 0.4 and 0.2, respectively based on the ridge trace plots (Figures 5.4 (d) and (e) show the ridge trace plots for the first and second projection vectors, respectively). The data points are separated into four clusters, but the separation of the clusters does not match the expected class separation. This might

be because minimization of the univariate kurtosis always tries to equally dichotomize the data points into two clusters. This happens for any projection directions, so the data are forced into four quadrants when two projection vectors are found in a stepwise way. This implies that when the data contain an even number of well-balanced clusters, univariate kurtosis will work well, but when the data have an odd number of clusters or unbalanced clusters, minimization of the univariate kurtosis may fail to reveal meaningful clusters. For data set 2 that contains three clusters, the failure for the univariate RPP to reveal meaningful clusters is not due to an inappropriate choice of the regularization coefficients, but because of the nature of the univariate kurtosis.

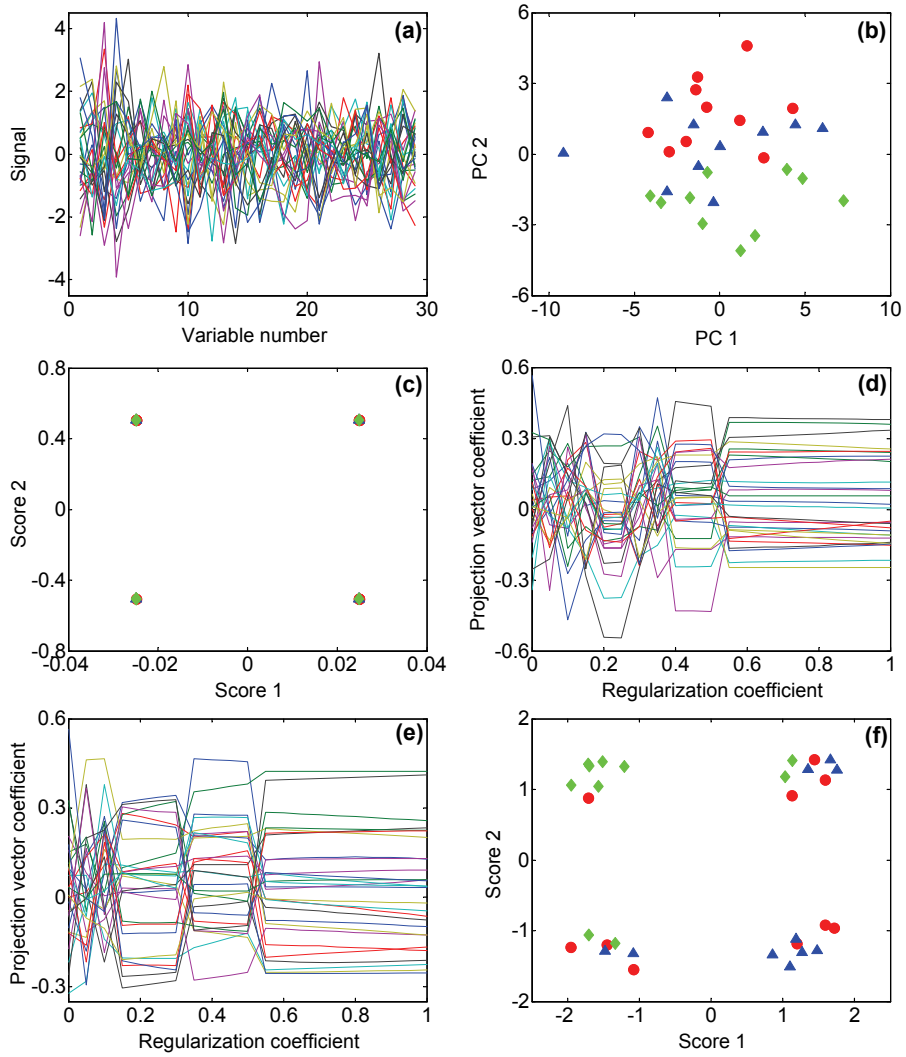

Figure 5.4 Results of data set 2 (PP methods: univariate approaches). (a) Sample data profiles, (b) PCA scores plot, (c) PP scores plot (stepwise univariate approach), (d) ridge trace plot for the 1st projection vector coefficients, (e) ridge trace plot for the 2nd projection vector coefficients, and (f) RPP scores plot (stepwise univariate approach).

Although the univariate kurtosis can be regarded as a special case of the multivariate kurtosis, the projection vectors obtained by the stepwise minimization of univariate kurtosis are generally different from those through the multivariate kurtosis. Figure 5.5 (a) shows the scores plot obtained from the PP using the multivariate kurtosis. As expected, the samples are forced to the edge of a circle and no class separation is observed. For the RPP method using the regularized multivariate kurtosis, the regularization coefficient, γ, was chosen as 0.25 based on the ridge trace plots for the two projection vectors shown in Figures 5.5 (b) and (c), respectively. Figure 5.5 (d) shows the scores plot obtained from minimization of the regularized multivariate kurtosis. Unlike the result of the univariate approach, the bivariate approach shows a clear separation of the three clusters that match the known clusters. This indicates that the multivariate kurtosis is less sensitive to unbalanced clusters. For the purpose of exploratory data analysis where the number of clusters in the data and the data structure are often unknown in advance, it is recommended to try both the stepwise univariate and multivariate (generally bivariate) approaches.

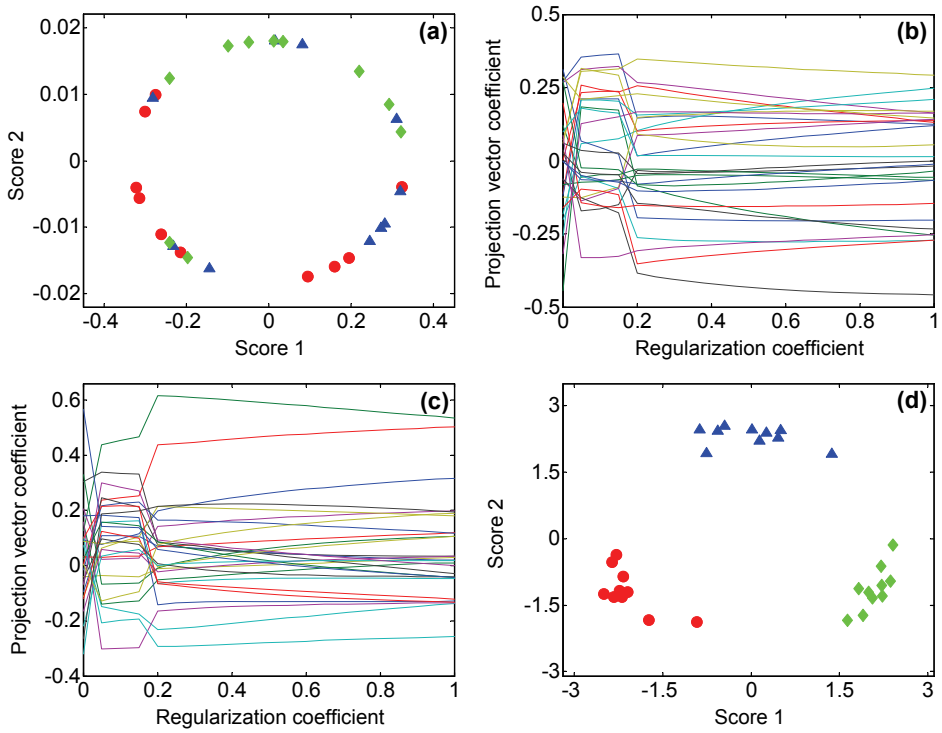

Figure 5.5 Results of data set 2 (PP methods: bivariate approaches). (a) PP scores plot (bivariate approach), (b) ridge trace plot for the 1st projection vector coefficients, (c) ridge trace plot for the 2nd projection vector coefficients, and (d) RPP scores plot (bivariate approach).

5.5.4 Group Data Set 2

Similar to the study in Section 5.5.2, 100 data sets were used to further evaluate the performance of the proposed RPP method in the case of three clusters. The regularization coefficient, γ, was also chosen as 0.5 for all the data sets without using the individual ridge trace plots for each of the data set. The generalized Fisher's discriminant value (F) [31] was also used to measure the quality of the cluster separation. Figure 5.6 shows the logarithms (base 10) of the generalized Fisher's discriminant values for the scores obtained by the five methods as described in Section 5.5.2. It can be seen that the RPP with bivariate (multivariate) approach gives higher F values for most of the data sets, indicating RPP using the regularized multivariate kurtosis has better performance than other methods. The RPP using the univariate kurtosis, however, does not give obviously better results than PCA or PP using the normal kurtosis. This is in accordance with the founding in data set 2 that univariate kurtosis does not work well in case of an odd number of clusters or unbalanced clusters.

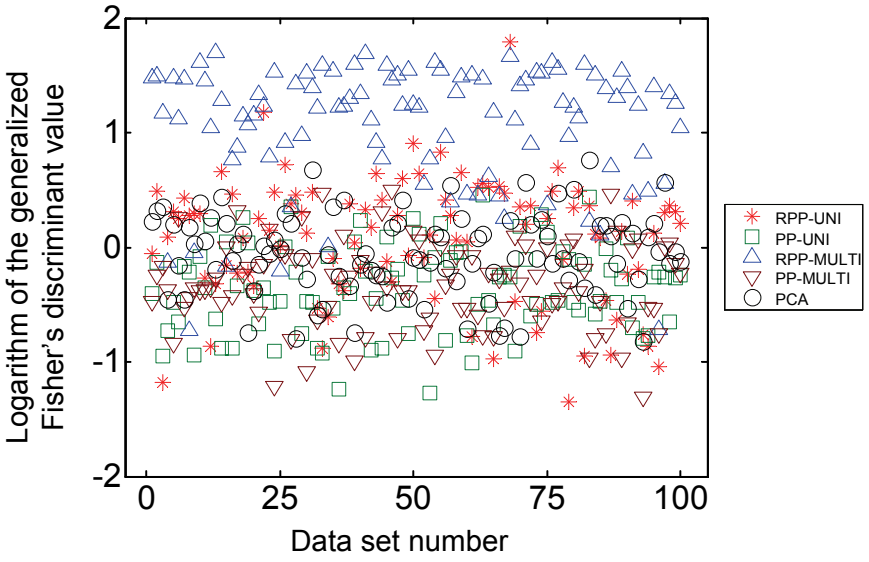

Figure 5.6 Plots of the logarithms of the generalized Fisher's discriminant values based on different methods for group data set 2.

5.6 Experimental Results

In Section 5.5, simulated data sets are used to demonstrate the performance of RPP over PP or PCA in the case of a small sample-to-variable ratio. However, this does not guarantee that RPP can give better results for experimental data because such data sets

often exhibit more complex structures. Therefore, three experimental data sets are used to evaluate the performance of RPP in this section.

5.6.1 Soybean Data

The soybean data were column mean-centered, but no scaling was performed. One reason for not performing scaling was that this data set included discrete ordered variables. The PCA scores plot is shown in Figure 5.7 (a) with the legend indicating the four types of

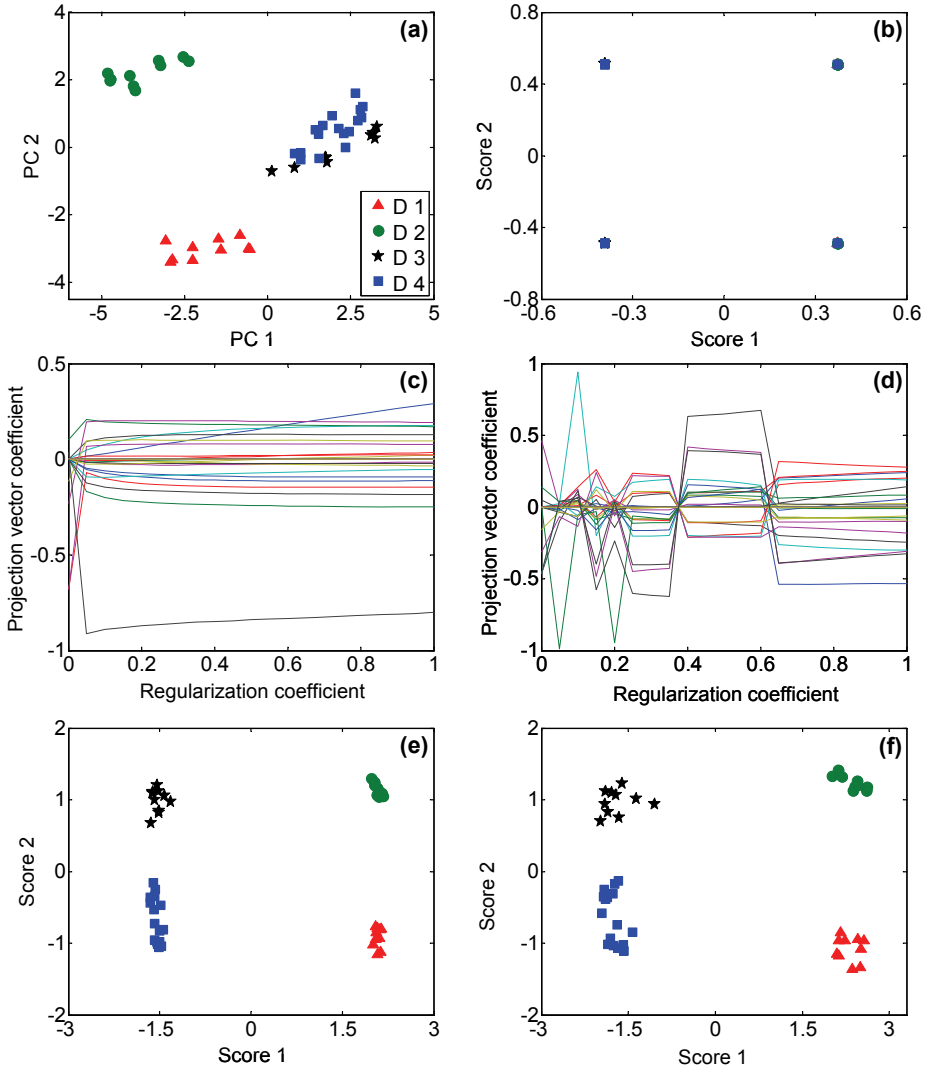

Figure 5.7 Results of soybean data (PP methods: univariate approaches). (a) PCA scores plot, (b) PP scores plot (stepwise univariate approach), (c) ridge trace plot for the 1st projection vector coefficients, (d) ridge trace plot for the 2nd projection vector coefficients, (e) RPP scores plot (stepwise univariate approach) with the 1st and 2nd regularization coefficients set to be 0.1 and 0.25, respectively, and (f) RPP scores plot (stepwise univariate approach) with the 1st and 2nd regularization coefficients both set to be 0.5.

soybean diseases. Three clusters can be clearly seen, but the samples from diseases 3 and 4 overlap. The scores plot of PP with stepwise univariate approach is shown in Figure 5.7 (b). As expected, the samples are divided into four distinct quadrants but no meaningful separation is obtained. The ranges of the scores are quite small, indicating the projection vectors are over-estimated. For RPP, the ridge trace plots for the first and second projection vectors are shown in Figures 5.7 (c) and (d), respectively. It can be seen that the projection vector coefficients for the first projection vector become smooth when $\gamma > 0.05$. Thus the regularization coefficient for the first projection vector was chosen as 0.1. Similarly, the regularization coefficient for the second projection vector was chosen as 0.25. The scores plot of RPP is shown in Figure 5.7 (e). It can be seen that the samples are clearly separated, which corresponds to the four types of soybean diseases. For comparison, Figure 5.7 (f) shows the scores plot when the two regularization coefficients were both arbitrarily set to 0.5. The scores plot is slightly different from that in Figure 5.7 (e), but a good separation is still observed. This indicates that the regularization coefficient covers a range in which the meaningful result can be obtained.

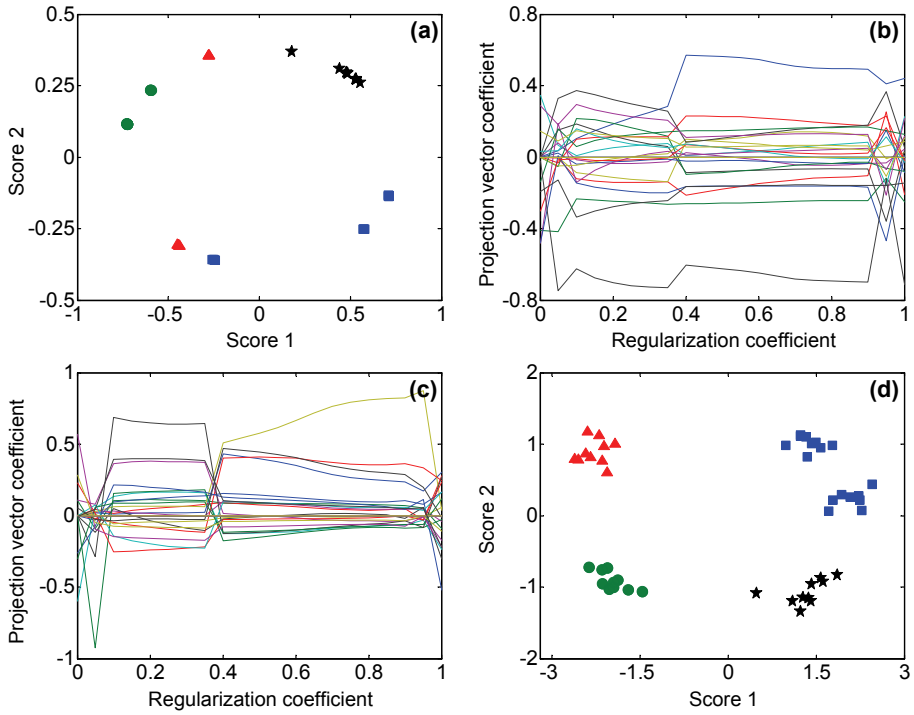

Figure 5.8 Results of soybean data (PP methods: bivariate approaches). (a) PP scores plot (bivariate approach), (b) ridge trace plot for the 1st projection vector coefficients, (c) ridge trace plot for the 2nd projection vector coefficients, and (d) RPP scores plot (bivariate approach).

The scores plot of PP with bivariate approach is shown in Figure 5.8 (a). As expected, the samples are located on the edge of the circle and the samples from the four classes are not well separated. Figures 5.8 (b) and (c) show the ridge trace plots for the two projection vectors. It can be seen that the projection vector coefficients have dramatic changes at three regions. As introduction of a larger regularization coefficient will lead to a larger bias, γ was chosen as 0.2, where the ridge trace plots show smooth patterns. The scores plot of RPP is shown in Figure 5.8 (d). It is clear that the samples with the four types of soybean diseases form four clusters that match the four classes very well. Compared with Figure 5.8 (a), it can be seen that the variances accounted by the first and second scores are much larger, indicating a different subspace has been found.

5.6.2 Glomerulonephritis Data

The original glomerulonephritis data were autoscaled in this work. Figure 5.9 (a) shows the PCA scores plot. The samples from the control group (healthy people) are clustered in a very small area but the samples from the diseased group spread more widely. There is a separation between the two classes of samples, but the separation is not very clear. The scores plot of PP with the stepwise univariate approach is shown in Figure 5.9 (b). Again, the samples occupy four distinct quadrants and the separation does not show meaningful information that matches the classes. The ridge trace plots for the first and second projection vector coefficients are shown in Figures 5.9 (c) and (d), respectively. For the first projection vector, when $\gamma > 0.15$, the coefficients become smooth. Thus γ was set to 0.2 for the first projection vector. The γ value for the second projection vector was set to 0.2 as well by observing the pattern of changes. The scores plot of RPP based on these settings is shown in Figure 5.9 (e). It can be seen that the samples from the control group and patient group are clearly separated along the first projection vector and match the class separation perfectly. The separation along the second projection vector is likely to be artificial, although the possibility that it corresponds to subclasses cannot be excluded. To examine how the regularization coefficients affect the results, the regularization coefficients were set to 0.5 for both projection vectors and the resulting scores plot is shown in Figure 5.9 (f). It can be seen that samples from the two classes are still well-separated except that one sample from the patient group falling into the control group.

This again indicates that the regularization coefficients can vary in a range and meaningful information can be revealed.

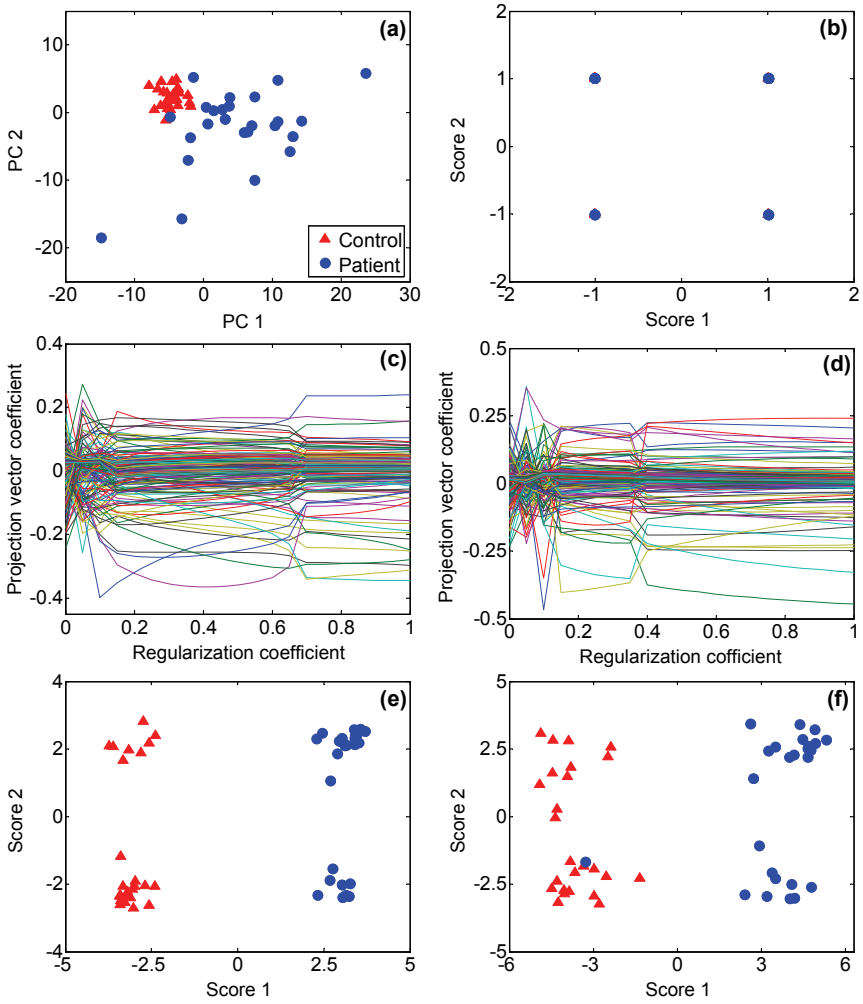

Figure 5.9 Results of glomerulonephritis data (PP methods: univariate approaches). (a) PCA scores plot, (b) PP scores plot (stepwise univariate approach), (c) ridge trace plot for the 1st projection vector coefficients, (d) ridge trace plot for the 2nd projection vector coefficients, (e) RPP scores plot (stepwise univariate approach) with the 1st and 2nd regularization coefficients both set to be 0.2, and (f) RPP scores plot (stepwise univariate approach) with the 1st and 2nd regularization coefficients both set to be 0.5.

The results of glomerulonephritis data from PP and RPP with the bivariate approach are shown in Figure 5.10. As expected, PP fails to give meaningful information as is shown by the mixing of classes in Figure 5.10 (a). The samples again are distributed at the edge of the circle due to "over-fitting". The ridge trace plots of the two projection vectors are shown in Figures 5.10 (b) and (c), respectively. The regularization coefficient was chosen as 0.25 since the plots in both figures become smooth when $\gamma > 0.2$. The scores

plot of multivariate RPP is shown in Figure 5.10 (d). The samples from the control group and patient group occupy the left and right parts of the circle, respectively. The samples from the two groups are largely separated with minor overlap (one sample). Also, the variances accounted by the two vectors are larger compared with those in Figure 5.10 (a).

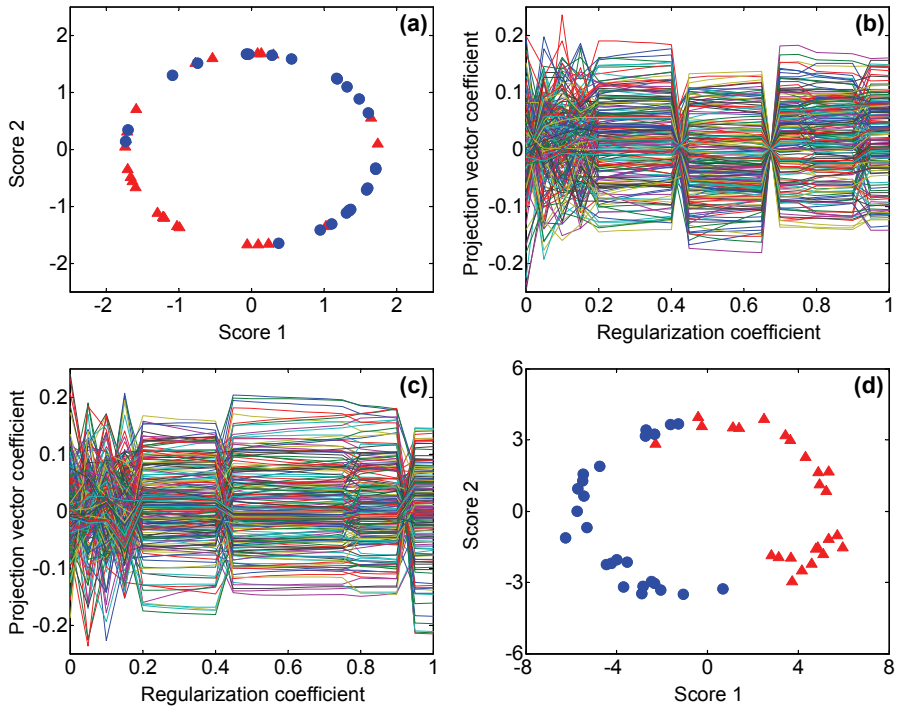

Figure 5.10 Results of glomerulonephritis data (PP methods: bivariate approaches). (a) PP scores plot (bivariate approach), (b) ridge trace plot for the 1st projection vector coefficients, (c) ridge trace plot for the 2nd projection vector coefficients, and (d) RPP scores plot (bivariate approach).

5.6.3 Cow Diet Data

The cow diet data were column mean-centered before PCA, PP and RPP were applied. The PCA scores plot is shown in Figure 5.11 (a) with the legend indicating the four different diet treatments. From the plot, the samples from different diet treatments exhibit some dispersion, but there is substantial overlap of the samples from different diet treatments. The scores plot for PP (stepwise univariate approach) is shown in Figure 5.11 (b). Again, PP fails to reveal the underlying data structure that matches the four different diet treatments. The ridge trace plots for the first and second projection vector coefficients are shown in Figures 5.11 (c) and (d), respectively. By examining the plots, the first and second regularization coefficients were set to 0.2 and 0.25, respectively. The result is shown in Figure 5.11 (e). Along the first projection vector, the samples with 0% and 15%

barley grain treatments are separated from those with 30% and 45% barley grain treatments, except one sample from the 30% treatment. The 0% and 15% barley grain treatments overlap each other, but the samples from the 30% and 45% treatments are well separated. Overall, the separation from RPP is much improved compared with those from PCA or PP. When the regularization coefficients were set to be 0.5 for both projection vectors, the scores are shown in Figure 5.11 (f). The scores plot is similar to that in Figure 5.11 (e), but obviously, a different subspace has been used. As expected, the variances accounted in Figures 5.11 (e) and (f) are much larger than those in Figure 5.11 (b).

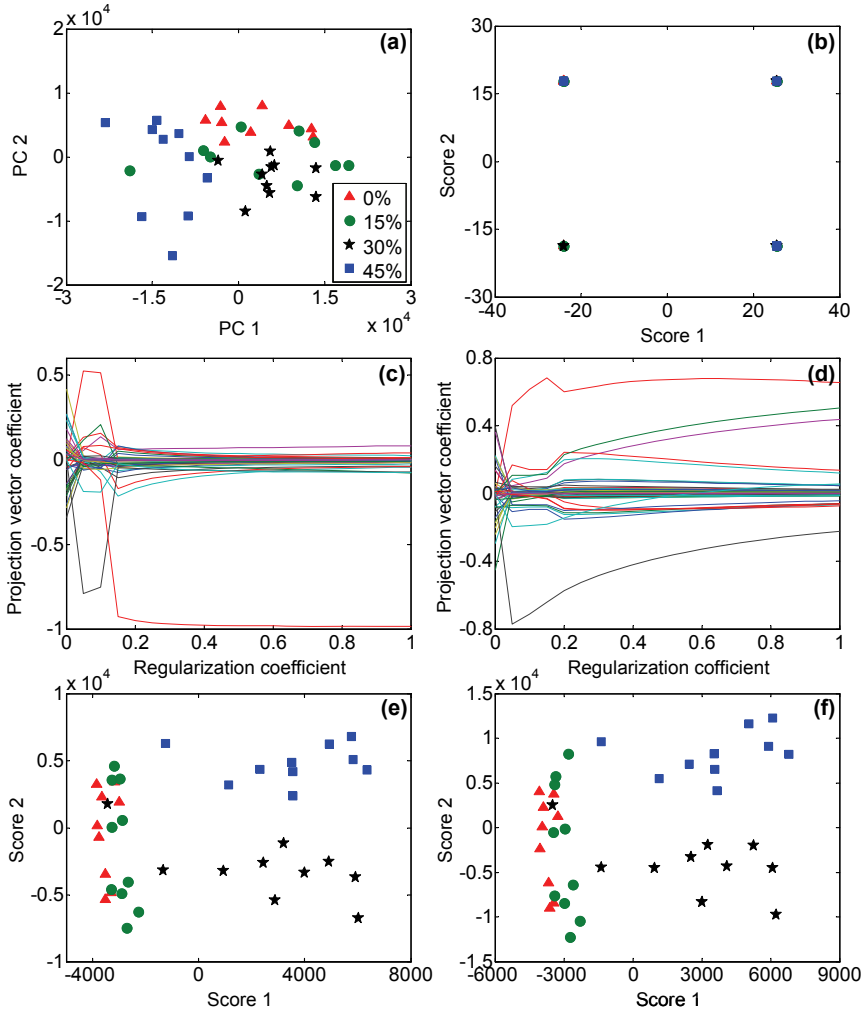

Figure 5.11 Results of cow diet data (PP methods: univariate approaches). (a) PCA scores plot, (b) PP scores plot (stepwise univariate approach), (c) ridge trace plot for the 1st projection vector coefficients, (d) ridge trace plot for the 2nd projection vector coefficients, (e) RPP scores plot (stepwise univariate approach) with the 1st and 2nd regularization coefficients set to 0.2 and 0.25, respectively, and (f) RPP scores plot (stepwise univariate approach) with the 1st and 2nd regularization coefficients both set to be 0.5.

The results from PP and RPP with bivariate approaches are shown in Figure 5.12. It can be seen in Figure 5.12 (a) that the samples again form a circle and PP fails to reveal class separation. Examination of the ridge trace plots in Figures 5.12 (b) and (c) suggests 0.3 for the regularization coefficient. The scores plot of RPP (bivariate approach) is shown in Figure 5.12 (d). The result is similar to that in Figures 5.11 (e) and (f) but the orientation is different. This is not surprising because the projection vectors for the multivariate approach are arranged to account for the variance of the projected data in a decreasing way. The direction to reveal clusters often does not match the direction to account for maximum variance. This is actually the reason why PP outperforms PCA.

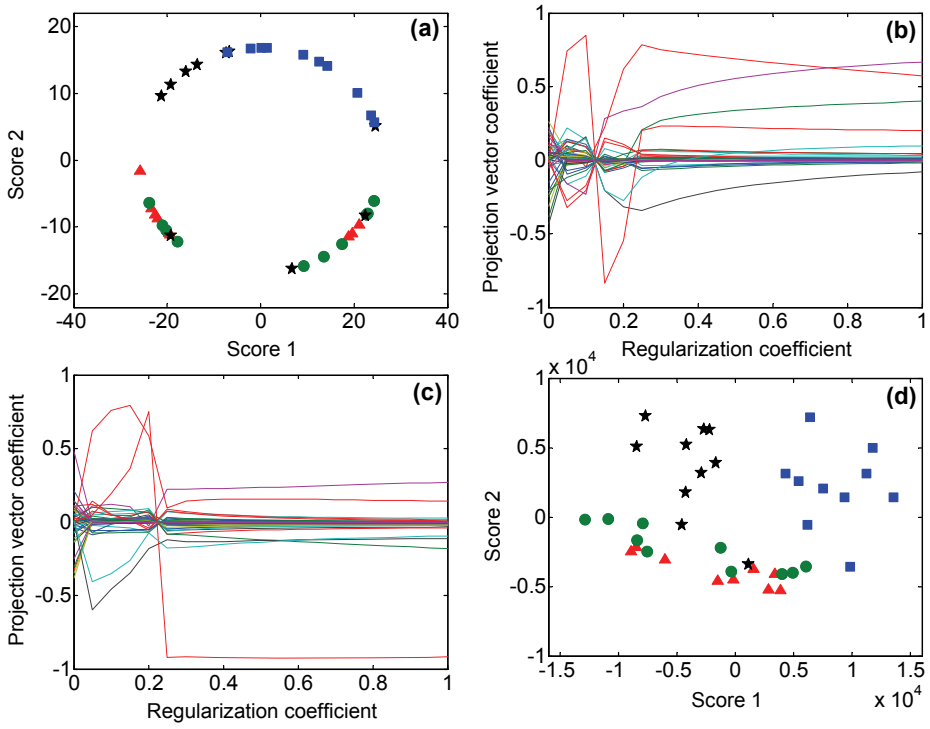

Figure 5.12 Results of cow diet data (PP methods: bivariate approaches). (a) PP scores plot (bivariate approach), (b) ridge trace plot for the 1st projection vector coefficients, (c) ridge trace plot for the 2nd projection vector coefficients, and (d) RPP scores plot (bivariate approach).

5.7 Discussion

The RPP proposed in this work aims to deal with data that have a small sample-to-variable ratio. It has been demonstrated by simulated and experimental data that RPP can give more meaningful information than PP. For minimization of kurtosis, if the regularization term $\lambda\mathbf{I}$ is ignored (Equations (5.10) and (5.22) for univariate kurtois and

multivariate kurtosis, respectively), the RPP objective function actually combines the effects of kurtosis and variance. Thus, sphering, which is mathematical transformation to make the projected data onto any direction have unit variance and has been a quite standard preprocessing step for most of the PP and ICA methods, is not recommended. This is because the information contained in the second order function (variance) disappears if the data are sphered. However, the commonly used method of scaling can be used to make the variables have compatible range.

Although RPP can, to some extent, solve the "over-fitting-like" problem in the case of data with a small sample-to-variable ratio, this does not exclude the variable compression method discussed in reference [3], where SVD or PCA was applied to the data and PP was applied to the dimensionality reduced data. It is often that the useful information is retained in the first few principal components of PCA. If the dimensionality reduction can effectively remove some variables but still retains the useful information, dimensionality reduction is still a good method to deal with data with a small sample-to-variable ratio. In fact, for the experimental data sets used in this work, dimensionality reduction followed by PP can also give meaningful results. Thus, a combination of using PCA, PP, and RPP to extract information from the data may be useful.

In this work, the regularization coefficient γ is determined by examining the ridge trace plots. The recommended range for γ is between 0 and 1. However, this does not mean γ can only be examined in this range. Depending on the data, a broader range can be explored. If some prior knowledge about the data is available, this may be helpful to determine the suitable size of γ. The scores plot may also help to determine a suitable γ. For the stepwise univariate approach, if γ is too small, the data tends to form four very tiny clusters located in four quadrants in a two-dimensional scores plot with very small variance accounted by the scores. For the multivariate approach, if γ is too small, the data in a two-dimensional scores plot tends to form a circle and account for very small variance as well. For some data sets, the scores plot may not be very sensitive to the choice of γ. If the ridge trace plots show several smooth patterns, different γ values can be tried to see if different meaningful information can be obtained.

It is also found that, for data with a small sample-to-variable ratio, if no regularization is performed, then the number of local minima is generally large. With the increase of the regularization coefficient γ, the number of local minima decreases. This removes some local minima caused by chance and increases the likelihood of finding the global minimum. Thus, the number of local minima may be another hint to determine the suitable size of the regularization coefficient γ.

It is worth emphasizing that, although RPP proposed in this work aims to deal with data with a small sample-to-variable ratio, one must be careful not to over-interpret the results. The result from RPP is exploratory and often gives useful information, but it should not be interpreted as definitive conclusion without proper validation. The results from RPP can be used to guide, but never aim to replace, further validation.

5.8 Conclusions

In exploratory analysis of multivariate data, PP can often give better results than PCA. However, today's chemical data often have many more variables than samples. This causes an "over-fitting-like" problem in the application of PP and meaningful information can not be extracted effectively. Dimensionality reduction by SVD before application of PP may be a solution to this problem, but it is subject to potential loss of useful information.

In this work, RPP is proposed as an alternative solution to deal with data that have a small sample-to-variable ratio. Since it does not use other dimensionality reduction methods such as SVD to remove information, more useful information may be retained. The RPP uses regularized kurtosis as a projection index and adapts the quasi-power methods for optimization. Examination of the ridge trace plot is recommended as a method to determine the suitable regularization coefficient. It has been illustrated through the use of simulated and experimental data that RPP can extract meaningful information that PP fails to obtain.

5.9 Bibliography

1. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Inc., 1977.
2. J. W. Tukey, We Need Both Exploratory and Confirmatory, *The American Statistician*, 34 (1980) 23-25.
3. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.
4. M. A. Golberg, and H. A. Cho, *Introduction to Regression Analysis*, WIT Press, 2004.
5. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, 2006.
6. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Inc., New York, 1958.
7. F. O'Sullivan, A Statistical Perspective on Ill-Posed Inverse Problems, *Statistical Science*, 1 (1986) 502-518.
8. D. M. Titterington, Common Structure of Smoothing Techniques in Statistics, *International Statistical Review*, 53 (1985) 141-170.
9. E. Hoerl, and R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, 12 (1970) 55-67.
10. E. Hoerl, and R. W. Kennard, Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics*, 12 (1970) 69-82.
11. D. W. Marquardt, and R. D. Snee, Ridge Regression in Practice, *The American Statistician*, 29 (1975) 3-20.
12. D. C. Montgomery, E. A. Peck, *Introduction to Linear Regression Analysis*, Second Edition, John Wiley & Sons, Inc., New York, 1992.
13. J. H. Friedman, Regularized Discriminant Analysis, *Journal of the American Statistical Association*, 84 (1989) 165-175.
14. Z. Zhang, G. Dai, and C. Xu, Regularized Discriminant Analysis, Ridge Regression and Beyond, *Journal of Machine Learning Research*, 11 (2010) 2199-2228.
15. T. C. Hsiang, A Bayesian View on Ridge Regression, *The Statistician*, 24 (1975) 267-268.
16. P. W. Wahl, and R. A. Kronmal, Discriminant Functions When Covariances Are Unequal and Sample Sizes Are Moderate, *Biometrics*, 33 (1977) 479-484.
17. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

18. R. Tibshirani, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, 58 (1996) 267-288.

19. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.

20. K. Yap, L. Guan, and J. Evans, Blind Adaptive Detection for CDMA Systems Based on Regularized Independent Component Analysis, *IEEE Global Telecommunications Conference*, San Antonio, 2001, pp. 249-253.

21. J. Sietsma, and R. J. F. Dow, Creating Artificial Neural Networks that Generalize, *Neural Networks*, 4 (1991) 67-69.

22. J. R. Magnus, and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Inc., New York, 1988, pp. 177.

23. K. B. Petersen, and M. S. Pedersen, *The Matrix Cookbook*, Version: Nov. 14, 2008, http://matrixcookbook.com. Last access on April 1, 2011.

24. D. Poole, *Linear Algebra: A Modern Introduction*, Brooks/Cole, 2003.

25. A. Frank and A. Asuncion, UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine, CA, 2010, http://archive.ics.uci.edu/ml. Last access on April 1, 2011.

26. R. S. Michalski, and R. L. Chilausky, Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Development an Expert System for Soybean Disease Diagnosis, *International Journal of Policy Analysis and Information Systems*, 4 (1980) 125-161.

27. N. G. Psihogios, R. G. Kalaitzidis, S. Dimou, K. I. Seferiadis, K. C. Siamopoulos, and E. T. Bairaktari, Evaluation of Tubulointerstitial Lesions' Severity in Patients with Glomerulonephritides: An NMR-based Metabonomic Study, *Journal of Proteome Research*, 6 (2007) 3760-3770.

28. Metaboanalyst: a web service for metabolomic data analysis, http://www.metaboanalyst.ca/MetaboAnalyst/faces/Docs/Format.jsp. Last access on April 10, 2011.

29. B. N. Ametaj, Q. Zebeli, F. Saleem, N. Psychogios, M. J. Lewis, S. M. Dunn, J. Xia, and D. S. Wishart, Metabolomics Reveals Unhealthy Alterations in Rumen Metabolism with Increased Proportion of Cereal Grain in the Diet of Dairy Cows, *Metabolomics*, 6 (2010) 583-594.

30. D. G. V. Emmanuel, S. M. Dunn, and B. N. Ametaj, Feeding High Proportions of Barley Grain Stimulates an Inflammatory Response in Dairy Cows, *Journal of Dairy Science*, 91 (2008) 606–614.

31. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.

Chapter 6: Conclusions

The past two decades have witnessed the rapid development of advanced analytical measurement techniques in chemistry and many other areas. The data obtained with these new techniques have become more and more complicated, characterized by the large volume and the complexity of the multivariate data. These characteristics generally imply more information in the data, but also mean extracting useful information from the data becomes more challenging.

As a highly interfacial discipline, chemometrics plays an important role in extracting useful information from chemical data. One of the primary applications of chemometrics to multivariate data is exploratory data analysis. When data are collected, the first step to examine the data is often exploratory, which generally involves dimensionality reduction and visualization of the extracted information in a low-dimensional space.

Performing effective dimensionality reduction to extract the useful information is never trivial. Many problems can lead to the failure of a dimensionality reduction method, such as a poorly defined objective function, the lack of efficient optimization algorithms, and complicated error structures. In the visualization of the information in a low-dimensional space, noisy data points may mask the useful information. Depending on the data structure and the purpose, different methods are needed, and there is no one panacea. The work presented in this thesis has been motivated to improve data processing methods for exploratory data analysis. The improvements presented in Chapters 2 to 5 are related, but each of them can be regarded as an independent achievement.

In chemistry, principal component analysis (PCA) is perhaps the mostly wide used method for exploratory data analysis. PCA can be regarded as a subspace modeling technique which is most effective when measurement errors are homoscedastic. However, heteroscedastic errors are common in multivariate data. In the case of significantly heteroscedastic noise, PCA becomes less effective in extracting information. Maximum likelihood principal component analysis (MLPCA), which has been developed to overcome the disadvantages of PCA in this case, has been applied successfully for other purposes, but its application in exploratory data analysis has previously not been explored.

This motivates the work reported in Chapter 2. Strategies for using MLPCA in exploratory data analysis to deal with data with significantly heteroscedastic errors were proposed. A new method to improve the visualization of the data in case of noisy measurements, referred to as partial transparency projection (PTP), was developed. These were demonstrated by successful applications in simulated and experimental data.

MLPCA was proposed in 1990's and has been used in different applications. However, this does not mean that further algorithmic improvements are not needed. One problem is that, in dealing with data where the underlying model has non-zero intercepts, MLPCA lacks an efficient optimization algorithm due to the complexity of the objective function, although the objective function has been well defined. This problem is addressed by the work presented in Chapter 3, where a new optimization algorithm was developed. The theory to develop the algorithm was provided and its performance was evaluated by simulated data.

Projection pursuit (PP) is another important method for exploratory data analysis, and can extract more useful information than PCA in many cases. However, it has not been widely used in chemistry and other areas. One of the major reasons is that easy and simple algorithms to optimize the complicated objective functions are not readily available. The work described in Chapter 4 provided new optimization algorithms, referred to as "quasi-power methods", to optimize kurtosis, which is used as a projection pursuit objective function. The algorithms have been successfully applied to simulated and real experimental data with obviously improved results in contrast to those obtained by PCA.

PP requires that the number of samples is much larger than the number of variables to give reliable information. However, data in chemistry and other areas often have fewer samples than variables. This problem may be addressed by first applying PCA to reduce the dimensionality of the data, but this incurs the risk of losing useful information. As an alternative method, a new projection pursuit method, called as regularized project pursuit (RPP), was proposed in Chapter 5. The utility of the proposed method was demonstrated with simulated and experimental data.

It is hoped that the achievements presented in the previous chapters can help chemists and other data analysts to extract useful information more effectively from multivariate data for exploratory data analysis.

References

Bibliography I

1. R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, Inc., New Jersey, 1998.
2. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Second Edition, John Wiley & Sons, Inc., New York, 1984.
3. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Inc., 1977.
4. J. W. Tukey, We Need Both Exploratory and Confirmatory, *The American Statistician*, 34 (1980) 23-25.
5. L. T. Fernholz, and S. Morgenthaler, A Conversation with John W. Tukey and Elizabeth Tukey, *Statistical Science*, 15 (2000) 79-94.
6. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, 2006.
7. D. C. Montgomery, and E. A, Peck, *Introduction to Linear Regression Analysis*, Second Edition, John Wiley & Sons, Inc., New York, 1992.
8. K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2 (1901) 559–572.
9. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part I), *The Journal of Educational Psychology*, 24 (1933) 417-441.
10. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part II), *The Journal of Educational Psychology*, 24 (1933) 498-520.
11. C. Spearman, “General Intelligence”, Objectively Determination and Measured, *The American Journal of Psychology*, 15 (1904) 201-292.
12. J. H. Friedman, and J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions and Computers*, 23 (1974) 881-890.
13. P. J. Huber, Projection Pursuit, *The Annals of Statistics*, 13 (1985) 435-475.
14. R. E. Bellman, *Dynamic Programming* (Republished version), Dover Publications, Inc., New York, 2003.
15. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, Maximum Likelihood Principal Component Analysis, *Journal of Chemometrics*, 11 (1997) 339-366.
16. P. D. Wentzell, Chapter 2.25: Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods, in S. D. Brown, R. Tauler, and B. Walczak (Editors): *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Elsevier Ltd., 2009.
17. W. A. Fuller, *Measurement Error Models*, Wiley, New York, 1987.

18. P. D. Wentzell, and S. Hou, Exploratory Data Analysis with Noisy Measurements, *Journal of Chemometrics,* 26 (2012) 264-281.

19. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.

20. J. D. Ingle, and S. R. Crouch, *Spectrochemical Analysis*; Prentice-Hall: Englewood Cliffs, New Jersey, 1988.

21. D. J. Bartholomew, Spearman and the Origin and Development of Factor Analysis, *British Journal of Mathematical and Statistical Psychology*, 48 (1995) 211-220.

22. S. A. Mulaik, Factor Analysis and Psychometrika: Major Developments, *Psychometrika*, 51 (1986) 23-33.

23. L. L. Thurstone, Multiple Factor Analysis, *Psychological Review*, 38 (1931) 406-427.

24. H. F. Kaiser, and J. Caffrey, Alpha Factor Analysis, *Psychometrika*, 30 (1965) 1-14.

25. R. I. Jennrich, and S. M. Robinson, A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 111-123.

26. K. G. Jöreskog, A General Approach to Confirmatory Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 183-202.

27. T. S. Rao, Canonical Factor Analysis and Stationary Times Series Models, *Sankhyā: The Indian Journal of Statistics, Series B*, 38 (1976) 256-271.

28. L. Stankov, Hierarchical Factoring Based on Image Analysis and Orthoblique Rotations, *Multivariate Behavioral Research*, 14 (1979) 339-353.

29. H. H. Harman, *Modern Factor Analysis*, Second Edition, The University of Chicago Press, Chicago, 1967.

30. E. R. Malinowski, *Factor Analysis in Chemistry*, Third Edition, John Wiley and Sons Inc., New York, 2002.

31. R. Cudeck, and R. C. MacCallum (Editors), *Factor Analysis at 100: Historical Development and Future Directions*, Lawrence Erlbaum Associates, Publishers, New Jersey, 2007.

32. K. G. Jöreskog, Factor Analysis by Least-Squares and Maximum-likelihood Methods, in K. Enslein, A. Ralston, and H. S. Wilf (Editors): *Statistical Methods for Digital Computers*, Volume III, John Wiley & Sons, Inc., 1977, pp. 125-153.

33. H. H. Harman, Minres Method of Factor Analysis, in K. Enslein, A. Ralston, and H. S. Wilf (Editors): *Statistical Methods for Digital Computers*, Volume III, John Wiley & Sons, Inc., 1977, pp. 154-165.

34. R. G. Brereton, *Applied Chemometrics for Scientists*, John Wiley & Sons Ltd., Chichester, England, 2007.

35. R. J. Adcock, A Problem in Least Squares, *The Analyst*, 5 (1878) 53-54.

36. G. Li, and Z. Chen, Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo, *Journal of American Statistical Association*, 80 (1985) 759-766.

37. M. Hubert, P. J. Rousseeuw, and S. Verboven, A Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60 (2002) 101-111.

38. M. Hubert, and S. Engelen, Robust PCA and Classification in Bioscience, *Bioinformatics*, 20 (2004) 1728-1736.

39. M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen, Bayesian Principal Component Analysis, *Journal of Chemometrics*, 16 (2002) 576-595.

40. C. M. Bishop, Bayesian PCA, in M. S. Kearns, S. A. Solla, and D. A. Cohn (Editors), *Advances in Neural Information Processing Systems*, MIT press, 1999, pp. 328-388.

41. B. Schölkopf, A. Smola, and K. R. Müller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, *Neural Computation*, 10 (1998) 1299-1319.

42. B. Schölkopf, A. Smola, and K. R. Müller, Kernel Principal Component Analysis, *Lecture Notes in Computer Science*, 1327 (1997) 583-588.

43. M. E. Tipping, and C. M. Bishop, Probabilistic Principal Component Analysis, *Technical Report NCRG/97/010*, Neural Computing Research Group, Aston University, 1997.

44. M. E. Tipping, and C. M. Bishop, Probabilistic Principal Component Analysis, *Journal of the Royal Statistical Society, Series B*, 21 (1999) 611-622.

45. S. Roweis, EM Algorithm for PCA and SPCA, in M. I. Jordan, M. J. Kearns, and S. A. Solla (Editors): *Advances in Neural Information Processing Systems*, MIT press, 1998, pp. 626-632.

46. C. Eckart, and G. Young, The Approximation of One Matrix by Another of Lower Rank, *Psychometrika*, 1 (1936) 211-218.

47. P. Horst, Sixty Years with Latent Variables and Still More to Come, *Chemometrics and Intelligent Laboratory System*, 14 (1992) 5-21.

48. D. T. Andrews, and P. D. Wentzell, Applications of Maximum Likelihood Principal Component Analysis: Incomplete Data Sets and Calibration Transfer, *Analytica Chimica Acta*, 350 (1997) 341-352.

49. P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Maximum Likelihood Multivariate Calibration, *Analytical Chemistry*, 69 (1997) 2299-2311.

50. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.

51. M. C. Jones, and R. Sibson, What is Projection Pursuit? *Journal of the Royal Statistical Society, Series A (General)*, 150 (1987) 1-37.

52. J. H. Friedman, Projection Pursuit, *Journal of the American Statistical Association*, 82 (1987) 249-266.

53. P. Hall, On Polynomial-Based Projection Indices for Exploratory Projection Pursuit, *The Annals of Statistics*, 17 (1989) 589-605.

54. C. Posse, An Effective Two-Dimensional Projection Pursuit Algorithm, *Communications in Statistics - Simulation and Computation*, 19 (1990) 1143-1164.

55. I. S. Yenyukov, Indices for Projection Pursuit, in E. Diday (Editors): *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science Publishers, New York, 1989, pp. 181-189.

56. J. N. Miller, and J. C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, Fifth Edition, Pearson Education Limited, 2005.

57. P. Common, Independent Component Analysis, A New Concept? *Signal Processing*, 36 (1994) 287-314.

58. J. V. Stone, *Independent Component Analysis: A Tutorial Introduction*, Cambridge, The MIT Press, 2004.

59. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., New York, 2001.

60. P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*; Springer, Berlin, 2004.

61. G. B. Dantzig, and M. N. Thapa, *Linear Programming 1: Introduction*, Springer, New York, 1997.

62. G. B. Dantzig, and M. N. Thapa, *Linear Programming 2: Theory and Extensions*, Springer, New York, 2003.

63. M. Mitchell, *An Introduction to Genetic Algorithms*, The MIT Press, Cambridge, 1998.

64. J. Stewart, *Calculus: Early Transcendentals*, Fourth Edition, Brooks/Cole Publishing Company, New York, 1999.

Bibliography II

1. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Inc., 1977.

2. K. Pearson, On Lines and Planes of Closest Fit to Systems of Points in Space, *Philosophical Magazine*, 2 (1901) 559–572.

3. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part I), *The Journal of Educational Psychology*, 24 (1933) 417-441.

4. H. Hotelling, Analysis of a Complex of Statistical Variables into Principal Components (Part II), *The Journal of Educational Psychology*, 24 (1933) 498-520.

5. R. A. Fisher, The Statistical Utilization of Multiple Measurements, *Annals of Eugenics*, 8 (1938) 376-386.

6. I. E. Frank, and J. H. Friedman, Classification: Oldtimers and Newcomers, *Journal of Chemometrics*, 3 (1989) 463-475.

7. M. Barker, and W. Rayens, Partial Least Squares for Discrimination, *Journal of Chemometrics*, 17 (2003) 166-173.

8. R. Rosipal, and N. Kramer, Overview and Recent Advances in Partial Least Squares, *Lecture Notes in Computer Science*, 3940 (2006) 34-51.

9. R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, Inc., New Jersey, USA, 1998.

10. C. J. C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2 (1998) 121-167.

11. M. R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, USA, 1973.

12. B. S. Everitt, *Cluster Analysis*, Third Edition, Edward Arnold, London, UK, 1993.

13. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, USA, 2006.

14. J. H. Friedman, and J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions and Computers*, 23 (1974) 881-890.

15. P. J. Huber, Projection Pursuit, *The Annals of Statistics*, 13 (1985) 435-475.

16. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.

17. M. Hubert, P. J. Rousseeuw, and S. A. Verboven, Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60 (2002) 101-111.

18. M. Hubert, and S. Engelen, Robust PCA and Classification in Bioscience, *Bioinformatics*, 20 (2004) 1728-1736.

19. M. A. Girshick, Principal Components, *Journal of the American Statistical Association*, 31 (1936) 519-528.

20. P. D. Wentzell, Chapter 2.25: Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods, in S. D. Brown, R. Tauler, and B. Walczak (Editors): *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Elsevier Ltd., 2009.

21. J. D. Ingle, and S. R. Crouch, *Spectrochemical Analysis*, Prentice-Hall, New Jersey, 1988.

22. P. D. Wentzell, T. K. Karakach, S. Roy, M. J. Martinez, C. P. Allen, and M. Werner-Washburne, Multivariate Curve Resolution of Time Course Microarray Data, *BMC Bioinformatics*, 7 (2006) 343.

23. T. K. Karakach, R. M. Flight, and P. D. Wentzell, Bootstrap Method for the Estimation of Measurement Uncertainty in Spotted Dual-Color DNA Microarrays, *Analytical and Bioanalytical Chemistry*, 389 (2007) 2125-2141.

24. T. K. Karakach, and P. D. Wentzell, Methods for Estimating and Mitigating Errors in Spotted, Dual-color DNA Microarrays, *OMICS: A Journal of Integrative Biology*, 11 (2007) 186-199.

25. P. Paatero , and U. Tapper, Analysis of Different Modes of Factor Analysis as Least Squares Fit Problems, *Chemometrics and Intelligent Laboratory Systems*, 18 (1993) 183-194.

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

27. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, Maximum Likelihood Principal Component Analysis, *Journal of Chemometrics*, 11 (1997) 339-366.

28. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.

29. D. T. Andrews, and P. D. Wentzell, Applications of Maximum Likelihood Principal Component Analysis: Incomplete Data Sets and Calibration Transfer, *Analytica Chimica Acta*, 350 (1997) 341-352.

30. P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Maximum Likelihood Multivariate Calibration, *Analytical Chemistry*, 69 (1997) 2299-2311.

31. S. K. Schreyer, M. Bidinosti, and P. D. Wentzell, Application of Maximum Likelihood Principal Components Regression to Fluorescence Emission Spectra, *Applied Spectroscopy*, 56 (2002) 789-796.

32. M. N. Leger, and P. D. Wentzell, Maximum Likelihood Principal Components Regression on Wavelet-Compressed Data, *Applied Spectroscopy*, 58 (2004) 855-862.

33. M. R. Keenan, Maximum Likelihood Principal Component Analysis of Time-of-Flight Secondary Ion Mass Spectrometry Spectral Images, *Journal of Vacuum Science & Technology A*, 23 (2005) 746-750.

34. L. Vega-Montoto, and P. D. Wentzell, Maximum Likelihood Parallel Factor Analysis, *Journal of Chemometrics*, 17 (2003) 237-253.

35. L. Vega-Montoto, H. Gu, and P. D. Wentzell, Mathematical Improvements to Maximum Likelihood Parallel Factor Analysis: Theory and Simulations, *Journal of Chemometrics*, 19 (2005) 216-235.

36. L. Vega-Montoto, and P. D. Wentzell, Mathematical Improvements to Maximum Likelihood Parallel Factor Analysis: Experimental Studies, *Journal of Chemometrics*, 19 (2005) 236-252.

37. M. Schuermans, I. Markovsky, P. D. Wentzell, and S. Van Huffel, On the Equivalence between Total Least Squares and Maximum likelihood PCA, *Analytica Chimica Acta*, 544 (2005) 254-267.

38. G. H. Golub, and C. F. Van Loan, An Analysis of the Total Least Squares Problem, *SIAM Journal on Numerical Analysis*, 17 (1980) 883-893.

39. S. Van Huffel, and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*; The Society for Industrial and Applied Mathematics, Philadelphia, 1991.

40. S. Van Huffel (Editors), *Recent Advances in Total Least Squares Techniques and Errors-in-Variables Modeling*, The Society for Industrial and Applied Mathematics, Philadelphia, 1997.

41. S. Van Huffel, and P. Lemmerling (Editors), *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, Kluwer Academic Publishers, Dordrecht, 2002.

42. P. Paatero, and U. Tapper, Positive Matrix Factorization: A Non-negative Factor Model with Optimal Utilization of Error Estimates of Data Values, *Environmetrics*, 5 (1994) 111-126.

43. P. Paatero, Least Squares Formulation of Robust Non-negative Factor Analysis, *Chemometrics and Intelligent Laboratory Systems*, 37 (1997) 23-35.

44. J. Kragten, Calculating Standard Deviations and Confidence Intervals with a Universally Applicable Spreadsheet Technique, *The Analyst*, 119 (1994) 2161-2165.

45. Y. Chen, E. R. Dougherty, and M. L. Bittner, Ratio-base Decisions and the Quantitative Anlaysis of cDNA Microarray Images, *Journal of Biomedical Optics*, 2 (1997) 364-374.

46. D. M. Rocke, and B. Durbin, A Model for Measurement Error for Gene Expression Arrays, *Journal of Computational Biology*, 8 (2001) 557-569.

47. M. J. Martinez, S. Roy, A. B. Archuletta, P. D. Wentzell, S. S. Anna-Arriola, A. L. Rodriguez, A. D. Aragon, G. A. Quinones, C. Allen, and M. Werner-Washburne. Genomic Analysis of Stationary-Phase and Exit in Saccharomyces Cerevisiae: Gene Expression and Identification of Novel Essential Genes, *Molecular Biology of the Cell*, 15 (2004) 5295-5305.

48. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.

49. A. Björck, and G. H. Golub, Numerical Methods for Computing Angles between Linear Subspaces, *Mathematics of Computation*, 27 (1973) 579-594.

50. S. Jiang, Angles between Euclidean Subspace, *Geometriae Dedicata*, 63 (1996) 113-121.

51. J. Miao, and A. Ben-Israel, On Principal Angles between Subspace in R^n, *Linear Algebra and Its Applications*, 171 (1992) 81-98.

Bibliography III

1. P. D. Wentzell, D. T. Andrews, D. C. Hamilton, K. Faber, and B. R. Kowalski, Maximum Likelihood Principal Component Analysis, *Journal of Chemometrics*, 11 (1997) 339-366.
2. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.
3. P. D. Wentzell, D. T. Andrews, and B. R. Kowalski, Maximum Likelihood Multivariate Calibration, *Analytical Chemistry,* 69 (1997) 2299-2311.
4. D. T. Andrews, and P. D. Wentzell, Applications of Maximum Likelihood Principal Component Analysis: Incomplete Data Sets and Calibration Transfer, *Analytica Chimica Acta*, 350 (1997) 341-352.
5. S. K. Schreyer, M. Bidinosti, and P. D. Wentzell, Application of Maximum Likelihood Principal Components Regression to Fluorescence Emission Spectra, *Applied Spectroscopy,* 56 (2002) 789-796.
6. L. Vega-Montoto, and P. D. Wentzell, Maximum Likelihood Parallel Factor Analysis, *Journal of Chemometrics*, 17 (2003) 237-253.
7. L. Vega-Montoto and P. D. Wentzell, Approaching the Direct Exponential Curve Resolution Algorithm from a Maximum Likelihood Perspective, *Analytica Chimica Acta*, 556 (2006) 383-399.
8. P. D. Wentzell, T. K. Karakach, S. Roy, M. J. Martinez, C. P. Allen, and M. Werner-Washburne, Multivariate Curve Resolution of Time Course Microarray Data, *BMC Bioinformatics*, 7 (2006) 343.
9. R. Tauler, M. Viana, X. Querol, A. Alastuey, R. M. Flight, P. D. Wentzell, and P. K. Hopke, Comparison of the Results Obtained by Four Receptor Modeling Methods in Aerosol Source Apportionment Studies, *Atmospheric Environment*, 43 (2009) 3989-3997.
10. P. D. Wentzell, and S. Hou, Exploratory Data Analysis with Noisy Measurements, *Journal of Chemometrics,* 26 (2012) 264-281.
11. R. J. Pell, M. B. Seasholtz, and B. R. Kowalski, The Relationship of Closure, Mean Centering and Matrix Rank Interpretation, *Journal of Chemometrics*, 6 (1992) 57-62.
12. M. B. Seasholtz, and B. R. Kowalski, The Effect of Mean Centering on Prediction in Multivariate Calibration, *Journal of Chemometrics*, 6 (1992) 103-111.
13. T. Iwata, and J. Koshoubu, Pretreatment of Spectral Data in PLS1 Quantitative Analysis, *Bunseki Kagaku*, 45 (1996) 85-89.
14. A. Lorber, K. Faber, and B. R. Kowalski, Local Centering in Multivariate Calibration, *Journal of Chemometrics*, 10 (1996) 215-220.
15. N. M. Faber, Mean Centering and Computation of Scalar Net Analyte Signal in Multivariate Calibration, *Journal of Chemometrics*, 12 (1998) 405-409.

16. T. J. Thurston, R. G. Brereton, D. J. Foord, and R. E. A. Escott, Principal Components Plots for Exploratory Investigation of Reactions Using Ultraviolet-Visible Spectroscopy: Application to the Formation of Benzophenone Phenylhydrazone, *Talanta*, 63 (2004) 757-769.
17. J. H. Kalivas, Learning from Procrustes Analysis to Improve Multivariate Calibration, *Journal of Chemometrics*, 22 (2008) 227-234.
18. D. Poole, *Linear Algebra: A Modern Introduction*, Brooks/Cole, 2003.
19. R. I. Jennrich, and S. M. Robinson, A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 111-123.
20. K. G. Jöreskog, A General Approach to Confirmatory Maximum Likelihood Factor Analysis, *Psychometrika*, 34 (1969) 183-202.
21. R. A. Johnson, and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fourth Edition, Prentice-Hall, Inc., New Jersey, 1998.
22. A. C. Rencher, *Methods of Multivariate Analysis*, John Wiley & Sons, Inc., New York, 1995.
23. J. Stewart, *Calculus: Early Transcendentals*, Fourth Edition, Brooks/Cole Publishing Company, New York, 1999.
24. J. R. Magnus, and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Inc., New York, 1988, pp. 177.
25. K. B. Petersen, and M. S. Pedersen, *The Matrix Cookbook*, Version: Nov. 14, 2008, http://matrixcookbook.com. Last access on Oct. 1, 2011.
26. N. M. Faber, Degrees of Freedom for the Residuals of Principal Component Analysis – A Clarification, *Chemometrics and Intelligent Laboratory System*, 93 (2008) 80-86.
27. M. A. Golberg, and H. A. Cho, *Introduction to Regression Analysis*, WIT Press, 2004.
28. S. Jiang, Angles between Euclidean Subspace, *Geometriae Dedicata*, 63 (1996) 113-121.
29. J. Miao, and A. Ben-Israel, On Principal Angles between Subspace in R^n, *Linear Algebra and Its Applications*, 171 (1992) 81-98.
30. A. Björck, and G. H. Golub, Numerical Methods for Computing Angles between Linear Subspaces, *Mathematics of Computation*, 27 (1973) 579-594.
31. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.

Bibliography IV

1. D. M. Glover, and P. K. Hopke, Exploration of Multivariate Chemical Data by Projection Pursuit, *Chemometrics and Intelligent Laboratory Systems*, 16 (1992) 45-59.

2. M. Hubert, P. J. Rousseeuw, and S. Verboven, A Fast Method for Robust Principal Components with Applications to Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 60 (2002) 101-111.

3. M. Daszykowski, I. Stanimirova, B. Walczak, and D. Coomans, Explaining a Presence of Groups in Analytical Data in Terms of Original Variables, *Chemometrics and Intelligent Laboratory Systems*, 78 (2005) 19-29.

4. J. H. Friedman, and J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions and Computers*, 23 (1974) 881-890.

5. J. B. Kruskal, Toward a Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding the Linear Transformation Which Optimizes a New "Index of Condensation", in R. C. Milton, and J. A. Nelder (Editors): *Statistical Computation*, Academic Press, New York, 1969.

6. J. B. Kruskal, Linear Transformation of Multivariate Data to Reveal Clustering, in R. N. Shepard, A. K. Romney, and S. B. Nerlove (Editors): *Multidimensional Scaling: Theory and Applications in the Behavioral Science*, Seminar Press, New York, 1972.

7. C. G. Posse, Projection Pursuit Discriminant Analysis for Two Groups, *Communications in Statistics - Theory and Methods*, 21 (1992) 1-19.

8. P. J. Huber, Projection Pursuit, *The Annals of Statistics*, 13 (1985) 435-475.

9. M. C. Jones, and R. Sibson, What is Projection Pursuit? *Journal of the Royal Statistical Society, Series A* (*General*), 150 (1987) 1-36.

10. J. H. Friedman, Projection Pursuit, *Journal of the American Statistical Association*, 82 (1987) 249-266.

11. P. Hall, On Polynomial-Based Projection Indices for Exploratory Projection Pursuit, *The Annals of Statistics*, 17 (1989) 589-605.

12. S. C. Morton, *Interpretable Projection Pursuit*, SLAC Report-355, Stanford Linear Accelerator Center, Standford University, California, 1989.

13. C. Posse, An Effective Two-Dimensional Projection Pursuit Algorithm, *Communications in Statistics - Simulation and Computation*, 19 (1990) 1143-1164.

14. I. S. Yenyukov, Indices for Projection Pursuit, in E. Diday (Editors): *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science Publishers, New York, 1989, pp. 181-189.

15. D. Peña, and F. J. Prieto, Cluster Identification Using Projections, *Journal of American Statistical Association*, 96 (2001) 1433-1445.

16. D. Peña, and F. J. Prieto, Multivariate Outlier Detection and Robust Covariance Matrix Estimation, *Technometrics*, 43 (2001) 286-310.

17. M. Hubert, Multivariate Outlier Detection and Robust Covariance Matrix Estimation: Discussion, *Technometrics*, 43 (2001) 303-306.

18. J. V. Stone, *Independent Component Analysis: A Tutorial Introduction*, Cambridge, The MIT Press, 2004.

19. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, Inc., New York, 2001.

20. N. Delfosse, and P. Loubaton, Adaptive Blind Separation of Independent Sources: A Deflation Approach, *Signal Processing*, 45 (1995) 59-83.

21. A. Hyvärinen, and E. Oja, A Neuron That Learns to Separate One Signal from a Mixture of Independent Sources, in *IEEE International Conference on Neural Networks,* Washington DC, 1996, pp. 62-67.

22. C. Croux, and A. Ruiz-Gazen, A Fast Algorithm for Robust Principal Components Based on Projection Pursuit, *COMPSTAS: Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 1996, pp. 211-217.

23. A. Hyvärinen, and E. Oja, A Fast Fixed-Point Algorithm for Independent Component Analysis, *Neural Computation*, 9 (1997) 1483-1492.

24. P. A. Tukey, and J. W. Tukey, Preparation; Prechosen Sequences of Views, in V. Barnett (Editors), *Interpreting Multivariate Data*, John Wiley, New York, 1981, pp. 189-213.

25. A. Hyvärinen, A Family of Fixed-Point Algorithms for Independent Component Analysis, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp. 3917-3920.

26. A. Hyvärinen, Fast and Robust Fixed-Point Algorithms for Independent Component Analysis, *IEEE Transactions on Neural Network*s, 10 (1999) 626-634.

27. P. A. Regalia, and E. Kofidis, Monotonic Convergence of Fixed-Point Algorithms for ICA, *IEEE Transactions on Neural Network*s, 14 (2003) 943-949.

28. V. Zarzoso, and P. Comon, Robust Independent Component Analysis by Iterative Maximization of the Kurtosis Contrast with Algebraic Optimal Step Size, *IEEE Transactions on Neural Network*s, 21 (2010) 248-261.

29. G. Brys, M. Hubert, and P. J. Rousseeuw, A Robustification of Independent Component Analysis, *Journal of Chemometrics*, 19 (2005) 364-375.

30. J. Stewart, *Calculus: Early Transcendentals*, Fourth Edition, Brooks/Cole Publishing Company, New York, 1999.

31. J. R. Magnus, and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Inc., New York, 1988, pp. 177.

32. K. B. Petersen, and M. S. Pedersen, *The Matrix Cookbook*, Version: Nov. 14, 2008, http://matrixcookbook.com. Last access on April 1, 2011.

33. D. Poole, *Linear Algebra: A Modern Introduction*, Brooks/Cole, 2003.

34. A. Höskuldsson, PLS Regression Methods, *Journal of Chemometrics*, 2 (1998) 211-228.

35. C. Posse, Tools for Two-Dimensional Exploratory Projection Pursuit, *Journal of Computational and Graphical Statistics*, 4 (1995) 83-100.

36. G. Nason, Three-Dimensional Projection Pursuit, *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 44 (1995) 411-430.

37. G. Nason, *Design and Choice of Projection Indices*, Ph. D. thesis, University of Bath, 1992.

38. K. V. Mardia, Measures of Multivariate Skewness and Kurtosis with Applications, *Biometrika*, 57 (1970) 519-530.

39. J. Johansson, Measuring Homogeneity of Planar Point-Pattern by Using Kurtosis, *Pattern Recognition Letters*, 21 (2000) 1149-1156.

40. S. N. Afriat, Orthogonal and Oblique Projections and the Characteristics of Pairs of Vector Spaces, *Mathematical Proceedings of the Cambridge Philosophy Society*, 53 (1957) 800-816.

41. J. Miao, and A. Ben-Israel, On Principal Angles between Subspaces in R^n, *Linear Algebra and Its Applications*, 171 (1992) 81-98.

42. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.

43. http://www.models.kvl.dk/. Last access on April 1, 2011.

44. J. Christensen, E. M. Becker, and C. S. Frederiksen, Fluorescence Spectroscopy and PARAFAC in the Analysis of Yogurt, *Chemometrics and Intelligent Laboratory Systems*, 75 (2005) 201-208.

45. T. K. Karakach, P. D. Wentzell, and J. A. Walter, Characterization of the Measurement Error Structure in 1D ^{1}H NMR Data for Metabolomics Study, *Analytica Chimica Acta*, 636 (2009) 163-174.

46. M. Forina, and C. Armanino, Eigenvector Projection and Simplified Non-Linear Mapping of Fatty Acid Content of Italian Olive Oils, *Annali di Chimica*, 72 (1982) 127-141.

47. M. Forina, and E. Tiscornia, Pattern Recognition Methods in the Prediction of Italian Olive Oil Origin by Their Fatty Acid Content, *Annali di Chimica*, 72 (1982) 143-155.

48. M. P. Derde, and D. L. Massart, Supervised Pattern Recognition: The Ideal Method? *Analytica Chimica Acta*, 191 (1986) 1-16.

49. J. Zupan, M. Novič, X. Li, and J. Gasteiger, Classification of Multicomponent Analytical Data of Olive Oils Using Different Neural Networks, *Analytica Chimica Acta*, 292 (1994) 219-234.

50. M. Daszykowski, B. Walczak, and D. L. Massart, A Journey into Low-Dimensional Spaces with Autoassociative Neural Networks, *Talanta*, 59 (2003) 1095-1105.

51. http://www2.ccc.uni-erlangen.de/publications/ANN-book/download/572oils.dat. Last access on April 1, 2011.

52. R. A. Horn, and C. A. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985, pp. 465.

Bibliography V

1. J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Inc., 1977.
2. J. W. Tukey, We Need Both Exploratory and Confirmatory, *The American Statistician*, 34 (1980) 23-25.
3. S. Hou, and P. D. Wentzell, Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index, *Analytica Chimica Acta*, 704 (2011) 1-15.
4. M. A. Golberg, and H. A. Cho, *Introduction to Regression Analysis*, WIT Press, 2004.
5. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science+Business Media, LLC, New York, 2006.
6. T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, Inc., New York, 1958.
7. F. O'Sullivan, A Statistical Perspective on Ill-Posed Inverse Problems, *Statistical Science*, 1 (1986) 502-518.
8. D. M. Titterington, Common Structure of Smoothing Techniques in Statistics, *International Statistical Review*, 53 (1985) 141-170.
9. E. Hoerl, and R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, 12 (1970) 55-67.
10. E. Hoerl, and R. W. Kennard, Ridge Regression: Applications to Nonorthogonal Problems, *Technometrics*, 12 (1970) 69-82.
11. D. W. Marquardt, and R. D. Snee, Ridge Regression in Practice, *The American Statistician*, 29 (1975) 3-20.
12. D. C. Montgomery, E. A. Peck, *Introduction to Linear Regression Analysis*, Second Edition, John Wiley & Sons, Inc., New York, 1992.
13. J. H. Friedman, Regularized Discriminant Analysis, *Journal of the American Statistical Association*, 84 (1989) 165-175.
14. Z. Zhang, G. Dai, and C. Xu, Regularized Discriminant Analysis, Ridge Regression and Beyond, *Journal of Machine Learning Research*, 11 (2010) 2199-2228.
15. T. C. Hsiang, A Bayesian View on Ridge Regression, *The Statistician*, 24 (1975) 267-268.
16. P. W. Wahl, and R. A. Kronmal, Discriminant Functions When Covariances Are Unequal and Sample Sizes Are Moderate, *Biometrics*, 33 (1977) 479-484.
17. V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.
18. R. Tibshirani, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B (Methodological)*, 58 (1996) 267-288.

19. P. D. Wentzell, and M. T. Lohnes, Maximum Likelihood Principal Component Analysis with Correlated Measurement Errors: Theoretical and Practical Considerations, *Chemometrics and Intelligent Laboratory System*, 45 (1999) 65-85.

20. K. Yap, L. Guan, and J. Evans, Blind Adaptive Detection for CDMA Systems Based on Regularized Independent Component Analysis, *IEEE Global Telecommunications Conference*, San Antonio, 2001, pp. 249-253.

21. J. Sietsma, and R. J. F. Dow, Creating Artificial Neural Networks that Generalize, *Neural Networks*, 4 (1991) 67-69.

22. J. R. Magnus, and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons, Inc., New York, 1988, pp. 177.

23. K. B. Petersen, and M. S. Pedersen, *The Matrix Cookbook*, Version: Nov. 14, 2008, http://matrixcookbook.com. Last access on April 1, 2011.

24. D. Poole, *Linear Algebra: A Modern Introduction*, Brooks/Cole, 2003.

25. A. Frank and A. Asuncion, UCI Machine Learning Repository. School of Information and Computer Science, University of California, Irvine, CA, 2010, http://archive.ics.uci.edu/ml. Last access on April 1, 2011.

26. R. S. Michalski, and R. L. Chilausky, Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Development an Expert System for Soybean Disease Diagnosis, *International Journal of Policy Analysis and Information Systems*, 4 (1980) 125-161.

27. N. G. Psihogios, R. G. Kalaitzidis, S. Dimou, K. I. Seferiadis, K. C. Siamopoulos, and E. T. Bairaktari, Evaluation of Tubulointerstitial Lesions' Severity in Patients with Glomerulonephritides: An NMR-based Metabonomic Study, *Journal of Proteome Research*, 6 (2007) 3760-3770.

28. Metaboanalyst: a web service for metabolomic data analysis, http://www.metaboanalyst.ca/MetaboAnalyst/faces/Docs/Format.jsp. Last access on April 10, 2011.

29. B. N. Ametaj, Q. Zebeli, F. Saleem, N. Psychogios, M. J. Lewis, S. M. Dunn, J. Xia, and D. S. Wishart, Metabolomics Reveals Unhealthy Alterations in Rumen Metabolism with Increased Proportion of Cereal Grain in the Diet of Dairy Cows, *Metabolomics*, 6 (2010) 583-594.

30. D. G. V. Emmanuel, S. M. Dunn, and B. N. Ametaj, Feeding High Proportions of Barley Grain Stimulates an Inflammatory Response in Dairy Cows, *Journal of Dairy Science*, 91 (2008) 606–614.

31. K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition, Academic Press, 1990.

Appendix: Copyright Permission Letters

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Jul 20, 2012

This is a License Agreement between Siyuan Hou ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	2953071453616
License date	Jul 20, 2012
Licensed content publisher	John Wiley and Sons
Licensed content publication	Journal of Chemometrics
Licensed content title	Exploratory data analysis with noisy measurements
Licensed content author	P. D. Wentzell,S. Hou
Licensed content date	Mar 30, 2012
Start page	264
End page	281
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Print and electronic
Portion	Full article
Will you be translating?	No
Order reference number	
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your Rightslink account (these are available at any time at http://myaccount.copyright.com)

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2. You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this licence must be completed within two years of the date of the grant of this licence (although copies prepared before may be distributed thereafter). The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired

thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

12. Any fee required for this permission shall be non-refundable after thirty (30) days from receipt.

13. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

14. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

15. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

16. This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

17. This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

Wiley Open Access Terms and Conditions

All research articles published in Wiley Open Access journals are fully open access: immediately freely available to read, download and share. Articles are published under the terms of the Creative Commons Attribution Non Commercial License. which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes. The license is subject to the Wiley Open Access terms and conditions: Wiley Open Access articles are protected by copyright and are posted to repositories and websites in accordance with the terms of the Creative Commons Attribution Non Commercial License. At the time of deposit, Wiley Open Access articles include all changes made during peer review, copyediting, and publishing. Repositories and websites that host the article are responsible for incorporating any publisher-supplied amendments or retractions issued subsequently.
Wiley Open Access articles are also available without charge on Wiley's publishing platform, **Wiley Online Library** or any successor sites.

Use by non-commercial users

For non-commercial and non-promotional purposes individual users may access, download, copy, display and redistribute to colleagues Wiley Open Access articles, as well as adapt, translate, text-

and data-mine the content subject to the following conditions:

• The authors' moral rights are not compromised. These rights include the right of "paternity" (also known as "attribution" - the right for the author to be identified as such) and "integrity" (the right for the author not to have the work altered in such a way that the author's reputation or integrity may be impugned).
• Where content in the article is identified as belonging to a third party, it is the obligation of the user to ensure that any reuse complies with the copyright policies of the owner of that content.
• If article content is copied, downloaded or otherwise reused for non-commercial research and education purposes, a link to the appropriate bibliographic citation (authors, journal, article title, volume, issue, page numbers, DOI and the link to the definitive published version on Wiley Online Library) should be maintained. Copyright notices and disclaimers must not be deleted.
• Any translations, for which a prior translation agreement with Wiley has not been agreed, must prominently display the statement: "This is an unofficial translation of an article that appeared in a Wiley publication. The publisher has not endorsed this translation."

Use by commercial "for-profit" organisations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee. Commercial purposes include:

• Copying or downloading of articles, or linking to such articles for further redistribution, sale or licensing;
• Copying, downloading or posting by a site or service that incorporates advertising with such content;
• The inclusion or incorporation of article content in other works or services (other than normal quotations with an appropriate citation) that is then available for sale or licensing, for a fee (for example, a compilation produced for marketing purposes, inclusion in a sales pack)
• Use of article content (other than normal quotations with appropriate citation) by for-profit organisations for promotional purposes
• Linking to article content in e-mails redistributed for promotional, marketing or educational purposes;
• Use for the purposes of monetary reward by means of sale, resale, licence, loan, transfer or other form of commercial exploitation such as marketing products
• Print reprints of Wiley Open Access articles can be purchased from: corporatesales@wiley.com

Other Terms and Conditions:

BY CLICKING ON THE "I AGREE..." BOX, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN

THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

v1.7

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500822387.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

ELSEVIER LICENSE
TERMS AND CONDITIONS

Jul 20, 2012

This is a License Agreement between Siyuan Hou ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Siyuan Hou
Customer address	550 University Avenue Charlottetown, PE C1A 4P3
License number	2953080669791
License date	Jul 20, 2012
Licensed content publisher	Elsevier
Licensed content publication	Analytica Chimica Acta
Licensed content title	Fast and simple methods for the optimization of kurtosis used as a projection pursuit index
Licensed content author	S. Hou,P.D. Wentzell
Licensed content date	17 October 2011
Licensed content volume number	704
Licensed content issue number	1–2
Number of pages	15
Start Page	1
End Page	15
Type of Use	reuse in a thesis/dissertation
Portion	full article
Format	both print and electronic
Are you the author of this Elsevier article?	Yes
Will you be translating?	No
Order reference number	

Title of your thesis/dissertation	Improved Projection Methods for Exploratory Data Analysis in Chemistry
Expected completion date	Aug 2012
Estimated size (number of pages)	200
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.0 USD / 0.0 GBP
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing

transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. **Website**: The following terms and conditions apply to electronic reserve and author websites:
Electronic reserve: If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if:
This license was made in connection with a course,
This permission is granted for 1 year only. You may obtain a license for future website posting,
All content posted to the web site must maintain the copyright information line on the bottom of each image,
A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com , and
Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. **Author website** for journals with the following additional clauses:

All content posted to the web site must maintain the copyright information line on the bottom of each image, and the permission granted is limited to the personal version of your paper. You are not allowed to download and

post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed edition to create an electronic version. A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx . As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier's online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article's Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. **Author website** for books with the following additional clauses: Authors are permitted to place a brief summary of their work online only. A hyper-text must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

19. **Website** (regular and for author): A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx. or for books to the Elsevier homepage at http://www.elsevier.com

20. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. **Other Conditions**:

v1.6

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500822402.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.
