An Exploration of Projection Pursuit Analysis and Independent Component Analysis as Alternatives for the Analysis of Multivariate Chemical Data

by

Chelsi C.R. Wicks

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

at

Dalhousie University
Halifax, Nova Scotia
August 2018

Table of Contents

List of Figures

Abstract

As multivariate data sets have become commonplace in chemical analysis, the need for new data analysis methods to extract different kinds of information has increased dramatically. Often methods are adopted and adapted from other fields of science to suit the needs of chemical data analysis. Two examples that have begun to gain interest for chemical applications are projection pursuit analysis (PPA) and independent component analysis (ICA).

PPA and ICA are both linear projection methods, but with seemingly different goals. PPA aims to locate "interesting" projections of the data to explore natural clusters that may exist among the samples. Since Gaussian distributions of the data are usually considered the least interesting, PPA often employs metrics which are able to measure deviations from Gaussian behaviour, such as the fourth statistical moment, kurtosis. In contrast, ICA methods aim to extract statistically independent source signals from mixtures of multiple components. Since statistical independence is difficult to measure, ICA often uses non-Gaussian behaviour as a surrogate for independence, establishing a relationship between these two methods.

This work presents a critical evaluation of key algorithms for PPA and ICA in applications related to clustering and signal extraction for chemical data. The relationship between PPA and ICA, which has been alluded to in the literature, is firmly established through theory and application. It is demonstrated through application to selected data sets that, while useful for certain kinds of problems, these methods are likely to have limited utility for signal extraction in chemistry, where the source signals are rarely statistically independent and have unpredictable distributions. In contrast, both methods are shown to be useful for clustering applications, but PPA is generally more powerful because of its ability to explore the full variable space. This is demonstrated through a case study where traditional exploratory data analysis methods fail due to a complex error structure in the data.

List of Abbreviations Used

ALS	Alternating least squares
BSS	Blind source separation
ECM	Error covariance matrix
EDA	Exploratory data analysis
ERNN	Elman recurrent neural networks
FDA	Fisher's discriminant analysis
FTIR	Fourier-transform infrared
GC-MS	Gas chromatography-mass spectrometry
HCA	Hierarchical cluster analysis
HOS	Higher order statistics
IC	Independent component
ICA	Independent component analysis
IR	Infrared
ISE	Ion selective electrode
JADE	Joint approximate diagonalization of eigenmatrices
KNN	k-nearest neighbours
kPPA	Kurtosis-based projection pursuit analysis
LDA	Linear discriminant analysis
LIBS	Laser induced breakdown spectroscopy
LS-SVM	Least squares support vector machines
MCR	Multivariate curve resolution
MCR-ALS	Multivariate curve resolution by alternating least squares
MI	Mutual information
MILCA	Mutual information least dependent component analysis
MLPCA	Maximum likelihood principal components analysis
MLR	Multiple linear regression
MSC	Multiplicative signal correction
NIR	Near infrared
NMR	Nuclear magnetic resonance
NN	Neural networks
PC	Principal component
PCA	Principal components analysis
PDF	Probability density function
PLS	Partial least squares
PLS-DA	Partial least squares discriminant analysis
PPA	Projection pursuit analysis
PPR	Projection pursuit regression
QESAR	Quantitative electronic structure-activity relationship
QSAR	Quantitative structure-activity relationship
RSD	Relative standard deviation
SNV	Standard normal variate
SPP	Sequential projection pursuit
XRF	X-ray fluorescence

Acknowledgements

These last few years have certainly not been easy, but I have learned a lot along the way. I certainly could not have gotten where I am without my amazing support system of family and friends (even if they've asked about my thesis a little too often). I would also like to thank my supervisor, Dr. Peter Wentzell, who has been patient and has given me the opportunity to learn about so much more than just my own research. Finally, I thank my readers for putting in the time and effort to read this document and provide their extremely valuable feedback.

Chapter 1 – Introduction

1.1 Introduction

Multivariate measurements are now commonplace in scientific data analysis and as the ability to acquire large amounts of data becomes more feasible, newer and more inventive methods are required to extract useful information. The increasing abundance and diversity of multivariate chemical measurements has been a constant motivation within the field of chemometrics to develop new tools to enhance the ability to extract information, often leveraging the unique characteristics of the measurement systems employed or the information target to improve on existing methods. Certain tools that were once foreign to chemistry, such as principal components analysis (PCA) and partial least squares regression (PLS), are now routinely applied to large data sets [1]. However, the quest continues for new and more powerful tools, especially to address evolving and challenging problems in areas such as hyperspectral imaging and metabolomics. New methods may involve an adaptation of existing techniques (e.g. improved data preprocessing), the adoption of strategies developed in other fields and transplanted to chemistry (such as support vector machines), or the rebirth of a previously little-used method due to relevant changes in implementation or application requirements. In some cases, the novelty of these approaches sustains them only temporarily or for niche applications in chemistry, while other methods become part of the mainstream of chemometrics and can even be transformative. In all cases, however, it is important to understand at a fundamental level how these methods work, how they differ from existing tools, and what particular advantages they bring to the analysis of chemical data, which often differs from other areas of application.

It is the goal of this work to investigate two multivariate data analysis methods that have recently been proposed as promising alternatives for the treatment of chemical data: projection pursuit analysis (PPA) [2] and independent component analysis (ICA) [3]. In particular, the objectives are: (1) to describe the underlying theory of these methods and the various algorithms used for their implementation, (2) to describe the relationships of these methods with more traditional approaches and with each other, (3) to provide an assessment of these methods for applications in exploratory data analysis (EDA) and multivariate curve resolution (MCR) using a variety of data sets, and (4) to provide a critical evaluation of PPA and ICA for chemical data analysis, emphasizing their strengths and limitations in the context of the selected applications.

This chapter begins with a brief overview of exploratory data analysis and multivariate curve resolution as important general problems in chemistry. This is followed by a short discussion of PPA and ICA as alternative approaches.

1.2 Exploratory Data Analysis

To understand the complex relationships among different sets of measurements (*e.g.* samples), simplification is often sought through visualization of high-dimensional data in low-dimensional spaces, sometimes referred to as exploratory data analysis (EDA) [1]. The goal of EDA is often to identify or confirm naturally occurring classes or groups within the data, so it is often closely associated with problems in classification, where the goal is to develop a rule or mathematical methods to classify unknown samples (e.g. healthy vs. diseased). Classification methods are typically divided into two categories: supervised and unsupervised [1]. Supervised methods use *a priori* knowledge of class membership to build classification models of the data, while unsupervised methods (which encompass EDA) look for naturally occurring clusters of the data without being told the classes. Supervised methods require caution and careful cross-validation to prevent cases of overfitting, particularly with high-dimensional data such as that often collected for chemical analysis. Unsupervised (EDA) methods can be used to explore data for naturally occurring separations to ensure sufficient information is present prior to building a classification model. This can be done completely blind for exploratory purposes or can be done under "semi-supervised" conditions where the algorithm is not given class information, but the user is able to evaluate the validity of the clustering using their knowledge of the classes present, often by visually inspecting the results.

Two unsupervised methods that are widely used for clustering are hierarchical cluster analysis (HCA) [1,4,5] and principal components analysis (PCA) [1,4–7]. HCA is a nonlinear mapping technique that renders the information about the distance among objects (samples) in a high-dimensional space into a two-dimensional representation known as a dendrogram. Figure 1.1 shows an example of a dendrogram from results that will be discussed in Chapter 3. Each terminal point on the x-axis represents a different sample, in

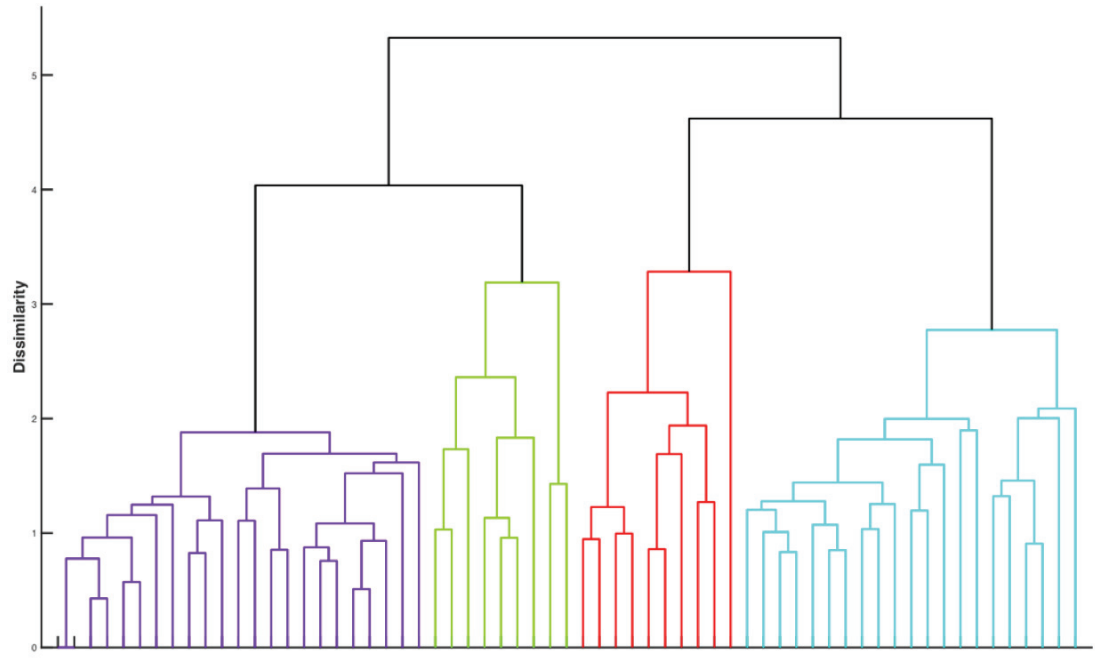

Figure 1.1 An example of a dendrogram. Each terminal end at the bottom corresponds to a sample and the colours indicate the class membership. (Data explained in Chapter 3)

this case samples of obsidian (a volcanic rock), where the colour coding represents different locations where it was collected. The length of vertical connections between samples indicates their proximity in a multidimensional space, in this case the space of metallic element abundance. In this way groupings of samples are easily visualized. Often dendrograms are used in conjunction with so-called heat maps to display the characteristics of variables, such as the expression levels of genes or proteins. Ideally the connections at higher levels in the dendrogram will be associated with the terminal nodes at the bottom to indicate that the samples closest to one another belong to the same group. While some groupings are evident in this example, in other cases a clear class separation may not be observed. This could be because such a separation is not possible or because the constraints of the data analysis method do not permit it to be observed. Alternative methods may be able to provide a different perspective.

PCA [1] is a linear projection technique expressed by the bilinear model given in Equation 1.1, where $\mathbf{X}$ (m samples by n variables) is the data matrix, $\mathbf{T}$ ($m \times m$, if $m < n$) is a matrix containing what are commonly known as the scores for each sample and the row vectors in the matrix $\mathbf{P}$ ($m \times n$, if $m < n$) are often referred to as the loadings or projection vectors.

$$\mathbf{X} = \mathbf{TP} \tag{1.1}$$

(Note: Throughout this thesis, uppercase bold symbols will be used to represent matrices, while lowercase bold symbols will be used to represent row or column vectors, unless otherwise specified. Scalars will be represented by italics.) PCA carries out an orthogonal rotation of the original variables (e.g. wavelength channels) to maximize the information contained in the lowest dimensions of the new space (i.e. the scores). Scatter plots of the first two or three columns of $\mathbf{T}$ represent a projection of the multivariate data into a two- or three-dimensional space while preserving the maximum amount of information about the relationships among objects. These plots, commonly known as scores plots, are often used to determine which samples group together and can therefore be considered to represent a cluster or class. As an example, Figure 1.2 shows a scores plot for the same data used in Figure 1.1, where the data matrix $\mathbf{X}$ consists of concentrations of metallic elements collected from m channels for n samples. Ideally, we would see a separation of objects in this space according to their class (different symbols, colours) which is seen here, though this is not always the case. PCA provides an alternative visualization of the data from HCA but may be subject to the same problems. The lack of a clear separation may be

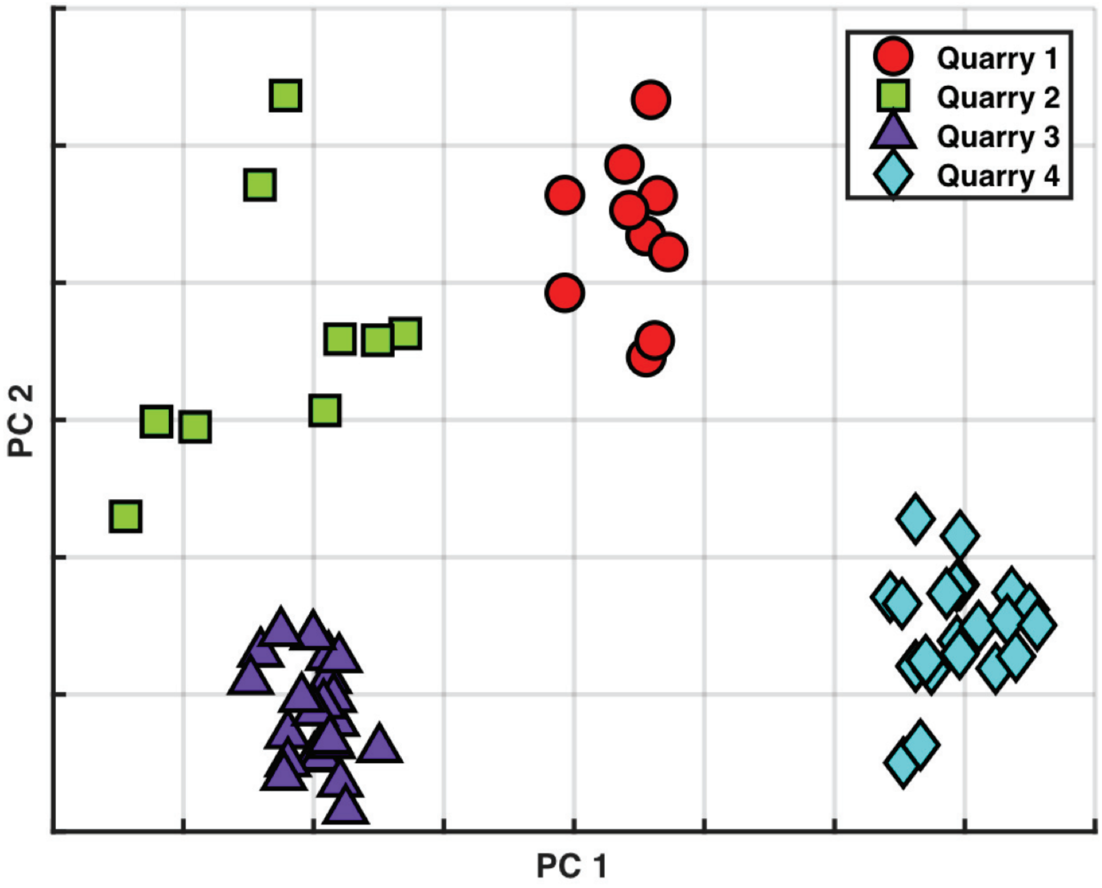

Figure 1.2 An example of a PCA scores plot. Each point represents a sample in the new 2-dimensional PCA space. The colours and symbols distinguish between the different classes (see legend) and are consistent with the colour coding in Figure 1.1.

the result of deficiencies in the data (insufficient information) or in the analysis method (insufficient criteria). Knowing which of these is the case is a critical question in chemometrics.

Both HCA and PCA are extensively used in chemical applications that include proteomics [8,9], metabolomics [10–12], food science [13,14], forensics [15–17], medical diagnostics [18,19], and threat detection [20]. An important goal of both techniques in these and other applications is to either identify or confirm groupings of samples that are consistent with external classifications that are based on other factors, such as disease state (medicine), geographic origin (food analysis), provenance (forensics) and biological species (chemotaxonomy). The widespread use of these tools is based, in part, on the fact that they are unsupervised methods, which means that the visualization uses no prior knowledge of the class structure. This is in contrast to supervised methods, such as partial least squares discriminant analysis (PLS-DA) [21,22], which actively employ class information to build a model and therefore require careful validation to avoid overfitting. Because no class information is employed in HCA and PCA, they have gained acceptance as suitable methods for hypothesis confirmation where the key question is whether the data contain sufficient information to distinguish different groups of samples, especially when the number of samples is small and the number of variables is large. This is often a critical question in research and can determine whether a line of inquiry continues or is abandoned. This accounts for the pervasive application of these methods across all areas of chemistry.

While HCA and PCA are powerful and useful techniques, they can be subject to serious limitations when applied to problems where the data do not meet certain criteria. HCA is based on the calculation of Euclidean distances among objects in higher dimensions, while PCA creates a subspace that maximizes the amount of variance retained in the data. Both of these methods are sensitive to the scale of the data, which means that variables which have a larger range will be weighted more heavily when mapping the high-dimensional data to lower dimensions, even if the information content is greater for variables with a smaller range. For example, a small mass spectral or nuclear magnetic resonance (NMR) peak that contains important information for the separation of classes may be eclipsed by larger peaks with a variability that does not correlate with class separation, resulting in a projection that does not reveal the critical relationships. In some cases, this problem may be mitigated by appropriate pretreatment of the data (*e.g.* variable scaling, log transformation) but this may give rise to other problems [4,23,24]. For example, scaling of variables that are predominately associated with noise (*e.g.* baseline regions) increases their influence in the mapping process even though they have no relevance in classification. This problem is further exacerbated by complex measurement noise structures which may include non-uniform error variance among variables (referred to as heteroscedastic noise) or correlated errors [25].

The principal weaknesses of HCA and PCA for unsupervised clustering with multivariate chemical data are: (1) lack of a criterion to distinguish meaningful chemical variance in a data set from the noise variance, and (2) a reliance on variance and distance metrics to develop interesting and useful projections of the data. A number of methods have been proposed as alternatives to HCA and PCA which use metrics that are less susceptible to the error structure of the data. A few of these methods will be demonstrated and discussed in the following chapters.

1.3 Multivariate Curve Resolution

Aside from projections and clustering, another common goal in chemical data analysis involves extracting underlying chemical profiles from samples containing mixtures of compounds through multivariate curve resolution (MCR). Such problems in mixture analysis are commonplace in chemistry. For example, one may obtain mixture spectra throughout the course of a chemical reaction and may be interested in obtaining the spectra of pure components, especially intermediates. Other examples include chromatography (where the goal is to extract the profiles of unresolved peaks) and environmental analysis (with the aim of identifying the profiles of individual pollution sources). More recently MCR has been applied to hyperspectral imaging where the goal is

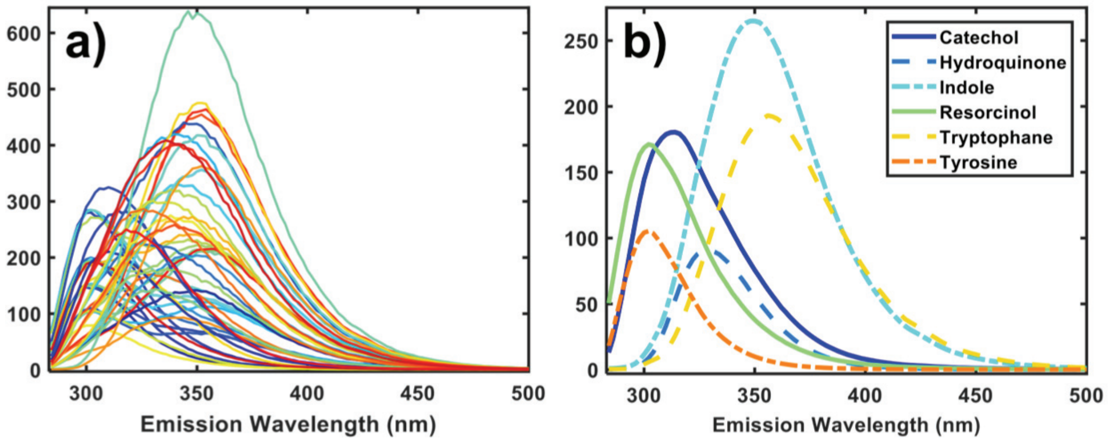

Figure 1.3 An example of the objective of multivariate curve resolution. a) Mixtures of signals. b) Pure compound spectra. (Data explained in Chapter 3)

to use spectral data at each pixel to locate particular objects (e.g. tumours, physical boundaries) within the image. Like PCA, multivariate curve resolution (MCR) can be expressed using a bilinear model (Equation 1.2), breaking the data, **D** $(m \times n)$, into matrices representative of the concentrations (contributions), **C** $(m \times p$, where p is the number of components), and profiles (e.g. pure component spectra), **S** $(n \times p)$, of the chemical components while also taking into account the errors present in the data, **E** $(m \times n)$**.**

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{1.2}$$

An example of this type of data is shown in Figure 1.3, where the objective would be to use the mixture spectra on the left (**D**) to extract the pure component spectra shown on the right (**S**) and the mixture concentrations (**C** - not shown). To obtain a unique solution (or a limited range of “feasible” solutions) for **S** and **C**, it is necessary to impose constraints on the solution, such as non-negativity and unimodality. Thus, instead of being solely based on variance like PCA, MCR methods are able to incorporate knowledge of the desired profiles into the optimization process. Although initially applied to UV-vis spectroscopy, MCR methods have since successfully been applied to a wide range of methods including chromatography [26], IR [27], NMR [28], hyperspectral imaging [29] and voltammetry [30].

Lawton and Sylvestre [31] have been credited as the first to look into fundamental MCR constraints, starting with simple non-negativity. Since then there have been many algorithms developed to perform MCR based on slightly different principles. The algorithm that has arguably become the most popular is multivariate curve resolution by alternating least squares (MCR-ALS) [32] which allows for a wide range of customizability in terms of profile constraints. This flexibility in functionality allows for the user to incorporate everything that is known about the data into the analysis such as non-negativity, unimodality and closure in the concentrations or profiles. Given the nature of the experimental system, many of these characteristics are known about the data and allow the algorithm to apply these restrictions to obtain more stable and reliable results. Due to this, the advantages of MCR-ALS for chemical data analysis are abundant and obvious.

1.4 Alternative Methods

Despite how well pre-existing methods may or may not perform, there is always interest in alternative methods. Since the nature of relying on variance alone is often a pitfall of PCA, there have been a number of suggestions made which aim to overcome this. One such method is maximum likelihood principal components analysis (MLPCA) which incorporates known error structures into the optimization in an effort to capture only the variance between samples rather than those due to other sources (e.g. sample preparation, instrumental error) [33]. The advantages of MLPCA have been demonstrated for a number of data analysis strategies, including clustering [34] and curve resolution [35], which are the principal focus of this work. However, MLPCA requires a reliable knowledge of the measurement error covariance matrix (ECM), which is often not available, so it is only included as a peripheral method in this work, appearing to demonstrate context in Chapter 4.

When clustering through exploratory data analysis is of interest, one alternative method that has been proposed is projection pursuit analysis (PPA) [36], which is similar to PCA in the sense that they are both projection methods based on bilinear decomposition of the data. However, instead of variance, PPA implements other metrics to evaluate the projected patterns of interest. Unlike PCA, which has a single mathematical description, PPA encompasses a philosophical approach of finding "interesting" projections of the data. This means that a variety of criteria and algorithms have been used, making comparisons more difficult. In addition to clustering, PPA has also been applied to other problems, such as regression. The present work will focus on kurtosis-based PPA (kPPA) applied to problems in clustering.

Another method that has generated a lot of interest recently is independent component analysis (ICA) [3]. Given its similarity in name to PCA it is not surprising that there are some relationships between the two methods. ICA (like PCA, MCR and PPA) also consists of a bilinear decomposition of the data, though instead of variance (PCA) or interesting projections (PPA) being the focus, the development of ICA was based around extracting independent signals from the data. Put another way, the application of PPA typically focuses on the relationships among the samples, while ICA normally emphasizes the characteristics of the source signals (e.g. pure component spectra). In this sense, ICA can be considered to be somewhat analogous to MCR, although it has also been used for other applications, such as clustering and regression.

The focal point of the current work will be to present a description of PPA and ICA through the context of their applications to chemical problems. The reader may wonder, given the many new methods developed in the field of chemometrics, why these two seemingly unconnected methods have been selected. In addition to being linear projection methods, there are several elements that connect these two techniques. While these techniques cannot be considered new (PPA was first proposed almost 50 years ago), their application to chemical systems has been intermittent, often originating from a few research groups, and their role in the analysis of chemical data is still not well-defined. The strengths and weaknesses of these techniques are often not fully appreciated, and the underlying mathematics is sometimes obscured by treatments that are overly anecdotal or, conversely, excessively theoretical. To further complicate matters, a variety of algorithms are employed for many aforementioned methods, making comparisons and discussions confusing. This work will attempt to clarify the algorithms most commonly applied to chemistry and illustrate their strengths and limitations through applications to chemical data, supporting or refuting some of the claims made in the literature. Finally, the relationship between these two methods, which has been implied but never fully elucidated in the literature, will be explored.

Chapter 2 aims to clarify some confusion surrounding PPA and ICA by starting with a general discussion of some of the theory behind these methods and a review of their applications in the analytical literature. To illustrate the strengths and weaknesses of each method, a few of the commonly used algorithms are applied to a number of selected data sets, both real and simulated in Chapter 3. Application to both exploratory data analysis (clustering) and curve resolution are considered. Chapter 4 then presents a comparison of PPA and ICA with traditional methods in a case where the measurement error structure of the data is a confounding issue. Finally, Chapter 5 draws some conclusions on the role and future of these methods in the analysis of chemical data.

Chapter 2 – Overview of Projection Pursuit Analysis and Independent Component Analysis

2.1 Introduction

In this chapter, the main goal is to provide some background for PPA and ICA in terms of their historical evolution, underlying theoretical principles, algorithms, implementations, and applications in chemistry to date. It is important to recognize that these are general methods with applications across all areas of science, so the treatment here will be necessarily limited in scope. Likewise, a comprehensive treatment of the theoretical foundations will be sacrificed for a more descriptive one which highlights only those algorithms which are most prominent in chemistry. The two methods are first discussed individually, with the connections between them described at the end of the chapter.

2.2 Projection Pursuit Analysis (PPA)

2.2.1 Background

Like PCA, PPA is an exploratory projection method that can be expressed by a bilinear model as given in Equation 2.1.

$$\mathbf{X} = \mathbf{T}_p\mathbf{V}_p + \mathbf{E}_p \tag{2.1}$$

In this equation, $\mathbf{X}$ $(m \times n)$ is the data matrix consisting of n responses (variables) for m objects (samples). For example, the rows of $\mathbf{X}$ may contain spectra measured at n channels for m different samples of, say, wine or blood plasma. For projection into a p-dimensional space (where p is normally 1, 2 or 3), $\mathbf{T}_p$ $(m \times p)$ represents the scores for each sample, $\mathbf{V}_p$ $(p \times n)$ is the matrix of $(1 \times n)$ loading vectors, and $\mathbf{E}_p(m \times n)$ is the matrix of residuals. In practice, PPA attempts to find a projection matrix $\mathbf{P}_p(n \times p)$ where

$$\mathbf{T}_p = \mathbf{X}\mathbf{P}_p \tag{2.2}$$

The matrix $\mathbf{P}_p$ contains the "projection vectors" in the columns and it is easily shown that $\mathbf{P}_p = \mathbf{V}^{\mathrm{T}}(\mathbf{V}\mathbf{V}^{\mathrm{T}})^{-1}$. The projection vectors (or projection directions) are typically orthogonal (although this is not required) and are often found sequentially by "deflating" the data (i.e. removing the variance associated with preceding vectors).

Rather than using variance like PCA, PPA searches for projection directions by optimizing a *projection index* which describes the inhomogeneity, or "interestingness", of the data such as the presence of clusters [36]. There have been a variety of projection indices developed which target certain distributions and groupings of interest within the data. Initially these projection indices tended to measure spread or local density of the points in the projected space, but more recently developed algorithms have moved more towards optimizing continuous functions.

In 1969 Kruskal was the first to suggest using lower-dimension projections to explore the structure of data [2] but the first successful application of PPA for clustering was by Friedman and Tukey [36] who coined the term “projection pursuit”. The sought projections were those that tended to produce many very small inter-point distances while maintaining the spread of the points, characteristic to the presence of clusters. This included manual projections and observations of results, followed by successive projections of clusters found in initial projections to evaluate further cluster breakdown. Friedman and Stuetzle later proposed projection pursuit regression (PPR) [37], a non-parametric regression method that searches for lower-dimensional regression surfaces containing a majority of data points and displaying underlying structures.

The key challenges in the implementation of PPA have been the identification of a suitable projection index to reflect “interestingness” and then finding ways to extract projection vectors to optimize that index. Huber [38], and Jones and Sibson [39] investigated aspects of PPA in the following years and came to a few key conclusions and generalizations. Friedman considered these generalizations and built upon his initial projection pursuit concept, specifically considering that normally distributed data is uninteresting and that projection indices should be “affine invariant”, or uninfluenced by the covariance structure of the data [40]. These conclusions point to projection indices that measure how data distributions stray from normality, rather than indicating a specific structure. This ties into the affine invariance criterion since normal distributions are characterized by the mean and standard deviation, or covariance, of the data.

With non-Gaussianity as the goal, entropy (more specifically information entropy) became a commonly discussed concept. The general equation for calculating the differential entropy of a population, x, is shown below in Equation 2.3, where $f(x)$ represents the probability density function.

$$h = -\int_{-\infty}^{\infty} f(x) log f(x) dx \tag{2.3}$$

According to information theory, a Gaussian distribution has the highest entropy for a given variance, so minimizing entropy should serve as a way to find non-Gaussian distributions. The optimization of entropy as a projection index has been explored in several works. Huber [38] and Friedman [40] performed PPA by minimizing entropy and used the least-normal projections to find clusters. Deflation by each subsequent projection resulted in orthogonal latent variables from PPA. In 1987 Jones and Sibson [39] developed a computationally efficient approach to implement an entropy-based PPA algorithm by first sphering the data and extracting interesting latent variables simultaneously. PPA in their approach was still limited to projections onto a one-, two-, or at most three-dimensional space, though three-dimensional plotting lacked the convenience it has today and could be time consuming and required specific software. High-dimensional data at the time consisted of fewer than 20 variables which made it impractical for many chemical applications (which can have thousands of variables), so projection pursuit was rarely used on chemical data until the early nineties.

2.2.2 PPA Algorithms

2.2.2.1 Sequential Projection Pursuit (SPP)

In 2000 Guo *et al* developed an entropy-based algorithm called sequential projections pursuit (SPP) which implements a genetic algorithm to sequentially search for projections [41]. The idea of information entropy has been mentioned already, however values of differential entropy are difficult to calculate. Instead of true entropy, Guo *et al* optimize Shannon entropy, expressed by Equation 2.4, for a discrete distribution.

$$\mathrm{H(X)} = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \tag{2.4}$$

A kernel density function is applied for the estimation of Shannon entropy and the data is deflated by each sequential latent variable, resulting in orthogonal projection vectors. These authors later proposed a feature selection method to select a subset of variables from the SPP solutions to preserve as much sample clustering information as possible [42]. Subsets of variables with retained inhomogeneous information (optimal consensus between the subsets and complete data set) could be selected using a genetic algorithm based on a Procrustes analysis algorithm.

2.2.2.2 Kurtosis-based Projection Pursuit Analysis

Another projection index suggested as a measure of non-Gaussian behaviour in the early days of PPA was kurtosis, the fourth statistical moment. Whereas the second moment (variance) measures the dispersion (spread) of a distribution and the third moment (skew) indicates symmetry, kurtosis is described as associated with the "tailedness" of a distribution, i.e. how much of the distribution resides in the tails [1]. For univariate data, the sample kurtosis, κ, is described by Equation 2.5, where x_i represents a measurement and N is the number of points.

$$\kappa = \frac{{}^{1}/_{N}\sum(x_i - \bar{x})^4}{\left({}^{1}/_{N}\sum(x_i - \bar{x})^2\right)^2} \tag{2.5}$$

For a Gaussian distribution, $\kappa = 3$. In PPA, the kurtosis of the scores projected onto each projection vector can be maximized or minimized to seek out non-Gaussian distributions.

Alternatively, the absolute value of the excess kurtosis ($\kappa - 3$) can be maximized, although this will likely favour a maximization of κ, which has a lower limit of unity.

A critical problem with the use of kurtosis-based PPA (kPPA) was the availability of an efficient algorithm to optimize the projection index. In 2011, Hou and Wentzell [43] presented a simple, efficient and fast "quasi-power" algorithm to perform exploratory PPA based on the optimization of kurtosis as a projection index with options for both univariate and multivariate kurtosis. The univariate kurtosis method optimizes the kurtosis of the scores on the projection vector ($\mathbf{t} = \mathbf{Xp}$, where $\mathbf{X}$ is column mean centered) according to Equation 2.6.

$$\kappa = \frac{{}^{1}\!/_{m} \sum_{i=1}^{m} (\mathbf{x}_i \mathbf{p})^4}{(\mathbf{p}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{p})^2} \tag{2.6}$$

Note that $\mathbf{x}_i$ represents a row vector from $\mathbf{X}$. For projections into spaces of dimensionality greater than 1, the optimization is applied sequentially after deflation of $\mathbf{X}$. For the simultaneous extraction of multiple dimensions, the multivariate kurtosis (for projection matrix $\mathbf{P}$) was defined as

$$K = m \sum_{i=1}^{n} \left\{ tr\left[(\mathbf{P}^{\mathrm{T}} \mathbf{X}^{\mathrm{T}} \mathbf{X} \mathbf{P})^{-1} \left(\mathbf{P}^{\mathrm{T}} \mathbf{x}_{\mathbf{i}} \mathbf{x}_{\mathbf{i}}^{\mathrm{T}} \mathbf{P} \right) \right] \right\} \tag{2.7}$$

Using this quasi-power method, kurtosis can be either maximized or minimized. Maximization of kurtosis is used to find sharp distributions with long tails (leptokurtic or super-Gaussian distributions) and can identify the presence of outliers in the data. Minimization is more commonly applied to clustering applications and is suited for finding flat or bimodal distributions (platykurtic or sub-Gaussian) which are useful for identifying clusters. This bimodal separation makes the algorithm ideally suited for binary separation of evenly populated groups in multiple dimensions. Additional notes and improvements have been made to the algorithm, including the implementation of a recentering approach which can improve results in cases where the classes are not evenly populated [44] or there is a small sample to variable ratio [45], as well as an algorithm that is able to handle complex-valued data [46].

The heart of the kPPA approach is the quasi-power method (qpkPPA), which is adopted from the power method used to estimate eigenvalues. The method is based on an iterative approach to finding the projection vectors which incorporates a learning algorithm. While the full details will not be presented here, the key equation for univariate kurtosis minimization is presented below as an example.

$$\mathbf{p}_{k+1} \leftarrow \left[\sum_{i=1}^{m}\left(\mathbf{p}_k^{\mathrm{T}}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\mathbf{p}_k^{\mathrm{T}}\right)\left(\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)\right]^{-1} (\mathbf{X}^{\mathrm{T}}\mathbf{X})\mathbf{p}_k \tag{2.8}$$

In this equation, k indicates the iteration number. Although is has only been available for a relatively short period, the kPPA algorithm has been shown to be a useful tool in exploratory data analysis.

2.2.3 Applications

2.2.3.1 Clustering

Given that PPA was developed with the purpose of finding “interesting” projections of the data, clustering has been the most common application. Exploratory applications of SPP have been sparse and have covered a variety of applications. Molecular similarity and diversity of chemical structures has been assessed by clustering and detecting inhomogeneities in FTIR spectra using SPP [47]. SPP has also been applied for characterization of the similarities of synthetic substances using electrospray ionization mass spectrometry data, and for the detection of outlying objects [48]. Other reports on applications of SPP include clustering of paracetamol formulations with the same synthesis pathway based on their trace-enriched chromatographic impurity profiles [49]. The most recent applications of SPP was by Alaerts in 2010, where it was applied to visualize the distinction of rhizomes from two Chinese herbs by analyzing chromatographic fingerprints under different extraction conditions [50].

In the search for appropriate chromatographic conditions to separate enantiomeric pharmaceuticals, Klerck *et al* [51] applied quasi-power kPPA for clustering of different enantioselective patterns of chiral polysaccharides based on supercritical fluid chromatographic systems. In the study, chromatographic systems with different chiral stationary and mobile phases were characterized by the enantioresolution of 29 racemates [51]. The quasi-power PPA algorithm has also been applied for unsupervised clustering for the forensic investigation of fraudulent documents by analyzing near and mid infrared hyperspectral images [52]. The authors were able to discriminate between different black pen inks on white and bank papers. A similar study was performed by Wentzell *et al* using FTIR spectra of blue pen inks on white paper [53]. The quasi-power PPA algorithm was able to clearly discriminate between multiple classes in both 2 and 3 dimensions.

2.2.3.2 Regression

Although the applications of PPA in analytical chemistry have been relatively sparse, the PPR algorithm has been used a number of times, primarily for applications to quantitative structure-activity relationship (QSAR) type problems. In 1996 Nguyen-Cong and Rode [54] applied PPR for quantitative electronic structure-activity relationship (QESAR) analysis of antibiotics and concluded that PPR could be a useful method for QSAR analysis when nonlinear relationships exist, which is not uncommon for these data types. Hassanzadeh *et al* [55] used the quasi-power kPPA algorithm as a compression method prior to applying radial basis function neural networks of solvatochromic descriptors for the prediction of gas chromatography retention behaviour of polychlorinated biphenyls. The compression was done by performing PPA and then applying the neural network on the scores obtained by PPA instead of the entire data set. PCA is often used for this purpose as well.

2.3 Independent Component Analysis (ICA)

2.3.1 Background

Though the origins of ICA have been disputed in the literature, a work by Herault and Ans [3] from 1984 is often credited with the first proposed concepts, while clarifications of these concepts were made by Comon in the early 1990s [56]. ICA was initially developed for the decomposition of temporal signals such as audio records with the main assumption being that the contribution signals were independent. Though many are familiar with the idea of independence there are many definitions of independence in varying scientific fields, many of which are relevant to data analysis.

The earliest references of ICA being applied in the chemical literature appears in 2000 when De Lathauwer *et al* provided an introduction to the method for use in signal processing [57]. Since then there have been approximately 156 publications applying ICA for chemical analysis, though De Lathauwer’s paper has been cited fewer than 100 times. More than 30 ICA algorithms have been reported in the literature [58] and they can generally be categorized into two groups based on their interpretation of independence: (1) methods that maximize the non-Gaussianity of the components and (2) methods that minimize the mutual information [58]. FastICA [59], joint approximate diagonalization of eigenmatrices (JADE) [60], and mutual information least-dependent component analysis (MILCA) [61] are examples of some of the algorithms most commonly used in the analytical chemistry literature.

2.3.2 Theory

The traditional ICA model used for blind source separation (BSS) can be represented as shown in Equation 2.9.

$$\mathbf{X} = \mathbf{AS} + \mathbf{E} \tag{2.9}$$

In this equation, the matrix $\mathbf{X}$ $(m \times n)$ consists of m different mixture signals of length n, $\mathbf{S}$ $(p \times n)$ is the matrix of p pure source signals that contribute to each measured signal in $\mathbf{X}$, $\mathbf{A}$ $(m \times p)$ is the mixing matrix which determines how the pure signals are mixed for each mixture in $\mathbf{X}$, and $\mathbf{E}$ $(m \times n)$ is the residual matrix. The goal of ICA is to determine the pure signals, $\mathbf{S}$, given only the information in $\mathbf{X}$. ICA is often formulated in terms of finding an unmixing matrix, $\mathbf{W}$ $(p \times n)$, which can be related to the pseudo-inverse of $\mathbf{A}$ as given in Equation 2.10.

$$\mathbf{S} = \mathbf{WX} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-\mathbf{1}}\mathbf{A}^{\mathrm{T}}\mathbf{X} \tag{2.10}$$

The similarities between the ICA and MCR models are immediately obvious. In fact, with the substitution of the contribution (concentration) matrix for $\mathbf{A}$ and the pure response (spectral) matrix for $\mathbf{S}$, the two models are equivalent. In both cases, unique solutions are not possible without additional constraints, and it is here that the two approaches differ. In MCR, constraints are consistent with chemical measurements (non-negativity, unimodality, etc.). In ICA, solutions are based on the independence of signals in $\mathbf{S}$.

ICA is often described in terms of a cocktail party scenario in which a number of digital audio recorders are placed around a room in which multiple conversations (the pure source signals) are taking place. Each device records a mixture of audio signal (determined by its distance from each conversation) over a period of time (determining the signal length, n). The m recordings make up the matrix **X** and the goal of ICA is to extract the individual conversations, **S**. To accomplish this goal, the assumption is made that the distributions are statistically independent from one another.

Statistical independence implies that knowing the value of one random variable does not convey information about the value of another. In the example above, it is reasonable to assume that knowing the magnitude of the audio signal from one conversation at a point in time tells us nothing about the signal from a separate conversation, so the variables can be considered independent. Independence can apply to both discrete and continuous variables, but the continuous case is more relevant for this discussion.

The definition of statistical independence between two continuous random variables involves the probability density functions (PDFs) of the variables. Considering two signals as random variables *x* and *y* with respective marginal PDFs $f_X(x)$ and $f_Y(y)$, then x and y are independent if and only if Equation 2.11 holds.

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \tag{2.11}$$

That is, their joint PDF is equivalent to the product of their respective individual PDFs. This principle extends to more random variables as well and the goal of ICA is to find source vectors in **S** that satisfy the model equation (Equation 2.9) while simultaneously maximizing the statistical independence of said vectors.

In practice, PDFs are difficult to calculate from finite populations such as measured signals, so often the independence of the signals is assessed using a variety of surrogate methods. Given that more than 30 algorithms have been reported for ICA [58], assessing them all is beyond the scope of this study. However, they can generally be divided into two categories based on their approach to estimating independence:

(i) methods that minimize the mutual information

(ii) methods that maximize the non-Gaussianity of components

Independent signals have defined values for other statistics as well, such as correlation. Independent random variables will have a correlation coefficient of zero (linearly independent), but it is important to note that this is a required but not sufficient condition to prove statistical independence. For this study, correlation has been used to illustrate when signals are not statistically independent, not to show their true independence.

2.3.2.1 Mutual information

Mutual information (MI) originates from information theory and is more closely related to the idea of statistical independence than to non-Gaussianity. Considering only two random variables, x and y, with marginal probability density function $f_x(x)$ and $f_y(y)$, and a joint PDF, $f_{xy}(x, y)$, the mutual information (MI) is given as

$$\mathrm{I}_{xy} = \iint f_{xy}(x, y) \log\left(\frac{f_{xy}(x, y)}{f_x(x) f_y(y)}\right) dxdy \tag{2.12}$$

The MI indicates how much information one random variable conveys about the other and has a value of zero for independent variables. Another related quantity that can be used is Shannon entropy, which is referred to as differential entropy for continuous random variables. For a variable x with probability density given by $f_x(x)$, this is given by

$$h_x(x) = -\int f_x(x) \log f_x(x)\, dx \tag{2.13}$$

The joint differential entropy of x and y is expressed by Equation 2.14 and is related to the MI by Equation 2.15 [61].

$$h_{xy} = -\iint f_{xy}\,(x, y) \log f(x, y) dxdy \tag{2.14}$$

$$I_{xy} = h_x + h_y - h_{xy} \tag{2.15}$$

Joint entropy, h_{xy}, is always larger than either of the individual entropies (h_x and h_y) and smaller than their sum [61]. For the larger values of joint entropy, the value of MI gets closer to zero, which means more independence of the variables [62]. The relationship between MI and individual and joint entropies can be extended to more than two dimensions [61].

Both the mutual information and joint entropy can be reliable ways to evaluate independence of signals, but their calculation is slow and relies on estimation of the PDF from local densities of point measurements. For this reason, alternative methods have been sought, as described in the next section.

2.3.2.2 Non-Gaussianity

The idea of using non-Gaussianity as a measure of independence comes from an interpretation of the Central Limit Theorem which states that the summation of independent random variables converges upon a normal distribution, regardless of the distributions of the initial random variables. Although the converse of this theorem is not necessarily true, it is a common generalization that the distributions of measured signals are often non-Gaussian in nature, which holds true for some data types (e.g. sound recordings) but not for others. By pairing this generalization with the central limit theorem, some researchers have concluded that non-Gaussianity can be used as a criterion for independence. This conjecture is made more ambiguous since Gaussian distributions are very well defined, but there are many ways in which a distribution can be non-Gaussian. The solution to this problem is often related to statistical moments such as kurtosis, or other non-linear functions for which Gaussian distributions have a defined value such as negentropy [59]. Examples of these statistical measures will be discussed when explaining the specific algorithms used.

It is important to note that ICA methods are based solely on the distribution of values within a signal sequence and do not consider the shapes of the signals themselves. Therefore, constraints such as unimodality (e.g. for chromatographic peaks) cannot be incorporated. Moreover, although using non-Gaussianity has become an important element in many ICA algorithms, it has a tenuous relation to statistical independence. Two Gaussian distributed signals can be independent as easily as two non-Gaussian signals. In addition, it is impossible to assess independence using a single distribution. The principle advantages of this type of algorithm are speed and simplicity.

2.3.3 ICA Algorithms

2.3.3.1 FastICA

There are many ICA algorithms readily available online to the public with the main difference between them being in the way they calculate or estimate the independence of the extracted components. Though there is a large selection of available algorithms, there are only a few which have been commonly used in the chemical literature. The algorithm seemingly used most often is FastICA which was developed by Hyvarïnen and Oja [59]. The algorithm estimates negentropy as a measure of independence and has been the algorithm of choice for problems of process monitoring.

Negentropy, like entropy, is a method of measuring non-Gaussianity and is defined as the difference between the entropy of a sample PDF, f_x, and that of a variable following a Gaussian distribution with the same variance, f_{Gauss}. It is always non-negative and is defined by Equation 2.16, where h represents the differential entropy and J is the negentropy.

$$J(f_x) = h(f_{Gauss}) - h(f_x) \tag{2.16}$$

The more non-Gaussian a distribution is the higher the value of negentropy. Negentropy values are difficult to calculate, so Hyvarïnen suggested a method of estimation shown in Equation 2.17 below.

$$J(y) \approx c[E\{G(y)\} - E\{G(v)\}]^2 \tag{2.17}$$

In this equation, E signifies the expectation, G can be any non-quadratic function, c is a constant, v is a variable following a Gaussian distribution with zero mean and unit variance, and y is a random variable assumed to have zero mean and unit variance. Although it is proposed that G could be any non-quadratic function, Hyvarïnen suggests three options expressed in Equations 2.18-2.20.

$$G_1(u) = \frac{1}{a_1} \log \cosh(a_1 u) \tag{2.18}$$

$$G_2(u) = -\frac{1}{a_2} e^{\left(\frac{-a_2 u^2}{2}\right)} \tag{2.19}$$

$$G_3(u) = \frac{1}{4} u^4 \tag{2.20}$$

In the above, a_1 is a constant between 1 and 2, a_2 is a constant approximately equal to 1 and u is a random variable. The default method used in the FastICA algorithm is G_3, which is based on kurtosis. Tong *et al* [63,64] developed a method of incorporating all three functions into their model to create a more optimal, robust solution. Rashid *et al* [65] have also explored improving the application of FastICA by incorporating a genetic algorithm to prevent the issue of getting stuck on local maxima which causes the algorithm to be inconsistent in its calculations.

2.3.3.2 JADE

The joint approximate diagonalization of eigenmatrices (JADE) algorithm [60] is another widely-used algorithm in the chemical literature and has been popularized, in particular, by Rutledge and colleagues [66–77]. The JADE algorithm estimates independence using the fourth order cumulants of the extracted signals, which differs from FastICA, but not drastically since the auto-cumulants are equivalent to the kurtosis. The cross-cumulants however provide insight about the relationship among multiple source signals, which is not something that FastICA measures. JADE also differs from FastICA in the spaces they search. While FastICA searches the entire given space for the components, JADE only considers rotations of the PCA loading vectors. Both approaches have their implications which the users should be aware of.

The JADE algorithm [60] has been applied to a variety of data types for curve resolution, including 3D front face fluorescence [68,71], Raman [72,78] and GC-MS [79–81] and has also been used to reduce dimensionality prior to regression [82]. JADE is a BSS algorithm that provides the most consistently reproducible and reasonable ICA solutions of all the available algorithms. Prior to ICA the JADE algorithm row-mean centers the data and whitens it, e.g. compresses the data to as many principal components (PCs) as there are requested independent components (ICs). The estimation of independence is then based on the construction of a fourth-order cumulant tensor from the whitened data.

Utilizing higher order statistics (HOS) in ICA, one can search for components that are as non-Gaussian as possible. JADE calculates a fourth-order cumulant tensor, which is a generalization of the variance-covariance matrix from second order statistics. The fourth order cumulants defines the fourth-order relationships among possible source signal vectors. For example, suppose we wish to evaluate three source vectors, $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ for non-Gaussian behaviour. The fourth-order cumulant tensor is created by taking all combinations of four vectors, $\mathbf{v}_i$, $\mathbf{v}_{\mathrm{j}}$, $\mathbf{v}_k$ and $\mathbf{v}_l$ such that i, j, k and l are varied between 1 and 3. Thus, the fourth-order cumulant tensor would have dimensions $(3 \times 3 \times 3 \times 3)$ and, assuming that the $\mathbf{v}$'s are row vectors of length n, the elements are calculated by Equation 2.21.

$$\mathbf{K}(i,j,k,l) = mean\left(\mathbf{v}_i \circ \mathbf{v}_j \circ \mathbf{v}_k \circ \mathbf{v}_l\right) - \sigma_{ij}\sigma_{kl} - \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk} \qquad (2.21)$$

Here $\mathbf{K}(i,j,k,l)$ is the element in position $(i,j,k,l)$of the cumulant tensor $\mathbf{K}$, "$\circ$" indicates the Hadamard, or element-wise, product and the σ_{xx} terms are the covariances between the corresponding vectors. Since JADE whitens the data before searching for independent components, the extracted vectors will be orthogonal and uncorrelated, making the covariance terms in Equation 2.21 equal to zero. When all four vectors are the same the result is called the auto-cumulant which is directly related to kurtosis and is contained in the superdiagonal elements of the fourth order tensor. When the vectors include at least two different vectors the results are cross-cumulants and are contained in the off-superdiagonal elements of the cumulant tensor.

The objective of the JADE algorithm is to find source vectors that maximize non-Gaussian behaviour. To do this, the original vectors (the principal components) are rotated in such a way as to diagonalize the fourth order cumulant tensor. This tends to maximize the auto-cumulants (super-diagonal elements) and minimize the cross-cumulants (off diagonal elements). Because the auto-cumulants represent the kurtosis values, JADE is essentially using kurtosis as a criterion to extract non-Gaussian vectors. In this way, JADE has similarities to kPPA and FastICA when a fourth-order contrast function is used. Unlike FastICA, however, JADE extracts the source vectors simultaneously rather than one at a time and does so very efficiently. In terms of kPPA, the relationship between the multivariate kurtosis and the fourth-order cumulant tensor has been derived as:

$$K_p = \sum_{i=1}^{p} \mathbf{K}(i,i,i,i) + 2\sum_{i=1}^{p-1}\sum_{j>i}^{p} \mathbf{K}(i,i,j,j) + p(p+2) \tag{2.22}$$

Here, $\mathbf{K}(x,x,x,x)$ is the fourth order cumulant calculated between the corresponding vectors as described in Equation 2.21 and p is the dimensionality of the multivariate space for which the multivariate kurtosis is being calculated. This draws direct lines between the cumulants employed by JADE and the multivariate kurtosis that can be used in kPPA, but it can be seen that the multivariate kurtosis only considers the auto-cumulants and even cross-cumulants, meaning the algorithms are still not identical.

2.3.3.3 MILCA

Another algorithm used frequently in the literature is mutual information least dependent components analysis (MILCA) [61]. MILCA calculates mutual information based on a complex nearest-neighbour algorithm and acknowledges that the extracted signals may not be completely independent, but rather as independent as possible while providing the lowest mutual information. This interpretation of statistical independence is more definitive than the non-Gaussian conjectures that FastICA and JADE are based on.

Mutual information (MI) is related to the joint probability density function of a set of random variables which can be calculated via a relationship with the joint entropy. If the joint entropy of continuous variables x and y is written as in Equation 2.14 then the mutual information of x and y can be expressed as is shown in Equation 2.15 [61]. Although these equations are relatively straightforward, there are two main complicating factors. First, because functional forms for the individual and joint PDFs are not available, MI (or entropy) must be estimated from the discrete data points available. A discussion of this methodology is beyond the scope of this thesis, but it can be appreciated that it is time consuming procedure since local densities must be estimated. Second, once an algorithm for calculating MI has been worked out, a procedure for minimizing MI through rotation of the source signals must be carried out. As a consequence, execution of the MILCA code can be quite slow. Nevertheless, it is the only one of the three commonly used algorithms that assesses true independence.

2.3.4 Applications

2.3.4.1 Clustering

Despite ICA being developed for the purpose of signal extraction, there have been many studies which feature ICA used as a clustering method. ICA has been applied for clustering and classification, or as a precursor step to other classification methods. Since ICA is often compared to PCA due to their bilinear matrix decomposition, it would seem appropriate to treat the ICA contributions in the same manner as the PCA scores by analyzing them visually or using them as input for other classification methods. Pereira *et al* [83] compared ICA to PCA and partial least squares (PLS) as a dimension reduction tool prior to linear discriminant analysis (LDA) and k-nearest neighbours (KNN) for the classification of wines using concentration profiles measure by HPLC with a photodiode array detector. The scores of the three compression methods were not analyzed beforehand, but the authors provided the rate of correct classification (%) for each method which showed that PCA and PLS produced more stable models when fewer components were retained, whereas ICA required many more components to attain adequate classification rates. Monakhova *et al* have also done studies involving the classification of wines (NMR) [76,84,85] as well as paints (XRF and IR) [84] and rice (NMR) [77]. For all studies, the authors provide PCA and ICA scores plots to illustrate the degree of separation between clusters. They have compared the classification abilities of ICA to LDA, PLS-DA and Fisher's discriminant analysis (FDA) [84,85] as well as comparing ICA to PCA for dimension reduction prior to LDA and FDA [76]. The results in these studies vary, with ICA often being on par with the other methods and no conclusive results showing it as superior. Aside from these studies, ICA has been applied for clustering electronic nose [86,87], laser induced breakdown spectroscopy (LIBS) [88] and proteomics [89] data and has been paired with other classification algorithms such as neural networks (NN) [90,91] and support vector machines (SVM) [87]. MILCA has been the most commonly used ICA algorithm for clustering [76,77,84,85] but FastICA [83,87–89] has also been applied.

2.3.4.2 Regression

ICA has often been applied for regression in the same way that PCA and PPA are used to reduce the dimensionality of data. ICA has been applied for variables compression prior to a variety of regression techniques. Wang *et al* [90,91] have applied NN to ion selective electrode (ISE) array data to create calibration models of multi-ion solutions. Both of their studies showed that using ICA prior to NN improved the prediction abilities of the models, though only one of the studies compared ICA as an alternative to PCA for this purpose [91]. Other regression methods that have been partnered with ICA include least squares-support vector machines (LS-SVM) [92], Elman recurrent neural networks (ERNN) [82], multiple linear regression (MLR) [93,94] and PLS [95].

2.3.4.3 Signal Extraction

There were no studies found where PPA was used for signal extraction. ICA however is often used for analyzing chemical data for pure signal extraction, analogous to multivariate curve resolution (MCR). ICA and MCR are similar methods in their goals but with ICA there is no information given to the algorithm about the types of signals that should be extracted (e.g. non-negative, unimodal). Another commonality of ICA and MCR is that the number of components must be decided before applying either method to ensure a reliable and accurate model. Although many publications report good results, the extracted signals are not frequently shown; instead researchers rely on statistical parameters to display their results, such as correlation to the known pure profiles or the Amari index which measures the amount of information remaining in the residuals when the data is reconstructed from the extracted signals [96,97]. There have been many studies comparing ICA to MCR methods with varying results being reported [72,78–81,96,98–100]. A study by Gaubert *et al* [72] concluded that ICA outperforms MCR, though the two methods were not treated equivalently. The number of components were calculated by different methods depending on which method was being used. Since the number of chemical components, or chemical rank, is fixed within a system and independent of the data analysis method, this should not be considered a fair evaluation and comparison of the two methods. There are multiple other studies which show MCR outperforming ICA, or

ICA being on par with MCR at best [72,78–81,96,98,100,101]. It has also been acknowledged that MCR is a more appropriate method for chemical data analysis since it has been designed to extract chemically meaningful information rather than independent signals [72,96]. ICA has been applied to many data types for this purpose including gas chromatography mass spectrometry (GC-MS) [74,79–81,102,103], voltammetry [102], NMR [67,76,104–106], Raman [72,78,107] and fluorescence [67,68,70,71,73,82].

JADE [68,71,72,78–82], FastICA [95,99–102,108–114] and MILCA [78,81,96,100,104] have all been applied for curve resolution type problems. Parastar *et al* [115] published a study in 2014 where they critically investigated whether ICA was appropriate for MCR and looked at MILCA specifically due to the MI criterion used rather than being based on non-Gaussianity. They concluded that ICA has a very limited range of applicability because chemical signals are not necessarily independent, but MILCA can produce equivalent results to MCR methods within the range of applicability.

2.3.4.4 Process Monitoring

Process monitoring is an area where ICA seems to have been applied often and involves building models based on data measured periodically by various sensors at multiple steps in a process. Some sensor responses follow a Gaussian distribution and are effectively modeled using PCA. However, many sensor responses do not behave in a Gaussian fashion so modeling the data with PCA may not be appropriate or effective. In these cases, ICA has been applied to better model the non-Gaussian behavior, and in many cases it has been shown to improve the resulting models [63–65,116–122]. The first studies demonstrating the use of ICA for process monitoring seem to have occurred around 2004 [122]. Since then there have been many methods suggested for improving this application of ICA including combining ICA models with PCA models [116,117], improving the optimization process using a genetic algorithm [65] and combining multiple ICA models that use different objective functions [63]. FastICA is often the algorithm used for process monitoring since it searches the multidimensional space to find sources that are non-Gaussian. This approach is appropriate for process monitoring for reasons discussed above but FastICA often has optimization issues which result in the algorithm reporting local maxima and providing inconsistent results.

2.4 Synopsis and Analysis of Methods

It is clear from the preceding discussion that, while PPA and ICA have not established themselves as mainstream methods, they have seen numerous applications in chemistry (perhaps just enough to confuse practitioners of their real utility). Moreover, it is apparent that neither PPA nor ICA refers to a single method, so broad statements about their applicability must be made with caution. This section attempts to summarize, based on the fundamental principles of each method, some of the key relationships. These will be further illustrated through applications in the following chapters.

A good starting point is a comparison of PPA, in particular kurtosis-based PPA (kPPA), with ICA, since it is evident that both FastICA and JADE employ kurtosis-based functions to extract “independent” signals. The similarities and differences between ICA and PPA have been mentioned briefly in many sources, but the comparisons have largely been vague and of dubious merit. Stone [123] states that PPA methods extract the projections one at a time while focusing only on maximizing the non-Gaussianity of the signals, contrasting with ICA which “typically” extracts all signals simultaneously. While the point that PPA focuses only on non-Gaussianity is fundamentally correct (since PPA only aims to find “interesting” projections of the data, where “interesting is often synonymous to “non-Gaussian”) there is no constraint on how many signals can be extracted at once. Hou *et al* [43] have even developed a multivariate kurtosis-based PPA algorithm which extracts multiple signals simultaneously. Additionally, FastICA is a popular ICA algorithm that gives the option to extract signals sequentially or simultaneously, with sequential extraction being the default option. Yet, despite FastICA having ICA in the name, Stone refers to it as a PPA method [123]. Hyvarïnen *et al* [124], the developers of the FastICA algorithm, describe how ICA algorithms can use non-Gaussianity as a measure of independence by (erroneous) relation to the central limit theorem. They state that, since PPA is “usually performed by finding the most non-Gaussian projections of the data” then PPA is the same as their explanation of estimating ICA models using measures of non-Gaussianity, therefore all the non-Gaussianity measures (e.g. kurtosis) and their corresponding ICA algorithms could also be referred to as PPA “indices” and algorithms.

One fundamental difference between ICA and PPA is quite straightforward. If we recall the relationship between scores and loadings, a critical distinction between ICA and PPA is simply that the non-Gaussianity measures are enforced on the scores for PPA and on the loadings for ICA. Although the PPA scores are not required to be uncorrelated there is a deflation step that normally performed during sequential extraction to ensure that subsequent projection vectors are not incorporating dimensions (and information) that have already been accounted for. Since the resulting vectors do not contain the same information, they tend to be uncorrelated as a result. Since these two characteristics (uncorrelated and non-Gaussian vectors) are the basis of many PPA and ICA algorithms, it is not hard to see why some believe PPA and ICA are the same. The definition of independence has many branches which allows for ambiguity among ICA algorithms. Algorithms based solely on the implied relationship between independence and the central limit theorem, optimizing non-Gaussianity rather than other estimates such as MI or joint PDFs, can be considered the same as PPA algorithms since PPA algorithms are concerned only with non-Gaussian distributions. ICA algorithms based on MI are unique from PPA algorithms since they are actually taking the information relationship into account rather than solely distributions.

Another important consideration in the application of these methods which is often overlooked is their relationship with PCA. For some ICA methods (JADE, MILCA), PCA is applied prior to the application of the procedure, with the PCA solution truncated to the number of factors corresponding to the number of independent components (source signals) to be extracted. This pretreatment, sometimes referred to as whitening or sphering, serves a number of purposes: (1) it reduces the size of the space to be examined, (2) it generates an initial set of source vectors that are orthogonal, but not necessarily statistically independent, and (3) it normalizes the variance of the data. However, this means that the source vectors produced by JADE and MILCA *are simply rotations of the original set of PCA loadings* and provide no new information. In other words, when these ICA methods are used for clustering or regression, they provide equivalent information as PCA carried out in the same space. Technically, FastICA does not have this constraint, although its application in the original variable space without PCA compression will generally produce unreliable results. On the other hand, kPPA extracts information which is often entirely different from what is obtained by PCA and is able to search a much larger space since it uses more limited variable compression or none at all.

Of the ICA algorithms described, only MILCA assesses true independence, but it is also generally the slowest. JADE is generally quite fast and produces more reproducible results than FastICA, which can be subject to local minima. The kPPA algorithm can also be fairly slow, depending on the size of the data set, and can also be prone to local minima.

When it comes to signal extraction, the applicability of ICA is arguable. Chemical signal distributions are not as stochastic in nature like the temporal signals that the algorithms were developed for and can be much shorter in length. Examples of these chemical distributions will be presented in the following chapters. More importantly, chemical signals are generally not independent, but often highly correlated (e.g. overlapped spectra). While their distributions are often non-Gaussian, this is usually not a good criterion for their extraction. Since MCR methods are specifically developed to find chemically meaningful information, it is often better to trust their results when extracting pure signals, while ICA should be used more as an exploratory tool. No matter the method it is always important to understand what is being done to the data and whether or not it is practical and appropriate.

Chapter 3 – Application and Analysis of PPA and ICA for Selected Data Sets

3.1 Introduction

In the previous chapter, the underlying principles of PPA and ICA, as well as their variants and relationships, were presented with the goal of highlighting their potential strengths and weaknesses in the analysis of chemical data. In this chapter, results are presented from the analysis of several data sets to support some of these assertions. The data sets have been carefully selected to represent a variety of situations and illustrate the characteristics of methods applied.

Both of the main applications of PPA and ICA, clustering and signal extraction, are considered in this chapter, with different data sets employed for each application category. The algorithms applied are those which have been most widely applied or appear to be the most useful in chemistry: kPPA (quasi-power method), JADE, FastICA and MILCA. In addition, PCA is also applied for clustering as a traditional method. For signal extraction, kPPA is applied to the columns of the data (i.e. $\mathbf{X}^{\mathrm{T}}$) rather than in the row direction as for clustering. This makes the application consistent with ICA methods. Likewise, for clustering, JADE is applied to the transpose of **X**.

3.2 Data sets

3.2.1 Signal Extraction Data sets

3.2.1.1 Acoustic Signals

To illustrate how ICA performs on the types of signals it was designed to handle, three open-sourced music files were selected as source signals and mixed together with simulated contributions. "*Kalimba*", "*Maid with the Flaxen Hair*" and "*Sleep away*" were the three digital music files chosen. Each file contained a matrix with 8 to 15 million variables and two audio vectors (stereo recordings). Only the first of the two audio vectors and the first 500,000 variables were used, corresponding to roughly 11 seconds of the recordings. The three resulting $1 \times 500{,}000$ source signals are shown in Figure 3.1 along

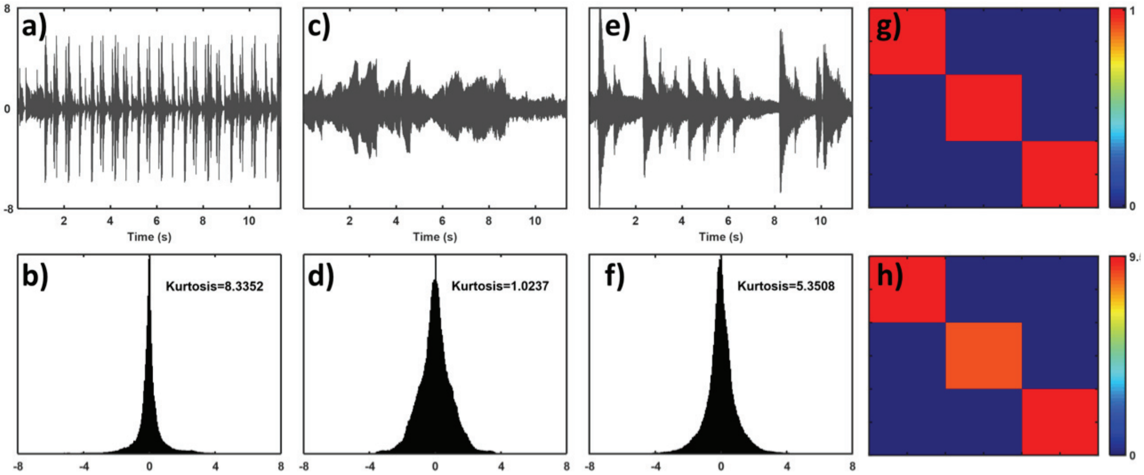

Figure 3.1 The original audio signals used with their distributions, correlations and mutual information estimation. a) "*Kalimba*" signal. b) Histogram for (a). c) "*Maid with the flaxen hair*" signal. d) Histogram for (c). e) "*Sleep away*" signal. f) Histogram for (e). g) Matrix showing the correlatations among the three source signals. h) Matrix showing the mutual information relationship among the three source signals, calculated using the mutual information estimation from the MILCA algorithm.

with their histograms and calculated excess kurtosis (excess kurtosis is the kurtosis calculated by Equation 2.5 adjusted by subtracting 3 to give a value of zero for a normal distribution). Note that all of the acoustic signals are super-Gaussian (leptokurtic) which reflects the dynamic character of musical sequences. Also shown are the correlations of the source signals (Figure 3.1g) and the mutual information as calculated by MILCA (Figure 3.1h), presented as 3×3 matrices, with the off-diagonal elements representing the correlations and MI between different signals. It is clear that the signals are both linearly independent (uncorrelated) and statistically independent (low MI).

The source signals were normalized to unit variance prior to being mixed using contributions generated from a uniform random distribution with range from zero to one,

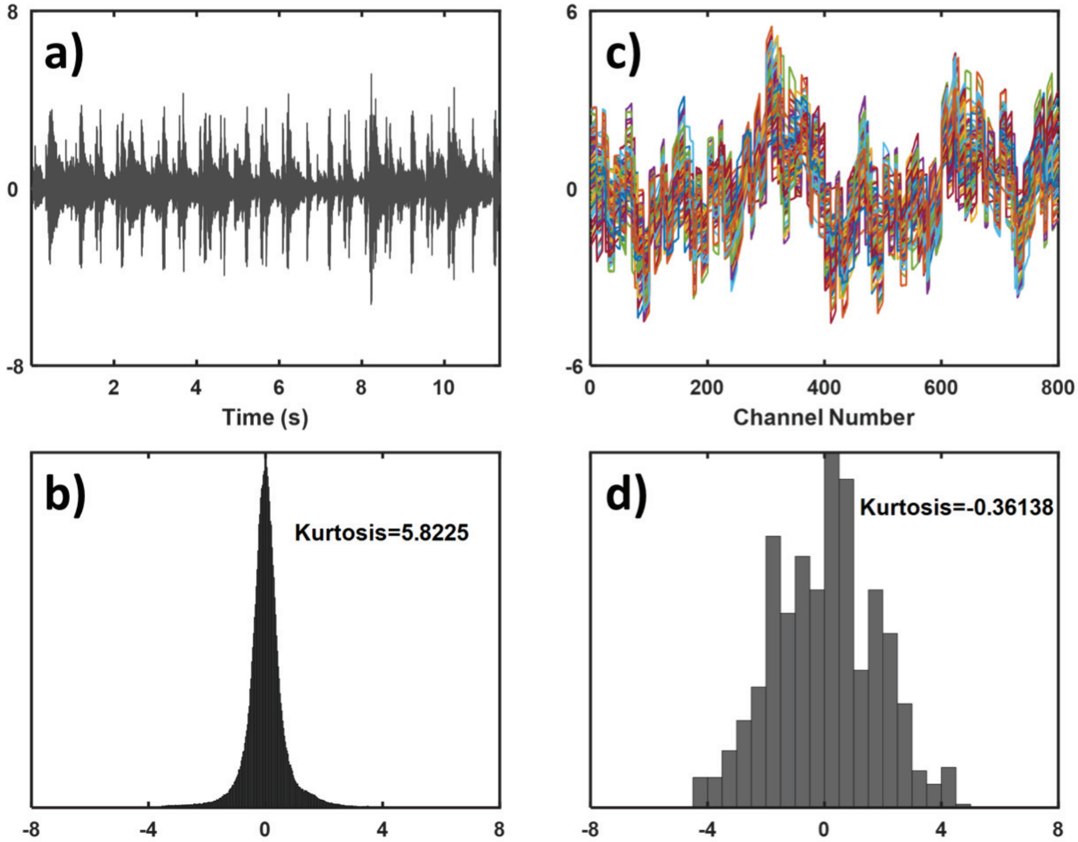

Figure 3.2 Examples of the data mixtures used for signal extraction. a) One of the acoustic signal mixtures. b) Distribution of the mixture in (a) with kurtosis displayed. c) The 100 mixtures used for the Rutledge simulation. d) Distribution of a Rutledge simulation mixture shown in (c).

creating 20 mixed audio signals. Noise was subsequently added, generated from a normal distribution with mean zero and standard deviation 0.01. The resulting data matrix was $20 \times 500{,}000$, containing the mixtures along the rows. Figure 3.2a shows a typical mixture signal and its histogram (Figure 3.2b). Note that the histogram displayed indicates that the mixture is still super-Gaussian. Although the mixture of non-Gaussian signals will tend towards Gaussian by the central limit theorem, the limited number of super-Gaussian signals considered here does not reach this limit.

3.2.1.2 Rutledge Simulation

Rutledge *et al* [69] used a simulated data set to show the utility of the JADE algorithm, so a similar data set was reproduced here for the purpose of comparing ICA methods and PPA. Although the parameters for Rutledge simulation were not clearly specified in the original work, the data were reconstructed as closely as possible. The data set was created using the Equation 3.1, where $\mathbf{X}$ is the simulated data, $\mathbf{C}$ is a matrix of contributions, $\mathbf{S}$ is a matrix of source signals and $\mathbf{E}$ is a matrix of white noise.

$$\mathbf{X} = \mathbf{CS} + \mathbf{E} \tag{3.1}$$

The source signals consist of two each of sawtooth, square and cosine waves calculated by Equations 3.2-3.4 where a is the amplitude and p is the period of each wave and x spans from 1 to 800.

$$s_{sqr} = a \csc\left(\frac{2\pi x}{p}\right) \left|\sin\left(\frac{2\pi x}{p}\right)\right| \tag{3.2}$$

$$s_{cos} = a \cos\left(\frac{2\pi x}{p}\right) \tag{3.3}$$

$$s_{saw} = \frac{-2a}{\pi} \arctan\left(\cot\left(\frac{\pi x}{p}\right)\right) \tag{3.4}$$

Equation 3.2 was used to calculate square waves with an amplitude of 1.00 and periods of 20 and 50, and Equation 3.3 was used to calculate cosine waves with an amplitude of 1.50 and periods of 160 and 320. Equation 3.4 was used to calculate two sawtooth waves with an amplitude of 1.75 and periods of 80 and 400. The resulting six pure sources were each contained in a row of $\mathbf{S}$ and are shown in Figure 3.3.

The contribution matrix, $\mathbf{C}$, was a 100×6 matrix calculated by taking values from a uniform random distribution ranging from 0 to 1. The noise matrix, $\mathbf{E}$, was a 100×800 matrix generated from a normal distribution with mean 0 and standard deviation 0.015 to resemble the signals shown by Rutledge [69]. The result was a 100×800 data matrix with the samples contained in the rows and maximum signal intensity around 6. The 100 mixtures are shown in Figure 3.2c. Figure 3.4 shows the distribution of source signals for the Rutledge simulation (a-f) and Figure 3.2d shows the distribution for a typical mixed signal. Note that, unlike the acoustic data, the source signals here are all sub-Gaussian

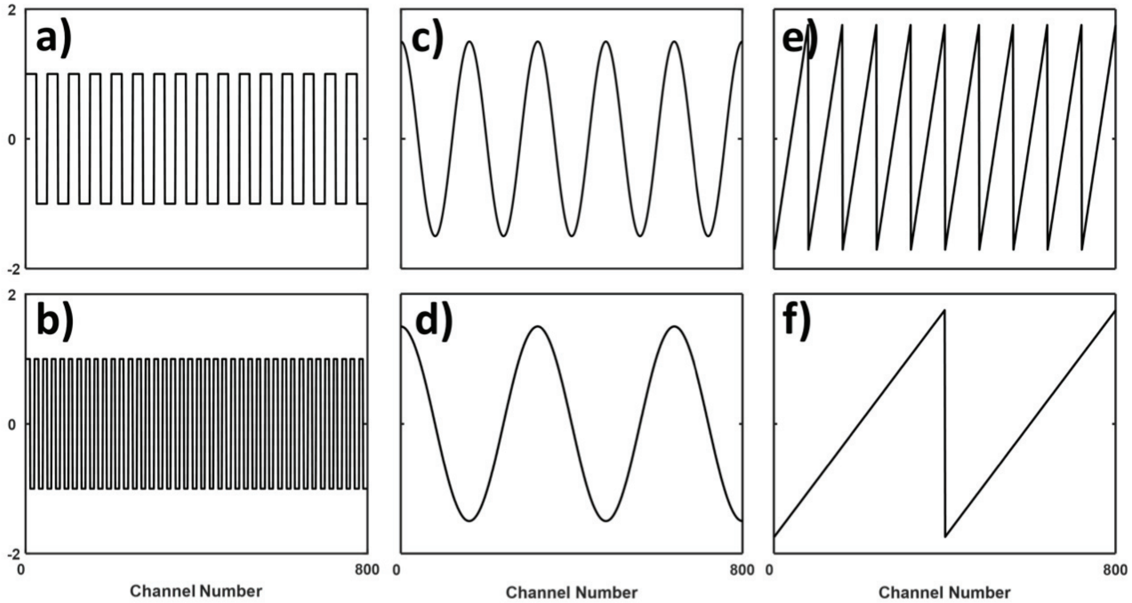

Figure 3.3 Source signals used for the Rutledge simulation. a)-b) Square waves. c)-d) Cosine waves. e)-f) Sawtooth waves.

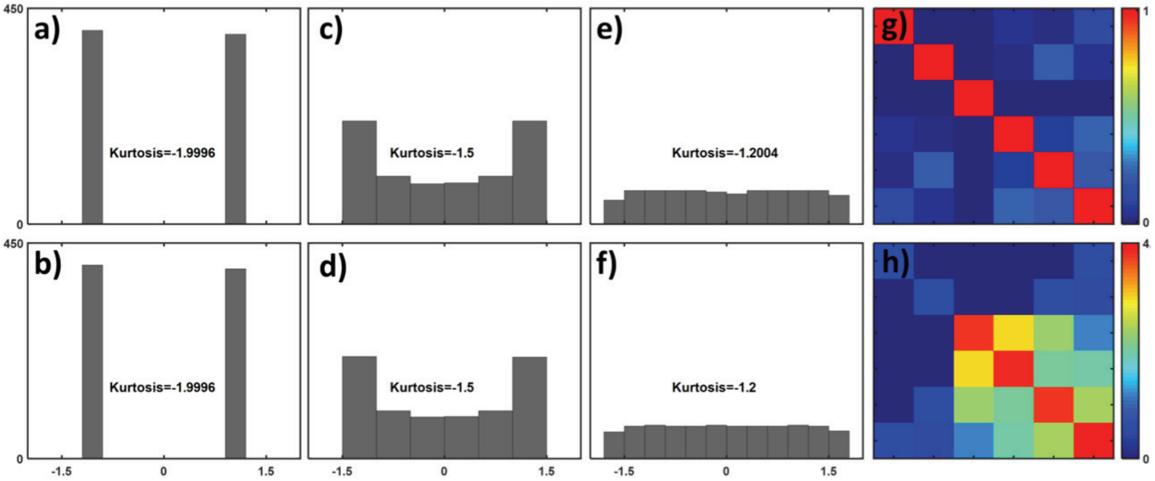

Figure 3.4 Distributions, mixtures and correlations of the source signals used in the Rutledge simulation. The excess kurtosis values of each source is displayed on each histogram. a) Histogram of square wave with amlpitude 1 and period 50. b) Histogram of square wave with amlpitude 1 and period 20. c) Histogram of cosine wave with amlpitude 1.5 and period 160. d) Histogram of cosine wave with amlpitude 1.5 and period 320. e) Histogram of sawtooth wave with amlpitude 1.75 and period 80. f) Histogram of sawtooth wave with amlpitude 1.75 and period 400. g) Matrix showing the correlation between the source signals. h) Matrix showing the MI between the source signals (calculated from the MILCA algorithm).

(platykurtic; excess kurtosis$<$ 0), but the mixed signal is closer to Gaussian. Figure 3.4g and 3.4h show the correlation and MI maps, which show that the signals are neither uncorrelated nor statistically independent.

3.2.1.3 Spectral Simulation

To illustrate a more chemically meaningful situation, three overlapping Gaussian curves were created to simulate the kind of shapes and overlap that are often seen in chemical data sets such as those derived from chromatographic and spectroscopic measurements. The Gaussian profile defined by Equation 3.5 was used, where **x** was a vector of integers from 1 to 100, σ was the standard deviation (set to 10 for all three curves), and c was the mean (chosen to be 40, 50 and 60).

$$f(\mathbf{x};\sigma,c) = e^{\frac{-(x-c)^2}{2\sigma^2}} \qquad (3.5)$$

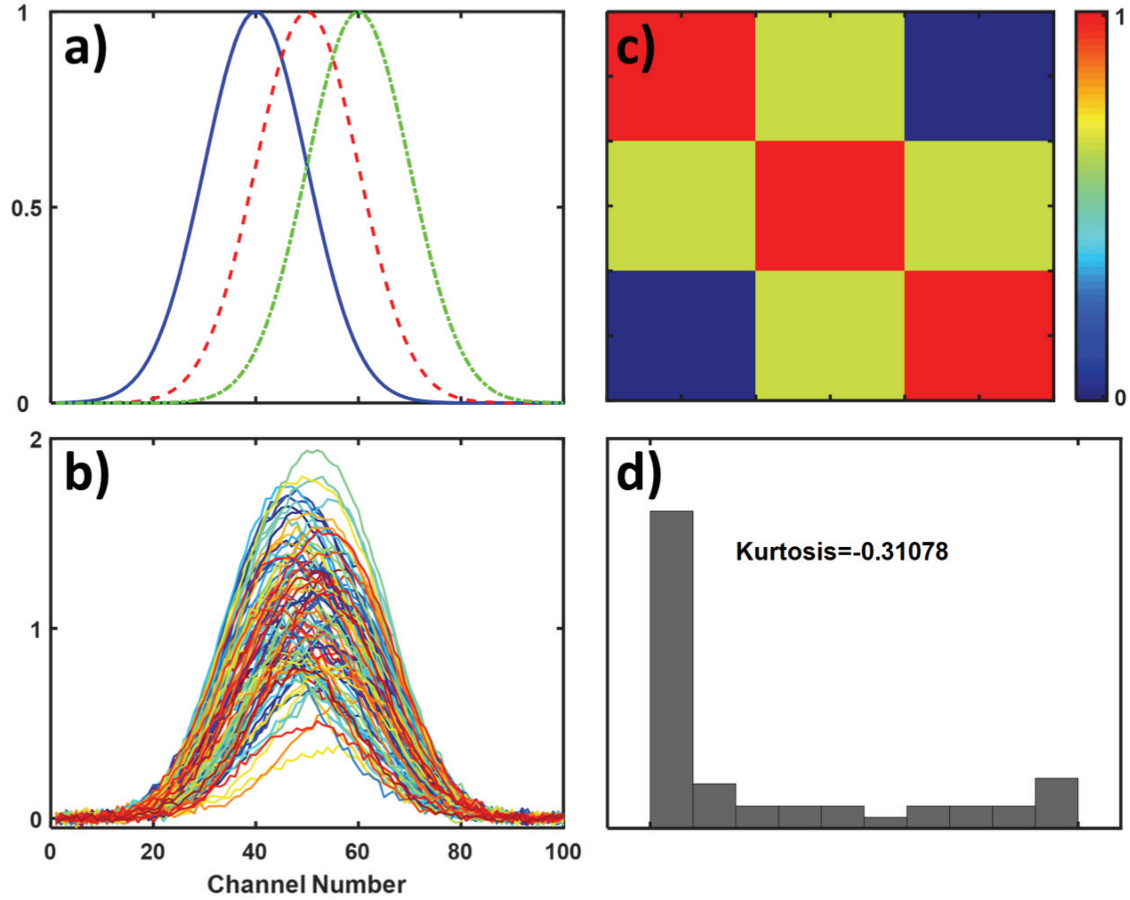

Figure 3.5 Source signals, distribution, mixtures and correlation of Gaussian simulated data. a) Source signals with mean 40 (solid blue), 50 (dashed red) and 60 (dashed-dotted green). b) 50 of the mixture spectra c) Matrix showing the correlations among the source signals. d) Histogram of the source signal distribution with excess kurtosis displayed.

A random concentration matrix, $\mathbf{C}$ (100×3), was generated from a uniform random distribution of values from 0 to 1 to simulate 100 random mixtures of the source signals. The source signals, $\mathbf{S}$ (3×100), were multiplied by $\mathbf{C}$ to generate the pure data matrix. White noise, $\mathbf{E}$ (100×100), was subsequently added with mean zero and standard deviation 0.015. The resulting data matrix, $\mathbf{X}$, had dimension 100×100, containing sample mixtures along the rows with maximum signal intensity around 2. The source signals and 50 of the mixtures are shown in Figure 3.5a and 3.5b respectively. The distribution of the source signals are very similar (since they are simply translated in time) and a typical histogram is shown in Figure 3.5d. The distributions are asymmetric (due to baseline regions) and exhibit sub-Gaussian characteristics. The correlation map in Figure 3.5c indicates that the source signals are not uncorrelated and (by implication) not statistically independent.

3.2.1.4 Fluorescence Data

The fluorescence data used here were used to represent an experimental chemical data set and were collected by Bro *et al* [125]. These consist of fluorescence emission spectra from mixtures of six different fluorophores: catechol, hydroquinone, indole, tryptophan, tyrosine and resorcinol. There are five replicate data sets with 405 samples per data set. The 405 samples can be divided into 12 separate data sets with specific design structures present for each one. The samples vary from blank samples to mixtures of the six fluorophores, with the maximum number of fluorophores per mixture being four. For this study, only samples containing at least two fluorophores were of interest, which encompassed 277 of the 405 samples. Using two of the replicate sets to maintain a high sample-to-variable ratio resulted in a total of 554 sample mixtures used. To simplify the problem to two dimensions, only emission spectra were selected from the original emission-excitation matrices, using an excitation wavelength of 275 nm since most samples reached a maximum emission intensity in that region. Once the emission spectra were selected, the region from 230-284 nm was removed to omit effects of scattering, consistent with one of the regions identified by the original authors and removed for the same reason. The result was a 554×109 data matrix. The pure spectrum for each fluorophore was calculated by taking the mean of each group of samples containing the individual fluorophore for the two replicate sets used. These mean pure spectra are shown

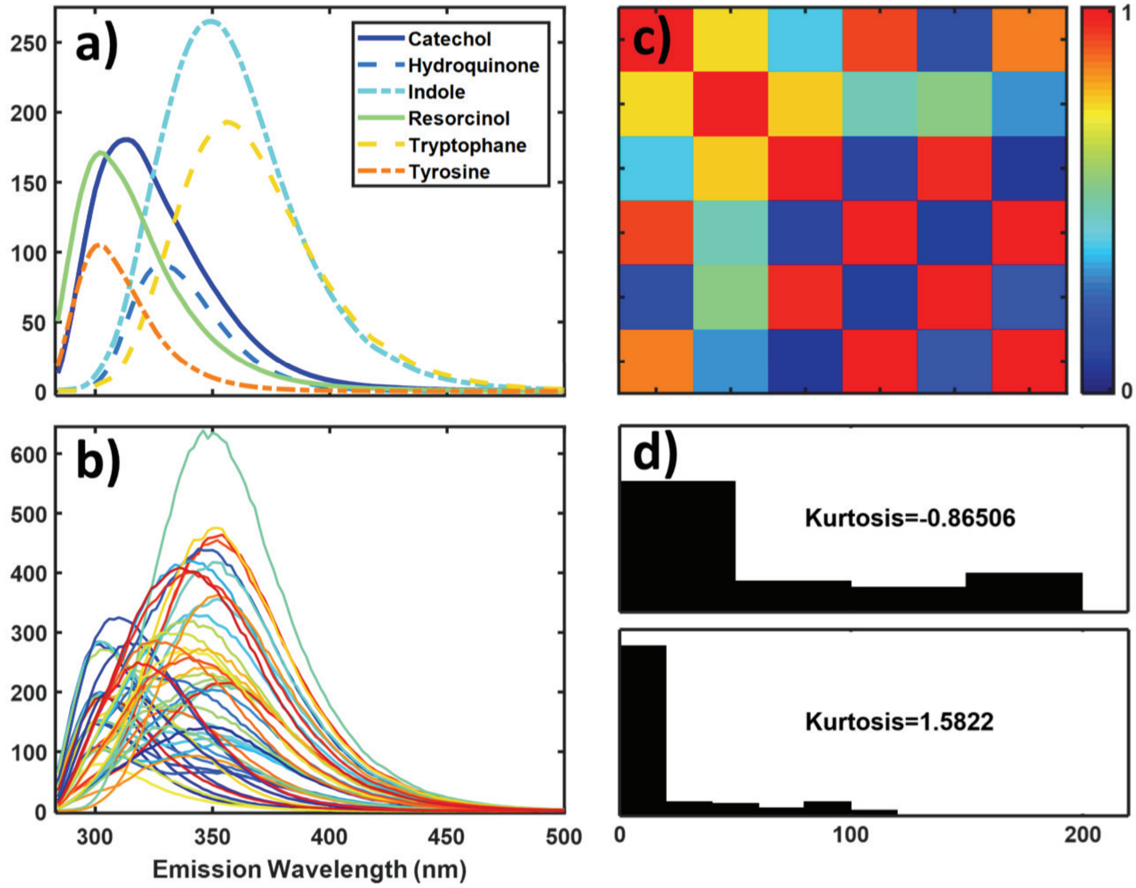

Figure 3.6 Pure emission profiles, mixture spectra, correlations and distributions of the fluorescence data. a) Pure emission spectra of the six fluorophores. b) Spectra of 50 of the mixtures used. c) A matrix showing the correlation among the pure source spectra. d) Histograms showing the distributions of tryptophan (top) and tyrosine (bottom) which have the lowest and highest kurtosis values of the sources respectively.

in Figure 3.6a alongside 50 of the mixtures used (Figure 3.6b). The correlation map in Figure 3.6c shows a significant correlation (and therefore lack of independence) among all of the pure component spectra. Figure 3.6d shows the distributions of two (extreme) spectra, which are similar to the spectral simulations in the previous section in that they are asymmetric and sub-Gaussian.

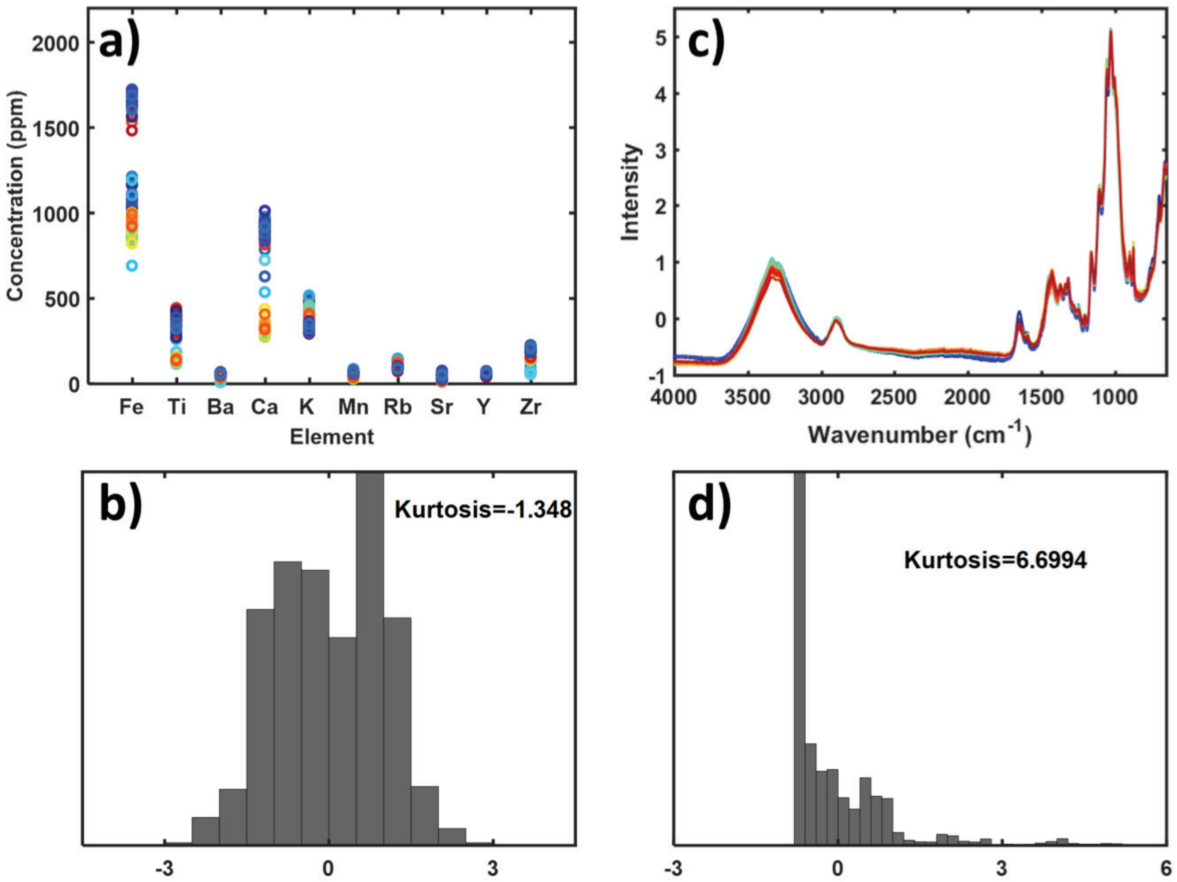

Figure 3.7 Examples of data used for clustering. a) Measured concentrations of 10 metallic elements for 63 samples of obsidian rock (no preprocessing). b) Histogram showing the distribution of data points for the data in (a) after mean centering. c) FTIR spectra of 239 ink samples (preprocessed by SNV). d) Histogram showing the distribution of one sample spectrum from (c).

3.2.2 Clustering Data sets

3.2.2.1 Obsidian Data

The obsidian data set relates to a study by Stevenson *et al* which aimed to track the movement of ancient native peoples in northern California [126]. A total of 63 samples of obsidian, a type of volcanic glass, were collected from four quarries at different locations. These were analyzed by x-ray fluorescence (XRF) for the content of ten metallic elements (Fe, Ti, Ba, Ca, K, Mn, Rb, Sr, Y and Zr) with mean concentrations ranging from about 30 to 1200 ppm. The goal for this data was to be able to apply exploratory data analysis methods to cluster samples based on geographical location. Prior to analysis, the data were column mean-centered. Figure 3.7a shows the distribution of the raw (non-preprocessed) elemental concentrations for the 63 samples and the distribution of the centered data is shown in Figure 3.7b. Note that all the data were used for this histogram to provide a better visual representation since there are fewer variables in this case than in the others presented in this chapter.

3.2.2.2 Ink Data

The ink data set [127] also relates to a classification problem but is larger than the obsidian data in terms of the number of samples and variables. The data were obtained as part of a study by Silva *et al* for the detection of fraudulent documents and consist of FTIR spectra of blue ink on paper taken over 3351 wavenumbers from 4000 to 650 cm^{-1}. The full data set contains 60 samples from each of 10 pen brands, resulting in 598 samples after outlier removal. For this study only four brands were selected (brands 6, 7, 8 and 9), resulting in 239 samples. Only four brands (classes) were selected because PPA methods are best suited to data that can be separated sequentially in a binary fashion. The spectra were preprocessed by the standard normal variate (SNV) to minimize multiplicative offset noise. The 239 preprocessed spectra are shown in Figure 3.7c accompanied by a histogram for one of the signals (following preprocessing) in Figure 3.7d.

3.3 Results and Discussion

3.3.1 Signal Extraction

The purpose of this section is to compare the PPA and ICA methods for their abilities to extract signals from bilinear data and thereby assess their value as alternatives to other curve resolution methods such as MCR-ALS. The range of data sets used are intended to span a variety of conditions, starting with acoustic data where the conditions of independence are clearly satisfied and progressing to a real chemical data set where such criteria are not met. Note that, since the intent is simply to establish if signal extraction can be successful, no comparisons with MCR-ALS are done.

3.3.1.1 Acoustic Signals

The acoustic data is an example of signals that should be statistically independent. The large number of data points and stochastic oscillations are features that contribute to the mutual independence among source signals. Figure 3.1g shows that these signals are uncorrelated which is not enough to prove independence but is a required condition. Figure 3.1h shows the mutual information calculated with the procedure used by the MILCA algorithm [61]. The mutual information estimation is a better indication of the signal independence and, since the MI among the signals is approximately zero, it is an indication that the signals are independent. Although independent, it is also commonly observed that sound data distributions tend to be leptokurtic given the oscillation and periods of silence. This is seen to be true for these signals by the distributions shown in Figure 3.1b, d and f. Due to this characteristic, when PPA was applied the kurtosis optimization was selected to maximize. There was no additional preprocessing performed before applying any of the methods.

The results of applying JADE, PPA, FastICA and MILCA to the sound mixtures are shown in Figure 3.8. In this instance, PPA was applied to the transpose of the matrix and kurtosis was maximized. The results of all methods appear to be consistent with the original signals and are indeed highly correlated with the original signals (correlation ≥ 0.9998). A number of points are worth noting about these results. First, as regards the extraction of signals with any bilinear method, there is an ambiguity of scale since the same reconstructed data can be obtained by reciprocal scaling of the signal and mixing matrices, so the results shown have been scaled to match the original signals for easy comparison. Second, there is a directional ambiguity, which means that an inversion (change in sign) of the extracted signal is an equally valid solution. Consequently, it is not unusual for the extracted signals to be "flipped" compared to the original (unless constraints are placed on the mixing matrix). Although this is hard to detect here due to the symmetry of the signals, close inspection of Figure 3.8i reveals that it is such a case. Finally, although the large number of points makes a detailed comparison of the actual signal profiles difficult (e.g. by overlay or visual comparison), they have been compared on an expanded scale. Very small differences in the signals are apparent as a consequence of noise, different optimization criteria and convergence criteria, but they are remarkably consistent, as evidenced by the high degree of correlation.

It is not surprising that the signal extraction is successful in this case given that the source signals meet the criteria for statistical independence and non-Gaussian behaviour. It is remarkable, however, that all of the methods were successful even though they were based on different objective functions and only one (MILCA) directly uses statistical independence as a criterion. Based on this success, it is easy to appreciate why JADE and

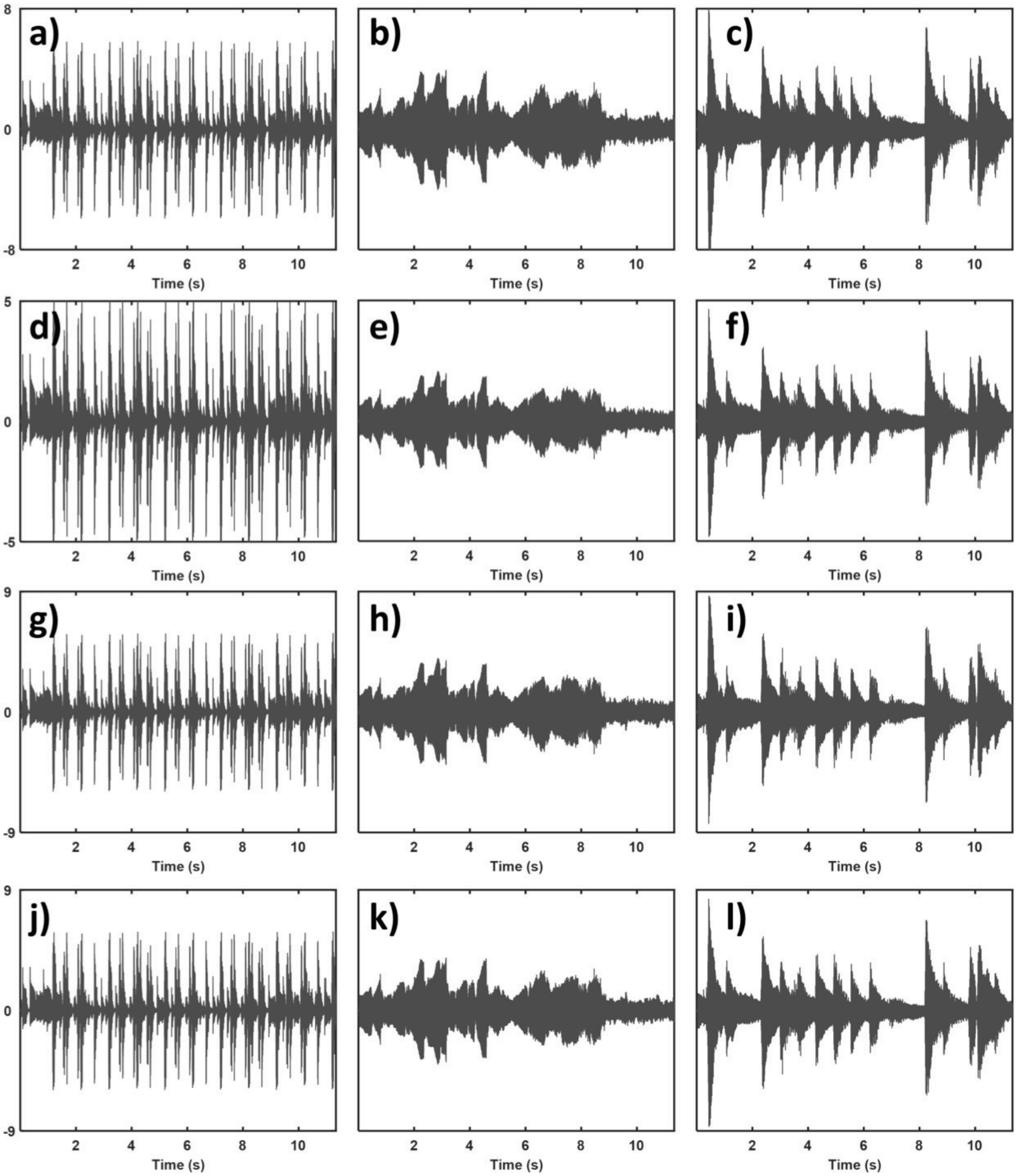

Figure 3.8 Results of applying PPA and ICA methods to the audio signal data a)-c) Results from applying JADE. d)-f) Results of applying PPA with kurtosis maximization. g)-i) Results of applying FastICA. j)-l) Results of applying MILCA.

FastICA are classified as ICA methods even though they do not directly use statistical independence as a criterion.

It is also worth commenting on the success of PPA, which was applied to the transpose of the matrix in this application. This clearly demonstrates a connection between ICA and PPA, which has been alluded to in the literature but never fully elucidated. This connection will become more apparent in the cases that follow.

3.3.1.2 Rutledge Simulation

The simulated data that were originally presented by Rutledge [69] were intended to facilitate a simple demonstration of the capabilities of the JADE algorithm using familiar and easily visualized functions. The choice of function is not entirely arbitrary, however. Although not completely uncorrelated, Figure 3.4g shows that certain pairs of signals are orthogonal or nearly so. Likewise, conditions for complete statistical independence are not met, but the definition is dubious in this case as the signals are not stochastic. In any case, JADE does not directly employ statistical independence, so this was not relevant in the original work. However, an important feature of these source signals is that they are all sub-Gaussian, which can be seen from the source histograms in Figure 3.4a-f which have their respective excess kurtosis values displayed.

Prior to applying any of the analysis methods the data were column mean centered, but no other processing was performed. When the data were analyzed by ICA methods (JADE, FastICA and MILCA), **X** was given as the data matrix and six ICs were requested from the algorithm since, for this case, we know exactly how many underlying sources there are. When applying PPA, the transpose of the data matrix, $\mathbf{X}^{\mathrm{T}}$, was used as the input to accommodate for the fact that JADE and PPA optimize kurtosis in opposite directions (e.g. scores vs. loadings). In addition, PPA was set to minimize kurtosis given our prior knowledge of the source signal distributions. The results of applying PPA to the mean centered data can be seen in the third column of Figure 3.9(m-r). The results from applying JADE can be seen in the second column of Figure 3.9(g-l). The results of applying MILCA and FastICA are shown in Figure 3.10g-l and 3.10m-r respectively. Correlation to the original signals is displayed in the figures as a numerical measure of similarity.

It can be seen in Figure 3.9g-l that JADE extracts signals very similar to the originals (a-f) (correlation displayed in the frames) although in many cases the signals are inverted (negative correlation). Since the inversion is a mathematical artifact related to the sign of the mixing matrix, it is not considered to reflect the quality of the extracted signals. The jaggedness observed in the signals, especially in signals j-l, is in part due to the fact that JADE forces the extracted signals to be orthogonal, even though all of the original signals were not. This is a consequence of the initial application of PCA and subsequent orthogonal rotation of the eigenvectors by JADE to diagonalize the fourth-order cumulant tensor. The success of JADE for this data set is due in large part to the non-Gaussian distributions of the source signals.

Because JADE and PPA are both kurtosis-based algorithms, one might expect to see similar results from the two. The PPA results in Figure 3.9m-r show some similar results to JADE (note the small differences in correlation) for some signals but become progressively more noisy as the signals become more Gaussian, to the point where the last signal is unrecognizable. The likely reason for this is that, unlike JADE, PPA does not apply PCA compression (also known as whitening of the data) prior to data analysis. As a consequence, it searches the entire data space (instead of only six dimensions) and is much

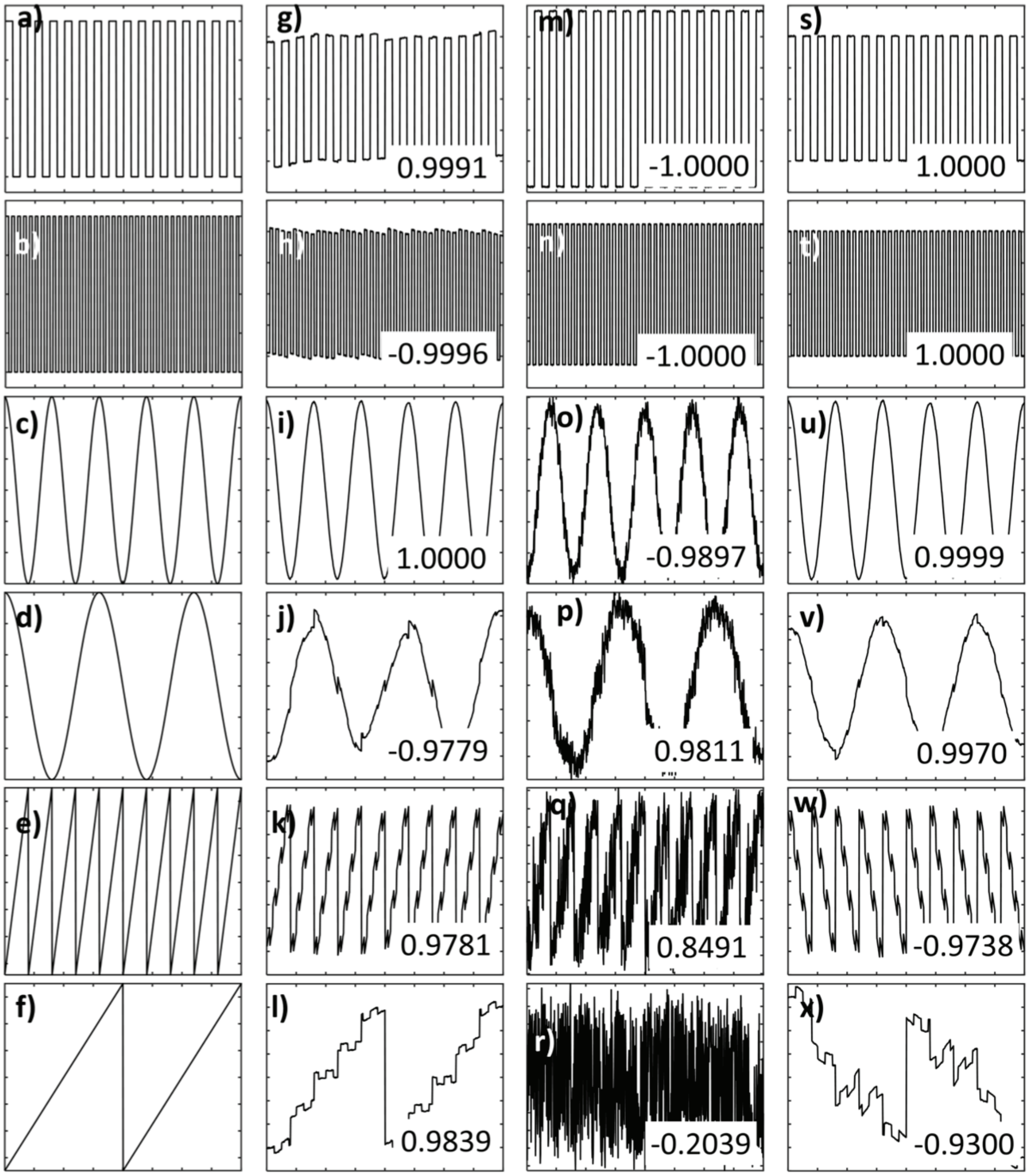

Figure 3.9 Original source signals and results of applying kPPA and ICA to the Rutledge simulation mixtures. Correlation to source signals displayed in lower right corner of each frame. a)-f) Original source signal. g)-l) Sources extracted by the JADE algorithm. m)-r) Sources extracted by kPPA using univariate kurtosis minimization on non-whitened, non-compressed data. s)-x) Sources extracted by kPPA using univariate kurtosis minimization on whitened, compressed data.

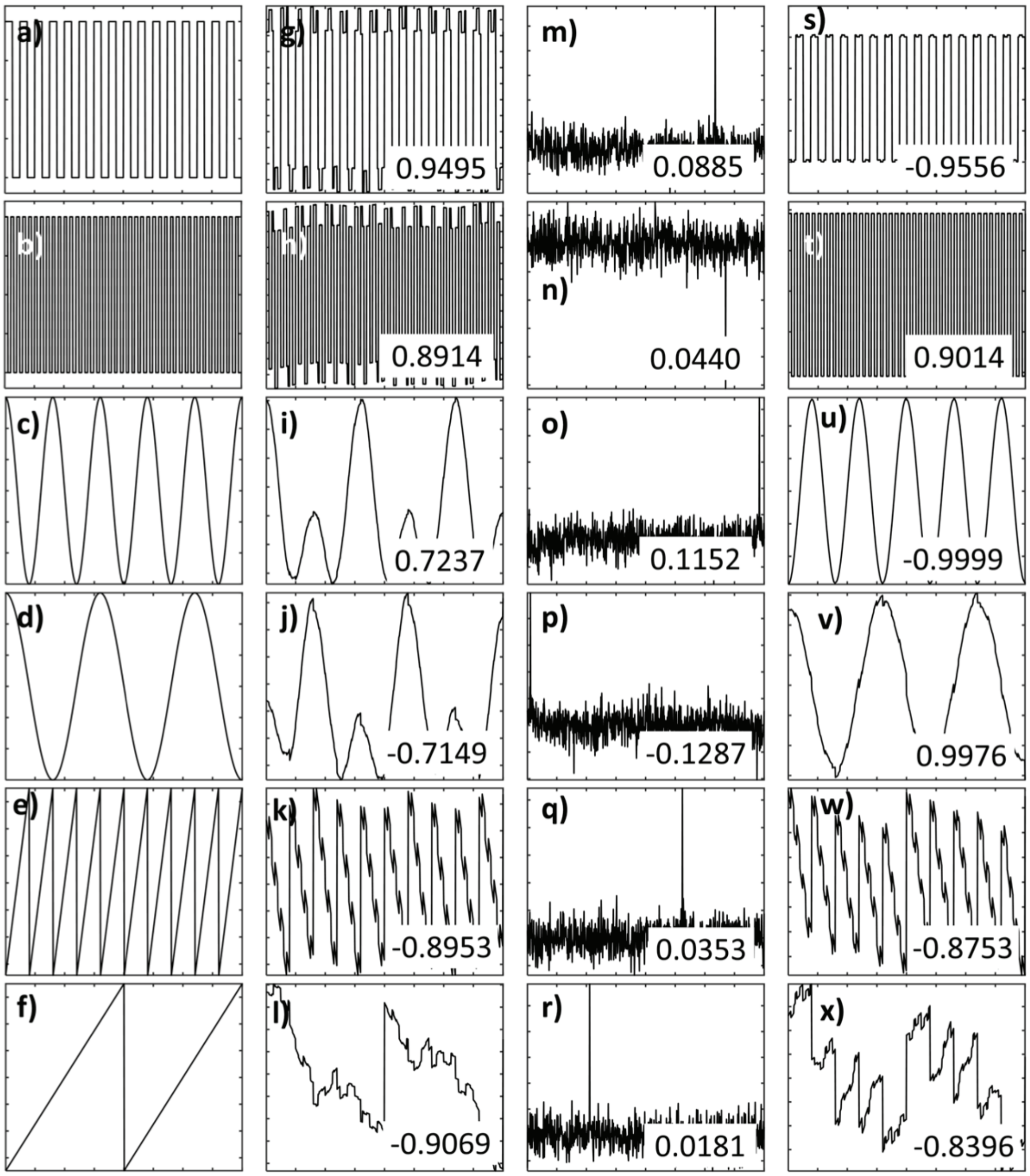

Figure 3.10 Original sources and results of applying ICA methods to the Rutledge simulation mixtures. Correlation to source signals displayed in lower right corner of each frame. a)-f) Original source signal. g)-l) Signals extracted by MILCA m)-r) Signals extracted by FastICA on non-whitened, non-compressed data. s)-x) Signals extracted by FastICA on whitened, compressed data.

more likely to be affected by random noise variation. To test this, the data were whitened in the same manner used by JADE followed by compression to six PCs prior to applying PPA. It should be noted that when JADE whitens the data it compresses the data down to as many signals as there are requested ICs (in this case six), however the PPA algorithm requires at least *n+1* input signals in order to output *n* vectors, so to combat this issue a vector of zeros was appended to the six whitened vectors. The results of PPA on the whitened data are shown in the fourth column of Figure 3.9(s-x) and clearly bare resemblance to the results obtained by JADE (except for inversion) which can be seen by comparing the correlation coefficients displayed in the figure. This once again establishes the close connection between JADE and kPPA.

Figure 3.10g-l shows the source signals extracted by MILCA. Although some similarities to the original signals (a-f) are apparent (some correlations greater than 0.9) there are particular problems with the sinusoidal signals extracted, which are linear combinations of the original signals. The reason for this is that MILCA applies very different criteria than the other methods. In addition to forcing signal orthogonality (again facilitated by pre-whitening by PCA) MILCA seeks rotations that minimize the mutual information, regardless of the distribution of the signals. These differences reinforce the idea that not all ICA methods are the same, even though the results with acoustic signals seemed to indicate otherwise.

The initial results with FastICA, shown in Figure 3.10m-r, indicate a complete failure of this method to extract meaningful signals. However, like PPA, the default application of FastICA searches the entire space, making convergence unreliable and subject to random noise. To combat this, the same approach was taken as with PPA and the data where whitened prior to applying FastICA. The results of FastICA on the whitened data are shown in Figure 3.10s-x and are strikingly similar to those extracted by PPA with comparable correlation coefficient for many of the signals. Both square waves extracted by PPA and FastICA on the whitened data resemble the original signals more closely than those extracted by JADE or MILCA (though the correlation coefficients do not reflect this) while the other four signals are nearly identical to JADE, with the larger sawtooth (bottom row of Figures 3.9 and 3.10) extracted by PPA and FastICA containing more contributions of the square waves than that extracted by JADE.

A number of conclusions can be drawn on the basis of these results. First, as one might expect, the kurtosis-based methods (kPPA, JADE, FastICA) produced similar, although not identical, results. The differences can be attributed to distinct variation in the algorithms; while all of the methods are founded on the idea of kurtosis, each approaches the problem somewhat differently. Second, it is clear that MILCA uses different criteria and therefore produced some results that were distinctly different. Third, it is clear that, for kPPA and FastICA, compression/whitening by PCA may be necessary for successful results. This is not inconsequential, since it limits the solutions to the PCA space, and this can have important implications for the clustering applications discussed later. Finally, it appears from these simulated data that these methods are a viable alternative for extracting signals from chemical data. However, the design of this data set was not arbitrary, and the true test requires data more typical of a chemical system.

3.3.1.3 Spectral Simulation

The high degree of signal overlap generated here results in sources that are not uncorrelated which is shown by the correlation map in Figure 3.5d. The yellow regions in

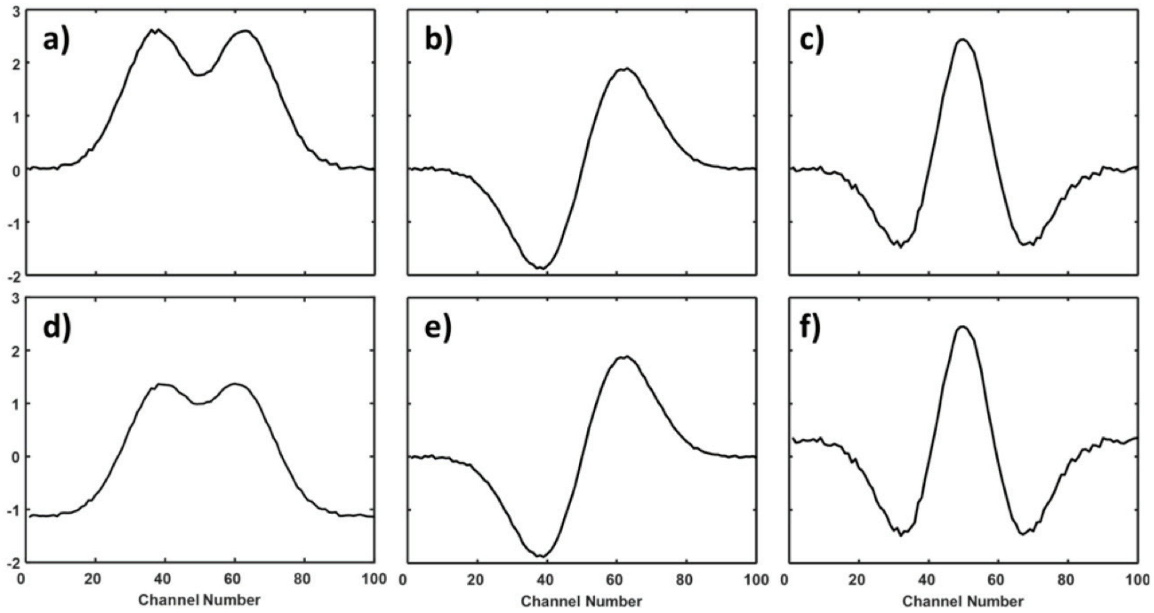

Figure 3.11 Results of applying JADE and PPA to simulated spectral data. a)-c) Signals extracted by JADE. d)-f) Signals extracted by PPA using univariate kurtosis minimization on whitened and compressed data.

the off-diagonal illustrates high correlation among adjacent signals. Since these signals are simply shifted versions of each other, they all have the same distribution and kurtosis which is shown in Figure 3.5b. The kurtosis values are much higher than those in the previous simulation, though still platykurtic, which is not always the case with chemical signals, especially those with a considerable amount of baseline signal which skews the data towards zero and hence creates leptokurtic distribution.

Given the results shown in the first two data sets, only one ICA method (JADE) will be applied going forward in this work. JADE was selected based on its performance in the other cases, as well as its overwhelming popularity in the literature and the fact that it is kurtosis based. Before applying JADE the data were mean centered, and prior to PPA the data were whitened according to the process described in the previous section since it has already been shown that whitening is an important factor when comparing these two methods. As was the case with the previous data sets, the data were transposed prior to the application of PPA and three components were requested from both algorithms. The results of both JADE and PPA on the simulated spectral data can be seen in Figure 3.11.

It is clear from inspection of Figure 3.11 that the extracted signals bear little resemblance to the original Gaussian-shaped spectra. Instead, the results are linear combinations of the original spectra. This could have been anticipated since the constraints of the algorithm force the extracted vectors to be rotations of the original PCA eigenvectors, which are orthogonal. Since the original source signals were not orthogonal, it is impossible for this approach to reproduce them. Although the results are not shown for FastICA or MILCA, they are similarly restricted. The JADE and PPA algorithms are also based on finding non-Gaussian distributions (note that it is important to distinguish the Gaussian *shape* of the source signals from the distribution of signal amplitudes). Notwithstanding the failure of the orthogonality constraint, there is no reason to anticipate if or how the distribution of source signals in chemical data will deviate from Gaussian behaviour. This is a second reason to question the successful application of ICA for the extraction of chemical signals. However, the similarity of the results for JADE and PPA in Figure 3.11 once again points to the close connection between these algorithms.

Based on the successful application of ICA for the acoustic data and Rutledge data, it is easy to see why it has been proposed as an alternative to MCR-ALS for the analysis of chemical data. However, this simulation clearly shows the limitations of this approach. This will be reinforced in the next section, which considers experimental data.

3.3.1.4 Fluorescence Data

The data presented so far have been the result of simulations which, while useful, have certain limitations. This section employs experimental data to further examine the capabilities of ICA and PPA for signal extraction.

The degree of overlap between the pure fluorescence emission spectra in this data set is even higher than that of the simulated data in the previous section. The correlation map in Figure 3.6c shows that the correlation among the source signals which is very high for combinations. This illustrates that the signals are neither uncorrelated nor statistically independent, which is not unusual for chemical data. The distributions of the source signals vary slightly with excess kurtosis values ranging from -0.87 to 1.5, which are not drastically different from the kurtosis of the normal distribution. The histograms of the spectra with the lowest and highest kurtosis values are shown in Figure 3.6d displaying the similarities between the two.

Prior to applying JADE and PPA, the data were column mean-centered. Additionally, prior to PPA the data were whitened and compressed using the method implemented by JADE. Since the kurtosis values of the source signals are a mix of both lepto- and platykurtic, both maximization and minimization approaches of PPA were applied.

The results of the methods applied are shown in Figure 3.12. It can be seen that none of the methods were able to extract accurate representations of the source signals. Something interesting to note is that the PPA results from kurtosis maximization very closely resemble the JADE results. The source signals are highly correlated with one another, again violating the assumptions of linear independence required for ICA. Further analysis of results shows that none of the results from JADE or PPA are highly correlated with the source signals, but the results of PPA using kurtosis maximization are highly correlated with the JADE results (correlation 0.9883-0.9998). Overall, none of these approaches have proven to be appropriate techniques for source signal extraction for this experimental chemical data. This reinforces the conclusions from the previous applications.

3.3.2 Clustering

In this section, the utility of ICA and PPA for clustering applications are investigated. In particular, the JADE and quasi-power PPA algorithms are explored because of the similar characteristics that have been observed in the earlier sections. PCA is also applied as a standard reference method for exploratory analysis.

3.3.2.1 Obsidian Data

The data were preprocessed by column mean-centering prior to all analysis methods. For both PCA and PPA the samples were contained along the rows of the data matrix since their objective functions (variance and kurtosis respectively) optimize in the scores space, but for JADE the transpose of the matrix was used to maintain an equivalent direction of optimization.

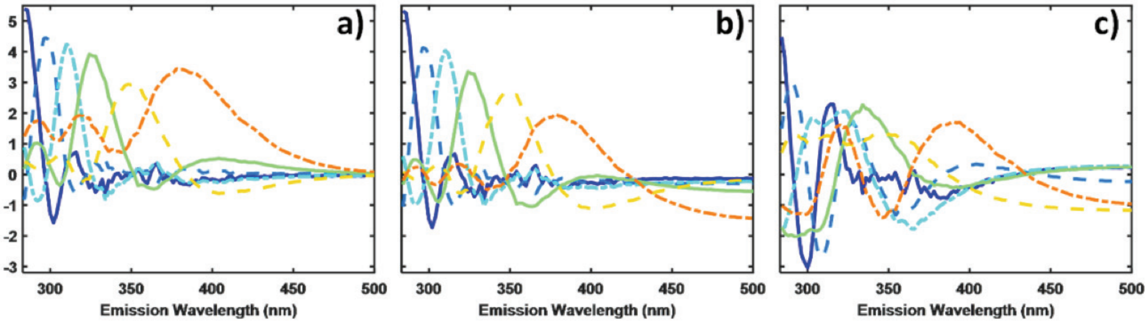

Figure 3.12 Results of applying JADE and kPPA to fluorescence data. a) Signals extracted by JADE. b) Signals extracted from kPPA using univariate kurtosis maximization. c) Signals extracted from kPPA using univariate kurtosis minimization.

Since the objective here is to observe the presence of any clusters, PCA was applied as well as PPA and ICA and the resulting scores are shown in Figure 3.13. The PCA results seen in Figure 3.13a show the presence of some clusters, with samples from quarries 1 and 2 overlapping. The result of PPA using stepwise univariate kurtosis is shown in Figure 3.13b which show quarries 1 and 4 overlapping as well as quarries 2 and 3, which is different than the overlap present in the PCA results. Given the relationship between

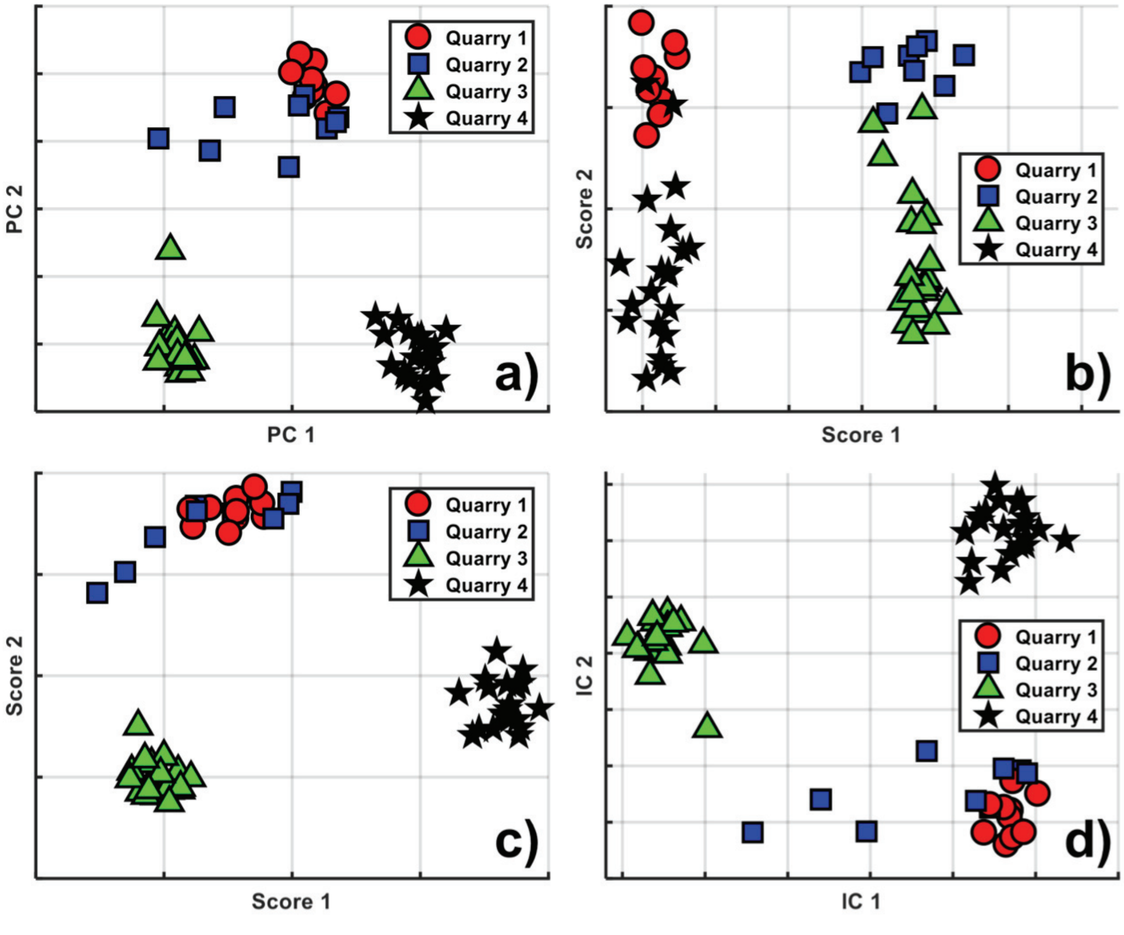

Figure 3.13 Results of applying different exploratory methods to the obsidian data. a) Scores plot obtained from applying PCA to the obsidian data. b) Scores plot obtained from applying kPPA with univariate kurtosis to the obsidian data. c) Scores plot obtained from applying kPPA with multivariate kurtosis to the obsidian data. d) IC plot obtained by applying JADE to the obsidian data and extracting two components.

multivariate kurtosis and fourth order cumulants given in Chapter 2, multivariate PPA was also applied to these data and the results are shown in Figure 3.13c. The results from multivariate PPA more closely resemble the PCA results than the univariate PPA, though the spread of samples from quarry 2 differs between PCA and PPA.

When applying JADE there were two ICs requested and those results are shown in Figure 3.13d. Looking at those results it can be seen that the JADE IC plot is merely a geometric transformation of the PCA scores, specifically a reflection and small rotation. This is to be expected given that JADE uses a PCA compression of the data and simply performs an orthogonal rotation of the PCA vectors. To try to achieve more interesting results than those acquired through PCA, JADE was applied and three ICs were extracted with all possible combinations displayed in Figure 3.14. The plot of IC 1 vs IC2 is similar but not identical to the results from when two ICs were extracted, but the other two plots show very different information, and both show a better separation of the four quarries than the initial results. What this shows is that, due to the simultaneous nature of IC extraction by JADE, the results will likely differ when different numbers of ICs are extracted, which is not the case with PCA or PPA. Additionally, requesting fewer ICs (2 or 3) will result in little to no extra information than what would be present through PCA since those PCA spaces are already easily analyzed.

In this example, none of the methods were able to completely resolve the four quarries, although the best results were perhaps those with JADE with three components (IC2 vs IC3). However, there are some subtle points that should be clarified. First, PPA searches the entire space available, so its results can be quite different from PCA, since it uses different criteria. In this case, the limited success of PPA is likely due to the unbalanced data, since the algorithm favours groups with equal numbers of samples. Some evidence of this is apparent from Figure 3.13b. On the other hand, ICA is limited to the PCA subspace used and simply rotates the data in that space. In other words, ICA does not provide more information than is available in the PCA subspace used, although the rotations may provide useful visual perspectives. From a practical point of view, the use of a large number of PCs with ICA can lead to a degradation in the quality of results. In summary, despite the similarities in the two algorithms demonstrated in the context of signal extraction, they can yield very different results when applied to problems in clustering.

3.3.2.2. Ink Data

Because PPA is applied to the entire space, it can encounter problems if the sample-to-variable ratio is low, so compression by PCA is often used to reduce the number of variables. This is different from the compression used by ICA, since it is unrelated to the number of projection vectors extracted.

In a previous study by Wentzell *et al* [53] involving the data used here, it was shown that variable compression by PCA prior to applying PPA can significantly improve results. For this data set in particular there were many regions of PCA compression that resulted in

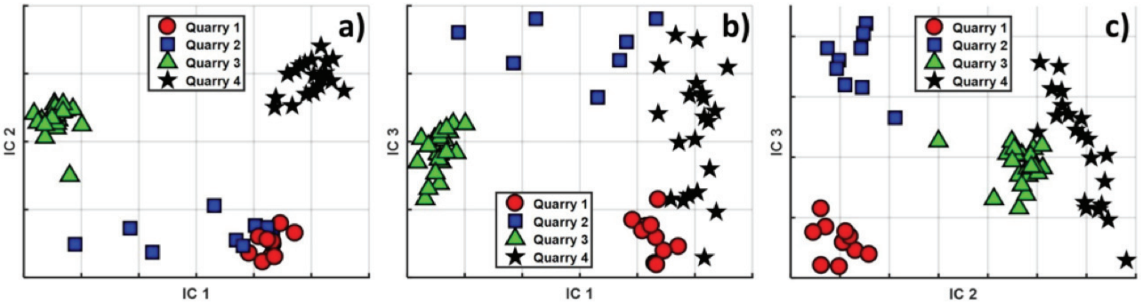

Figure 3.14 Results of requesting more components from JADE. a) IC1 vs IC2 b) IC1 vs IC3 d) IC2 vs IC3

the clustering of the four classes. Based on those results, the data were compressed to 24 PCs for this study. In other words, PCA was performed on the data and the first 24 scores were retained and used as the input for both PPA and JADE. For PPA the 239x24 scores matrix was used while for JADE this matrix was transposed to ensure the proper dimensionality of optimization.

The previous study displayed the ability of univariate PPA to separate the four classes and those results have been recreated here in Figure 3.15c. A better comparison to JADE

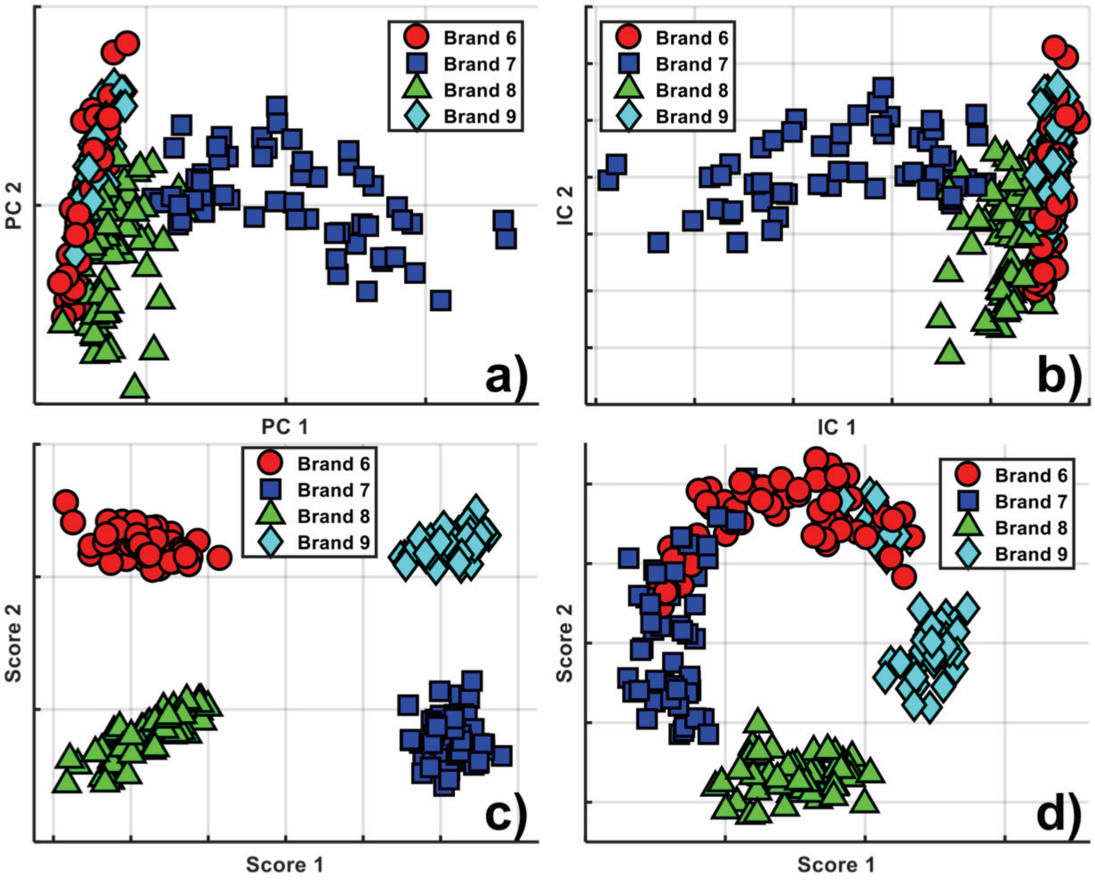

Figure 3.15 Results of applying PCA, JADE and kPPA to the ink data. a) Scores plot obtained by PCA. b) Result of applying JADE and requesting 2 ICs. c) Scores plot obtained from kPPA using univariate kurtosis minimization. d) Scores plot obtained from kPPA using multivariate kurtosis minimization.

for this application might be the multivariate kurtosis approach of PPA which extracts vectors simultaneously. The results of this approach are shown in Figure 3.15d.

Unlike the obsidian data, PCA is not able to find clusters present in the ink samples (Figure 3.15a), but univariate PPA is able to separate them clearly. The multivariate PPA is able to push the classes in separate directions but there is a lot of overlap among classes. The result of extracting two ICs only is presented in Figure 3.15b which appears to simply be a mirror image of the adjacent PCA results. As before, this is a direct consequence of the two-dimension PCA compression prior to ICA. The results of extracting three ICs are shown in Figure 3.16 and it can be seen that, in this more complex example, two of the results resemble those of two ICs but the arrangement of samples has clearly changed. This illustrates an important property of the JADE method: that the solutions are not nested. In other words, the components extracted will not be the same if a different number of components is extracted. This is not true with univariate PPA.

Although JADE may have some advantages when applied to cluster analysis, its principle drawback is that it is constrained to the dimensions of the PCA subspace used. If projected into two or three dimensions, this means that the same results can be visualized with PCA. The use of more components may produce informative rotations, but this is a

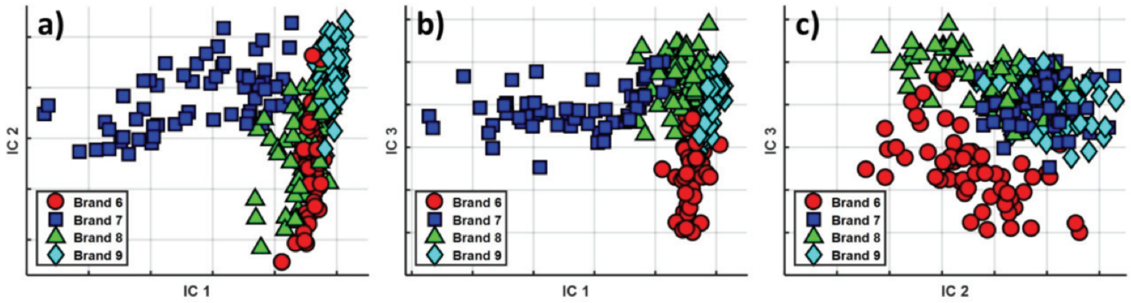

Figure 3.16 Result of requesting more components from JADE for the ink data. a) IC1 vs IC2 b) IC1 vs IC3 d) IC2 vs IC3

trial and error process which could require picking different numbers of components and visualizing the results pair-wise. In contrast, PPA explores the entire available space and seeks the most interesting subspace projections on the first pass. The success of this strategy is clearly illustrated in this example.

3.4 Summary

PPA and ICA are both exploratory tools that have existed for decades and have recently been incorporated into chemical analysis. While the exploratory goals of PPA are quite general and many projection indices can be used depending on the desired projection, ICA was developed with very specific goals in mind, namely extracting independent signals. The main hurdle when it comes to ICA methods is the way in which independence is estimated. Since many ICA algorithms make the leap to non-Gaussianity to estimate independence the distinction between PPA and ICA becomes blurred. It has been shown here that when kurtosis is used as the measure of non-Gaussianity, both PPA and ICA algorithms achieve similar results. This fact is true despite that the sequential extraction used by PPA differs from the simultaneous extraction used by JADE.

The similarities between PPA and JADE regarding signal extraction are immediately obvious, but the criteria used are not sufficient for chemical signals. The data explored here had varying results, with success achieved only for non-chemically meaningful data. The main issues with the application of ICA methods to the extraction of chemical source signals (e.g. spectra, chromatograms) is that such signals (where application is warranted) are usually neither uncorrelated nor statistically independent, and assumptions about non-Gaussianity are normally dubious or ambiguous. Thus, this method can be expected to fail in the general case. Moreover, incorporating non-negativity is a problem with chemical signals among other things. There has been a study exploring the effects of preprocessing on ICA results [108], specifically derivatization of the signals, which concluded that higher derivatives provide better results. However, it is doubtful that these would fully resolve the limitations of ICA for signal extraction.

With regard to clustering, the target number of components is not as clear as with signal extraction. PPA has the flexibility of being able to extract the same results independently of the number of components requested, as well as being able to select minimization or maximization of your projection index which provides more control over the desired results. The JADE algorithm lacks the ability to explore the full space and is limited to the selected PCA subspace, making it difficult to obtain consistent and useful results. Despite this, it can still provide interesting results, but the variability requires a varying number of components to be extracted during multiple runs of the algorithm in order to explore the possibilities of interesting projections.

Chapter 4 – The Role of Measurement Errors in Exploratory Data Analysis: A Case Study[1]

4.1 Introduction

The central theme of this thesis up to this point has been an examination of the fundamental and applied aspects of PPA and ICA in the context of chemical data, with a particular emphasis on clustering and signal extraction. In addition to a critical evaluation of the methods themselves, an attempt has been made to show the relationship between kPPA and some forms of ICA, especially JADE. A question at this point is whether or not these methods are useful alternatives to existing methods such as PCA, HCA and MCR-ALS. This question can be considered in two parts: (1) do the methods achieve their goals?, and (2) do they solve problems that cannot be addressed by traditional methods?

In terms of curve resolution (signal extraction), it is clear from Chapter 3 that ICA and PPA as applied do not meet the basic threshold of providing a correct answer except

[1] This chapter is based on a published article [149] for which the thesis author was a main contributor, performing the data analysis, interpretation and discussion of the results. The thesis author did not participate in sample collection or recording measurements.

in special cases. Despite a considerable volume of literature on this topic, this outcome could have been anticipated. Chemical signals are not, in general, statistically independent, and so they fail to meet the basic criterion for ICA. Ironically, however, two of the three ICA algorithms examined, as well as PPA, do not look for statistical independence, but rather non-Gaussian distributions with an orthogonality constraint. These methods are still (generally) not useful for chemical signals which do not adhere to distributional characterization and are not (usually) linearly independent (orthogonal). Therefore, the utility of these methods for curve resolution can be readily dismissed.

The case of clustering is another matter, however. It was shown in Chapter 3 that PPA can provide information not available from PCA because it uses different projection criteria. PCA is based on variance and assumes that the largest source of variance in the data is the between class variance. When this is not the case PCA may fail to reveal the anticipated class structure. The same is true for HCA, which is based on similar metrics. Although preprocessing of the data may help, in many cases another method may be required.

The motivation for this study is the discrimination of different wood species for the purpose of fraud detection which has been explored in a few other studies [128,129]. The goal of this chapter is to examine in more detail a particular case where traditional methods (PCA, HCA) fail and alternative methods are more effective. For the NIR spectral data presented here, the problem arises (at least in part) from the structure of the measurement errors, which are both heteroscedastic and correlated. This is demonstrated through an examination of the error structure and application of MLPCA, which results in better clustering. The application of kPPA and ICA (JADE) are then examined as alternatives that do not require a knowledge of the error structure.

The chapter begins with a brief review of PCA and HCA as traditional methods that are applied to clustering problems. The concept of MLPCA is then introduced, with a discussion of measurement error structure as a necessary prerequisite. These methods are then applied to a problem in the classification of wood species using NIR spectra and the performance of PPA and ICA as alternatives is assessed.

4.2 Background

4.2.1 PCA and HCA

Since PCA and HCA are widely used techniques, and were introduced in Chapter 1, only a brief overview will be provided here to place them in the context of the other methods to be applied, and the reader is referred to more detailed treatments in standard texts [1,4,6]. Recall that if the measurement data are represented by the matrix **X** (m x n) consisting of n variables (measurement channels) for m objects (samples), then the PCA decomposition results in an orthogonal rotation of the original variable space such that the data matrix can be represented as:

$$\mathbf{X} = \mathbf{TP} \tag{4.1}$$

where **T** (m x p) is the scores matrix, giving the coordinates of the objects in the new space, and the rows of **P** (p x n) represent the eigenvectors (or loadings), which define the rotation directions of the original space (*i.e.* the linear combinations of the original variables giving rise to the new variables). The dimension p will be the smaller of m or n, and defines the mathematical rank of **X**. There are an infinite number of possible rotations of the original space, but PCA provides the solution which maximizes the variance accounted for by each subsequent dimension (principal component or factor). The q-dimensional estimation of the data is given by:

$$\hat{\mathbf{X}}_q = \mathbf{T}_q\mathbf{P}_q \tag{4.2}$$

where $q \leq p$, and $\mathbf{T}_q$ (m x q) and $\mathbf{P}_q$ (q x n) are the truncated scores and loadings matrices, consisting of the first q columns of **T** and the first q rows of **P** (this truncation is the process performed by the JADE and MILCA algorithms). For a given q, the decomposition minimizes the sum of squared residuals, SSR_q:

$$SSR_q = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_{ij} - \hat{x}_{ij}\right)^2 \tag{4.3}$$

Equivalently, this maximizes the amount of total variance retained in $\hat{\mathbf{X}}$. If q is chosen to be 2 or 3, plotting the columns of $\mathbf{T}_q$ against one another is a scores plot. Ideally, this yields an optimal visualization of the relationships among objects.

In HCA, the concept is to measure the distances among objects in the data set and group the objects (rows of **X**) that are closest together. Starting with the same matrix, **X**, the Euclidean distance between each pair of objects, i and j, is first calculated according to:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}\left(x_{ik} - x_{jk}\right)^2} \tag{4.4}$$

This leads to a symmetric distance matrix, **D** (m x m) with diagonal elements of zero. In the next step, the two objects with the shortest distance are identified and combined to form a new object which replaces the former objects, and a new distance matrix is calculated. Because the new object is a combination of the original objects, there are a variety of options (called linkage methods) to represent the new distance, such as using the average distance to the group or the distance to the nearest original object, but these will not be discussed in detail here. This process is then repeated, incrementally reducing the number of objects present at each iteration until only a single connection remains to be made. The hierarchy of these connections is finally displayed as a tree structure (a dendrogram) with the relationships between objects represented as a chain of branch points where the vertical height of each branch point represents the distance between the connected objects (a measure of "dissimilarity"). Those objects (most often samples) emanating from a common branch point are considered to be most closely related (a cluster) with their similarity related to the height of the branch point.

While they are different approaches, both PCA and HCA are based on measuring the squared differences among objects. These differences include both chemical variations and measurement noise. Both methods are designed to provide an optimal representation of the chemical variance when the measurement noise is independent and identically distributed with a normal distribution, often referred to as *iid* normal noise. This means that it is assumed that all of the measurements in the data set have the same error variance and there are no relationships among the errors for different variables (*i.e.* they are uncorrelated). While this is an implicit assumption in many data analysis methods (*e.g.* univariate regression), it is violated more often than not and can lead to suboptimal results [25,130–132].

4.2.2 Measurement Error Structures

For univariate measurements, the uncertainty can be fully described by the error variance, σ^2, of the measurement. For multivariate measurements, it is necessary not only to provide the measurement variance for each variable, σ_i^2, but also the covariance between measurement channels, σ_{ij}. Chemical measurement vectors are often heteroscedastic, meaning that different elements of the vector can exhibit different error variance. This non-uniform variance arises naturally from the measurement system [130–133]. For example, fundamental counting statistics, governed by the Poisson distribution, give rise to what is often referred to as shot noise, where the error standard deviation is proportional to the square root of the signal intensity. Such noise may be limiting in spectroscopic or mass spectrometric measurements where the signal amplitude is low. Proportional noise, where σ is proportional to the magnitude of the signal, is also commonly observed and typically associated with variations in a light source or ion source. Likewise, many measurement systems exhibit noise that is highly correlated. This includes baseline offset noise and multiplicative offset noise, the latter of which is typically the limiting noise source in NIR reflectance spectroscopy [*vide infra*], arising from variations in path length due to sample heterogeneity. Low frequency noise, also known as pink noise or 1/*f* noise, also falls into this category and is sometimes referred to as source flicker noise or drift noise in the context of analytical measurements [133–139].

A common method to characterize multivariate measurement errors is the error covariance matrix (ECM) [25,33,130,131]. If we consider a measurement vector, $\mathbf{x}\ (1 \times n)$, which is an observation of a true (error-free) vector, $\mathbf{x}^{\mathrm{o}}$, the error vector, **e**, is defined as the difference between these vectors, $\mathbf{e} = \mathbf{x} - \mathbf{x^0}$. The error covariance between measurement channels i and j of the vector is defined as the expectation of the product of the corresponding errors:

$$\sigma_{ij} = E\left(e_i \cdot e_j\right) = \lim_{N \to \infty} \frac{\sum (x_i - x_i^o)\left(x_j - x_j^o\right)}{N} \qquad (4.5)$$

Here the summation is over multiple realizations of measurement vector **x** and x_i and x_j are elements of that measurement vector. When $i = j$, the corresponding quantity is the error variance, signified as $\sigma_i{}^2$ rather than σ_{ii}. The collection of all of these error covariances is described by the ECM ($\boldsymbol{\Sigma}$) defined as the outer product of the expectation of the error vectors:

$$\boldsymbol{\Sigma} = E(\mathbf{e}^{\mathbf{T}} \cdot \mathbf{e}) = E[(\mathbf{x} - \mathbf{x}^{\mathrm{o}})^{\mathrm{T}}(\mathbf{x} - \mathbf{x}^{\mathrm{o}})] \qquad (4.6)$$

The ECM is a symmetric ($n \times n$) matrix, where the diagonal elements represent the error variance of each of the *n* variables and the off-diagonal represents the error covariances of the corresponding elements. The ECM is one of the most complete ways to describe the errors in a vectorial measurement with stationary characteristics. A related method is the error correlation matrix, **R**, which normalizes the off-diagonal elements by their corresponding standard deviations such that:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \cdot \sigma_j} \qquad (4.7)$$

This removes the effects of scale (diagonal elements are unity) and allows more direct visualization of correlation. Errors with $\rho_{ij} = 1$ are perfectly correlated.

In practice, the true measurement vector is unknown, so the experimental ECM is normally estimated by making replicate observations of the measurement vector and subtracting the sample mean vector ($\bar{\mathbf{x}}$). If r experimental replicates of the measurement vector (*e.g.* a spectrum) are made, the ECM is estimated as:

$$\boldsymbol{\Sigma}_{expt} = \frac{1}{(r-1)} \sum_{k=1}^{r} (\mathbf{x}_k - \bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{x}_k - \bar{\mathbf{x}}) \qquad (4.8)$$

It should be noted that the definition of the replicate is very important in this context, since it needs to capture all of the sources of variation one wishes to consider as measurement errors. Consequently, the ECM can be quite different depending on whether it is to include, for example, only technical replication or also sampling variability.

The ECM estimated by the replication procedure above is likely to be quite noisy itself when the number of replicates is relatively small [25,130] and therefore of limited practical utility. Two approaches are commonly used to improve the quality of the ECM. The first is to pool (average) the ECMs obtained for different measurement vectors, each with a limited number of replicates[130]. This results in an averaging effect that leads to a smoother ECM but makes the implicit assumption that measurement vectors for different samples have the same ECM. While not strictly valid, this assumption is reasonable where measurements exhibit similar characteristics. The second approach is to develop an empirical model of the ECM [130,131,140]. For many kinds of measurements, the ECM can be represented using a model characteristic of that particular technique using a limited number of parameters. Where this can be done, the result is a smoother, more reliable ECM that can be calculated separately for each measurement vector.

Knowledge of the measurement error characteristics through the ECM is key to improving data analysis methods since it allows better extraction of the chemical variance from the associated noise variance. By implicitly describing the information associated with each measurement, the ECM allows more optimal results to be obtained.

4.2.3 Maximum Likelihood Principal Components Analysis (MLPCA)

MLPCA was developed as a tool to provide better subspace estimation for multivariate data when assumptions of *iid* normal errors are no longer valid [25,33,141]. It can be viewed as a more generalized form of PCA in which the ECM is incorporated into the decomposition procedure to yield a more optimal solution. Rather than simply minimizing the residual variance of the truncated q-dimensional solution, MLPCA uses a weighted objective function that attempts to match the residual variance to the characteristics of the ECM for each measurement vector. The approach is analogous to using weighted least squares in univariate regression. The specific objective function used depends on the complexity of the error structure and there are six general categories, ranging in complexity from the trivial case of *iid* normal errors (where MLPCA and PCA are equivalent) to general error heteroscedasticity and correlation that can exist within both the rows and columns of a data matrix. One of the most common implementations is where error correlation exists only within the rows of a data matrix. Under these conditions, the objective function to be minimized is defined as:

$$S_{obj}^2 = \sum_{i=1}^{m} (\mathbf{x}_i - \hat{\mathbf{x}}_i)^{\mathrm{T}} \mathbf{\Sigma}_i^{-1} (\mathbf{x}_i - \hat{\mathbf{x}}_i) \tag{4.9}$$

Here, $\mathbf{x}_i$ represents measurement vector i (row i of $\mathbf{X}$), $\hat{\mathbf{x}}_i$ is the estimate of the vector based on the MLPCA decomposition, and $\mathbf{\Sigma}_i$ is the ECM for the vector. In the general case, this objective function is optimized by an alternating least squares (ALS) algorithm, but in the special case where $\mathbf{\Sigma}_i$ is the same for all row vectors, a direct solution can be obtained through rotation and scaling of the original data. Another difference between MLPCA and PCA is that, where PCA uses an orthogonal projection of the measurement vector onto the subspace to obtain the scores vector ($\mathbf{t}_q = (\mathbf{P}_q \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$), MLPCA employs a maximum likelihood projection:

$$\mathbf{t}_q = \mathbf{x}\mathbf{\Sigma}^{-1}\mathbf{P}_q^{\mathrm{T}}\left(\mathbf{P}_q\mathbf{\Sigma}^{-1}\mathbf{P}_q^{\mathrm{T}}\right)^{-1} \tag{4.10}$$

This oblique projection uses the information in the error covariance matrix to ensure that the projection uses the measurements in $\mathbf{x}$ that minimize the uncertainty in the low dimensional projection.

In principle, MLPCA should result in the optimal subspace estimation assuming that the intrinsic dimensionality of the data (also called the pseudorank, q) and the ECM are exactly known. In practice, q is often uncertain and only an estimated ECM is available, so this can limit the optimality of the solution. There can also be complications from rank deficiency of the ECM (which needs to be inverted) when it is estimated from a limited number of replicates, although there are strategies to address this [141,142]. Despite these limitations, however, MLPCA has demonstrated superior performance to PCA in a variety of applications ranging from multivariate calibration [143,144] to curve resolution [35,145].

4.2.4 Projection Pursuit Analysis (PPA)

An inherent limitation of PCA and HCA is an assumption that the largest source of chemical variation in a data set is associated with the characteristic we are interested in, specifically, in the current context, the classification of samples into two or more groups. For example, in the detection of a disease state, it is hoped that the dominant source of difference is in a set of chemical compounds that are associated with the presentation of the disease, often referred to as biomarkers. However, the differences among these compounds may be obscured by other natural variations in the data set, resulting in an exploratory visualization that does not reveal clustering according to the anticipated characteristics. To overcome these limitations, it is necessary to use visualization methods that do not rely solely on variance metrics.

As discussed in previous chapters, the concept of projection pursuit was first advanced nearly five decades ago and is based on the idea of looking for linear projections of the multivariate data that are interesting based on a measure of "interestingness" as quantified by a projection index. The historical evolution and fundamental principles of PPA have been outlined in Chapter 2, with some applications presented in Chapter 3. In particular, kPPA based on the quasi-power algorithm has been shown to be an effective clustering tool when other methods fail [43–46,52,53]. In this chapter, kPPA is applied to a data set for which replicate measurements are available, allowing its direct comparison to MLPCA and providing evidence that kPPA can address circumstances where measurement error structure is the complicating factor.

4.2.5 Independent Component Analysis (ICA)

Like PPA and PCA, ICA is a projection method which decomposes the data into a bilinear model. Where PCA and PPA commonly use the terms "scores" and "loadings" to describe the two matrices, ICA models are instead described as "contributions" and "signals", respectively. Despite the terminology difference, the contributions and signals represent the same information as scores and loadings and will be compared as such. Although ICA was initially developed for signal extraction, it has since been used in other applications including cluster analysis [77,100,146]. Considering the relationship between kurtosis and cumulants, PPA and ICA will be applied equivalently for this study.

For this application, only the JADE algorithm was evaluated as an ICA technique since it has been widely applied for clustering applications and is the most closely related to PPA. The algorithm is applied in the way it was applied for clustering in Chapter 3, that is applying it to $\mathbf{X}^{\mathrm{T}}$, in order to target independent scores rather than independent loadings.

4.3 Experimental

4.3.1 Computational Aspects

All calculations were carried out within the MatLab programming environment (Mathworks, Natick, MA). Programs for carrying out MLPCA and PPA were written in-house and are available from the corresponding author, as are the data.

4.3.2 Data set

The data set which was employed for this part of the research involves the classification of four different types of tropical wood species by near-infrared (NIR) reflectance spectroscopy. The data were obtained from a study by Brazilian researchers whose aim was to develop classification models to allow the rapid discrimination of mahogany from other similar species. Exploratory data analysis with clustering was undertaken to validate the hypothesis that the data contained sufficient information to build a classification model.

There are several characteristics of this data set that made it ideal for a more focused study. First, traditional exploratory methods (PCA, HCA) had limited success at revealing the anticipated class structure. Second, the acquired data was well-suited to sequential binary classification by kPPA since it consisted of four balanced classes. Finally, the availability of replicate data allowed calculation of the ECM, which opened the possibility of making comparisons with MLPCA and establishing the error structure as a complicating factor in the exploratory analysis.

4.3.3 Species Selection

The broad objective of this research, of which this study is a part, is the development of instrumental methods to distinguish wood species, with a particular emphasis on discriminating high value species such as mahogany [128,129]. Species were selected based on the book "Similar woods to mahogany (*Swietenia macrophylla* King): An illustrated key for anatomical field identification" [147], edited by the Brazilian Forest Service. From the 15 species listed, the three species that were the most difficult to distinguish, based on the appearance and macroscopic wood characteristics, were chosen for this study. These were *Carapa guianensis*, *Cedrela odorata*, and *Micropholis melinoniana*, along with mahogany itself, *Swietenia macrophylla*.

4.3.4 Sampling and Sample Preparation

Each sample of crabwood (*C. guianensis*), cedar (*C. odorata*) and curupixa (*M. melinoniana*) was obtained from an individual disk located at the base of a tree trunk. These species (*S. macrophylla*) were collected in authorized forestry exploitation areas in Para state, Brazil. Mahogany samples were obtained from tips of seized boards coming from the state of Mato Grosso, Brazil. Altogether, 108 solid samples were measured, 26 being of crabwood, 28 of cedar, 29 of curupixa and 25 of mahogany. Besides alleged species, all samples were identified by a wood anatomist of the Forest Products Laboratory in Brasilia, registered as FPBw in the Index Xylariorum [148].

Samples were dried in open air conditions and cut into blocks of approximately 2 cm^3 with oriented faces according to wood growth directions. Surfaces were made uniform with 80 grit sandpaper.

4.3.5 Acquisition of Spectra

Samples were measured by coauthors on a handheld spectrometer, Phazir RX (Polychromix). Four replicate spectra were obtained for each sample, two on each radial face, measured on distinct spots, resulting in a total of 432 spectra. Spectra were measured in the diffuse reflectance mode between 939.5 and 1796.6 nm with 9 nm of resolution. Resultant spectra, shown in Figure 4.1, consisted of 100 data points per spectrum and were converted to log(1/R) scale (where R is the response) for the data analysis. Figure 4.1(a)

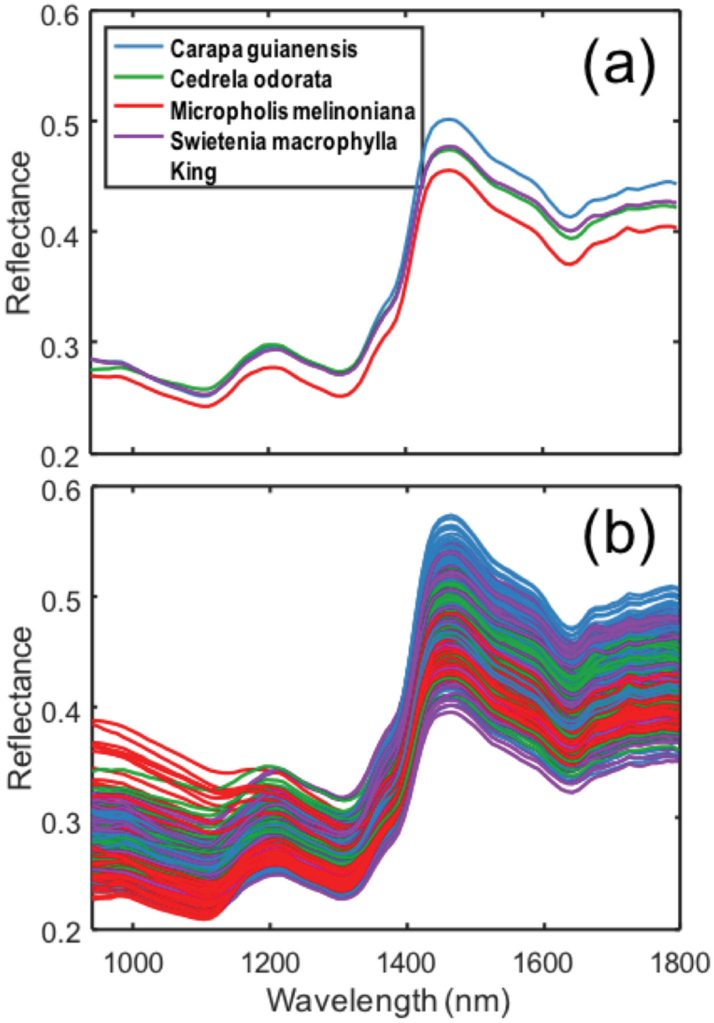

Figure 4.1 Near-infrared (NIR) reflectance spectra of wood samples. a) Mean spectra of four species, as indicated in the legend. b) Full set of 432 spectra.

shows the mean spectrum for each of the four species, while Figure 4.1(b) shows all 432 spectra, with each of the replicates displayed individually.

4.4 Results and Discussion

4.4.1 PCA and HCA of NIR Spectra

It is clear from Figure 4.1 that spectra of the four species exhibit a strong similarity and that the variation between individual samples is quite large, making the discrimination of the four classes a challenging problem. To determine if the usual data visualization methods would be able to distinguish the classes, PCA and HCA were applied to the NIR spectra. To improve the quality of the measurements and simplify the visualization, the mean of the four replicate spectra were used for each sample, resulting in a data matrix of 108 samples by 100 wavelength channels. Initially, only column mean centering was applied to the data. The paired scores plots for the first four principal components from PCA are presented in Figure 4.2, where the different species are represented by different symbols as indicated in the legend. Based on the distribution of samples in the scores plots, there is no apparent separation of the species based on the NIR spectra. While there is some suggestion of separation of classes 2 and 4 (*C. odorata* and *S. macrophylla*) using the third and fourth PCs, there is still strong overlap and no clear clustering is evident. Higher PCs did not improve separation.

In many cases of multivariate analysis, it is necessary to preprocess data to obtain satisfactory results, so a variety of common preprocessing methods were employed here to see if the class separation could be improved. These included autoscaling, multiplicative signal correction (MSC) and the standard normal variate (SNV). MSC and SNV are widely used in NIR spectroscopy to account for multiplicative offset noise [19,23]. None of the methods implemented resulted in any improvement of the PCA results.

The application of HCA did not improve on these results. HCA was implemented through algorithms in the Statistics and Machine Learning Toolbox of MatLab. A variety of linkage and preprocessing options were applied to both the full set of 432 spectra and the set of 108 sample mean spectra. Results for the latter are shown in Figure 4.3, with the symbols and the color of the bottom branches representing the species. The results shown are for an average distance calculation and mean-centering as the only preprocessing. While some local groupings are evident in the tree structure, no consistent clusters are

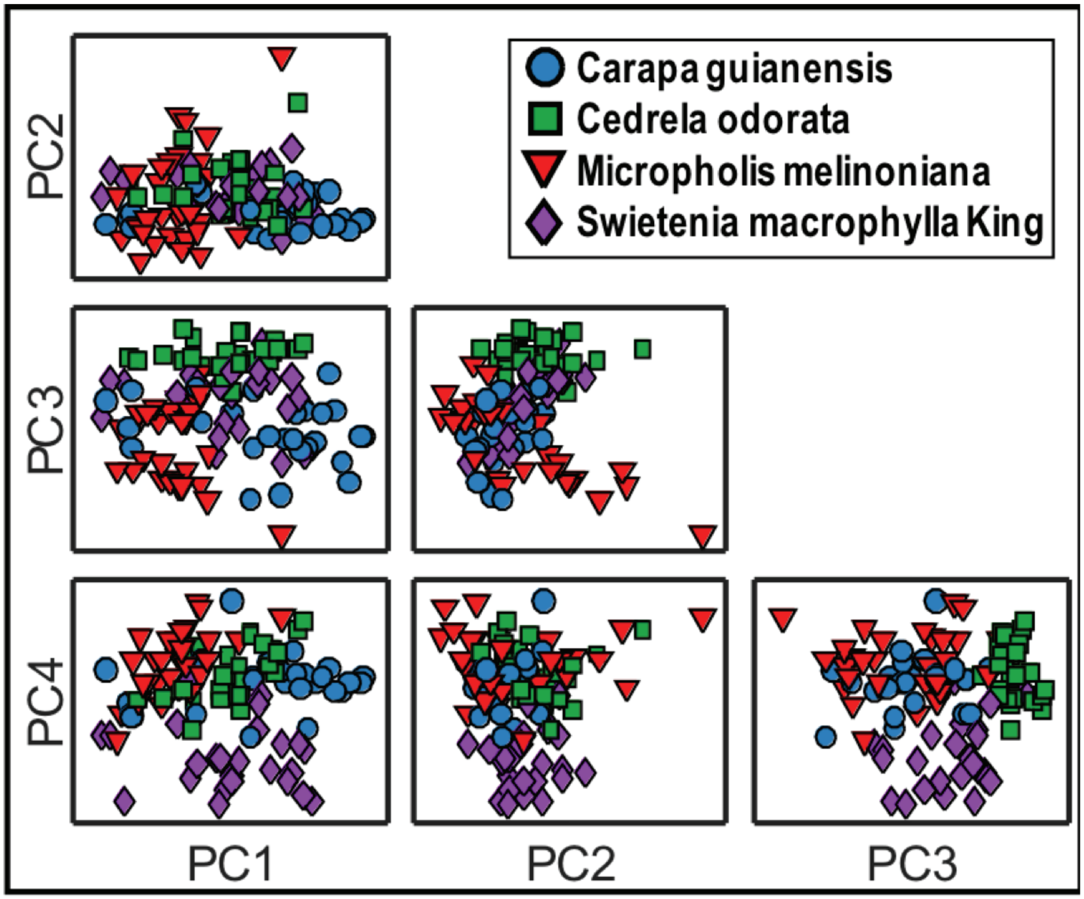

Figure 4.2 Paired scores plots from principal components analysis of sample mean spectra after column mean-centering, with species identified as in the legend.

observed that would strongly support the hypothesis that the species can be distinguished on the basis of their NIR spectra. Although different preprocessing and linkage options produced changes in the tree structure, similar irregular class distributions were observed in all cases with no strong evidence of clusters related to species.

4.4.2 Error Structure of NIR Spectra

A central premise of this work is that exploratory analysis by HCA and PCA can be adversely affected by non-*iid* error structures. It is therefore necessary to examine the measurement error characteristics of the NIR spectra which are the focus of this study. For each of the 108 samples examined, replicate spectra were obtained from four different physical locations and should reflect the within-sample error variance. On the basis of these four replicates, an ECM can be calculated for each sample using Equation 4.8, leading to 108 individual ECMs.

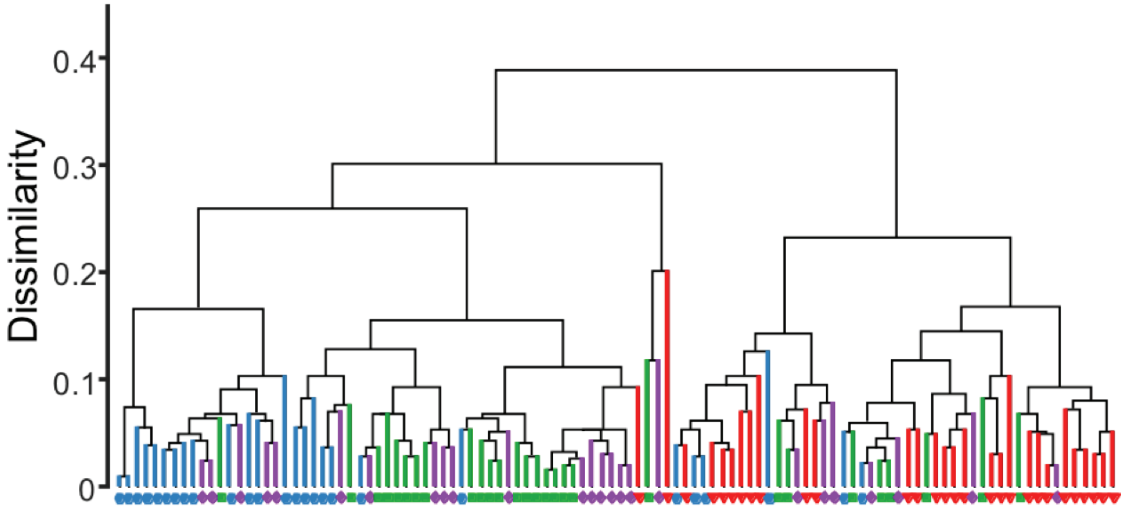

Figure 4.3 Dendrogram resulting from hierarchical clustering of sample mean spectra after mean-centering. Species are color coded in the same manner as Figures 4.1 and 4.2. Average distance was used in the clustering algorithm.

Unfortunately, ECMs calculated on the basis of a small number of replicates are very noisy and unreliable [140]. For example, an error variance estimated from four replicates is expected to have a relative standard deviation (RSD) of about 82%. This high level of noise in the individual ECMs makes their visual interpretation difficult and precludes their use in any advanced data analysis strategy. One solution to this problem, as noted earlier, is to pool (average) individual sample ECMs. This is valid in cases where the spectral characteristics of the samples are very similar, and therefore the improved precision gained by pooling outweighs any between-sample differences. The similarity of the spectra in this study is evident from Figure 4.1, so pooling was a viable option.

Initial pooling of the ECMs was carried out within each of the four species investigated. This was done as a preliminary evaluation to confirm the similarity of the ECMs within each group prior to global pooling, which is normally done. It was anticipated that the four ECMs would show very similar characteristics which were consistent with NIR spectra. While this was true for three of the groups (classes 1, 2 and 4), the remaining group (class 3) was distinctly different from the others, as shown in Figure 4.4. For classes 1, 2 and 4 (Figures 4.4(a), (b) and (d)), the ECMs are typical for NIR spectra [25,130] showing heteroscedastic noise (non-uniform variance) along the diagonal and, more importantly, structured covariance (off-diagonal elements) that is consistent with offset and multiplicative offset noise. The latter is a dominant noise source in NIR reflectance measurements, arising from differences in the effective path length of scattered photons caused by changes in the scattering characteristics of the sampled region. The result is a shift in the spectral intensity proportional to its magnitude (hence the term multiplicative offset noise). The direction of the shift from the mean is random, but is consistent within a spectrum, leading to highly correlated noise in which variance and covariance are directly related to the signal magnitude, as is evident in Figures 4.4(a), (b), and (d). Figure 4.4(c) is anomalous in this regard, however. While the correlated and heteroscedastic noise is still evident, the magnitude of the measurement errors is largest in the shorter wavelength regions, where the signal is the lowest. This is confirmed through an examination of Figure 4.1(b), which shows a substantially greater variation of *M. melinoniana* in this region. The reason for the anomalous behavior of the third class is unclear, but it may be due to different physical properties that lead to different scattering characteristics by these samples.

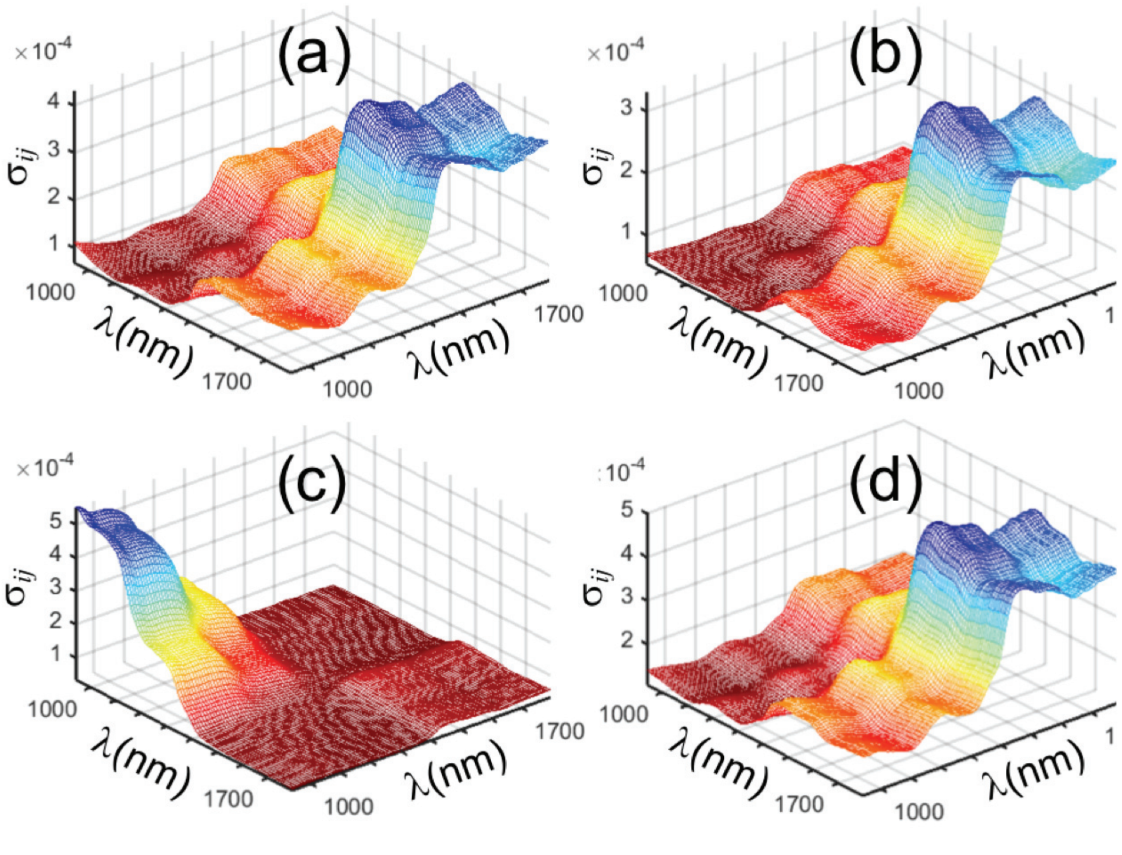

Figure 4.4 Pooled error covariance matrices (ECMs) of the NIR spectra for each of the four species examined: a) Class 1: *Carapa guianensis*, b) Class 2: *Cedrela odorata*, c) Class 3: *Micropholis melinoniana*, d) Class 4: *Swietenia macrophylla*.

Although the differences observed above suggest that a global pooling of ECMs from each class may not be representative of all samples, global pooling was nevertheless carried out and the results are shown in Figure 4.5(a). As anticipated, the globally pooled ECM (calculated using a weighted average reflecting the number of samples in each group) reflects the characteristics of the dominant classes (1, 2 and 4) but with a higher variance/covariance in the short wavelength region due to the contribution of class 3. Despite its inaccuracy in its universal representation of all errors, the globally pooled ECM can still give improved results because it is still superior to the assumption of *iid* normal errors which is made in the usual applications of HCA and PCA. This is further explored in the section that follows. Also shown in Figure 4.5(b) is the error correlation matrix (**R**) corresponding to the ECM (**Σ**) in Figure 4.5(a), calculated using Equation 4.7. The error correlation matrix removes the effects of the magnitude of the error which are evident in the ECM, showing only how they are related. Figure 4.5(b) shows almost perfect correlation (same direction and relative magnitude change in the errors) within three regions (<1148 nm, 1166-1343 nm, >1395 nm), but a smaller degree of correlation between these regions. This is typical for offset/multiplicative offset noise in NIR spectra and shows a strong interdependence of measurement errors.

4.4.3 MLPCA of NIR Spectra

For a matrix of chemical measurements, the intrinsic rank (pseudorank, chemical rank) is defined as the dimensionality of the space needed to account for all of the chemical variation in the absence of measurement error, and for linear systems this is typically equal to the number of independently observable chemical components. When the intrinsic rank is well-defined and the ECMs of the measurement vectors are accurately known, MLPCA should yield the optimal estimate of the chemical subspace. For exploratory data analysis, however, decomposition by MLPCA is only guaranteed to provide an optimal visualization of the data when the intrinsic rank is equal to the dimensionality of the space into which the data are projected (called the projection rank), which can only be realized when the intrinsic rank is less than or equal to three [34]. In cases where the intrinsic rank exceeds the projection dimensionality, the advantages of MLPCA are less certain, but its application may provide a more useful visual projection of the data than PCA. In general, a definitive

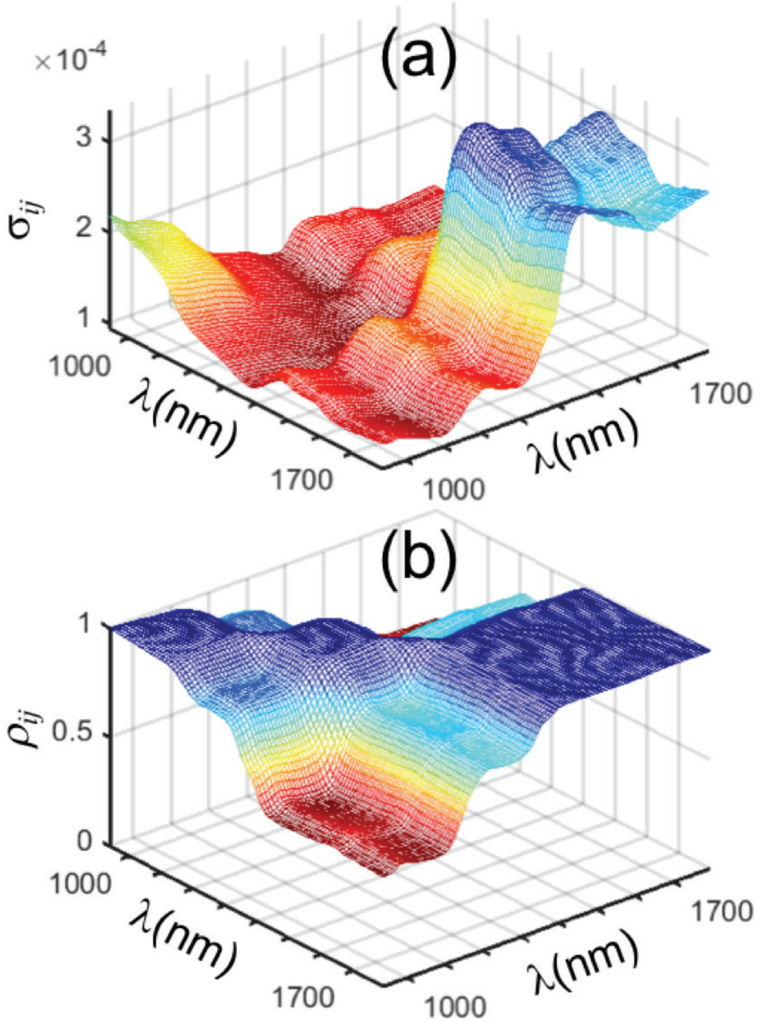

Figure 4.5 Pooled error structure of NIR spectra based on 108 samples: a) Pooled error covariance matrix. b) Pooled error correlation matrix.

determination of the intrinsic rank (q) is difficult, so the application of MLPCA is typically carried out using different values to assess the projections empirically.

The application of MLPCA requires a specification of the data matrix, the corresponding ECMs, and the dimensionality of the subspace to be estimated. Based on the results of the previous section, which showed that the ECMs were not homogeneous among the different sample classes, it was decided to assign the ECM for each measurement vector based on its class membership (species), using the pooled ECM for the corresponding class. This error structure is representative of Case E for the MLPCA algorithms [25,141] for general row-correlated errors. The objective function in this case is given by Equation 4.9 and is minimized through the ALS method. The data matrix consisted of 108 rows corresponding to the sample mean spectra (column mean-centered) and an initial rank of two was selected. Although the ALS algorithm is slower than the direct solution which can be obtained when all of the ECMs can be assumed to be the same, it is considered to be more reliable when this assumption is violated, and the execution time was only about 20 s in this case.

The scores plot obtained through the application of MLPCA(E) (with a specified rank of 2) in this manner is shown in Figure 4.6(a). The results show a clear clustering of the samples into separate groups corresponding to the individual species, with the exception of one point from class 3 (*M. melinoniana*; it is noted that this does not correspond to the extreme point in the upper left panel of Figure 4.2). This supports the hypothesis that there is sufficient information in the NIR spectra to distinguish among the four species. More importantly, in the context of the current work, it supports the broader hypothesis that the visualization of data by PCA can be impeded by non-*iid* measurement error structures and that this problem can be mitigated through the application of MLPCA. By incorporating information about the measurement error variance and covariance into the decomposition of the data, MLPCA can more effectively separate the variability originating from chemical differences from that arising from measurement noise, thereby giving a more useful picture of the relationships among samples.

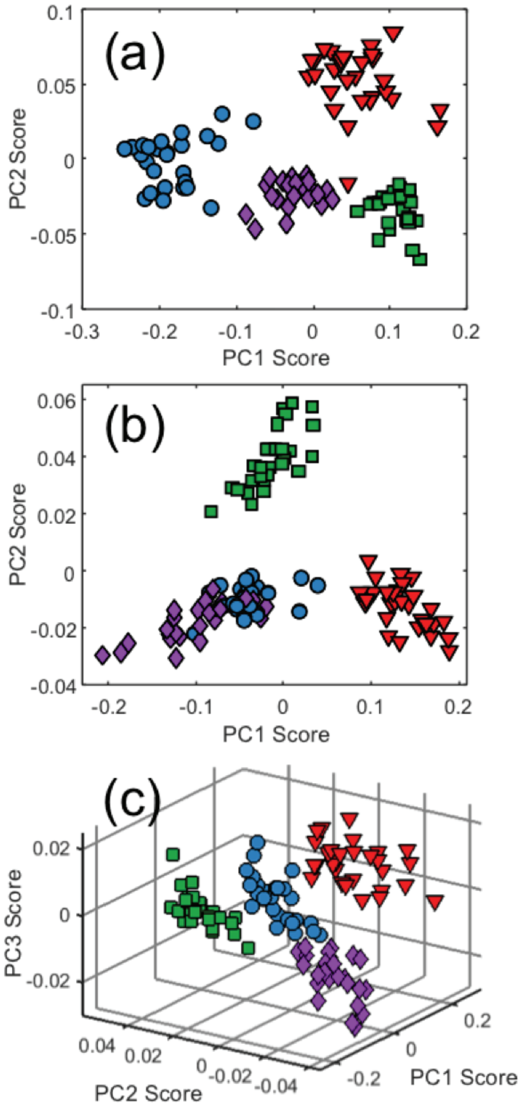

Figure 4.6 Scores plots from maximum likelihood principal components analysis (MLPCA) of NIR spectra. a) Rank 2 MLPCA results using class-specific ECMs. b) Rank 2 MLPCA results using a global average ECM. c) Rank 3 MLPCA results using a global average ECM. Symbols correspond to the legend in Figure 4.2.

A potential argument that can be made to counter the conclusions drawn above is that, by defining the ECM according to class membership, indirect information related to class membership is being provided to the MLPCA algorithm and therefore biasing the outcome. This is a legitimate argument, since a truly unsupervised method should not include any information that could indirectly be associated with class membership. While it cannot be concluded that the results in Figure 4.6(a) are biased, this possibility cannot be excluded, so further evidence is needed. There are three possible ways to exclude bias. The first would be to provide an individual ECM for each sample based on its replicates. However, since there are only four replicates measured for each sample, the ECMs would be unreliable, as well as rank deficient due to the small number of replicates (rank = 3). Under these circumstances, anomalously small variances (due to limited replication) tend to drive the optimization, giving excessive weight to a few samples. This was confirmed by using the individual ECMs, resulting in a scores plot with a tight central cluster and a few dispersed samples (results not shown). A second possibility is to use a parameterized model for the ECM developed from multiple samples [130,140]. This can then be employed to calculate individual ECMs with greater reliability. In this case, however, it is clear that the same model could not be applied to all samples due to the differing characteristics of one of the classes. The third option would be to employ the globally pooled ECM, shown in Figure 4.5(a), to all of the samples. Although it is expected that the MLPCA solution obtained in this way would be suboptimal, it eliminates the possibility of bias and may produce projections superior to PCA.

To implement this third option, MLPCA (Case D, common row covariance) was applied to the 108 sample mean spectra (column mean-centered) using the global pooled

ECM with a specified rank of two and three. The scores plot for the rank two solution is shown in Figure 4.6(b). This result shows a clear separation of classes 2 and 3 (*C. odorata* and *M. melinoniana*) but strong overlap of the other two classes. However, the three-dimensional projection, presented in Figure 4.6(c), shows a distinct separation of all four classes. As might be expected, the separation observed here is not as clear as for Figure 4.6(a) since a common ECM is erroneously assumed for all samples, but the results are far more informative than PCA. These results also exclude the possibility of an unintended bias and support the premise that class information can be more clearly extracted by incorporating measurement error information into the data analysis.

It should be noted that, in all of these cases, higher rank MLPCA solutions were also investigated. Separation of classes was still observed with increasing dimension, although the quality was diminished in the case of MLPCA(E), and slightly improved in the case of MLPCA(D) (results not shown).

4.4.4 PPA of NIR Spectra

A weakness of both methods investigated so far (HCA, PCA) is that they rely on an assumption that the dominant sources of chemical variance are associated with the classes of interest, but even when error variance is altered by preprocessing in an effort to reduce it, other sources of chemical variance may eclipse the factors of interest. For example, in biological samples, variation in chemical species among individuals in a population or due to diurnal rhythms may mask smaller effects of interest. Projection pursuit approaches can avoid this problem by examining other criteria to obtain the optimal low dimensional projection.

In this work, kurtosis-based PPA was implemented using a stepwise univariate algorithm and orthogonal scores, with a two-dimensional projection space. This algorithm uses a stepwise procedure that first minimizes the univariate kurtosis along one projection dimension, optimally resulting in a binary separation of the data. After “deflation” of the data to remove the extracted dimension, the process is repeated in an attempt to provide a binary separation in subsequent dimensions, ultimately resulting in scores and loadings of selected dimensions analogous to PCA (although the loadings are not required to be orthogonal in this case). In this application, all 432 spectra (mean centered) were employed, since PPA works best when the ratio of samples to variables is high. The algorithm uses a nonlinear optimization method that is significantly slower than PCA, and random initial starting points are used to ensure a global optimum. In this implementation, 1000 initial guesses were used and the execution time was about 20 min.

The scores plot resulting from PPA of the raw data is shown in Figure 4.7(a) and shows clear clustering of the four species, although there are a few samples that are grouped incorrectly. For a more direct comparison with earlier figures (Figure 4.2 and 4.6), the 108 sample mean spectra have been projected into the same subspace and exhibit no overlap, as might be expected due to the smaller error variance. It is important to note that no class information was provided implicitly or explicitly to the algorithm, so the natural clustering on the basis of species was discovered solely on the basis of the observed spectra, supporting the hypothesis that the multivariate information available in the NIR spectra can be used to distinguish among the classes. No preprocessing of the data was necessary other than column mean-centering, and no measurement error information was provided to the algorithm.

Although PPA is an extremely powerful tool for exploratory studies, it is not without its limitations. Current algorithms are best suited for balanced data sets (approximately equal numbers of samples in each class) with more samples than variables, and are most effective for 2, 4 or 8 classes. Ongoing work is directed at removing some of these limitations.

4.4.5 ICA of NIR Spectra

To illustrate another alternative method to traditional PCA and HCA, ICA was applied to the data using the JADE algorithm. JADE searches for a rotation that optimizes the set of all fourth order cumulants to minimize information shared in each direction. The relationship between fourth order cumulants and kurtosis is clear which draws lines

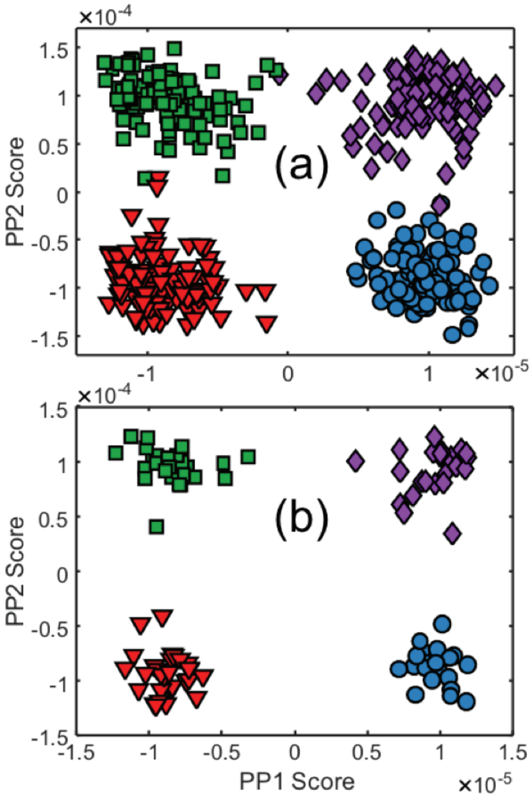

Figure 4.7 Scores plots for the projection pursuit analysis (PPA) of NIR spectra. a) Scores plot from the analysis of all 432 spectra. b) Scores plot resulting from the projection of sample means into the same space. Symbols correspond to the legend in Figure 4.2.

between JADE and PPA, however JADE also has strong ties to PCA since PCA compression is done prior to optimization within the algorithm. This compression step results in a very fast algorithm, but also creates its own set of complications, as discussed in earlier chapters. Due to these connections to both PPA and PCA it was unclear which set of the data would be most appropriate for a valid comparison so both data subsets were analyzed. That is, the full set of all 432 mean centered spectra, and alternatively the 108 mean-centered mean spectra. There was little variation in the results so for simplicity only the results of the 108 mean spectra are shown.

Unlike PCA and PPA, the rank selection is integral to the results obtained by JADE (i.e. the results for 2 components are not consistent with those for 3 components). To maintain consistency with the PCA analysis, a 4 component models was calculated. The scores plots resulting from these analyses are shown in Figure 4.8. Figure 4.8 shows all paired scores plots obtained from a 4 component model where it can be seen in some combinations that class 4 (*S. macrophylla*) partially separates from the other groups, and in other cases class 1 (*C. guianensis*) partially separates as well. Despite this, there are no combinations lacking high levels of overlap between the classes.

Despite the connections drawn between PPA and ICA it is clear from these results that the algorithms produce very different results and the ICA results resemble more closely the results of PCA. This is not surprising, since ICA solutions generated in this manner are constrained to orthogonal rotations of the PCA subspace and therefore do not provide new information.

4.5 Summary

The results of this study can be summarized by five main conclusions. First, even when chemical information related to classification is present in a data set, traditional exploratory methods such as HCA and PCA may be incapable of revealing it. This is demonstrated by juxtaposing Figures 4.2, 4.3 and 4.8, which show no clear organization of the samples, with Figures 4.6 and 4.7, which clearly show division of the samples based on biological species. A second conclusion, derived from the results shown in Figure 4.6, is that inclusion of measurement error information into the data analysis, in this case through the application of MLPCA, can greatly improve the visualization of chemical information by more effectively separating the chemical variation from the noise variance. It can be further inferred from this that the limiting factor in the effective implementation

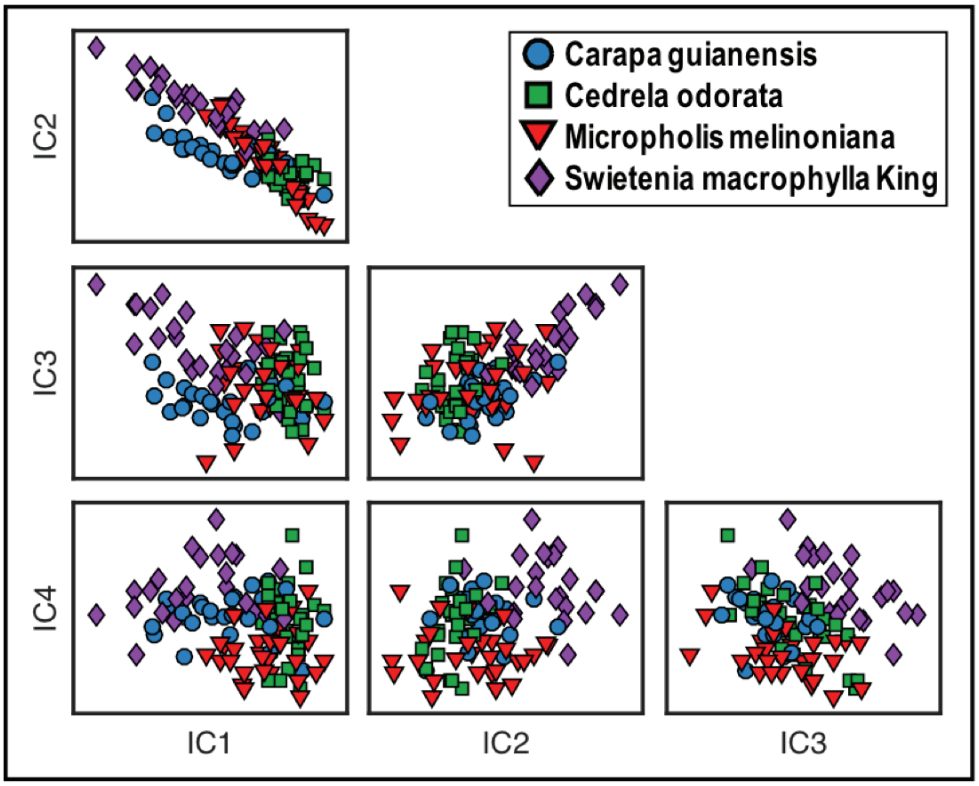

Figure 4.8 Paired IC plots from applying JADE to sample mean spectra after column mean-centering, with species identified as in the legend.

of PCA (and likely HCA) was the presence of heteroscedastic and correlated errors (*i.e.* a non-*iid* error structure), suggesting that a better understanding of measurement errors should be a key component in the analysis of any multivariate data set. Fourth, it was clearly demonstrated through Figure 4.7 that the implementation of data visualization methods such as PPA that do not rely strictly on variance as a criterion for low dimensional projection could be extremely beneficial in studies involving multivariate data. Although PPA was able to separate the classes, ICA (Figure 4.8) was not able to cluster the points for reasons that have been discussed in earlier chapters. Finally, with regard to the specific experimental data employed in this work, there is clear evidence that NIR spectroscopy has the capability to distinguish similar species of wood using the procedures described.

Many areas of modern scientific discovery are initiated by testing an initial hypothesis that a complex multivariate data set contains information relevant to a desired goal, such as disease detection or forensic classification. Such studies often involve a limited number of samples and a large number of variables. While supervised classification methods (by design) are well-suited to building classification models, they are poorly suited to test an initial hypothesis based on limited samples due to their need for extensive validation. Unsupervised (exploratory) methods play a key role in this workflow, since they do not have such strict validation requirements, but are currently limited to two dominant techniques, HCA and PCA. As demonstrated here, these methods can fail to reveal important information in certain circumstances, and failure to support an initial hypothesis can impede the advance of research. Therefore, it is important to expand the toolbox available to researchers for exploratory analysis, and the alternative methods described here, ICA and PPA, are two approaches that can contribute in this regard. Although expanding the toolbox is a key component to moving forward with data analysis, understanding these methods in order to apply them optimally also plays a key role in the future success of data analysis.

Chapter 5 – Conclusion

As the complexity and abundance of multivariate chemical data increases, so does the demand for new methods capable of handling and extracting the relevant information from the data. As convenient as it may seem to explore (and exploit) the application of methods developed in other scientific fields, the characteristics of chemical signals are not always consistent with the types of data for which these methods are equipped to handle. This does not mean that alternative methods should not be tested, but rather that their fundamentals should be understood before broad application occurs. Sometimes there are ways to adjust or alter these methods to better suit the properties and structure of chemical data.

One major application area in chemometrics is in the extraction of source signals, better known as curve resolution, from mixture data. These applications, which include the analysis of mixtures in chromatography, chemical reaction studies, and environmental monitoring, have matured with the development of MCR-ALS, which incorporates known characteristics of the chemical data, such as non-negativity and unimodality to constrain the range of linear solutions. On the surface, ICA presents an attractive alternative for source signal extraction, since (depending on the algorithm) it uses only statistical independence or non-Gaussian point distributions as a criterion. However, the point distribution of the underlying source signals is highly variable (and nearly impossible to predict in most cases), making it an impractical criterion to use when attempting to extract source signals. Another common feature of chemical signals, which drives the motivation for curve resolution methods, is that chemical signals often overlap one another. As it has been shown, high degrees of overlap in chemical signals results in high correlation between these signals, meaning they are not orthogonal and certainly not statistically independent. Given these features, it is not practical to apply ICA methods to curve resolution problems based on the assumption that certain properties, (such as non-Gaussian point distributions, orthogonality or statistical independence), are present. ICA methods encompass a wide variety of algorithms built around these same properties and should not be applied blindly in cases where they are not valid.

Exploratory data analysis for clustering is another area where methods developed in other fields are often applied to chemical data. PCA is perhaps the most widely applied method and is based on a fundamental matrix decomposition developed in mathematics and adopted by statisticians for visualizing and exploring data. Despite its success, it is a variance-based method which can result in sub-optimal extraction of information in cases where the between class variance does not dominate, such as when error structures in chemical data breach the *iid* noise assumptions made by PCA. Alternatives to PCA can succeed in cases where PCA fails often by taking a different approach to finding underlying data structures. Two such methods described in this work are MLPCA (which compensates for known data error structure) and PPA (which seeks interesting point distributions). Both MLPCA and PPA also have drawbacks (which have been outlined in this work), but perform exceedingly well in cases where their assumptions are valid and required conditions are met. The principal advantage of PPA, however, is that it requires no prior information about the measurement error structure.

In this work, ICA (specifically the JADE algorithm) was compared to both MLPCA and PPA in clustering applications, with the latter comparison being of greatest interest because of the common underlying principles. One reason MLPCA and PPA are able to outperform PCA in some cases, is that they are not constrained to a low-dimensional PCA subspace. Though it has been shown that PPA can benefit from PCA compression when the number of variables greatly exceeds the number of samples, the size of the subspace remains large and it is not a forced constraint on the algorithms. Since JADE forces a low dimensional compression from the start, the resulting *p* ICs extracted are restricted to the PCA subspace alone, which leaves little to no potential for discovering new structures or information beyond what would be seen from the original PCA results. Even if a larger number of ICs are extracted (>3), an exhaustive search of every pair-wise scores plot would have to be investigated. Though this is not to say that interesting results could not be found, it is a tedious approach for exploration when other methods exist with potentially more straightforward ways of measuring the quality of results. However, current ICA algorithms may serve as a starting point to develop modified techniques that would overcome this weakness.

In summary, a number of primary conclusions can be drawn on the basis of this work. First, the close relationship between kurtosis-based PPA and distribution-based ICA algorithms, particularly JADE, has been established. This connection has been vague in the past, but the methods have been shown to be tied together through the kurtosis criterion and a simple transposition of the input matrices. Although there are subtle differences, establishing this link helps to simplify our understanding of the multitude of methods currently applied. A second major assertion of this study is that, despite a number of claims to the contrary, there is no basis to believe that ICA is a suitable method for signal extraction (curve resolution) for chemical data. Although the main ICA algorithms are based on different principles (independence, non-normality), there is no reason to believe that any of the criteria are consistent with the nature of most chemical signals. This does not exclude the possibility of niche applications, but the fundamental lack of mutual orthogonality would preclude the application of ICA (and, by implication PPA) on first principles. Finally, kurtosis-based PPA has once again been demonstrated to be a valuable addition to the toolkit for cluster analysis, often succeeding where traditional methods fail. Based on the fundamental similarities to PPA, it was anticipated that some of the ICA methods might also show potential for clustering, but basic constraints on current algorithms limit their ability to effectively search the full variable space. Despite the recent appeal of ICA for chemical applications, a better understanding of these tools is needed to develop them into useful tools for chemistry.

Given the amount of time and consideration put into the process of collecting chemical data, from experimental design and sample preparation to making the measurements themselves, it is essential that the right tools are applied for data analysis. Knowledge of the underlying processes and chemical properties is essential to interpreting the results of an experiment, and there should be an equivalent amount of understanding incorporated when selecting the appropriate analysis methods. This work has attempted to extend such an understanding to new analysis tools in a way that will allow chemists to use them more effectively.

References

[1] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics : A textbook, 1st ed., Elsevier Science, 1988. doi:10.1002/cem.1180020409.

[2] J.B. Kruskal, Toward A Practical Method Which Helps Uncover the Structure of a Set of Multivariate Observations by Finding a Linear Transformation Which Optimizes a New "Index of Condensation," in: R.C. Milton, J.A. Nelder (Eds.), Statical Comput., 1st ed., Academic Press, New York, USA, 1969: pp. 427–440.

[3] J. Herault, B. Ans, Réseau de neurones à synapses modifiables: décodage de messages sensoriels composites par apprentissage non supervisé et permanent, Comptes Rendus Des Séances l'Académie Des Sci. Série 3, Sci. La Vie. 299 (1984) 525–528.

[4] K.R. Beebe, R.J. Pell, M.B. Seasholtz, Chemometrics: A Practical Guide, Wiley, New York, USA, 1998.

[5] T.P.E. Auf der Heyde, Analyzing chemical data in more than two dimensions: A tutorial on factor and cluster analysis, J. Chem. Educ. 67 (1990) 461. doi:10.1021/ed067p461.

[6] E.R. Malinowski, Factor Analysis in Chemistry, 3rd ed., Wiley, New York, USA, 2002.

[7] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. doi:10.1039/c3ay41907j.

[8] A. Meyer-Baese, J. Wildberger, U. Meyer-Baese, C.L. Nilsson, Data analysis techniques in phosphoproteomics, Electrophoresis. 35 (2014) 3452–3462.

doi:10.1002/elps.201400219.

[9] W. Peng, Y. Zhang, R. Zhu, Y. Mechref, Comparative membrane proteomics analyses of breast cancer cell lines to understand the molecular mechanism of breast cancer brain metastasis, Electrophoresis. 38 (2017) 2124–2134. doi:10.1002/elps.201700027.

[10] J.G.M. Pontes, A.J.M. Brasil, G.C.F. Cruz, R.N. De Souza, L. Tasic, NMR-based metabolomics strategies: plants, animals and humans, Anal. Methods. 9 (2017) 1078–1096. doi:10.1039/c6ay03102a.

[11] L. Yi, N. Dong, Y. Yun, B. Deng, D. Ren, S. Liu, Y. Liang, Chemometric methods in data processing of mass spectrometry-based metabolomics: A review, Anal Chim Acta. 914 (2016) 17–34. doi:S0003-2670(16)30186-6 [pii]\r10.1016/j.aca.2016.02.001 [doi].

[12] M.M.W.B. Hendriks, F.A. va. Eeuwijk, R.H. Jellema, J.A. Westerhuis, T.H. Reijmers, H.C.J. Hoefsloot, A.K. Smilde, Data-processing strategies for metabolomics studies, TrAC - Trends Anal. Chem. 30 (2011) 1685–1698. doi:10.1016/j.trac.2011.04.019.

[13] L. Laghi, G. Picone, F. Capozzi, Nuclear magnetic resonance for foodomics beyond food analysis, Trac-Trends Anal. Chem. 59 (2014) 93–102. doi:10.1016/j.trac.2014.04.009.

[14] J.M. Bosque-Sendra, L. Cuadros-Rodríguez, C. Ruiz-Samblás, A.P. de la Mata, Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data-A review, Anal. Chim. Acta. 724 (2012) 1–11. doi:10.1016/j.aca.2012.02.041.

[15] C. Martín-Alberca, F.E. Ortega-Ojeda, C. García-Ruiz, Analytical tools for the analysis of fire debris. A review: 2008-2015, Anal. Chim. Acta. 928 (2016) 1–19. doi:10.1016/j.aca.2016.04.056.

[16] C.K. Muro, K.C. Doty, J. Bueno, L. Halámková, I.K. Lednev, Vibrational spectroscopy: Recent developments to revolutionize forensic science, Anal. Chem. 87 (2015) 306–327. doi:10.1021/ac504068a.

[17] M. Calcerrada, C. García-Ruiz, Analysis of questioned documents: A review, Anal. Chim. Acta. 853 (2015) 143–166. doi:10.1016/j.aca.2014.10.057.

[18] O.J. Old, L.M. Fullwood, R. Scott, G.R. Lloyd, L.M. Almond, N.A. Shepherd, N. Stone, H. Barr, C. Kendall, Vibrational spectroscopy for cancer diagnostics, Anal. Methods. 6 (2014) 3901–3917. doi:10.1039/c3ay42235f.

[19] H.J. Byrne, P. Knief, M.E. Keating, F. Bonnier, Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells, Chem. Soc. Rev. 45 (2016) 1865–1878. doi:10.1039/C5CS00440C.

[20] M.J. Kangas, R.M. Burks, J. Atwater, R.M. Lukowicz, P. Williams, A.E. Holmes, Colorimetric Sensor Arrays for the Detection and Identification of Chemical Weapons and Explosives, Crit. Rev. Anal. Chem. 47 (2017) 138–153. doi:10.1080/10408347.2016.1233805.

[21] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding, Anal. Chim. Acta. 879 (2015) 10–23. doi:10.1016/j.aca.2015.02.012.

[22] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: Taking the

magic away, J. Chemom. 28 (2014) 213–225. doi:10.1002/cem.2609.

[23] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem. 28 (2009) 1201–1222. doi:10.1016/j.trac.2009.07.007.

[24] J. Engel, J. Gerretzen, E. Szymańska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing?, TrAC - Trends Anal. Chem. 50 (2013) 96–106. doi:10.1016/j.trac.2013.04.015.

[25] P.D. Wentzell, Measurement errors in multivariate chemical data, J. Braz. Chem. Soc. 25 (2014) 183–196. doi:10.5935/0103-5053.20130293.

[26] R. Gargallo, R. Tauler, F. Cuesta-Sánchez, D.L. Massart, Validation of alternating least-squares multivariate curve resolution for chromatographic resolution and quantitation, TrAC - Trends Anal. Chem. 15 (1996) 279–286. doi:10.1016/0165-9936(96)00048-9.

[27] C. Ruckebusch, L. Duponchel, B. Sombret, J.P. Huvenne, J. Saurina, Time-Resolved Step-Scan FT-IR Spectroscopy: Focus on Multivariate Curve Resolution, J. Chem. Inf. Comput. Sci. 43 (2003) 1966–1973. doi:10.1021/ci034094i.

[28] J. Jaumot, V. Marchán, R. Gargallo, A. Grandas, R. Tauler, Multivariate Curve Resolution Applied to the Analysis and Resolution of Two-Dimensional [1 H, 15 N] NMR Reaction Spectra, Anal. Chem. 76 (2004) 7094–7101. doi:10.1021/ac049509t.

[29] S. Piqueras, L. Duponchel, R. Tauler, A. De Juan, Resolution and segmentation of hyperspectral biomedical images by Multivariate Curve Resolution-Alternating Least Squares, Anal. Chim. Acta. 705 (2011) 182–192.

doi:10.1016/j.aca.2011.05.020.

[30] M. Esteban, C. Ariño, J.M. Díaz-Cruz, M.S. Díaz-Cruz, R. Tauler, Multivariate curve resolution with alternating least squares optimisation: A soft-modelling approach to metal complexation studies by voltammetric techniques, TrAC - Trends Anal. Chem. 19 (2000) 49–61. doi:10.1016/S0165-9936(99)00184-3.

[31] W.H. Lawton, E.A. Sylvestre, Self modeling curve resolution, Technometrics. 13 (1971) 617–33. doi:10.1080/00401706.1971.10488823.

[32] R. Tauler, Multivariate curve resolution applied to second order data, Chemom. Intell. Lab. Syst. 30 (1995) 133–146. doi:10.1016/0169-7439(95)00047-X.

[33] P.D. Wentzell, D.T. Andrews, D.C. Hamilton, K. Faber, B.R. Kowalski, Maximum likelihood principal component analysis, J. Chemom. 11 (1997) 339–366. doi:10.1002/(SICI)1099-128X(199707)11:4<339::AID-CEM476>3.0.CO;2-L.

[34] P.D. Wentzell, S. Hou, Exploratory data analysis with noisy measurements, J. Chemom. 26 (2012) 264–281. doi:10.1002/cem.2428.

[35] P.D. Wentzell, T.K. Karakach, S. Roy, M.J. Martinez, C.P. Allen, M. Werner-Washburne, Multivariate curve resolution of time course microarray data, BMC Bioinformatics. 7 (2006). doi:10.1186/1471-2105-7-343.

[36] J.H. Friedman, J.W. Tukey, Projection Pursuit Algorithm for Exploratory Data-Analysis, IEEE Trans. Comput. C 23 (1974) 881–890. doi:10.1109/T-C.1974.224051.

[37] J.H. Friedman, W. Stuetzle, Projection Pursuit Regression, J. Am. Stat. Assoc. 76 (1981) 817–823. doi:10.2307/2287576.

[38] P.J. Huber, Projection Pursuit, Ann. Stat. 13 (1985) 435–475.

doi:10.1214/aos/1176349519.

[39] M.C. Jones, R. Sibson, What is Projection Pursuit, J. R. Stat. Soc. Ser. A-Statistics Soc. 150 (1987) 1–36. doi:10.2307/2981662.

[40] J.H. Friedman, Exploratory Projection Pursuit, J. Am. Stat. Assoc. 82 (1987) 249–266. doi:10.2307/2289161.

[41] Q. Guo, W. Wu, F. Questier, D.L. Massart, C. Boucon, S. de Jong, Sequential projection pursuit using genetic algorithms for data mining of analytical data, Anal. Chem. 72 (2000) 2846–2855. doi:10.1021/ac0000123.

[42] Q. Guo, W. Wu, D.L. Massart, C. Boucon, S. de Jong, Feature selection in sequential projection pursuit, Anal. Chim. Acta. 446 (2001) 85–96. doi:10.1016/S0003-2670(01)01000-5.

[43] S. Hou, P.D. Wentzell, Fast and simple methods for the optimization of kurtosis used as a projection pursuit index, Anal. Chim. Acta. 704 (2011) 1–15. doi:10.1016/j.aca.2011.08.006.

[44] S. Hou, P.D. Wentzell, Re-centered kurtosis as a projection pursuit index for multivariate data analysis, J. Chemom. 28 (2014) 370–384. doi:10.1002/cem.2568.

[45] S. Hou, P.D. Wentzell, Regularized projection pursuit for data with a small sample-to-variable ratio, Metabolomics. 10 (2014) 589–606. doi:10.1007/s11306-013-0612-z.

[46] S. Hou, P.D. Wentzell, C.B. Riley, Simple methods for the optimization of complex-valued kurtosis as a projection index, J. Chemom. 29 (2015) 224–236. doi:10.1002/cem.2700.

[47] V. Schoonjans, F. Questier, Q. Guo, Y. Van der Heyden, D.L. Massart, Assessing

molecular similarity/diversity of chemical structures by FT-IR spectroscopy, J. Pharm. Biomed. Anal. 24 (2001) 613–627. doi:10.1016/S0731-7085(00)00437-4.

[48] S. Caetano, T. Decaestecker, R. Put, M. Daszykowski, J. Van Bocxlaer, Y. Vander Heyden, Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors, Anal. Chim. Acta. 550 (2005) 92–106. doi:10.1016/j.aca.2005.06.069.

[49] M. Dumarey, A.M. van Nederkassel, I. Stanimirova, M. Daszykowski, F. Bensaid, M. Lees, G.J. Martins, J.R. Desmurs, J. Smeyers-Verbeke, Y. Vander Heyden, Recognizing paracetamol formulations with the same synthesis pathway based on their trace-enriched chromatographic impurity profiles, Anal. Chim. Acta. 655 (2009) 43–51. doi:10.1016/j.aca.2009.09.050.

[50] G. Alaerts, M. Merino-Arevalo, M. Dumarey, B. Dejaegher, N. Noppe, N. Matthijs, J. Smeyers-Verbeke, Y. Vander Heyden, Exploratory analysis of chromatographic fingerprints to distinguish rhizoma chuanxiong and rhizoma ligustici, J. Chromatogr. A. 1217 (2010) 7706–7716. doi:10.1016/j.chroma.2010.10.010.

[51] K. De Klerck, Y. Vander Heyden, D. Mangelings, Exploratory data analysis as a tool for similarity assessment and clustering of chiral polysaccharide-based systems used to separate pharmaceuticals in supercritical fluid chromatography, J. Chromatogr. A. 1326 (2014) 110–124. doi:10.1016/j.chroma.2013.12.052.

[52] J.F.Q. Pereira, C.S. Silva, A. Braz, M.F. Pimentel, R.S. Honorato, C. Pasquini, P.D. Wentzell, Projection pursuit and PCA associated with near and middle infrared hyperspectral images to investigate forensic cases of fraudulent

documents, Microchem. J. 130 (2017) 412–419. doi:10.1016/j.microc.2016.10.024.

[53] P.D. Wentzell, S. Hou, C.S. Silva, C.C. Wicks, M.F. Pimentel, Procrustes rotation as a diagnostic tool for projection pursuit analysis, Anal. Chim. Acta. 877 (2015) 51–63. doi:10.1016/j.aca.2015.03.006.

[54] V. Nguyen-Cong, B.M. Rode, Quantitative electronic structure-activity relationships of pyridinium cephalosporins using nonparametric regression methods, Eur. J. Med. Chem. 31 (1996) 479–484. doi:10.1016/0223-5234(96)85168-3.

[55] Z. Hassanzadeh, P. Ebrahimi, M. Kompany-Zareh, R. Ghavami, Radial basis function neural networks based on projection pursuit approach and solvatochromic descriptors: single and full column prediction of gas chromatography retention behavior of polychlorinated biphenyls, J. Chemom. 30 (2016) 589–601. doi:10.1002/cem.2822.

[56] P. Comon, Independent component analysis, A new concept?, Signal Processing. 36 (1994) 287–314. doi:10.1016/0165-1684(94)90029-9.

[57] L. De Lathauwer, B. De Moor, J. Vandewalle, An introduction to independent component analysis, J. Chemom. 14 (2000) 123–149. doi:10.1002/1099-128X(200005/06)14:3<123::AID-CEM589>3.0.CO;2-1.

[58] D. Jouan-Rimbaud Bouveresse, D.N. Rutledge, Independent Components Analysis: Theory and Applications, in: Data Handl. Sci. Technol., 30th ed., Elsevier, 2016. doi:10.1016/B978-0-444-63638-6.00007-3.

[59] A. Hyvarinen, E. Oja, Independent component analysis: algorithms and

applications, Neural Networks. 13 (2000) 411–430. doi:10.1016/S0893-6080(00)00026-5.

[60] J.F. Cardoso, A. Souloumiac, Blind Beamforming for Non-Gaussian Signals, Iee Proceedings-F Radar Signal Process. 140 (1993) 362–370.

[61] H. Stogbauer, A. Kraskov, S.A. Astakhov, P. Grassberger, Least-dependent-component analysis based on mutual information, Phys. Rev. E. 70 (2004) 66123–66141. doi:10.1103/PhysRevE.70.066123.

[62] A.J. Bell, T.J. Sejnowski, An Information-Maximization Approach to Blind Separation and Blind Deconvolution, Neural Comput. 7 (1995) 1129–1159. doi:10.1162/neco.1995.7.6.1129.

[63] C. Tong, T. Lan, X. Shi, Soft sensing of non-Gaussian processes using ensemble modified independent component regression, Chemom. Intell. Lab. Syst. 157 (2016) 120–126. doi:10.1016/j.chemolab.2016.07.006.

[64] C. Tong, A. Palazoglu, X. Yan, Improved ICA for process monitoring based on ensemble learning and Bayesian inference, Chemom. Intell. Lab. Syst. 135 (2014) 141–149. doi:10.1016/j.chemolab.2014.04.012.

[65] M.M. Rashid, J. Yu, A new dissimilarity method integrating multidimensional mutual information and independent component analysis for non-Gaussian dynamic process monitoring, Chemom. Intell. Lab. Syst. 115 (2012) 44–58. doi:10.1016/j.chemolab.2012.04.008.

[66] F. Ammari, R. Bendoula, D.J.-R. Bouveresse, D.N. Rutledge, J.-M. Roger, 3D front face solid-phase fluorescence spectroscopy combined with Independent Components Analysis to characterize organic matter in model soils, Talanta. 125

(2014) 146–152. doi:10.1016/j.talanta.2014.02.049.

[67] F. Ammari, C.B.Y. Cordella, N. Boughanmi, D.N. Rutledge, Independent components analysis applied to 3D-front-face fluorescence spectra of edible oils to study the antioxidant effect of Nigella sativa L. extract on the thermal stability of heated oils, Chemom. Intell. Lab. Syst. 113 (2012) 32–42. doi:10.1016/j.chemolab.2011.06.005.

[68] R. Saad, D.J.-R. Bouveresse, N. Locquet, D.N. Rutledge, Using pH variations to improve the discrimination of wines by 3D front face fluorescence spectroscopy associated to Independent Components Analysis, Talanta. 153 (2016) 278–284. doi:10.1016/j.talanta.2016.03.023.

[69] D.N. Rutledge, D.J.-R. Bouveresse, Independent Components Analysis with the JADE algorithm, Trac-Trends Anal. Chem. 50 (2013) 22–32. doi:10.1016/j.trac.2013.03.013.

[70] F. Ammari, D.J.-R. Bouveresse, N. Boughanmi, D.N. Rutledge, Study of the heat stability of sunflower oil enriched in natural antioxidants by different analytical techniques and front-face fluorescence spectroscopy combined with Independent Components Analysis, Talanta. 99 (2012) 323–329. doi:10.1016/j.talanta.2012.05.059.

[71] R. Garcia, A. Boussard, L. Rakotozafy, J. Nicolas, J. Potus, D.N. Rutledge, C.B.Y. Cordella, 3D-front-face fluorescence spectroscopy and independent components analysis: A new way to monitor bread dough development, Talanta. 147 (2016) 307–314. doi:10.1016/j.talanta.2015.10.002.

[72] A. Gaubert, Y. Clement, A. Bonhomme, B. Burger, D.J.-R. Bouveresse, D.

Rutledge, H. Casabianca, P. Lanteri, C. Bordes, Characterization of surfactant complex mixtures using Raman spectroscopy and signal extraction methods: Application to laundry detergent deformulation, Anal. Chim. Acta. 915 (2016) 36–48. doi:10.1016/j.aca.2016.02.016.

[73] A. Kassouf, M. El Rakwe, H. Chebib, V. Ducruet, D.N. Rutledge, J. Maalouly, Independent components analysis coupled with 3D-front-face fluorescence spectroscopy to study the interaction between plastic food packaging and olive oil, Anal. Chim. Acta. 839 (2014) 14–25. doi:10.1016/j.aca.2014.06.035.

[74] A. Kassouf, J. Maalouly, H. Chebib, D.N. Rutledge, V. Ducruet, Chemometric tools to highlight non-intentionally added substances (NIAS) in polyethylene terephthalate (PET), Talanta. 115 (2013) 928–937. doi:10.1016/j.talanta.2013.06.029.

[75] A. Kassouf, A. Ruellan, D.J.-R. Bouveresse, D.N. Rutledge, S. Domenek, J. Maalouly, H. Chebib, V. Ducruet, Attenuated total reflectance-mid infrared spectroscopy (ATR-MIR) coupled with independent components analysis (ICA): A fast method to determine plasticizers in polylactide (PLA), Talanta. 147 (2016) 569–580. doi:10.1016/j.talanta.2015.10.021.

[76] Y.B. Monakhova, R. Godelmann, T. Kuballa, S.P. Mushtakova, D.N. Rutledge, Independent components analysis to increase efficiency of discriminant analysis methods (FDA and LDA): Application to NMR fingerprinting of wine, Talanta. 141 (2015) 60–65. doi:10.1016/j.talanta.2015.03.037.

[77] Y.B. Monakhova, D.N. Rutledge, A. Rossmann, H.-U. Waiblinger, M. Mahler, M. Ilse, T. Kuballa, D.W. Lachenmeier, Determination of rice type by H-1 NMR

spectroscopy in combination with different chemometric tools, J. Chemom. 28 (2014) 83–92. doi:10.1002/cem.2576.

[78] M.R. Almeida, D.N. Correa, J.J. Zacca, L.P. Lima Logrado, R.J. Poppi, Detection of explosives on the surface of banknotes by Raman hyperspectral imaging and independent component analysis, Anal. Chim. Acta. 860 (2015) 15–22. doi:10.1016/j.aca.2014.12.034.

[79] X. Domingo-Almenara, A. Perera, N. Ramirez, N. Canellas, X. Correig, J. Brezmes, Compound identification in gas chromatography/mass spectrometry-based metabolomics by blind source separation, J. Chromatogr. A. 1409 (2015) 226–233. doi:10.1016/j.chroma.2015.07.044.

[80] S. Ghaheri, S. Masoum, A. Gholami, Resolving of challenging gas chromatography-mass spectrometry peak clusters in fragrance samples using multicomponent factorization approaches based on polygon inflation algorithm, J. Chromatogr. A. 1429 (2016) 317–328. doi:10.1016/j.chroma.2015.12.003.

[81] H. Seifi, S. Masoum, S. Seifi, Performance assessment of chemometric resolution methods utilized for extraction of pure components from overlapped signals in gas chromatography-mass spectrometry, J. Chromatogr. A. 1365 (2014) 173–182. doi:10.1016/j.chroma.2014.08.095.

[82] L. Gao, S. Ren, Integrating Independent Component Analysis with Artificial Neural Network to Analyze Overlapping Fluorescence Spectra of Organic Pollutants, J. Fluoresc. 22 (2012) 1595–1602. doi:10.1007/s10895-012-1100-y.

[83] A.C. Pereira, M.J. Carvalho, A. Miranda, J.M. Leca, V. Pereira, F. Albuquerque, J.C. Marques, M.S. Reis, Modelling the ageing process: A novel strategy to

analyze the wine evolution towards the expected features, Chemom. Intell. Lab. Syst. 154 (2016) 176–184. doi:10.1016/j.chemolab.2016.03.030.

[84] Y.B. Monakhova, A.M. Tsikin, S.P. Mushtakova, Independent Components Analysis as an Alternative to Principal Component Analysis and Discriminant Analysis Algorithms in the Processing of Spectrometric Data, J. Anal. Chem. 70 (2015) 1055–1061. doi:10.1134/S1061934815090117.

[85] Y.B. Monakhova, R. Godelmann, A. Hermann, T. Kuballa, C. Cannet, H. Schaefer, M. Spraul, D.N. Rutledge, Synergistic effect of the simultaneous chemometric analysis of H-1 NMR spectroscopic and stable isotope (SNIF-NMR, O-18, C-13) data: Application to wine analysis, Anal. Chim. Acta. 833 (2014) 29–39. doi:10.1016/j.aca.2014.05.005.

[86] T. Aguilera, J. Lozano, J.A. Paredes, F.J. Alvarez, J.I. Suarez, Electronic Nose Based on Independent Component Analysis Combined with Partial Least Squares and Artificial Neural Networks for Wine Prediction, Sensors. 12 (2012) 8055–8072. doi:10.3390/s120608055.

[87] P. Jia, F. Tian, Q. He, S. Fan, J. Liu, S.X. Yang, Feature extraction of wound infection data for electronic nose based on a novel weighted KPCA, Sensors and Actuators B-Chemical. 201 (2014) 555–566. doi:10.1016/j.snb.2014.04.025.

[88] J. Lasue, R.C. Wiens, T.F. Stepinski, O. Forni, S.M. Clegg, S. Maurice, C. Team, Nonlinear mapping technique for data visualization and clustering assessment of LIBS data: application to ChemCam data, Anal. Bioanal. Chem. 400 (2011) 3247–3260. doi:10.1007/s00216-011-4747-3.

[89] M. Zhang, P. Tong, W. Wang, J. Geng, Y. Du, Proteomic profile analysis and

biomarker discovery from mass spectra using independent component analysis combined with uncorrelated linear discriminant analysis, Chemom. Intell. Lab. Syst. 105 (2011) 207–214. doi:10.1016/j.chemolab.2011.01.007.

[90] L. Wang, D. Yang, C. Fang, Z. Chen, P.J. Lesniewski, M. Mallavarapu, R. Naidu, Application of neural networks with novel independent component analysis methodologies to a Prussian blue modified glassy carbon electrode array, Talanta. 131 (2015) 395–403. doi:10.1016/j.talanta.2014.08.010.

[91] L. Wang, D. Yang, Z. Chen, P.J. Lesniewski, R. Naidu, Application of neural networks with novel independent component analysis methodologies for the simultaneous determination of cadmium, copper, and lead using an ISE array, J. Chemom. 28 (2014) 491–498. doi:10.1002/cem.2599.

[92] Y. Shao, Y. He, Visible/Near Infrared Spectroscopy and Chemometrics for the Prediction of Trace Element (Fe and Zn) Levels in Rice Leaf, Sensors. 13 (2013) 1872–1883. doi:10.3390/s130201872.

[93] A. Al-Mbaideen, M. Benaissa, Coupling subband decomposition and independent component regression for quantitative NIR spectroscopy, Chemom. Intell. Lab. Syst. 108 (2011) 112–122. doi:10.1016/j.chemolab.2011.05.012.

[94] Z. Xiaobo, Z. Jiewen, M. Holmes, M. Hanpin, S. Jiyong, Y. Xiaopin, L. Yanxiao, Independent component analysis in information extraction from visible/near-infrared hyperspectral imaging data of cucumber leaves, Chemom. Intell. Lab. Syst. 104 (2010) 265–270. doi:10.1016/j.chemolab.2010.08.019.

[95] Y. Lu, C. Du, C. Yu, J. Zhou, Determination of Nitrogen in Rapeseed by Fourier Transform Infrared Photoacoustic Spectroscopy and Independent Component

Analysis, Anal. Lett. 48 (2015) 1150–1162. doi:10.1080/00032719.2014.976872.

[96] X. Zhang, R. Tauler, Measuring and comparing the resolution performance and the extent of rotation ambiguities of some bilinear modeling methods, Chemom. Intell. Lab. Syst. 147 (2015) 47–57. doi:10.1016/j.chemolab.2015.08.005.

[97] Y.B. Monakhova, S.A. Astakhov, A. Kraskov, S.P. Mushtakova, Independent components in spectroscopic analysis of complex mixtures, Chemom. Intell. Lab. Syst. 103 (2010) 108–115. doi:10.1016/j.chemolab.2010.05.023.

[98] Y. Gut, M. Boiret, L. Bultel, T. Renaud, A. Chetouani, A. Hafiane, Y.-M. Ginot, R. Jennane, Application of chemometric algorithms to MALDI mass spectrometry imaging of pharmaceutical tablets, J. Pharm. Biomed. Anal. 105 (2015) 91–100. doi:10.1016/j.jpba.2014.11.047.

[99] J. Li, J. Gao, H. Li, X. Yang, Y. Liu, Study of the synthesis mechanism of 4-amino-3,5-dimethyl pyrazole by fibre optic in-line FT-IR spectroscopy combined with independent component analysis, Anal. Methods. 6 (2014) 4305–4311. doi:10.1039/c4ay00334a.

[100] Y.B. Monakhova, A.M. Tsikin, S.P. Mushtakova, M. Mecozzi, Independent component analysis and multivariate curve resolution to improve spectral interpretation of complex spectroscopic data sets: Application to infrared spectra of marine organic matter aggregates, Microchem. J. 118 (2015) 211–222. doi:10.1016/j.microc.2014.10.001.

[101] L. Cui, Z. Ling, J. Poon, S.K. Poon, H. Chen, J. Gao, P. Kwan, K. Fan, A parallel model of independent component analysis constrained by a 5-parameter reference curve and its solution by multi-target particle swarm optimization, Anal. Methods.

6 (2014) 2679–2686. doi:10.1039/c3ay42196a.

[102] L. Gorski, W.W. Kubiak, M. Jakubowska, Independent Components Analysis of the Overlapping Voltammetric Signals, Electroanalysis. 28 (2016) 1470–1477. doi:10.1002/elan.201501089.

[103] S. Masoum, H. Seifi, E.H. Ebrahimabadi, Characterization of volatile components in Calligonum comosum by coupling gas chromatography-mass spectrometry and mean field approach independent component analysis, Anal. Methods. 5 (2013) 4639–4647. doi:10.1039/c3ay40451j.

[104] Y.B. Monakhova, S.P. Mushtakova, Application of MATLAB package for the automation of the chemometric processing of spectrometric signals in the analysis of complex mixtures, J. Anal. Chem. 71 (2016) 759–767. doi:10.1134/S1061934816060125.

[105] Y.B. Monakhova, M. Betzgen, B.W.K. Diehl, H-1 NMR as a release methodology for the analysis of phospholipids and other constituents in infant nutrition, Anal. Methods. 8 (2016) 7493–7499. doi:10.1039/c6ay02063a.

[106] I. Toumi, B. Torresani, S. Caldarelli, Effective Processing of Pulse Field Gradient NMR of Mixtures by Blind Source Separation, Anal. Chem. 85 (2013) 11344–11351. doi:10.1021/ac402085x.

[107] M. Boiret, D.N. Rutledge, N. Gorretta, Y.-M. Ginot, J.-M. Roger, Application of independent component analysis on Raman images of a pharmaceutical drug product: Pure spectra determination and spatial distribution of constituents, J. Pharm. Biomed. Anal. 90 (2014) 78–84. doi:10.1016/j.jpba.2013.11.025.

[108] M. Toiviainen, F. Corona, J. Paaso, P. Teppola, Blind source separation in diffuse

reflectance NIR spectroscopy using independent component analysis, J. Chemom. 24 (2010) 514–522. doi:10.1002/cem.1316.

[109] W. Windig, M.R. Keenan, Homeopathic ICA: A simple approach to expand the use of independent component analysis (ICA), Chemom. Intell. Lab. Syst. 142 (2015) 54–63. doi:10.1016/j.chemolab.2015.01.003.

[110] B. Debrus, P. Lebrun, J.M. Kindenge, F. Lecomte, A. Ceccato, G. Caliaro, J.M.T. Mbay, B. Boulanger, R.D. Marini, E. Rozet, P. Hubert, Innovative high-performance liquid chromatography method development for the screening of 19 antimalarial drugs based on a generic approach, using design of experiments, independent component analysis and design space, J. Chromatogr. A. 1218 (2011) 5205–5215. doi:10.1016/j.chroma.2011.05.102.

[111] B. Debrus, P. Lebrun, E. Rozet, T. Schofield, J.K. Mbinze, R.D. Marini, S. Rudaz, B. Boulanger, P. Hubert, A New Method for Quality by Design Robust Optimization in Liquid Chromatography, LC-GC Eur. 26 (2013) 2–8.

[112] X. Li, J. Hansen, X. Zhao, X. Lu, C. Weigert, H.-U. Haering, B.K. Pedersen, P. Plomgaard, R. Lehmann, G. Xu, Independent component analysis in non-hypothesis driven metabolomics: Improvement of pattern discovery and simplification of biological data interpretation demonstrated with plasma samples of exercising humans, J. Chromatogr. B-Analytical Technol. Biomed. Life Sci. 910 (2012) 156–162. doi:10.1016/j.jchromb.2012.06.030.

[113] Y.B. Monakhova, S.P. Mushtakova, S.S. Kolesnikova, S.A. Astakhov, Chemometrics-assisted spectrophotometric method for simultaneous determination of vitamins in complex mixtures, Anal. Bioanal. Chem. 397 (2010) 1297–1306.

doi:10.1007/s00216-010-3623-x.

[114] M. Pietroletti, S. Mattiello, F. Moscato, F. Oteri, M. Mecozzi, One Step Ultrasound Extraction and Purification Method for the Gas Chromatographic Analysis of Hydrocarbons from Marine Sediments: Application to the Monitoring of Italian Coasts, Chromatographia. 75 (2012) 961–971. doi:10.1007/s10337-011-2172-6.

[115] H. Parastar, M. Jalali-Heravi, R. Tauler, Is independent component analysis appropriate for multivariate resolution in analytical chemistry?, TrAC - Trends Anal. Chem. 31 (2012) 134–143. doi:10.1016/j.trac.2011.07.010.

[116] J. Huang, X. Yan, Dynamic process fault detection and diagnosis based on dynamic principal component analysis, dynamic independent component analysis and Bayesian inference, Chemom. Intell. Lab. Syst. 148 (2015) 115–127. doi:10.1016/j.chemolab.2015.09.010.

[117] J. Huang, X. Yan, Double-step block division plant-wide fault detection and diagnosis based on variable distributions and relevant features, J. Chemom. 29 (2015) 587–605. doi:10.1002/cem.2743.

[118] H. Jiang, Q. Chen, Development of Electronic Nose and Near Infrared Spectroscopy Analysis Techniques to Monitor the Critical Time in SSF Process of Feed Protein, Sensors. 14 (2014) 19441–19456. doi:10.3390/s141019441.

[119] P.P. Odiowei, Y. Cao, State-space independent component analysis for nonlinear dynamic process monitoring, Chemom. Intell. Lab. Syst. 103 (2010) 59–65. doi:10.1016/j.chemolab.2010.05.014.

[120] M.A.A. Rad, M.J. Yazdanpanah, Designing supervised local neural network

classifiers based on EM clustering for fault diagnosis of Tennessee Eastman process, Chemom. Intell. Lab. Syst. 146 (2015) 149–157. doi:10.1016/j.chemolab.2015.05.013.

[121] Y. Yang, Y. Chen, X. Chen, X. Liu, Multivariate industrial process monitoring based on the integration method of canonical variate analysis and independent component analysis, Chemom. Intell. Lab. Syst. 116 (2012) 94–101. doi:10.1016/j.chemolab.2012.04.013.

[122] H. Albazzaz, X.Z. Wang, Statistical process control charts for batch operations based on independent component analysis, Ind. Eng. Chem. Res. 43 (2004) 6731–6741. doi:10.1021/ie049582+.

[123] J. V Stone, Independent Component Analysis: A Tutorial Introduction, The MIT Press, Cambridge, Massachusetts, 2004. doi:10.1007/978-3-642-81500-3_1.

[124] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., New York, USA, 2001. doi:10.1002/0471221317.

[125] R. Bro, Å. Rinnan, N.M. Faber, Standard error of prediction for multilinear PLS 2. Practical implementation in fluorescence spectroscopy, Chemom. Intell. Lab. Syst. 75 (2005) 69–76. doi:10.1016/j.chemolab.2004.04.014.

[126] D.P. Stevenson, F.H. Stross, R.F. Heizer, An evaluation of x-ray fluorescence analysis as a method for correlating obsidian artifacts with source loacation, Archaeometry. 13 (1971) 17–25. doi:10.1111/j.1475-4754.1971.tb00026.x.

[127] C.S. Silva, F. de S.L. Borba, M.F. Pimentel, M.J.C. Pontes, R.S. Honorato, C. Pasquini, Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis, Microchem. J. 109 (2013) 122–127.

doi:10.1016/j.microc.2012.03.025.

[128] E.M. Oliveira, J.W.B. Braga, A.F. da Costa, E.M. Oliveira, J.W.B. Braga, A.F. da Costa, Discrimination between similar woods by molecular fluorescence and partial least squares, Quim. Nova. 38 (2015) 1176–1180. doi:10.5935/0100-4042.20150127.

[129] T.C.M. Pastore, J.W.B. Braga, V.T.R. Coradin, W.L.E. Magalhães, E.Y.A. Okino, J.A.A. Camargos, G.I.B. De Muñiz, O.A. Bressan, F. Davrieux, Near infrared spectroscopy (NIRS) as a potential tool for monitoring trade of similar woods: Discrimination of true mahogany, cedar, andiroba, and curupixá, Holzforschung. 65 (2011) 73–80. doi:10.1515/HF.2011.010.

[130] M.N. Leger, L. Vega-Montoto, P.D. Wentzell, Methods for systematic investigation of measurement error covariance matrices, Chemom. Intell. Lab. Syst. 77 (2005) 181–205. doi:10.1016/j.chemolab.2004.09.017.

[131] T.K. Karakach, P.D. Wentzell, J.A. Walter, Characterization of the measurement error structure in 1D 1H NMR data for metabolomics studies, Anal. Chim. Acta. 636 (2009) 163–174. doi:10.1016/j.aca.2009.01.048.

[132] P.D. Wentzell, A.C. Tarasuk, Characterization of heteroscedastic measurement noise in the absence of replicates, Anal. Chim. Acta. 847 (2014) 16–28. doi:10.1016/j.aca.2014.08.007.

[133] J.D.J. Ingle, S.R. Crouch, Spectrochemical analysis, 1st ed., Prentice Hall, 1988.

[134] A.T. Ince, J.G. Williams, A.L. Gray, Noise in inductively coupled plasma mass spectrometry: Some preliminary measurements, J. Anal. At. Spectrom. 8 (1993) 899–903. doi:10.1039/ja9930800899.

[135] Y. Hayashi, R. Matsuda, Deductive Prediction of Measurement Precision from Signal and Noise in Liquid Chromatography, Anal. Chem. 66 (1994) 2874–2881.

[136] Y. Hayashi, R. Matsuda, Deductive Prediction of Measurement Precision from Signal and Noise in Fluorometry, Anal. Sci. 11 (1995) 929–934.

[137] C.M. Van Vliet, Responsivity and noise in illustrative solid-state chemical sensors, Sensors Actuators B. Chem. 24 (1995) 6–16. doi:10.1016/0925-4005(95)85006-6.

[138] Y. Hayashi, R. Matsuda, R.B. Poe, Measurement precision and 1/f noise in analytical instruments, J. Chromatogr. A. 722 (1996) 157–167.

[139] C.R. Mittermayr, B. Lendl, E. Rosenberg, M. Grasserbauer, The application of the wavelet power spectrum to detect and estimate 1/f noise in the presence of analytical signals, Anal. Chim. Acta. 388 (1999) 303–313. doi:10.1016/S0003-2670(99)00083-5.

[140] P.D. Wentzell, C.S. Cleary, M. Kompany-Zareh, Improved modeling of multivariate measurement errors based on the Wishart distribution, Anal. Chim. Acta. 959 (2017) 1–14. doi:10.1016/j.aca.2016.12.009.

[141] P.D. Wentzell, Other Topics in Soft-Modeling: Maximum Likelihood-Based Soft-Modeling Methods, in: S.D. Brown, R. Tauler, B. Walczak (Eds.), Compr. Chemom., 2nd ed., Elsevier, 2010: pp. 507–558. doi:10.1016/B978-044452701-1.00057-0.

[142] P.D. Wentzell, M.T. Lohnes, Maximum likelihood principal component analysis with correlated measurement errors: Theoretical and practical considerations, Chemom. Intell. Lab. Syst. 45 (1999) 65–85. doi:10.1016/S0169-7439(98)00090-2.

[143] P.D. Wentzell, D.T. Andrews, B.R. Kowalski, Maximum likelihood multivariate calibration, Anal. Chem. 69 (1997) 2299–2311. doi:10.1021/ac961029h.

[144] S.K. Schreyer, M. Bidinosti, P.D. Wentzell, Application of maximum likelihood principal components regression to fluorescence emission spectra, Appl. Spectrosc. 56 (2002) 789–796. doi:10.1366/000370202760076857.

[145] R. Tauler, M. Viana, X. Querol, A. Alastuey, R.M. Flight, P.D. Wentzell, P.K. Hopke, Comparison of the results obtained by four receptor modelling methods in aerosol source apportionment studies, Atmos. Environ. 43 (2009) 3989–3997. doi:10.1016/j.atmosenv.2009.05.018.

[146] I. Nistor, M. Cao, B. Debrus, P. Lebrun, F. Lecomte, E. Rozet, L. Angenot, M. Frederich, R. Oprean, P. Hubert, Application of a new optimization strategy for the separation of tertiary alkaloids extracted from Strychnos usambarensis leaves, J. Pharm. Biomed. Anal. 56 (2011) 30–37. doi:10.1016/j.jpba.2011.04.027.

[147] V.T.R. Coradin, J.J.A. Camargos, L.F. Marques, E.R. de Silva Jr, Madeiras similares ao mogno (Swietenia macrophylla King.): chave ilustrada para identificação anatômica em campo, Serviço Florestal Brasileiro, Brasilia, 2009.

[148] W.L. Stern, Index Xylariorum: Institutional Wood Collections of the World, IAWA Bull. 9 (1988) 209–210.

[149] P.D. Wentzell, C.C. Wicks, J.W.B. Braga, L.F. Soares, T.C.M. Pastore, V.T.R. Coradin, F. Davrieux, Implications of measurement error structure on the visualization of multivariate chemical data: Hazards and alternatives, Can. J. Chem. 96 (2018) 738–748. doi:10.1139/cjc-2017-0730.